Studies in Computational Intelligence 231

Mislav Grgic
Kresimir Delac
Mohammed Ghanbari (Eds.)

Recent Advances in Multimedia Signal Processing and Communications

Springer

Mislav Grgic, Kresimir Delac, and Mohammed Ghanbari (Eds.)

Recent Advances in Multimedia Signal Processing and Communications

Studies in Computational Intelligence, Volume 231

Mislav Grgic, Kresimir Delac and
Mohammed Ghanbari (Eds.)

Recent Advances in Multimedia Signal Processing and Communications

Springer

Mislav Grgic
University of Zagreb
Faculty of Electrical Engineering and Computing
Department of Wireless Communications
Unska 3/XII
HR-10000 Zagreb
Croatia
E-mail: mislav.grgic@fer.hr

Kresimir Delac
University of Zagreb
Faculty of Electrical Engineering and Computing
Department of Wireless Communications
Unska 3/XII
HR-10000 Zagreb
Croatia
E-mail: kdelac@ieee.org

Mohammed Ghanbari
School of Computer Science and Electronic Engineering
University of Essex
Wivenhoe Park
Colchester CO4 3SQ
UK
E-mail: ghan@essex.ac.uk

ISBN 978-3-642-02899-1 e-ISBN 978-3-642-02900-4

DOI 10.1007/978-3-642-02900-4

Studies in Computational Intelligence ISSN 1860-949X

Library of Congress Control Number: Applied for

Typeset & *Cover Design:* Scientific Publishing Services Pvt. Ltd., Chennai, India.

Printed in acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

Preface

The rapid increase in computing power and communication speed, coupled with computer storage facilities availability, has led to a new age of multimedia applications. Multimedia is practically everywhere and all around us we can feel its presence in almost all applications ranging from online video databases, IPTV, interactive multimedia and more recently in multimedia based social interaction. These new growing applications require high-quality data storage, easy access to multimedia content and reliable delivery. Moving ever closer to commercial deployment also aroused a higher awareness of security and intellectual property management issues.

All the aforementioned requirements resulted in higher demands on various areas of research (signal processing, image/video processing and analysis, communication protocols, content search, watermarking, etc.). This book covers the most prominent research issues in multimedia and is divided into four main sections: i) content based retrieval, ii) storage and remote access, iii) watermarking and copyright protection and iv) multimedia applications.

Chapter 1 of the first section presents an analysis on how color is used and why is it crucial in nowadays multimedia applications. In chapter 2 the authors give an overview of the advances in video abstraction for fast content browsing, transmission, retrieval and skimming in large video databases and chapter 3 extends the discussion on video summarization even further. Content retrieval problem is tackled in chapter 4 by describing a novel method for producing meaningful segments suitable for MPEG-7 description based on binary partition trees (BPTs). Chapter 5 deals with object recognition in cluttered scenes. Chapter 6 presents pose-invariant face recognition system important for human-computer interaction, virtual collaboration, video indexing etc. In chapter 7 the authors analyze the problem of intuitive and naturally feeling interfaces for interacting with multimedia content.

In the second section various aspects of storing, accessing, browsing, retrieval and delivery of content are addressed. Chapter 8 is an overview of super-resolution techniques used in multidimensional signal processing. Chapter 9 deals with technologies that enable efficient handling of multimedia content through motion models for scalable and high definition video coding. Chapter 10 describes recent advances in video transcoding and processing architectures for multimedia content adaptation (even to lower bit rates) for better browsing and access. Chapters 11 and 12 deal with communication issues, the former with the transmission of wavelet coded images and the latter with communication at very low bitrates, using the compression of facial videos as an application example. In chapter 13 the authors present the next step in IPTV evolution - Next Generation Networks and highlight the standardization efforts in this area. Chapter 14 examines HDTV characteristics as well as the influence of video coding tools on high definition video quality.

The third section deals with security and intellectual property management issues through watermarking and encryption techniques. Chapter 15 gives an exhaustive introduction into the area by formulating the encryption problem in general and further by formulating the watermarking problem. It then continues describing a method for robust watermarking in fractional wavelet domain. Chapter 16 gives a brief overview of multimedia encryption and chapter 17 presents a method for content-based image replication detection. A novel image watermarking algorithm in discrete wavelet transform domain for stereo image coding is presented in chapter 18 and a reversible watermarking for 3D cameras in chapter 19. Chapters 20 gives an overview of digital audio watermarking systems and of existing systems that use different audio watermarking methods.

The fourth and final section of the book starts with a survey of music structure analysis techniques (like genre recognition, summarization, search, streaming and compression) for music applications, given in chapter 21. Chapter 22 presents one virtual learning environment and explores how such a narrative learning environment facilitates information extraction and emergent narrative mechanisms. Why art is a perfect testbed for computer vision and signal processing research is answered in chapter 23. The chapter walks us through a few projects and immerses the reader into a virtual world. Chapter 24 concludes the book with a survey of image processing techniques in yet another application area - digital mammography.

We strongly feel that this unique combination of surveys on past research, basic theoretical analyses and novel state-of-the-art research developments will make this book a valuable contribution to anyone's shelf.

May 2009

Mislav Grgic
Kresimir Delac
Mohammed Ghanbari

Contents

Section I

Section II

Section III

Section IV

Color in Multimedia

Rastislav Lukac

Abstract. Color plays an important role in the human perception and interpretation of the visual world. It is therefore not surprising that in many application areas manufacturers and consumers have been losing interest in conventional grayscale imaging and have been turning instead to its information-richer, color-driven counterpart. An explosive growth in the diversity of image and video processing solutions developed in the past decade has resulted, among others, in a number of commercial products for digital imaging and multimedia applications where color provides crucial information for both human observers and data processing machines. Methods for representing and using color for the purpose of image and video acquisition, processing, analysis, storage, displaying, printing, quantitative manipulation and image quality evaluation in such applications are surveyed in this chapter.

1 Introduction

Resent advances in color theory, hardware and software, computer vision, digital imaging, multimedia, graphics, biomedicine and telecommunications are making color image processing a daily necessity in practical life. In these, as well as many other application areas, color image processing has become commonplace as consumers choose the convenience of color imaging over traditional gray-scale imaging. To utilize color as a visual cue, an appropriate representation of the color signal is needed in order to specify object colors and effectively acquire, manipulate, store and display the image data.

This chapter focuses on color imaging in multimedia applications. Namely, the chapter describes popular methods for representing color and makes a connection between these methods and a color perception mechanism of a human visual

Rastislav Lukac
Epson Edge, 3771 Victoria Park Avenue, Toronto, Ontario, M1W 3Z5, Canada
e-mail: `lukacr@ieee.org`
`http://www.colorimageprocessing.com`

M. Grgic et al. (Eds.): Rec. Advan. in Mult. Sig. Process. and Commun., SCI 231, pp. 1–25.
springerlink.com

system. Furthermore, it is shown how these methods can assist in the design of image and video acquisition, processing, analysis, storage and performance evaluation solutions, as they can be used to define and discriminate colors, judge similarity between colors and identify color categories for a number of applications.

To facilitate the discussions on the role of color in human perception, Section 2 presents color fundamentals, including color vision basics and associated digital signal representations. Building on these fundamentals, Section 3 describes use of color representations in various digital imaging and multimedia tasks. Namely, Section 3.1 focuses on image acquisition and displaying whereas Section 3.2 targets color printing. In Section 3.3, the reader's attention is shifted to the problem of color discrimination using popular measures for evaluation of distances and similarities among color vectors. This is followed by the discussions on color-driven image processing and analysis in Section 3.4 and compression and encoding in Section 3.5. Quantitative manipulation and image quality evaluation solutions are reviewed in Section 3.6. Finally, this chapter concludes with Section 4 by summarizing main digital color imaging ideas for multimedia processing and its applications.

2 Color Basics

Color is a psycho-physiological sensation [1] used by the human observers to sense the environment and understand its visual semantics. It is commonly interpreted as a perceptual result of light interacting with the spectral sensitivities of the photoreceptors on the retina of the human eye. Visible light occupies a small portion of the electromagnetic spectrum. This portion is commonly referred to as a visible spectrum having wavelengths, usually denoted with the symbol λ and measured in the units of nanometers (nm), ranged from approximately 380 nm to 700 nm [2]. Different colors correspond to electromagnetic waves of different wavelengths, such as $\lambda \approx 400\ \mathrm{nm}$ for blue, $\lambda \approx 550\ \mathrm{nm}$ for green, and $\lambda \approx 700\ \mathrm{nm}$ for red. Wavelengths of $\lambda < 400\ \mathrm{nm}$ and $\lambda > 700\ \mathrm{nm}$ correspond, respectively, to an ultraviolet and infrared spectrum.

2.1 Human Visual System

The retina contains two types of photoreceptors, termed as rods and cones, responsible for sensing the visual steers [3]. The rods, usually around 120 million distributed all over the retina, are highly sensitive to light but insensitive to color. Therefore, they are very useful for scotopic vision, which is vision in low light conditions, where no color is usually seen and only shades of gray can be perceived. The cones, usually around 7 million localized at a place called the fovea, are less sensitive than rods and are responsible for the color sensitivity. Thus, they are greatly useful in photopic vision, which refers to vision under typical light conditions, where the perception mechanism completely relies on the less-sensitive cones as the highly-sensitive rods become saturated and do not contribute to vision. Finally, for certain

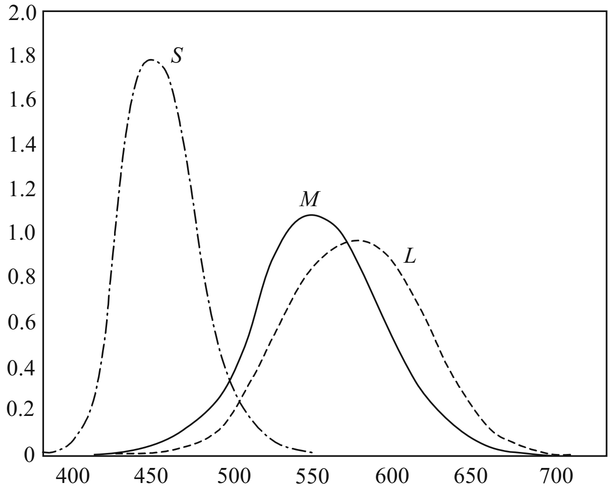


Fig. 1 Estimated effective sensitivities of the cones. The horizontal axis shows wavelengths in nanometers.

illumination levels, which can be seen as gradual changes from scotopic to photopic vision, the perception is based on both rods and cones.

It is estimated that humans are capable of resolving about 10 million color sensations [4]. The ability of the human eye to distinguish colors is based on the three types of cones which differ in their spectral sensitivities due to different spectral absorption characteristics of their photosensitive pigments. The three types of color-receptive cells — commonly called S, M, and L cones for their respective sensitivity to short, middle, and long wavelengths — yield three signals based on the extent to which each type of cones is stimulated. Namely, as shown in Fig. 1, they are most sensitive to light with wavelengths of approximately 420 nm for S cones, 534 nm for M cones, and 564 nm for L cones, which respectively correspond to light perceived by humans as violet, green, and yellowish-green. The response of color-receptive cells can be accurately modelled by a linear system defined by the cone spectral sensitivities. Note that two spectrally different lights that have the same L, M, and S cone responses give the same color sensation — this well-known phenomenon is termed in the literature as metamers.

Color cannot be specified without an observer and therefore it is not an inherent feature of an object. The attribution of colors to objects or lights is a specific instance of the so-called stimulus error wherein a sensation experienced by an observer is identified with the stimulus causing the sensation [3]. Given an object, any desired color stimulus can be theoretically produced by manipulating the illumination. Thus, the color stimulus depends on both the illumination and the object, and is obtainable as the color-signal spectrum via the product of the illumination spectrum and the surface reflectance spectrum. This also suggests that both the illumination and the object reflectance are equally important for the perception of color [5].

2.2 Color Representation

Since the perception of color depends on the response of three types of cones, any visible color can be represented using three numbers called tristimulus values [6]. This trichromatic property of the human visual system suggests that there exist three color primaries or colorimetrically independent light sources such that the color of any primary cannot be visually matched by a linear combination of the two others. The numerical representation of a particular color can thus be specified by a three-component vector within a three-dimensional coordinate system with color primaries lying on its axes. The set of all such vectors constitutes a color space.

Two color spaces, CIE-RGB and CIE-XYZ, were standardized by the Commission Internationale de l'Éclairage (CIE). The RGB color space was derived based on color matching experiments which aim at finding a match between a color obtained through an additive mixture of color primaries and a color sensation [7]. Fig. 2 shows the resulting color matching functions. A number of tasks, such as image storage, processing, analysis and visualization is done in the RGB color space, which will be discussed in detail later. The XYZ color space was obtained through the transformation from $\bar{r}(\lambda)$, $\bar{g}(\lambda)$, $\bar{b}(\lambda)$ to $\bar{x}(\lambda)$, $\bar{y}(\lambda)$, $\bar{z}(\lambda)$ color matching functions. This transform can be formally written as follows [6], [8]:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.49 & 0.31 & 0.20 \\ 0.17697 & 0.81240 & 0.01063 \\ 0 & 0.01 & 0.99 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{1}$$

As can be seen in Fig. 3, the $\bar{x}(\lambda)$, $\bar{y}(\lambda)$, $\bar{z}(\lambda)$ color matching functions avoid negative values, which was one of the requirements in the design of physical measuring

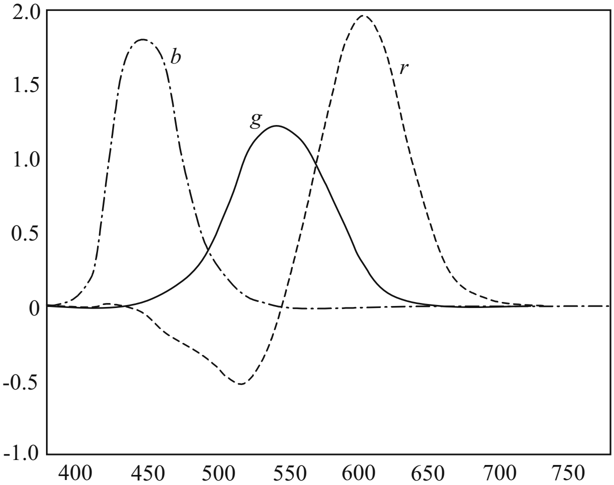


Fig. 2 CIE $\bar{r}(\lambda)$, $\bar{g}(\lambda)$, $\bar{b}(\lambda)$ color matching functions. The horizontal axis shows wavelengths in nanometers.

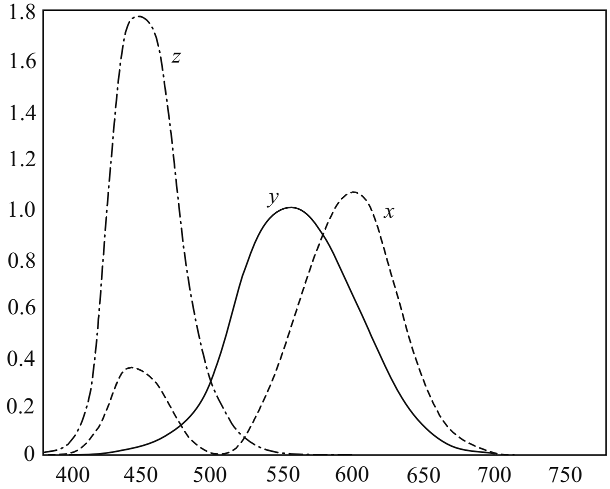


Fig. 3 CIE $\bar{x}(\lambda)$, $\bar{y}(\lambda)$, $\bar{z}(\lambda)$ color matching functions. The horizontal axis shows wavelengths in nanometers.

devices. The Y component correspond to the luminance whereas X and Z do not correspond to any perceptual attributes. The XYZ color space is device-independent and thus very useful in color management or generally applications where consistent color representation across devices with different characteristics is required.

As already discussed, both the illumination and the object reflectance are factors in the perception of color [4]. Using $f_I(\lambda)$ and $f_R(\lambda)$ as physical measurements of illumination and surface reflectance, respectively, a color response can be characterized by the following values [3], [4]:

$$\begin{aligned} X &= K \int_{\lambda} f_I(\lambda) f_R(\lambda) \bar{x}(\lambda)\,\mathrm{d}\lambda \\ Y &= K \int_{\lambda} f_I(\lambda) f_R(\lambda) \bar{y}(\lambda)\,\mathrm{d}\lambda \\ Z &= K \int_{\lambda} f_I(\lambda) f_R(\lambda) \bar{z}(\lambda)\,\mathrm{d}\lambda \end{aligned} \tag{2}$$

where K is a normalization factor calculated as follows:

$$K = 100 / \int_{\lambda} f_I(\lambda) \bar{y}(\lambda)\,\mathrm{d}\lambda \tag{3}$$

so that $Y = 100$ for a perfect diffuser and $f_R(\lambda) = 1$ for a perfect reflector.

For effective visualization on two-dimensional media, a three-component color vector can be normalized by dividing each of its component by the sum of all its

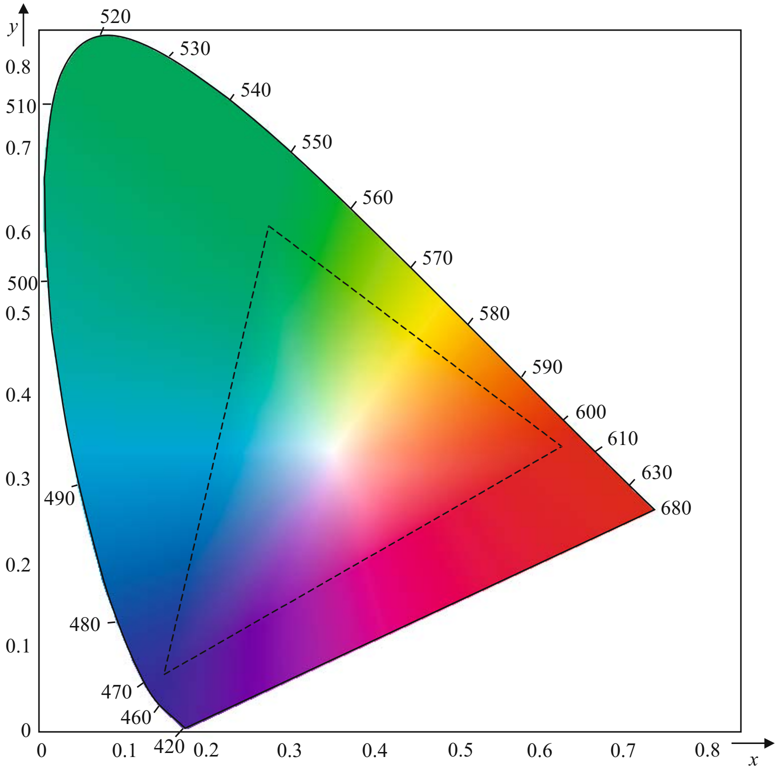


Fig. 4 CIE xy chromaticity diagram

components. By doing so, color vectors are normalized for changes in intensity and mapped to a unit plane which indicates their chromaticity [6]. Therefore, the coordinates of the normalized vectors are often termed as chromaticity coordinates and a plot of colors on the unit plane using these coordinates is referred to as a chromaticity diagram. Since normalized components sum up to unity, any normalized component can be expressed using two others. For instance, $z = 1 - x - y$ for components:

$$x = \frac{X}{X+Y+Z}, \quad y = \frac{Y}{X+Y+Z}, \quad z = \frac{Z}{X+Y+Z} \tag{4}$$

which constitute a normalized version of the XYZ tristimulus vector.

Typical diagrams plot only two chromaticity coordinates along orthogonal axes [3]. Fig. 4 shows the CIE xy chromaticity diagram which is commonly employed both to specify the gamut of human vision and describe the possible range of colors that can be produced by an output device. The gamut is convex in shape and its outer curved boundary denotes the spectral locus, with wavelengths shown in nanometers. Colors within the locus signify all visible chromaticities and as can be seen, they correspond to non-negative values of chromaticity coordinates. All colors that can be formed by mixing two colors located at any two points on the chromaticity diagram lie on a straight line connecting those two points. Colors obtained by mixing three or more colors are located, respectively, within a triangle or a higher-order shape determined by points corresponding to those source colors. The gamut of any device is a subset of the gamut of human vision due to the inability of various devices to precisely represent all visible chromaticities. The triangle in Fig. 4 depicts an example of such gamut limitations.

3 Applications of Color Representations

A number of color models were introduced to achieve the desired performance of digital imaging solutions and/or provide convenient color representations for the purpose of image data manipulation, processing, storage and inspection. Devices such as cameras, scanners, printers and displays operate in color spaces which are predefined and fixed in hardware by the manufacturers. To match image colors among these devices, the best profile for each such device has to be selected by the user. On the other hand, when designing an image processing, analysis or image quality evaluation method or when using various image editing software, it is possible to choose any color space as a working space. Some popular color spaces are reviewed in what follows.

3.1 Image Acquisition and Displaying

The RGB space (Fig. 5) is the most popular color space today because it provides a reasonable resolution, range and depth of color reproduction while being efficiently implementable on various hardware platforms. It is a device-dependent color space, as it models the output of physical devices rather than human visual perception. Without color management, different devices detect or reproduce the same RGB color vector differently since color imaging elements vary among manufacturers or even the response of these elements vary within the same device due to ageing.

To offer compatibility to a wide range of consumer devices, the specification of advanced variants of the RGB space may require the white point, gamma correction curve, dynamic range and viewing conditions [9]. The most well-known solution is the so-called standardized RGB (sRGB) color space [10] which is commonly used in the today's image acquisition and display devices as well as the Internet, some printing systems, device drivers and operating systems for personal computers. This color space provides an alternative to the current color management strategies through the utilization of a simple and robust device independent color definition, offering good quality and backward compatibility with minimum transmission and system overhead. Given an XYZ color vector whose components range from zero to one and whose reference white is the same as that of the RGB system, the conversion to sRGB values starts as follows:

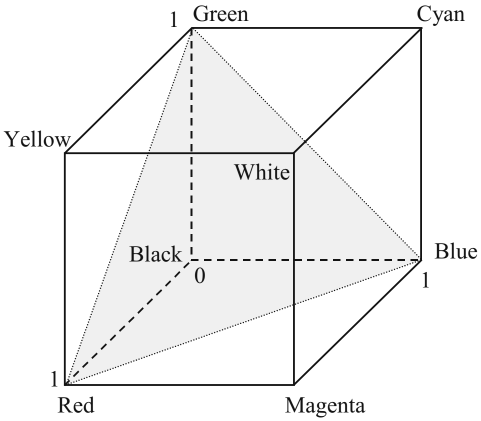


Fig. 5 A red-green-blue (RGB) color space. The gray surface is the so-called Maxwell triangle which denotes the chromaticity plane in the RGB cube.

$$\begin{bmatrix} R \\ G \\ B \end{bmatrix} = \begin{bmatrix} 3.2410 & -1.5374 & -0.4986 \\ -0.9692 & 1.8760 & 0.0416 \\ 0.0556 & -0.2040 & 1.0570 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \tag{5}$$

Since RGB encoding usually does not support values outside the nominal range, both negative values and values exceeding one are clipped to zero and one, respectively. Each of the clipped R, G and B components undergoes gamma correction which is implemented as follows:

$$f(\tau) = \begin{cases} 1.055\tau^{1/2.4} - 0.055 & \text{if } \tau > 0.00304 \\ 12.92\tau & \text{otherwise} \end{cases} \tag{6}$$

where τ denotes the uncorrected color component. Finally, gamma-corrected components $f(R)$, $f(G)$ and $f(B)$ are multiplied by 255 to obtain their corresponding values in standard eight bits per channel encoding.

Since R, G and B components range in the same interval, color vectors defined in an RGB coordinate system can be meaningfully represented by a scalar value $\sqrt{R^2+G^2+B^2}$ denoting their magnitude and a vector $[R,G,B]^T/\sqrt{R^2+G^2+B^2}$ denoting their orientation [11]. These two descriptors are important for human perception as they indirectly indicate luminance and chrominance properties of RGB colors. For instance, vectors with similar orientations have similar chrominance characteristics as they point to the same chromaticity region on a unit sphere of the vector space.

Because of the additive nature of the RGB space, colors are obtained by combining the three primaries through their weighted contributions [7]. A simplified concept of additive color mixing is depicted in Fig. 6. Black and white correspond, respectively, to no contribution and the maximum contributions of the red, green and blue primaries. Any pure secondary color is formed by maximum contributions

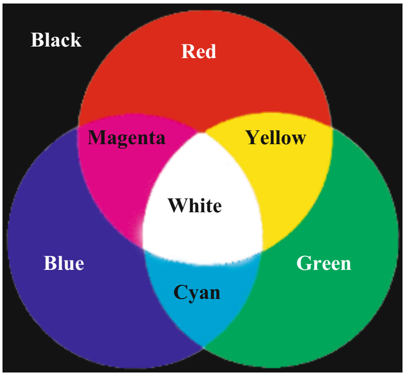
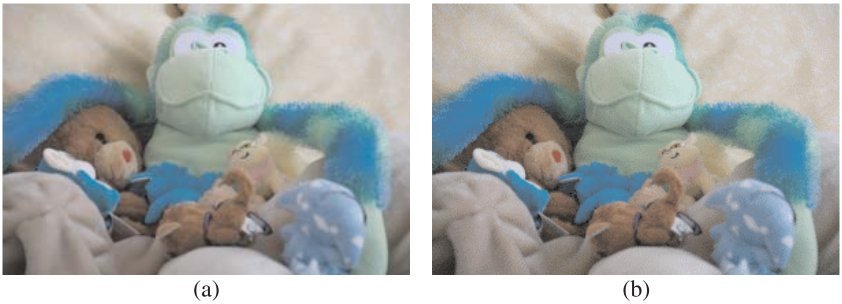


Fig. 6 Additive mixing of colors

of two primary colors. Namely, cyan is obtained as a mixture of green and blue, magenta as a mixture of red and blue, and yellow as a mixture of red and green. When the contribution of one of the primaries is much higher than that of two others, the resulting color is a shade of that dominant primary. When two primaries contribute greatly, the result is a shade of the corresponding secondary color. Equal contributions of all three primaries give a shadow of gray.

Despite varying performance characteristics of different imaging systems, the design of any image acquisition and display device aims to mimic the characteristics of the human visual system. A digital camera comprises one or more image sensors whose photosensitive cells are covered by a color filter to overcome their monochromatic nature and capture the visual scene in color. Spectral sensitivities of such filters usually correspond to that of the S, M, and L cones [4]. There are several camera architectures [12], they differ in use of spectrally selective filters. The so-called three-sensor devices [13], [14] use a beam splitter to separate incoming light onto three optical paths, each of which have its own red, green or blue color filter and sensor for sampling the filtered light. Imaging solutions based on X3 technology [15], [16] are designed based on the fact that the penetration of light in silicon depends on its wavelength. They employ a layered image sensor which directly captures the complete color information at each photosensitive cell through the blue, green and red filters which are stacked vertically and ordered according to their spectral sensitivities. Finally, in the so-called single-sensor devices [17], [18], each photosensitive cell of the image sensor has its own, usually red, green or blue filter to acquire one measurement only. The two missing color components are determinable from the measurements stored in spatially adjacent cells through the color interpolation process often referred to as demosaicking [18], [19].

Display devices, on the other hand, visualize color using three separated light sources [20], arranged as either one or three sources per displaying unit. Using color filters to select desired wavelengths, these light sources correspond, respectively, to the red, green and blue component of a color pixel and their spatial separation becomes indistinguishable for a human eye from a certain viewing distance. From



(a) (b)

Fig. 7 Color-mapped imaging: (a) full-color image with 47488 distinct colors and (b) its color-mapped version with 256 distinct colors

this distance, the three color components are perceived as a solid color due to their additive mixing. A color image is displayed using a rectangular array of such units placed on a screen surface. This concept is adopted in various display systems, such as those based on cathode ray tube [21], liquid crystal [22], plasma [23] and light emitting diode [24] technologies.

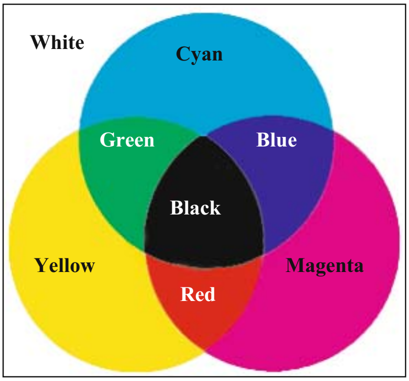
Despite the fact that modern imaging systems use eight bits for displaying a color component and eight or even more bits per color component for processing the color data, color images are sometime converted to their indexed representation [25], [26] in order to allow backward compatibility with older hardware which may be unable to display millions of colors simultaneously. Each pixel location in the indexed image (Fig. 7) has a pointer to a color palette of fixed maximum size, usually 256 colors. Such indexed image representations are efficiently implemented using a look-up table (LUT). In addition to display devices, indexed representations are used for storing purposes, particularly in the Graphics Interchange Format (GIF) [27] and the Portable Network Graphics (PNG) [28] format.

3.2 Printing

Unlike displays which emit light to visualize color information by mixing the three additive primaries, paper or other typical materials used in printing absorb or reflect specific wavelengths. Since white is a natural color of most materials used in printing as a background, as opposed to black which constitutes the background color in most displays and corresponds to the absence of light, printing processes subtract various degrees of red, green and blue from white light to produce desired colors [3], [7]. Such subtractive characteristics are achieved by mixing the cyan (C), magenta (M) and yellow (Y) pigments; this concept is shown in Fig. 8. As can be further seen in Fig. 5, each of these so-called secondary colors complements one primary color, which can be formally expressed as $C = 1 - R$, $M = 1 - G$ and $Y = 1 - B$, [29].

Black is realizable by fully combining the three secondary colors, however, deeper black tones are often produced using a black (K) pigment as the fourth color.

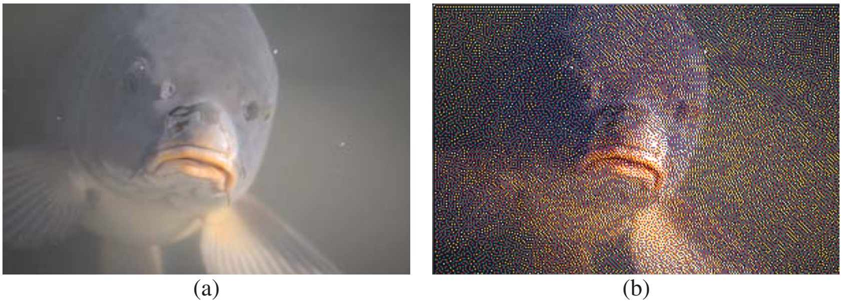
Fig. 8 Subtractive mixing of colors



Typical color printing devices therefore use four colors of ink in accordance to the subtractive CMYK color model. Advanced systems often use six colors, usually with the addition of light cyan and magenta inks to CMYK, in order to expand their gamut and produce higher visual quality of printouts.

The inks are placed on the substrate in layers of dots. Using the so-called halftoning process [29], [30], the image being printed is represented by the density of the dots to simulate the intensity levels and create the illusion of color depth (Fig. 9). Color tones not available in the palette of ink pigments are approximated by a diffusion of base colors. The human eye perceives this diffusion as a mixture or spatial average of the colors within it.

The conversion from RGB to CMYK values is not straightforward because both these are device-dependent spaces. To visually match printed colors with the colors that appear on the display and preserve details and vibrancy of the picture, conversions are usually done through color management using the International Color Consortium (ICC) device profiles [31], [32]. These conversions, however, are approximate due to very different gamuts of RGB and CMYK spaces.


(a)

(b)

Fig. 9 Image representation from a printing perspective: (a) continuous-tone image and (b) its halftone version

3.3 Color Discrimination

A number of image processing and analysis operations are based on some form of color discrimination. An ability of a processing solution to discriminate among colors can be achieved through quantification of the difference or similarity between two color vectors $\mathbf{x}(i) = [x_1(i), x_2(i), x_3(i)]^T$ and $\mathbf{x}(j) = [x_1(j), x_2(j), x_3(j)]^T$ where $x_k(\cdot)$ denotes the kth component of the vector $\mathbf{x}(\cdot)$, for example, $k = 1$ for an R component, $k = 2$ for a G component, and $k = 3$ for a B component. The choice of the measure used for such quantification plays a significant role in color discrimination [33]. This approach is applicable to color vectors expressed in any color space, however, it may not produce completely meaningful results in situations when color space coordinates have significantly different definite ranges.

In the magnitude domain, the difference between $\mathbf{x}(i)$ and $\mathbf{x}(j)$ is commonly measured using the Minkowski metric [11], [34]:

$$d(\mathbf{x}(i), \mathbf{x}(j)) = \|\mathbf{x}(i) - \mathbf{x}(j)\|_L = c \left(\sum_{k=1}^{3} \xi_k |x_k(i) - x_k(j)|^L \right)^{1/L} \tag{7}$$

where the non-negative scaling parameter c is a measure of the overall discrimination power, the exponent L defines the nature of the distance metric, and the weighting coefficient ξ_k, for $\sum_k \xi_k = 1$, denotes the proportion of attention allocated to the k-th component. Depending on the setting of the above parameters [11], [35], the so-called city-block distance

$$\|\mathbf{x}(i) - \mathbf{x}(j)\|_1 = \sum_{k=1}^{3} |x_k(i) - x_k(j)| \tag{8}$$

and the Euclidean distance

$$\|\mathbf{x}(i) - \mathbf{x}(j)\|_2 = \sqrt{\sum_{k=1}^{3} (x_k(i) - x_k(j))^2} \tag{9}$$

can be expressed as special cases of the Minkowski metric. Another special case is the so-called chess-board distance which corresponds to $L \to \infty$ and is determinable as the maximum value of $\{|x_k(i) - x_k(j)|; \text{ for } 1 \leq k \leq 3\}$.

In the directional domain, the difference between $\mathbf{x}(i)$ and $\mathbf{x}(j)$ is commonly measured using the normalized inner product [11], [33]:

$$s(\mathbf{x}(i), \mathbf{x}(j)) = \frac{\mathbf{x}(i)\mathbf{x}(j)^T}{|\mathbf{x}(i)||\mathbf{x}(j)|} \tag{10}$$

which corresponds to the cosine of the angle between $\mathbf{x}(i)$ and $\mathbf{x}(j)$. This equation defines a similarity measure which returns a large value when its inputs are similar and converges to zero if its inputs are dissimilar. Since similar colors have similar orientations in a three-dimensional color space such as the RGB space, the difference in orientation of the two vectors can be correspondingly quantified as follows [11], [33]:

$$\mathrm{A}(\mathbf{x}(i), \mathbf{x}(j)) = \arccos\left(\frac{\mathbf{x}(i)\mathbf{x}(j)^T}{|\mathbf{x}(i)||\mathbf{x}(j)|}\right) \tag{11}$$

where $\mathrm{A}(\cdot, \cdot)$ denotes the angular distance which outputs the angle between color vectors $\mathbf{x}(i)$ and $\mathbf{x}(j)$.

Both magnitude and orientation characteristics of the color vectors can be combined in the design of a generalized measure which can provide a robust solution to the problem of similarity quantification between two vectors. Using the degree of common content $C_{i,j}$ in relation to the total content $T_{i,j}$ of $\mathbf{x}(i)$ and $\mathbf{x}(j)$ constitutes a model of content-based similarity measures [36]:

$$s(\mathbf{x}(i), \mathbf{x}(j)) = \frac{C_{i,j}}{T_{i,j}} \tag{12}$$

Different commonality and totality concepts utilized within this framework can provide different similarity measures, such as

$$s(\mathbf{x}(i), \mathbf{x}(j)) = w_1 \left(\frac{\mathbf{x}(i)\mathbf{x}(j)^T}{|\mathbf{x}(i)||\mathbf{x}(j)|}\right) w_2 \left(1 - \frac{||\mathbf{x}(i)| - |\mathbf{x}(j)||}{\max\left(|\mathbf{x}(i)|, |\mathbf{x}(j)|\right)}\right) \tag{13}$$

where w_1 and w_2 are tunable weights. Other examples of content-based similarity measures can be found in [33], [37].

Color discrimination approaches have been found particularly useful in image denoising, enhancement, data normalization, edge detection and segmentation. Processing solutions used in these application areas usually deal with a population of color vectors, such as those located within a supporting window sliding over the entire image [11]. In order to discriminate among color vectors in the population $\Omega = \{\mathbf{x}(i);\ \text{for}\ i = 1, 2, ..., N\}$, each vector $\mathbf{x}(i)$ can be represented by a scalar value

$$D_i = \sum_{j=1}^{N} d(\mathbf{x}(i), \mathbf{x}(j)) \ \text{ or } \ D_i = \sum_{j=1}^{N} s(\mathbf{x}(i), \mathbf{x}(j)) \tag{14}$$

which corresponds to the aggregated distances or the aggregated similarities. According to the so-called order-statistic concept [38], [39], [40] extended for vector data, the set of all such values can be ranked in order to reveal relations of the color vectors in the data population Ω. The ordered sequence $D_{(1)} \leq D_{(2)} \leq \ldots \leq D_{(n)} \leq \ldots \leq D_{(N)}$ of values D_i, for $i = 1, 2, ..., N$, implies the same ordering of the corresponding vectors $\mathbf{x}(i) \in \Omega$ as follows [33], [37]:

$$\mathbf{x}_{(1)} \leq \mathbf{x}_{(2)} \leq ... \leq \mathbf{x}_{(n)} \leq ... \leq \mathbf{x}_{(N)} \tag{15}$$

where $\mathbf{x}_{(n)}$, for $1 \leq n \leq N$, denotes the so-called nth vector order statistics.

For the aggregated distances used as an ordering criterion, vectors which diverge greatly from the data population usually appear in higher indexed locations in the

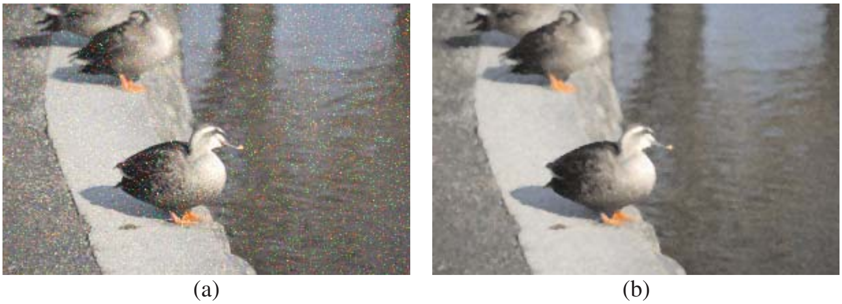


Fig. 10 Noise filtering using the lowest vector order statistics: (a) noisy color image and (b) its filtered version

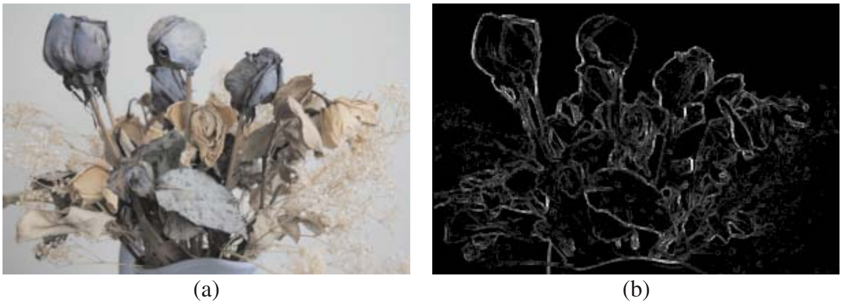
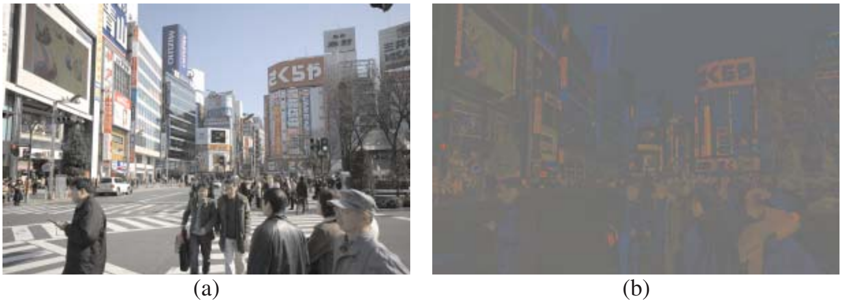


Fig. 11 Edge detection using the highest and lowest vector order statistics: (a) color image and (b) its corresponding edge map

ordered sequence. Therefore, popular noise filtering solutions [11], [37] output the lowest ranked vector in order to produce a robust estimate (Fig. 10). Vectors associated with lower ranks can also be used to perform normalization of a color image since they are most representative of the local image area. Edge detection and segmentation solutions [33], [37] often employ both the lowest and the highest vector order statistics in order to determine object boundaries and localize transitions among color regions (Fig. 11). Note that for the orderings based on the aggregated similarities, lower ranks usually correspond to outlying vectors whereas higher ranks are occupied by the vectors which are most typical for a given data population.

3.4 Image Processing and Analysis

Color images are commonly processed and analyzed as RGB data because the RGB space is the natural space for acquiring, storing and displaying the color image data. Moreover, since operating directly on RGB color vectors requires no conversion from or to the working color space, significant computational savings can be achieved.



(a) (b)

Fig. 12 Image representation using: (a) RGB values and (b) normalized RGB values

Another important aspect is that a natural color image usually exhibits significant correlation among its R, G and B planes, and many algorithms are designed to take advantage of these spectral characteristics. For example [41], such algorithms have found their application in the areas of noise filtering, edge detection, interpolation, demosaicking, feature extraction, pattern recognition and object tracking.

The drawback of the RGB color space is that the image data is quite sensitive to the changes in illumination [42]. To overcome this problem, RGB images are often converted to a normalized RGB space as follows:

$$r = \frac{R}{R+G+B}, \quad g = \frac{G}{R+G+B}, \quad b = \frac{B}{R+G+B} \tag{16}$$

These normalized components denote chromaticity of color vectors and constitute a chromaticity plane termed as the Maxwell triangle (Fig. 5) [43]. Since they are independent of illumination to some extent, the normalized RGB space is quite often used in the design of various pattern recognition and object tracking solutions, such as those for face detection / recognition [42] and surveillance [44] applications. Moreover, since any normalized component is a function of two others, using the normalized RGB space allows to reduce the amount of computations and memory. The latter was found particularly useful in color histogram-based tasks such as image analysis and retrieval, as the determination of a color histogram using the normalized RGB values reduces the dimensionality of the histogram from three to only two dimensions. Finally, as demonstrated in Fig. 12, normalized RGB representations do not correspond to the perceptual meaning of color RGB data. To partially compensate for it, a luminance component is usually used in conjunction with two normalized components.

To obtain some relationship with the human visual system, many hue-oriented color representations were introduced. Examples of color spaces based on such representations include hue saturation value (HSV), hue saturation lightness (HSL), hue saturation intensity (HSI) and hue chroma intensity (HCI) spaces. These spaces, however, have no visual meaning until three-component vectors expressed in these spaces are converted back to the RGB space [7]. The rationale behind a family of hue-oriented color representations is explained on the example of the HSV color space.

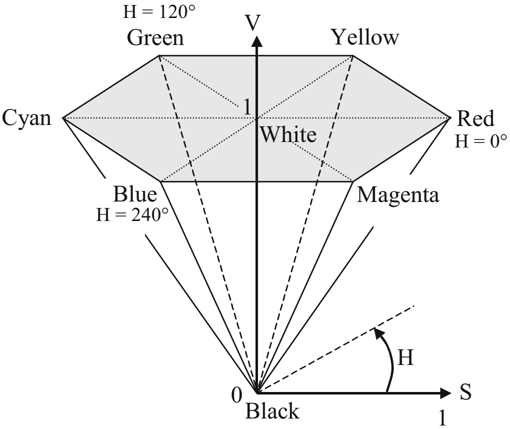


Fig. 13 HSV color space

As shown in Fig. 13, the HSV coordinate system is cylindrical and can be represented by the hexcone model [33], [45]. The hue component (H) characterizes the blend of the three components, thus giving an indication of the spectral composition of a color. It is measured by the angle around the vertical axis, ranging from 0 to 360 degrees, with red corresponding to zero degrees. The saturation component (S) ranges from zero to one and indicates how far a color is from a gray by referring to the proportion of pure light of the dominant wavelength. Finally, the value component (V) defines the relative brightness of the color, with the minimum value of zero corresponding to black and the maximum value of one corresponding to white. The conversion from RGB to HSV values is defined as follows:

$$\begin{aligned} H &= \begin{cases} \eta & \text{if } B \leq G \\ 360 - \eta & \text{otherwise} \end{cases} ; \eta = \arccos\left(\frac{(\Delta_{RG} + \Delta_{RB})/2}{\sqrt{\Delta_{RG}^2 + \Delta_{RB}\Delta_{GB}}}\right) \\ S &= \frac{\max(R,G,B) - \min(R,G,B)}{\max(R,G,B)} \\ V &= \frac{\max(R,G,B)}{255} \end{aligned} \tag{17}$$

where $\Delta_{RG} = R - G$, $\Delta_{RB} = R - B$ and $\Delta_{GB} = G - B$. The H component is undefined and the S component is zero whenever the input RGB vector is a pure shade of gray, which happens for $R = G = B$.

Hue-oriented color spaces were found useful in image segmentation [33] and facial image analysis [42]. An example depicted in Fig. 14 shows the result of human skin segmentation in the HSV space. However, hue-oriented color spaces as well as other color spaces already presented in this chapter are not the only option in the design of image processing, analysis and pattern recognition algorithms for color images. As discussed in next few sections, there exist various luminance-chrominance

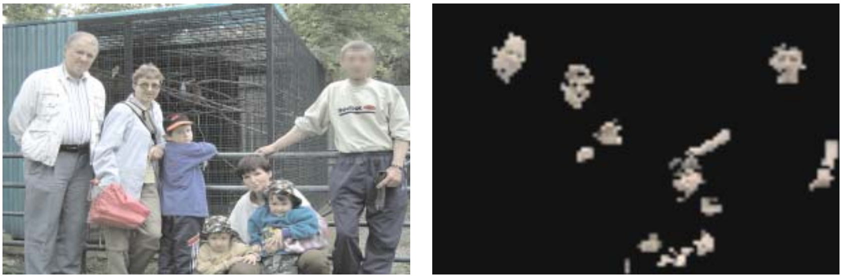

(a) (b)

Fig. 14 Human skin segmentation: (a) original image and (b) segmented image

representations and perceptually uniform spaces which exhibit unique characteristics and can therefore reveal additional information on color and its properties.

3.5 Image Compression and Encoding

A few color transforms are commonly employed for the purpose of image coding. Since an RGB model is defacto a standard color model used in image acquisition and display devices, a number of today's popular file formats, such as the Bitmap (BMP) file format [46], the Graphic Interchange Format (GIF) [47], and the Tagged Image File Format (TIFF) [48], store color images as RGB data. In these formats, each of the three color components is represented by the equal bit depth and each color channel by the equal spatial resolution. Such representations, however, may not be ideal from the efficiency point of view because the human visual system is most sensitive to green and less sensitive to red and blue light.

To account for the above characteristics of human perception, advanced file formats, such as the Joint Photographic Experts Group (JPEG) format [49], [50] for digital images and the Moving Picture Experts Group (MPEG) format [51] for digital video, first convert an RGB color image to some other color space and then encode the image contents by processing those transformed pixel values. Most image and video coding standards use various luminance-chrominance representations of image signals for encoding due to the fact that such representations are analogous to the receptive field encoding at the ganglion cells in the human retina [52]. Luminance refers to the perceived brightness whereas chrominance is specified by hue which characterizes the color tone and saturation which denotes the color pureness.

The two most popular luminance-chrominance representations of image signals are based on YUV and $\mathrm{YC}_b\mathrm{C}_r$ triplets. The YUV representation (Fig. 15) is used in several composite color television systems, such as the National Television System Committee (NTSC) [53] system, the Phase Alternating Line (PAL) system [54] and Séquentiel couleur a mémoire (SECAM) system [55]. A luminance component Y and two chrominance components U and V are obtainable from an RGB triplet as follows:

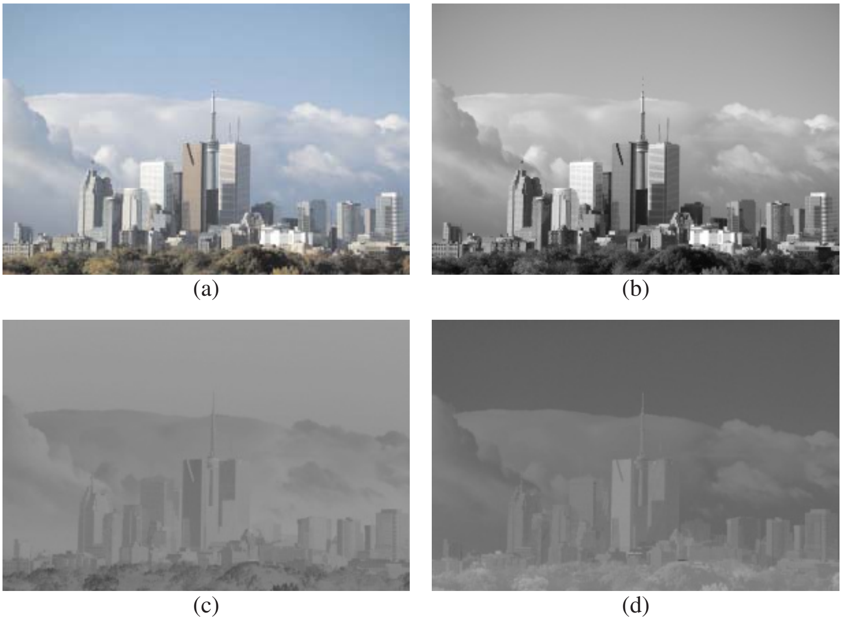

(a) (b)
(c) (d)

Fig. 15 Luminance-chrominance representation of a color image: (a) original image, (b) Y luminance plane, (c) U chrominance plane, and (d) V chrominance plane

$$\begin{bmatrix} Y \\ U \\ V \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ -0.147 & -0.289 & 0.436 \\ 0.615 & -0.515 & -0.100 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{18}$$

Alternatively, U and V can be expressed using Y as $U = 0.492(B - Y)$ and $V = 0.877(R - Y)$.

The YC$_b$C$_r$ representation is used in digital image and video coding. The conversion formula from RGB to YC$_b$C$_r$ values is defined as follows:

$$\begin{bmatrix} Y \\ C_b \\ C_r \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ -0.169 & -0.331 & 0.500 \\ 0.500 & -0.419 & -0.081 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{19}$$

where chrominance components C_b and C_r can be equivalently expressed using Y as $C_b = 0.564(B - Y)$ and $C_r = 0.713(R - Y)$.

Since both U and C_b represent a scaled version of the difference between the blue signal and the luminance signal, they also indicate the extent to which the color deviates from gray toward blue. In a similar way, both V and C_r indicate the extent to which the color deviates from gray toward red, as they can be expressed as a scaled version of the difference between the red signal and the luminance signal. It should be also noted that the conversion from RGB to the above luminance-chrominance representations often involves gamma correction in order to approximate a nonlinear response of display systems to the input voltage signal [56].

The amount of data required to represent a digital image in any color space can be reduced by image compression [57]. Achieving simultaneously high compression rates and good visual quality of compressed images has been a driving factor behind many progresses in that field. Better understanding of the human visual system, particularly its tolerance to a modest amount of loss, has allowed the design of efficient compression procedures. Another important observation used in encoding is related to the luminance-chrominance representation of image signals; namely, that the human visual system is less sensitive to high frequencies in chrominance than in luminance [52], [56]. Therefore, the spatial resolution of chrominance planes can be reduced compared to that of luminance without observing any visual degradation in the image. Typically, chrominance planes are subsampled by a factor ranging from two to four which allows allocating reduced bit rates for chrominance information coding while maintaining the desired visual quality.

3.6 Quantitative Manipulation / Quality Evaluation

The CIE-XYZ color space is perceptually highly nonuniform, meaning that equal Euclidean distances or mean square errors expressed in this space do not equate to equal perceptual differences. It is therefore inappropriate for quantitative manipulations involving color perception [58]. Despite its rare use in image processing applications, the CIE-XYZ color space has a significant role in image processing since other color spaces can be derived from it through mathematical transforms [33]. For example, more perceptually uniform representations, such as CIE u, v, CIELuv and CIELab, are derived from XYZ values. Namely, CIE u and v values are obtainable as follows:

$$u = \frac{4X}{X + 15Y + 3Z}, \quad v = \frac{9Y}{X + 15Y + 3Z} \tag{20}$$

and can be used to form a chromaticity diagram which corresponds better to the characteristics of human perception than the CIE xy chromaticity diagram. It can therefore serve as a good indicator of the gamut limitations of output devices, that is, limitations in the range of colors that are physically obtainable by an output device [59].

The transformations from CIE-XYZ to CIELuv and CIELab color spaces [8] use a reference point — expressed as the X_n, Y_n, Z_n CIE tristimulus values of the reference white under the reference illumination — in order to account for adaptive characteristics of the human visual system [7]. The CIELab transformation is defined by

$$\begin{aligned} L^* &= 116 f(Y/Y_n) - 16 \\ a^* &= 500\left(f(X/X_n) - f(Y/Y_n)\right) \\ b^* &= 200\left(f(Y/Y_n) - f(Z/Z_n)\right) \end{aligned} \tag{21}$$

whereas the CIELuv values can be obtained via

$$\begin{aligned} L^* &= 116 f(Y/Y_n) - 16 \\ u^* &= 13L^*\,(u - u_n) \\ v^* &= 13L^*\,(v - v_n) \end{aligned} \tag{22}$$

The terms $u_n = 4X_n/(X_n + 15Y_n + 3Z_n)$ and $v_n = 9Y_n/(X_n + 15Y_n + 3Z_n)$ correspond to the reference point $[X_n, Y_n, Z_n]^T$ while u and v are determinable via Eq. (20) and correspond to the color vector $[X, Y, Z]^T$ to be mapped to these perceptually uniform color spaces. In an attempt to account for different characteristics of the human visual system in normal illumination and low light levels, $f(\cdot)$ is defined as follows [8]:

$$f(\gamma) = \begin{cases} \gamma^{1/3} & \text{if } \gamma > 0.008856 \\ 7.787\gamma + 16/116 & \text{otherwise} \end{cases} \tag{23}$$

Both CIELuv and CIELab belong to a family of opponent color spaces [60]. As depicted in Fig. 16, u^* and a^* coordinates represent the difference between red and green whereas v^* and b^* coordinates represent the difference between yellow and blue. This arrangement is based on the assumption that a color cannot be both red and green, or blue and yellow. A zero or close to zero value of these components gives a neutral or near neutral color. The term L^* represents the lightness of a color vector, and it ranges from 0 to 100 which correspond to black and white, respectively. The CIELab model is used in Adobe PostScript (level 2 and level 3) and for color management as the device independent model of the ICC device profiles [32]. The CIELuv model assists in the registration of color differences experienced with lighting and displays [59].

In both these color spaces, perceptual differences between two colors $[X_1, Y_1, Z_1]^T$ and $[X_2, Y_2, Z_2]^T$ are measurable using the Euclidean distance as follows [7]:

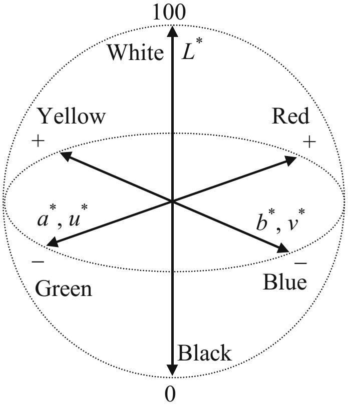


Fig. 16 CIELab and CIELuv color spaces

$$\Delta E^*_{uv} = \sqrt{(L^*_1 - L^*_2)^2 + (u^*_1 - u^*_2)^2 + (v^*_1 - v^*_2)^2} \tag{24}$$

$$\Delta E^*_{ab} = \sqrt{(L^*_1 - L^*_2)^2 + (a^*_1 - a^*_2)^2 + (b^*_1 - b^*_2)^2} \tag{25}$$

In psychovisual experiments, a value of ΔE equal to unity represents a just noticeable difference (JND) in either of these two color models [61]. In pictorial scenes, JND values are much higher in order to account for the complexity of visual information.

The differences in the CIELab space can be equivalently expressed as follows [7], [59]:

$$\Delta E^*_{ab} = \sqrt{(\Delta L^*_{ab})^2 + (\Delta H^*_{ab})^2 + (\Delta C^*_{ab})^2} \tag{26}$$

where $\Delta L^*_{ab} = L^*_1 - L^*_2$ denotes the difference in lightness, ΔH^*_{ab} defined by

$$\Delta H^*_{ab} = \sqrt{(a^*_1 - a^*_2)^2 + (b^*_1 - b^*_2)^2 - (\Delta C^*_{ab})^2} \tag{27}$$

denotes a measure of hue difference, and ΔC^*_{ab} defined by

$$\Delta C^*_{ab} = C^*_1 - C^*_2 = \sqrt{(a^*_1)^2 + (b^*_1)^2} - \sqrt{(a^*_2)^2 + (b^*_2)^2} \tag{28}$$

denotes the difference in chroma. For the sake of completeness note that the hue angle is defined as $h^*_{ab} = \arctan(b/a)$.

An advanced formulation of the ΔE measure weights the hue and chroma components by a function of chroma [62]:

$$\Delta E^*_{94} = \sqrt{\left(\frac{\Delta L^*_{ab}}{k_L S_L}\right)^2 + \left(\frac{\Delta H^*_{ab}}{k_H S_H}\right)^2 + \left(\frac{\Delta C^*_{ab}}{k_C S_C}\right)^2} \tag{29}$$

where $S_L = 1$, $S_H = 1+0.015\sqrt{C^*_1 C^*_2}$, $S_C = 1+0.045\sqrt{C^*_1 C^*_2}$, and $k_L = k_H = k_C = 1$ for reference conditions. A newer, more sophisticated version of the ΔE measure can be found in [63], [64]. Other formulations are known under the names S-CIELab [65] which employs spatial filtering to provide an improved measure for pictorial scenes and ST-CIELab [66] which accounts for the spatiotemporal nature of digital video.

4 Conclusion

This chapter surveyed popular methods for representing and using color in digital imaging and multimedia applications. Most of these methods specify color by three numerical components, which is similar to collecting the responses from three types of color-receptive cells by the human visual system. Such three-component representations have an important role in acquisition, processing, storage and inspection of color data. They can be predefined in imaging device hardware by its manufacturer or chosen as a working space by the algorithm designer or the image editing software user.

Well-designed color representations allow effective color manipulation in various image processing, analysis and quality evaluation tasks. Given the vectorial nature of color data, a number of measures can be used to discriminate colors and judge their similarities by taking into account the length or the orientation of color vectors or operating on both the magnitude and directional domains simultaneously. The existence of numerous color representations with varying characteristics suggests that these methods constitute an indispensable tool for modern digital imaging and multimedia systems which use color as a cue for better understanding of visual information by both humans and computing machines.

References

[1] Gonzalez, R., Woods, R.E.: Digital image processing, 3rd edn. Prentice Hall, Reading (2007)

[2] Susstrunk, S.: Image formation. In: Peres, M.R. (ed.) Focal encyclopedia of photography, 4th edn., pp. 382–388. Focal Press / Elsevier, Burlington / Massachusetts (2007)

[3] Sharma, G.: Color fundamentals for digital imaging. In: Sharma, G. (ed.) Digital color imaging handbook, pp. 1–113. CRC Press / Taylor & Francis, Boca Raton / Florida (2002)

[4] Lam, E.Y., Fung, G.S.K.: Automatic white balancing in digital photography. In: Lukac, R. (ed.) Single-sensor imaging: Methods and applications for digital cameras, pp. 267–294. CRC Press / Taylor & Francis, Boca Raton / Florida (2008)

[5] Giorgianni, E., Madden, T.: Digital color management. Addison Wesley, Reading (1998)

[6] Wyszecki, G., Stiles, W.S.: Color science, concepts and methods, quantitative data and formulas, 2nd edn. John Wiley, New York (1982)

[7] Trussell, H.J., Saber, E., Vrhel, M.: Color image processing. IEEE Signal Processing Magazine 22, 14–22 (2005)

[8] Hunt, R.W.G.: Measuring colour, 3rd edn. Fountain Press (1998)

[9] Susstrunk, S., Buckley, R., Swen, S.: Standard RGB color spaces. In: Proceedings of The Seventh Color Imaging Conference: Color Science, Systems, and Applications, Scottsdale, Arizona, pp. 127–134 (1999)

[10] Stokes, M., Anderson, M., Chandrasekar, S., Motta, R.: A standard default color space for the internet - sRGB, Technical report (1996), http://www.w3.org/Graphics/Color/sRGB.html

[11] Lukac, R., Smolka, B., Martin, K., Plataniotis, K.N., Venetsanopulos, A.N.: Vector filtering for color imaging. IEEE Signal Processing Magazine 22, 74–86 (2005)

[12] Lukac, R., Plataniotis, K.N.: Single-sensor camera image processing. In: Lukac, R., Plataniotis, K.N. (eds.) Color image processing: Methods and applications, pp. 363–392. CRC Press / Taylor & Francis, Boca Raton / Florida (2006)

[13] Sharma, G., Trussell, H.J.: Digital color imaging. IEEE Transactions on Image Processing 6, 901–932 (1997)

[14] Adams, J., Parulski, K., Spaulding, K.: Color processing in digital cameras. IEEE Micro 18, 20–30 (1998)

[15] Lyon, R.F., Hubel, P.M.: Eying the camera: Into the next century. In: Proceedings of the IS&TSID Tenth Color Imaging Conference, Scottsdale, Arizona, pp. 349–355 (2002)

[16] Hubel, P.M., Liu, J., Guttosh, R.J.: Spatial frequency response of color image sensors: Bayer color filters and Foveon X3, Technical Report ID 6502 San Antonio, USA (2002)

[17] Parulski, K., Spaulding, K.E.: Color image processing for digital cameras. In: Sharma, G. (ed.) Digital color imaging handbook, pp. 728–757. CRC Press / Taylor & Francis, Boca Raton / Florida (2002)

[18] Lukac, R.: Single-sensor imaging: Methods and applications for digital cameras. CRC Press / Taylor & Francis, Boca Raton / FLorida (2008)

[19] Gunturk, B.K., Glotzbach, J., Altunbasak, Y., Schaffer, R.W., Murserau, R.M.: Demosaicking: Color filter array interpolation. IEEE Signal Processing Magazine 22, 44–54 (2005)

[20] Vrhel, M., Saber, E., Trussell, H.J.: Color image generation and display technologies. IEEE Signal Processing Magazine 22, 23–33 (2005)

[21] Sharma, G.: LCDs versus CRTs-color-calibration and gamut considerations. Proceedings of the IEEE 90, 605–622 (2002)

[22] Kawamoto, H.: The history of liquid-crystal displays. Proceedings of the IEEE 90, 460–500 (2002)

[23] Uchiike, H., Hirakawa, T.: Color plasma displays. Proceedings of the IEEE 90, 533–539 (2002)

[24] Forrest, S., Burrows, P., Thompson, M.: The dawn of organic electronics. IEEE Spectrum 37, 29–34 (2000)

[25] Battiato, S., Lukac, R.: Color-mapped imaging. In: Furth, B. (ed.) Encyclopedia of multimedia, 2nd edn., pp. 83–88. Springer, New York (2008)

[26] Marques, O.: Image data representations. In: Furth, B. (ed.) Encyclopedia of multimedia, 2nd edn., pp. 323–328. Springer, New Work (2008)

[27] Miano, J.: Compressed image file formats. ACM Press / Addison-Wesley Professional (1999)

[28] Roleof, G.: PNG: The definitive guide. 2nd edn. (2003), `http://www.libpng.org/pub/png/book/`

[29] Monga, V., Damera-Venkata, N., Evans, B.L.: Color image halftoning. In: Lukac, R., Plataniotis, K.N. (eds.) Color image processing: Methods and applications, pp. 157–183. CRC Press / Taylor & Francis, Boca Raton / Florida (2006)

[30] Hains, C., Wang, S.G., Knox, K.: Digital color halftones. In: Sharma, G. (ed.) Digital color imaging handbook, pp. 385–490. CRC Press / Taylor & Francis, Boca Raton / Florida (2002)

[31] Giorgianni, E.J., Madden, T.E., Spaulding, K.E.: Color management for digital imaging systems. In: Sharma, G. (ed.) Digital color imaging handbook, pp. 239–268. CRC Press / Taylor & Francis, Boca Raton / Florida (2002)

[32] Sharma, A.: ICC color management: Architecture and implementation. In: Lukac, R., Plataniotis, K.N. (eds.) Color image processing: Methods and applications, pp. 1–27. CRC Press / Taylor & Francis, Boca Raton / Florida (2006)

[33] Plataniotis, K.N., Venetsanopoulos, A.N.: Color image processing and applications. Springer, Berlin (2000)

[34] Nosovsky, R.M.: Choice, similarity and the context theory of classification. Journal of Experimental Psychology Learning, Memory Cognition 10, 104–114 (1984)

[35] Duda, R.O., Hart, P.E., Stork, D.G.: Pattern classification and scene analysis, 2nd edn. John Wiley, Danvers (2000)

[36] Plataniotis, K.N., Androutsos, D., Venetsanopoulos, A.N.: Adaptive fuzzy systems for multichannel signal processing. Proceedings of the IEEE 87, 1601–1622 (2000)

[37] Lukac, R., Plataniotis, K.N.: A taxonomy of color image filtering and enhancement solutions. In: Hawkes, P.W. (ed.) Advances in imaging and electron physics, vol. 140, pp. 187–264. Elsevier / Academic Press, San Diego / California (2006)

[38] Barnett, V.: The ordering of multivariate data. Journal of Royal Statistical Society A 139, 318–354 (1976)

[39] Hardie, R.C., Arce, G.R.: Ranking in RP and its use in multivariate image estimation. IEEE Transactions on Circuits and Systems for Video Technology 1, 197–208 (1991)

[40] Pitas, I., Venetsanopoulos, A.N.: Order statistics in digital image processing. Proceedings of the IEEE 80, 1892–1919 (1992)

[41] Lukac, R., Plataniotis, K.N.: Color image processing: Methods and applications. CRC Press / Taylor & Francis, Boca Raton / Florida (2006)

[42] Martinkauppi, J.B., Pietikainen, M.: Facial skin color modeling. In: Jain, A.K., Li, S.Z. (eds.) Handbook of Face Recognition, pp. 113–136. Springer, New York (2005)

[43] Gomes, J., Velho, L.: Image processing for computer graphics. Springer, New York (1997)

[44] Foresti, G.L., Mahonen, P., Regazzoni, C.S.: Multimedia video-based surveillance systems: Requirements, issues, and solutions. Kluwer / Springer (2000)

[45] Guan, L., Kung, S.Y., Larsen, J.: Multimedia image and video processing. CRC Press, Boca Raton (2001)

[46] Graphics file dormats, `http://www.digicamsoft.com/bmp/bmp.html`

[47] Graphic interchange format. CompuServe Inc., `http://www.w3.org/Graphics/GIF/spec-gif89a.txt`

[48] Tagged image file format, Adobe Systems, `http://partners.adobe.com/public/developer/tiff/index.html`

[49] ISO/IEC, Information technology - digital compression and coding of continuous-tone still images: Requirements and guidelines. International Standard 10918-1, ITU-T Recommendation T.81 (1994)

[50] ISO/IEC, Information technology - JPEG, image coding system. International Standard 15444-1, ITU Recommendation T.800 (2000)

[51] ISO/IEC, Information technology - coding of audio-visual object - part 2: Visual. International Standard 14496-2, MPEG-4 (1999)

[52] Alleysson, D., de Lavarene, B.C., Susstrunk, S., Herault, J.: Linear minimum mean square error demosaicking. In: Lukac, R. (ed.) Single-sensor imaging: Methods and applications for digital cameras, pp. 213–237. CRC Press / Taylor & Francis, Boca Raton / Florida (2008)

[53] Recommendation ITU-R BT.470-7 Conventional analog television systems. International Telecommunication Union (1998)

[54] Recommendation ITU-R BT.470-6 Conventional television systems. International Telecommunications Union (1998)

[55] World analogue television standards and waveforms, `http://www.pembers.freeserve.co.uk/World-TV-Standards/Colour-Standards.html#SECAM-IV`

[56] Argyropoulos, S., Boulgouris, N.V., Thomos, N., Kompatsiaris, Y., Strintzis, M.G.: Coding of two-dimensional and three-dimensional color image sequences. In: Lukac, R., Plataniotis, K.N. (eds.) Color image processing: Methods and applications, pp. 503–523. CRC Press / Taylor & Francis, Boca Raton / Florida (2006)

[57] Wang, Y., Ostermann, J., Zhang, Y.: Video processing and communications. Prentice Hall, Upper Saddle River (2002)

[58] Poynton, C.A.: A technical introduction to digital video. Prentice Hall, Toronto (1996)

[59] Susstrunk, S.: Colorimetry. In: Peres, M.R. (ed.) Focal encyclopedia of photography, 4th edn., pp. 388–393. Focal Press / Elsevier, Burlington / Massachusetts (2007)

[60] Kuehni, R.G.: Color space and its divisions: Color order from antiquity to the present. Wiley-Interscience, Hoboken (2003)

[61] Rogers, D.F., Earnshaw, R.E.: Computer graphics techniques: Theory and practice. Springer, New York (2001)

[62] CIE publication No 116, Industrial colour difference evaluation. Central Bureau of the CIE (1995)

[63] Luo, M.R., Cui, G., Rigg, B.: The development of the CIE 2000 colourdifference formula: CIEDE 2000. Color Research and Applications 26(5), 340–350 (2001)

[64] Luo, M.R., Cui, G., Rigg, B.: Further comments on CIEDE 2000. Color Research and Applications 27, 127–128 (2002)

[65] Wandell, B.: S-CIELAB: A spatial extension of the CIE L*a*b* DeltaE color difference metric (1998), `http://white.stanford.edu/~brian/scielab/`

[66] Tong, X., Heeger, D.J., van den Branden, L., Christian, J.: Video quality evaluation using ST-CIELAB. In: Proceedings of SPIE Human Vision and Electronic Imaging IV, vol. 3644, pp. 185–196 (1999)

Advances in Video Summarization and Skimming

Richard M. Jiang, Abdul H. Sadka, and Danny Crookes

Abstract. This chapter summarizes recent advances in video abstraction for fast content browsing, skimming, transmission, and retrieval of massive video database which are demanded in many system applications, such as web multimedia, mobile multimedia, interactive TV, and emerging 3D TV. Video summarization and skimming aims to provide an abstract of a long video for shortening the navigation and browsing the original video. The challenge of video summarization is to effectively extract certain content of the video while preserving essential messages of the original video. In this chapter, the preliminary on video temporal structure analysis is introduced, various video summarization schemes, such as using low-level features, motion descriptors and Eigen-features, are described, and case studies on two practical summarization schemes are presented with experimental results.

1 Introduction

With the explosion of multimedia database due to the widespread internet and wireless multimedia technology, the management of vast video contents demands an automatic summarization to abstract the most concerned contents or useful information from the massive visual data set [1-3]. Recent advances [1-3] in this area have successfully generated a number of commercialized products, such as VideoCollage developed by Microsoft [4] and VideoSue by IBM [5].

Video summarization refers to creating an excerpt of a digital video, which must satisfy the following three principles:

- The video summary must contain high priority entities and events from the video;
- The summary should be free of redundancy;
- The summary itself should display rational degrees of continuity.

The challenge of video summarization is to effectively extract primary contents of the video while preserving its original story [1-3].

Richard M. Jiang
Computer Science, Loughborough University, Loughborough, UK
e-mail: M.Jiang@lboro.ac.uk

Abdul H. Sadka
Electronics & Computer Engineering, Brunel University, West London, UK

Danny Crookes
Computer Science, Queen's University Belfast, Belfast, UK

M. Grgic et al. (Eds.): Rec. Advan. in Mult. Sig. Process. and Commun., SCI 231, pp. 27–50.
springerlink.com

Video abstraction is a technique that abstracts video content into the representation of a compact manner. There are basically two types of video abstraction: static video summarization and dynamic video skimming. Static video summarization is a process that selects a set of salient images called key frames to represent the video content, while dynamic video skimming represents the original video in the form of a short video clip. Video abstraction forms a key ingredient in a practical video content management system, as the generated key frames and skims provide users an efficient way to browse or search video content. With the proliferation of digital videos, this process will become an indispensable component to any practical content management system. A video summary can be played without the worry of timing issues. Moreover, the extracted key frames could be used for content indexing and retrieval.

Usually, from the viewpoint of users, a video skim may provide a more smart choice since it contains both audio and motion information that makes the abstraction more natural, interesting and informative for story narration, while static video summarization may provide a glance of video contents in a more concise way. Both ways are actually similar to each other, while temporal structure analysis is the shared basis in their technical implementation, as shown in Figure 1.

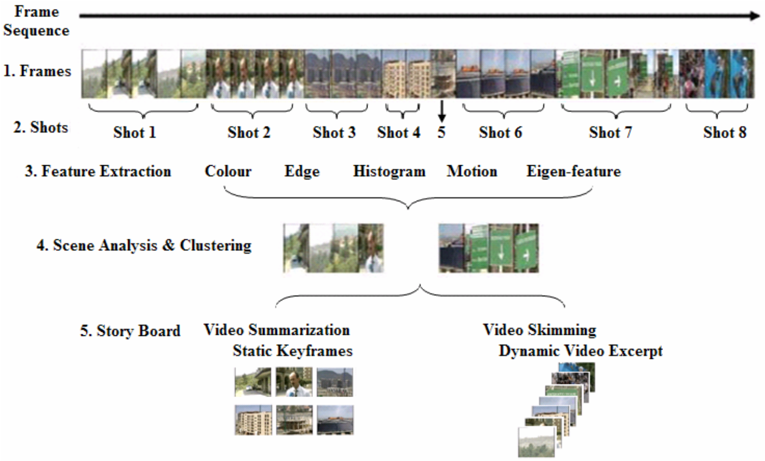


Fig. 1 Static video summarization and dynamic video skimming

1.1 Video Summarization

Based on how a key frame is extracted for video summarization, existing work in this area can be divided into three categories: sampling based, shot based, and segment based. Most of the earlier summarization work belongs to the first class. In these work, key frames were uniformly sampled from the original video, which demonstrate the simplest way to extract key frames. The MiniVideo [6] and the video magnifier [7] systems are such two examples. However, such an arrangement may be difficult to capture the meaningful video content, especially when the scene changes frequently.

By adapting to dynamic video content to extract key frames, more sophisticated work has then been developed. In view of the fact that a shot is defined as a video sequence taken from a continuous period, a simple and straightforward way for video summarization is to take one or several key frames from each shot using low-level features, such as color and motion. A typical approach provided in [8] uses a sequential manner via thresholding to extracts key frames. Other schemes may include the use of color clustering, texture analysis, shape classification, global motion, or gesture analysis [9-11].

While being aware of the impossibility of using simple key frames to effectively represent the underlying dynamic contents in videos, scientists have searched for an alternative method using a synthesized panoramic image, namely the *mosai*c, to represent the shot content. Along this path, various styles of mosaics such as static background mosaics and synopsis mosaics have been proposed in [12, 13]. A combined use of mosaic images and key frames has also been investigated in [14].

Various mathematical models applied to the summarization procedure have also been investigated. For example, the video sequence can be represented by the symbol of a feature curve in a high-dimensional feature space with key frames corresponding to the points on the curve [15].

For above-segment scene-level summarization, a number of clustering-based schemes to extract representative frames at the higher scene-level have also been reported. In these schemes, segments are first generated from frame clustering, and the frames that are closest to the centroid of each qualified segment are chosen as key frames [16, 17]. In the work reported in [18] on video summarization at the scene level, the shot structure of a given video is detected, all shots are classified into a group of clusters using a time-constrained clustering algorithm, and meaningful story units, such as dialogues and actions, are extracted with selected representative images for each story unit to represent its component shot clusters.

Other schemes based on techniques such as hierarchical frame clustering [19], fuzzy classification [20], sophisticated temporal frame sampling [21], singular value decomposition, and principle component analysis have also been tried with promising results.

1.2 Video Skimming

Video skimming can usually be partitioned into two classes: summary oriented and highlight oriented. A summary-oriented skim provides users a summarized version [22] while keeping the mainframe structure of the original video. On the other way, only a few interesting parts of the original video can be selected in the highlight-oriented skim, which is typical in movie trailers and sports highlights [23].

How to map human perception into an automated abstraction process is a challenging topic. It is a subjective and difficult procedure to define which video segments to be highlighted. Hence, most current video skimming work is summary-oriented.

One simple approach is to compress the original video by speeding up frame rate using time compression technique, as reported in [24]. Similarly Amir et al. [25] applied an audio time-scale modification scheme. However, these techniques suffered from the limitation in the maximum time compression ratio of 1.5–2.5. Once the compression factor goes above this range, the total quality degrades quickly.

Specific events such as the presence of camera motion and human faces have also attracted researcher's attentions. Nam and Tewfik [26] generated skims by using a dynamic sampling scheme, where subshots segmented from the video were associated with motion intensity indices that were quantized into bins of predefined sampling rates, and key frames were sampled from each sub-shot based on the assigned rate to provide users a narrative storyboard for motion contents and events. Similar techniques are also presented in [27, 28].

It has also researched recently to use some special features to generate skims for domain-specific video data. A skimming system for news videos was presented in [29] by detecting the specific contents, such as commercials and "anchor persons". It filtered out commercials using audio cues, and then detected anchor persons using Gaussian mixture models. Following this, since a news story is usually led and summarized by an anchor person, the skim can be generated by gluing all video parts that contain anchor persons.

The VidSum project [30] applied a presentation structure to map low-level signal events to semantically meaningful events to produce skims of their regular weekly forum. Some research efforts on generating skims for sports videos [42] have also been reported by using audio-visual cues to identify exciting highlights such as football touchdowns and soccer goals.

Figure 2 exhibits the diagram of a three-layer example system for video skimming. In this system, low-level features are seperated and temporal video segmentation is performed at the first layer. At the second layer, mid- to high-level semantic features are derived, which can be achieved by using techniques such as face detection, audio classification, video text recognition, and semantic event detection. The third layer assembles these summaries into the final abstract according to user specifications.

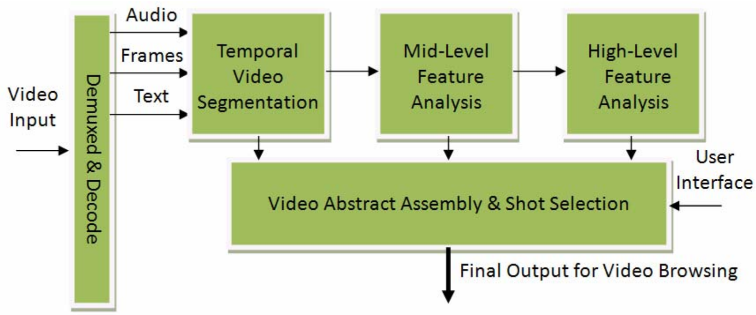

Video Input
Demuxed & Decode
Audio
Frames
Text
Temporal Video Segmentation
Mid-Level Feature Analysis
High-Level Feature Analysis
User Interface
Video Abstract Assembly & Shot Selection
Final Output for Video Browsing

Fig. 2 A three layer video abstraction system

2 Video Temporal Structure Analysis

To provide video summarization and skimming, the primary task is to perform proper video structure analysis to find out contents or frames of user's specific interest for video browsing. Since scene changes almost always happen on a shot change, temporal video segmentation or shot boundary detection is a useful step for video summarization and skimming, while temporal structure analysis also provides convenient jump points for video browsing. The abstract of the video can then be the extraction of a set of either independent frames or short sequences from a video. This can be used as a substitute for the whole video for the purposes of indexing, comparison, and categorization. It is also especially useful for video browsing [3].

The results of video temporal structure analysis and condensed video representation do not need to be immediately mapped to the above applications; they may alternatively be stored as XML metadata and used when they are demanded. This can be attained by using standards such as MPEG-7 [31], which provide appropriate specifications to catch up the concepts of shots, scenes, and various types of video abstraction.

2.1 Preliminary on Temporal Video Segmentation

In temporal video segmentation, a video is completely segmented into a series of disjointed scenes, which are subsequently segmented into a sequence of shots. The concept of temporal video segmentation [32] is not new, as it dates back to the early days of motion pictures, well before the emerging of computers, when motion picture specialists always segmented their works into hierarchical segments due to the heavy weights of film rolls.

The terminology "scene", originating in the theater, is even much older than motion pictures. A scene refers to a specific spectacle that is spatially cohesive, which may be revisited several times in a video. On the other hand, shots are defined as the temporally consistent sequence that originates from the same camera take, where the camera images in a continuous run.

While the automatic summarization of a video into scenes ranges from very difficult to intractable, video segmentation into shots has been exactly defined and quantified by observable features of the video stream, because video content within a shot is usually continuous as a natural result of the continuity of both the physical scene and the camera parameters (such as zoom, focus, motion). As a result, in principle, the detection of a shot change between two adjacent frames simply requires the computation of an appropriate similarity metric. Nevertheless, this straightforward thought has three major barriers.

The first problem as the most obvious one is how to define a metric to measure the continuity for the video in such a way that it is discriminant enough to find meaningful scene changes but robust to gradual changes in term of camera parameters, lighting, and physical scene content. The simplest way to do so is to extract one or several visual or audio features from each frame and define dissimilarity functions in the feature space.

The second difficulty is how to decide the range of values of the continuity metric that corresponds to a shot change. This is not easy, since the features within certain shots can vary dramatically in some shots while keep small variation in other shots. A number of decision methods for temporal video segmentation have been researched, such as adaptive thresholds and statistical detection methods.

The third and most difficult complication is the fact that not all shot transitions are so abrupt to be easily detected. Shot changes can usually be categorized into the following classes:

- Cut: This is the typical case of abrupt changes, where two adjacent frames apparently belong to two different consecutive shots, respectively.
- Wipe: In this case, the appearing and disappearing shots coexist in different spatial regions of the intermediate video frames, and the region occupied by the latter grows while the regions of the former diminishes.
- Dissolve: This means that the scene of the disappearing shot fades out into the scene of the appearing shot through a number of overlapped frames, where the intensity of each pixel slips into the next from the previous scene.
- Fade: This case has two steps: In first step, the disappearing shot fades out into a blank frame; in second step, the blank frame fades into the appearing shot.

The challenges in detecting all above shot transitions have been intensively researched in the competition of various proposed schemes in TRECVID benchmark test. Figure 3 gives several examples of such elusive transitions.

Obviously, any temporal segmentation scheme has to address above three complications, which leads to three aspects: feature selection, metric threshold, and transition awareness, which bring challenges to the design of various algorithms for temporal video segmentation.

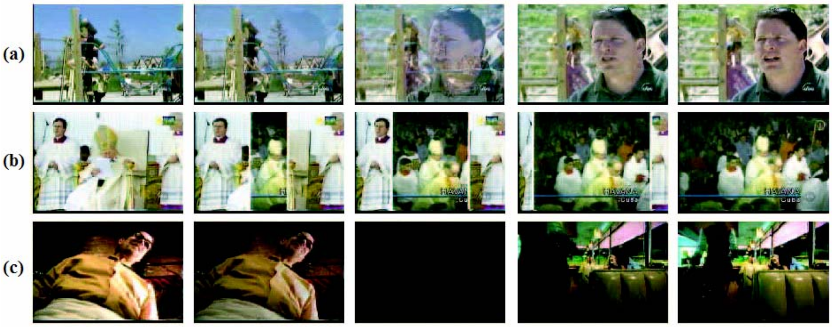


Fig. 3 Examples of gradual transitions from the TRECVID 2003 Corpus: (a) Dissolve; (b) Fade; (c) Wipe

2.2 Algorithms for Temporal Video Segmentation

As mentioned above, temporal video segmentation works by extracting one or more features from a video frame. An algorithm can then use different methods to detect shot structure from these features. Since there are many different ways the above components can be combined, in order to provide a more insight view of different classes of algorithms, we present below the different choices that can be made for each component, along with their advantages and disadvantages. These can then be combined more or less freely to design different temporal video segmentation algorithms, as we will demonstrate in section 3 and 4.

1) Feature Selection for Measurable Metrics

Almost all temporal video segmentation algorithms reduce the computational complexity of processing vast pixel-level video signals by extracting a small number of features in each video frame [33]. Such features include:

- Luminance/Color: The simplest feature that can be used to characterize a frame is its average color or luminance. Usually hue saturation value (HSV) is considered as a more robust choice to provide a suitable color space [34-35].
- Histogram of Luminance/color: The grayscale or color histogram is a richer feature to be able to represent a frame, which has the advantages as discriminant and easy to compute [36]. By combining with spatial distribution, spatial-colour histogram may be a more precise one.
- Motion: Motion is a primary cue for detecting video events. It is usually coupled with other features for temporal video segmentation, since by itself it can be highly discontinuous within a shot.
- Image edges: Edge information [37] is another choice for characterizing a frame in video sequence. The advantage of this feature is that it is related to the human visual perception of a scene. Its disadvantages may include its computational cost and noise sensitivity.
- Fourier/Cosine/Wavelet Transform coefficients*:* These are conventional ways to describe the image information, which may contain higher rank image information.

Though these features can be used independently for temporal video analysis, recent advances exhibit an increasing tendency of combining them together to achieve better performance. In the following section 3, the combined use of colour and motion is introduced as a case study.

2) Measuring Method of Scene Changes

After a feature (or a set of features) computed on each frame has been defined, a temporal segmentation algorithm needs to detect where these frames exhibit discontinuity. To attain this objective, a proper decision method is required to evaluate a metric expressing the similarity or dissimilarity of the features computed on adjacent frames. This can be done in the following ways [33]:

- Static thresholding: As the most basic decision method, this method compares similarity metrics against fixed thresholds [41]. It performs well only if the thresholds are manually adjusted for each video sequence.
- Adaptive thresholding: By vary the threshold depending on a statistic average of the feature difference metrics, as stated in [38] and [40], it provides a better and simple solution for automatic temporal segmentation.

- Probabilistic detection: More rigorously, this method [34,35] provides a solution by modeling the pattern of specific types of shot transitions and performing optimal scene change estimation according to presupposed specific probability distributions for the feature difference metrics in each kind of scene transitions.
- Trained classifier: By introducing machine learning algorithms, this method is to formulate the problem as a classification task of two classes, namely “shot change’’ and “no shot change’’ [39].

Usually, adaptive thresholding is considered as a simple and effective method in practical applications, and thus more preferred in generic applications to achieve a wide coverage of massive video database.

3 Video Summarization Using Colour and Motion

As it is described in the previous section, feature selection is the first step to develop a video summarization scheme. Feature selection largely relies on the purpose of the application. For example, sports video summarization usually focuses more on the motion of foreground athletes and balls [42], while generic-purpose applications may concern more about background scenes. In most cases, motion and colour are two primary cues to measure the scene change and summarise the video sequence.

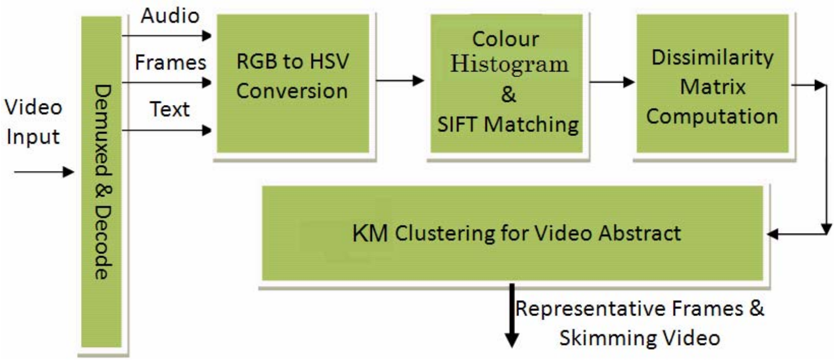

Video Input
Demuxed & Decode
Audio
Frames
Text
RGB to HSV Conversion
Colour Histogram & SIFT Matching
Dissimilarity Matrix Computation
KM Clustering for Video Abstract
Representative Frames & Skimming Video

Fig. 4 Video abstraction using HSV histogram

The approach developed by Lienhart [39] detects dissolves with a trained classifier, operating on either YUV/HSV color histograms, magnitude of directional gradients, or edge-based contrast. Cohen et al. [43] represented object extraction and tracking data using an attributed tracking model. Naturally, a more sophisticated scheme that we can be easily inferred from the above works is the combination of colour and motion together as selected features for video summarization and skimming.

This section introduces such a combined video summarization scheme using both colour and motion, as shown in Figure 4, where the summarization system decodes the input video, implements the RGB-to-HSV conversion, computes the colour histogram, estimates the motion using SIFT (scale-invariant feature tracking) matching [44], generates the dissimilarity matrix, and performs K-means clustering for video abstraction.

3.1 Feature Selection for Video Summarization

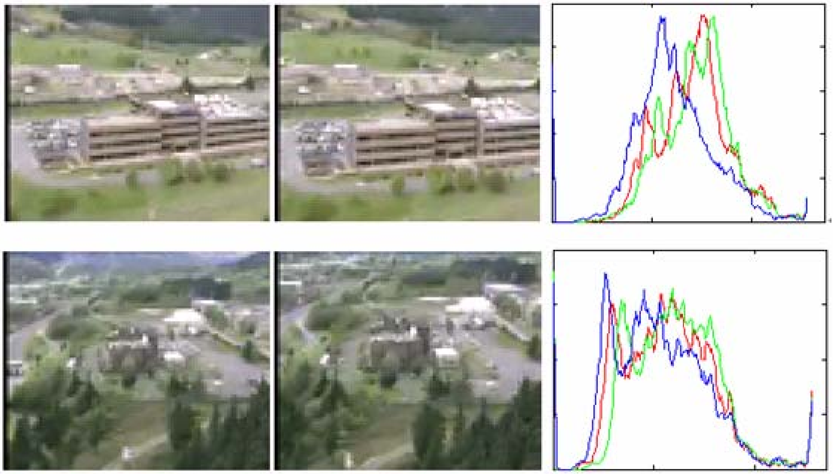
As stated above, two kinds of visual features (colour and motion) are taken into account in this case study: colour histograms and motion vectors. The motion can be estimated between two image frames, and colour histogram can be obtained by statistic counting of all pixels in a frame. In order to make an easy demonstration, here, each video sequence is first segmented into non-overlapped blocks, and the motion and colour histograms of each segment are subsequently computed. In the case study, the entire video sequence is sliced into a number of uniform segments, each having 30 frames.

For each frame, three-channel colour histograms are formulated in the hue, saturation and value (HSV) colour-space, which is a common and standard parameterization. The colour histogram is obtained by statistically counting all pixels in all frames in a given segment. The discontinuity value between two segments is then defined as the Sum of the Absolute Difference (SAD) between their histograms. Figure 5 illustrates an example of computation of HSV histograms over two different shots in a test video.

The motion features can be computed for each frame within a shot using the well-established scale invariant feature transforms (SIFT) algorithm [32]. In fact, this motion estimation is the moving distance of feature points across two image frames. The SIFT features are local and based on the appearance of the object at particular interest points, and are invariant to image scale and rotation. The SIFT algorithm consists of three steps as follows:

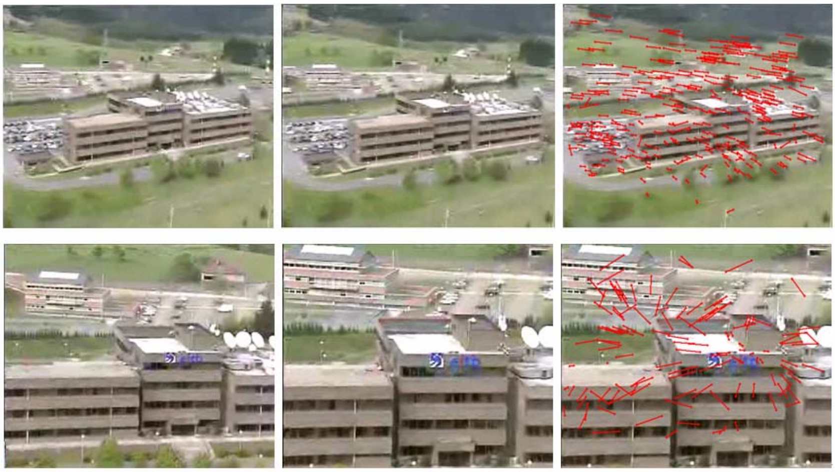
- Difference of Gaussian (DoG) is performed at different scales. Then local maxima/minima of the DoG images are selected as candidate key points.
- The standard keypoint descriptor used by SIFT is created by sampling the magnitudes and orientations of the image gradient in the patch around the keypoint, and building smoothed orientation histograms to capture the important aspects of the patch.
- A 4×4 array of histograms, each with 8 orientation bins, captures the rough spatial structure of the patch. This 128-element vector is then normalized to unit length and thresholded to remove elements with small values to obtain the SIFT descriptor.

Once the SIFT feature points are found in two image frames respectively, a sum-of-absolute-differences (SAD) method is applied to match these feature points between two frames. Finally, a motion vector can be estimated according to the displacements between two groups of corresponding features over the two frames. Figure 6



(a) (b) (c)

Fig. 5 HSV histogram of two segments in the example video: (a) and (b) are the start and end frames of two different segments; (c) shows their histograms in terms of H, S and V



(a) (b) (c)

Fig. 6 Feature matching using SIFT: (a) and (b) two neigh-boring image frames; (c) Displacements of corresponded points

illustrates an example of using SIFT to match two groups of image features. Technical details in the use of SIFT approach for motion estimation can be found in Ref.[44].

3.2 Dissimilarity Matrix between Segments

In this case study, video summarization is based on the use of pair-wise inter-segment dissimilarities. The dissimilarity matrix describes pair-wise distance in feature space between two arbitrary segments, which is a symmetrical matrix with the (n, m)-th element that represents the measured distance between the n-th and m-th segments. The underlying basic concept is to utilize the extracted colour histogram and motion features in an attempt to establish a dissimilarity matrix to describe inter-segment relationship. However, since some frames can significantly differ in colour while having small camera motion, the contributions of these different features must be fused into one indication for overall evaluation. Thus, a combinatorial form with adaptive weights to both features is required.

The motion activity within a segment can be quantified by the average value of all SIFT-based affine motion vectors of all frames within this segment (each segment has 30 frames), which can be denoted as f^n. The distinction between two segments can be obtained as $||f^m- f^n||$. Afterwards, the colour and motion estimates are normalized against the overall shots, which lead to dissimilarities Q_H (colour histogram) and Q_M (motion), respectively. The simplest way for metric combination can use the following linear equation to formulate the combined dissimilarity matrix:

$$Q_{Fusion}(m,n) = w_1 Q_H(m,n) + w_2 Q_M(m,n) \qquad (1)$$

Where Q_{Fusion} (·) is the dissimilarity element considering the fusion of both colour histogram and motion, and w_1 and w_2 are the weights of features.

Figure 7 denotes an example of dissimilarity matrix using the example video with 50 segments from a total number of 1500 image frames (each segment has 30 frames), where different colours represent different dissimilarity levels.

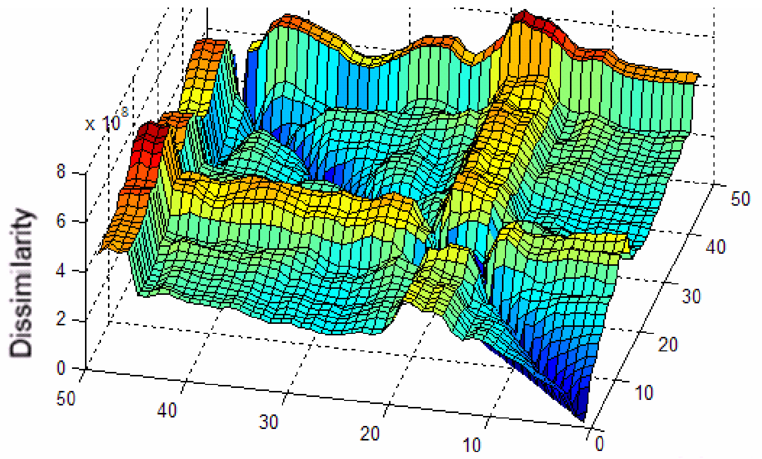


Fig. 7 The dissimilarity between segments

3.3 Video Summarization Using K-Mean Clustering

Among various clustering algorithms, k-mean clustering (KMC) is a practical and easy method for this kind of problems. KMC is a clustering algorithm to partition n data set into k clusters (where k < n), which is very similar to the expectation-maximization (EM) algorithm for mixtures of Gaussians in that they both attempt to find the centers of clusters in the data.

To attain the target of clustering, KMC algorithms are based on minimization of the following objective function:

$$Z = \sum_{j=1}^{K} \sum_{x_i \in S_j}^{N} \left\| x_i - c_j \right\|^2 \tag{2}$$

Where there are *k* clusters S_j, i = 1, 2, ..., *K*, and c_j is the centroid or mean point of all the points x_i. ||*|| is any norm denoting the similarity between any measured data and the cluster centre.

In order to find an appropriate solution to the above equation, a cluster number must be provided before the KMC iteration starts. Assuming $\{c_i\}^{(k)}$ is the set of $K^{(k)}$ initial cluster centers for KCM clustering, to obtain an optimal result of cluster centers $\{c_i\}^{(k+1)}$ that matches the optimization target in the clustering model Ω, maximum likelihood estimation can be obtained through Expectation-Maximization iteration. The iteration procedure can be summarized as follows:

- Assume there are initially k=2 clusters in the data set. The KMC algorithm is applied to $\{x_i\}$, resulting in two cluster centres with corresponding coordinates.
- Euclidean distances between the cluster centres $\{c_i\}^{(m)}$ are computed. If their mean distance are greater than the predefined thresholds $\{T^{(k)}\}$, the optimal cluster number *k* is increased to *k*+1; Otherwise, the iteration is terminated, and take *k*=*m* as our final result.
- Taking similar iterations like this, one is able to determine a cluster number that brings us a Euclidean distance less than the threshold. This cluster number is what we anticipate.

At the final, the optimal cluster number is obtained by the above iteration procedure.

Figure 8 illustrates some problems in the video summary of the test video by the above clustering method, where we can find there are several key frames that are duplicated (e.g. interview scenario). One reason for this redundancy is due to the significant illumination changes within these particular segments. From this test, we can also see that using motion and HSV colour histogram descriptors only are not enough to capture the variation along video sequence. In section 4, we will introduce a video summarization scheme using Eigen-features which may have better discriminative power for video summarization.

Overall experiments are conducted on five videos of the video database [51-52], the results are listed in Table 1. Recall and precision are employed as the measure for performance evaluation. Here, recall rate and precision are defined as,

$$\text{Recall Rate} = \frac{K_P}{K_A} \qquad \text{Precision} = \frac{K_P}{K_P + K_N} \tag{3}$$

Where, K_P is the number of correctly detected representative frames, K_A is the number of all real key frames, and K_N is the number of wrong detected representative frames. In addition, skim rates will be utilized to evaluate the length of the generated abstracts from the videos, which is equivalent to the number of the extracted frames divided by the number of the entire video frames.

From Table 1, it can be seen that the proposed video summarization algorithm has been well performed on this small database, leading to a recall of 0.53, a precision of 0.60 and a skimming rate of 18% on average.

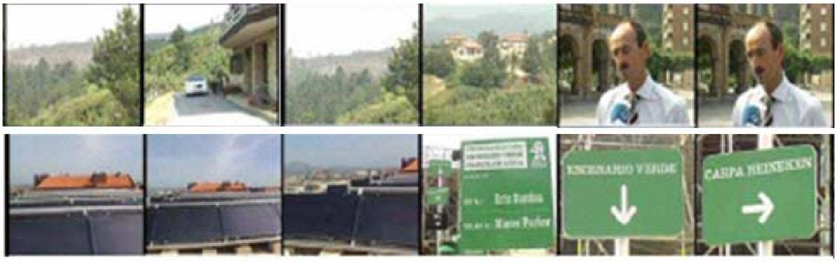


Fig. 8 Redundancy in detected key frames using motion & colour

Table 1 Statistics of Video Summary

Video No.	Recall	Precision	Skimming Rate
1	0.56	0.63	8.51%
2	0.75	0.80	20.25%
3	0.45	0.56	17.88%
4	0.37	0.51	16.21%
5	0.52	0.47	31.75%

4 Video Summarization Using Eigen-Features

The first step in video summarization is to build a proper model to measure the dissimilarity between frames for video temporal structure analysis. As described in the above sections, major features commonly applied to measure the dissimilarity include motion descriptors [45, 46], colour histogram [36], and Eigen-features [47-50], etc..

Recently, singular value decomposition (SVD) or latent semantic analysis (LSA) [47-50] emerges as an attractive computational model for video summarization because Eigen-features are usually the most representative and discriminative features for frame comparison. It can also put all frames into a balanced comparison, and thus the overall video structure can be hierarchically organized with adaptive thresholds. Cernekova et al. [41] proposed to perform LSA on the RGB color histograms of each frame to reduce feature vector dimensionality, in order to make the feature space more apparently discriminative. Others apply LSA approach directly on the image sequence [48].

However, this approach is computationally intensive since it operates directly on video frames. A solution for this problem, as introduced in this section, is to select a set of reference frames to form an Eigen space to measure intra-frame dissimilarity indirectly. In this section, we give an example of this kind of schemes to use Eigen-features.

4.1 Latent Semantic Analysis

Latent semantic analysis (LSA) was born out of text retrieval and uses at its core singular value decomposition (SVD) [47-50]. Given a sparse $m\times n$ term-document matrix Δ, an SVD decomposition, $\Delta = \Phi\Sigma\Psi^T$, can be performed, normally through a QR decomposition, which yields $\Phi(m\times n)$, the term-concept matrix, $\Sigma(n\times n)$, a diagonal matrix containing the singular values in decreasing order, and $\Psi^T(n\times n)$, the concept-document matrix.

Normally, a form of dimension reduction is then applied, often referred to as rank lowering, where only the top k singular values are retained. This dimension reduction has the effect of resulting zero valued entries in the original matrix A to become non-zero.

To apply LSA to image sequence, considering we have N frames, which can be represented as a $W\times H$-dimensional vector f_n, and in total, we can give a feature matrix for this frame sequence as,

$$F= [f_1, f_2, \ldots, f_N] \tag{4}$$

Here, every f_n is the n-th frame image. LSA, also known as Karhunen-Loeve methods [41, 47-50], is a typical method to analyze the discriminant information hidden in such feature matrices.

In LSA approach, the covariance matrix of can be obtained by

$$\Delta = \{f_i - f_m\}, f_m = \frac{1}{N}\sum_N f_i \tag{5}$$

Given $f_i \in \mathbb{R}^h$, a linear function can be found to reflect it to $\mathbb{R}^l$,

$$\Sigma\Sigma^T = \Phi\Delta\Delta^T\Phi^T \tag{6}$$

Where Φ is called Eigen vectors in linear algebra. With this linear projection in Eigen space $\Phi \in \mathbb{R}^l$, we have Eigen projection matrix W_{LSA},

$$W_{LSA} = \Delta\Phi \tag{7}$$

Here, W_{LSA} resembles the discriminant information hidden in the feature matrix F.

The distance of a frame i from another frame k can be calculated as,

$$d_k = \| W_{LSA} f_i - W_{LSA} f_k \| \tag{8}$$

Where, $\|\cdot\|$ is Hilbert-Schmidt norm, and d_k is the Mahalanobis distance that expresses the dissimilarity between two images i and k.

4.2 Hierarchical Video Temporal Structure Summarization

Latent semantic analysis (LSA) projects the frames into its Eigen subspace, where LSA is applied to extract a subspace in which the variance is maximized. Its objective function is as follows:

$$\max_{W} \sum_{i=1}^{n} (y_i - \bar{y})^2, \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{9}$$

However, the direct application of LSA is usually very compute-intensive. For a video with 10000 frames, LSA need to perform singular value decomposition (SVD) of a 10000×10000 matrix to obtain its projection matrix W_{LSA}.

Considering N frames selected from the video stream with a regular interval Δ (Here we use Δ=200), we have a set of frames as,

$$\Lambda = [\Lambda_1, \Lambda_2, \ldots, \Lambda_K] \tag{10}$$

From equations in the above section, we can obtain the Eigen projection matrix $W_{LSA}{}^T$ from Λ. As stated in the previous section, a given image x_i can then be projected into this reference subspace as y_i. The distance of the image x_i from a reference image Λ_k can be calculated as,

$$d_k = \| W_{LSA}{}^{\mathrm{T}} x_i - W_{LSA}{}^{\mathrm{T}} \Lambda_k \| = \| y_i - y_k \| \tag{11}$$

Considering the dissimilarity between a given image and a set of the selected frames $\{\Lambda_k\}$, there may forms a dissimilarity vector D^i,

$$D^i = \{d_1, d_2, \ldots, d_K\} \tag{12}$$

Thus, any frame can be simply featured by its similarity projection D^i in the dissimilarity subspace $\mathbb{R}^K$ provided by reference frames. Figure 9 is the projected result of 1000 frames of a test video in the first two dimensions of the reference-based Eigen space, where about 300 reference frames are selected from 6000 frames. With the use of the selected reference frame set, the matrix size for SVD computation is reduced dramatically from N×N to N/200×N/200.

With this subspace projection, the dissimilarity between any two frames x_i and x_j can be computed by their distance in this reference subspace,

$$ds^{\{i,j\}} = \| D^i - D^j \| \tag{13}$$

In video temporal structure analysis, the most useful information is the dissimilarity between neighbored frames i and (i+1), namely $ds^{\{i,\, i+1\}}$. Figure 10 is the computed neighborhood dissimilarity of the test video.

With the above overall dissimilarity measure, it is not difficult to estimate the hierarchical video temporal structure. To detect video shot boundary, rather than using any predefined threshold, we apply an adaptive threshold obtained from the dissimilarity distribution ds,

$$d_{TH} \to \arg\max_{d_{TH}} \left(Q(ds, d_{TH}) \right) \tag{14}$$

Where, Q is the cost function to estimate the optimal threshold values that can be adaptive to the video. A frame with its neighborhood dissimilarity ds greater than

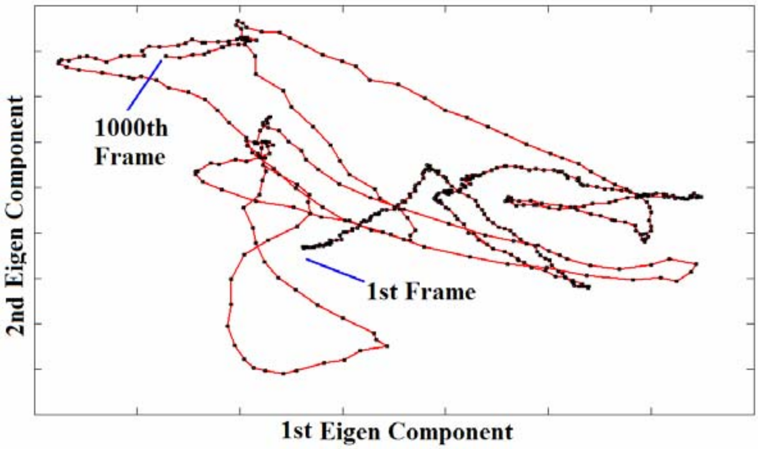


Fig. 9 Projection of frames in the reference-based Eigen space

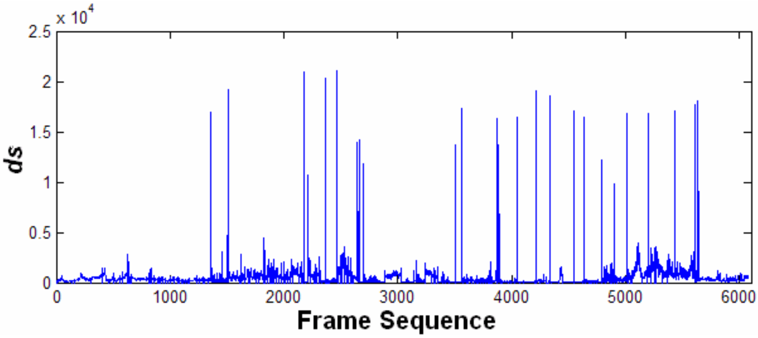


Fig. 10 Dissimilarity distribution *ds* between neighbored frames in reference-based Eigen subspace

d_{TH} can be defined as a shot boundary, and the middle frame between two shot boundary frames is defined as the key frame of this shot.

After the shot-level video segment structure is determined, its hierarchical sub-shot structure can be extracted subsequently. The basic scheme is similar to the KMC procedure in section 3.3. The only difference is, here we use all intra-shot frames in one shot to generate the dissimilarity matrix measured in Eigen subspace for this shot. After the intra-shot dissimilarity matrix is obtained, KMC approach is then applied to this matrix, to find out most representative sub-shot key frames that cannot be represented well by shot-level key frames.

On the other side, the higher-level scenes can be similarly summarized by comparing the shot-level key frames and using k-means clustering to find the most representative scene-level key frames from all shot-level key frames.

From the above description, we can see that both sub-shot level and above-shot scene level summarization use the same mathematic model, k-means clustering, to find out most representative frames. The difference is their targets: sub-shot summarization uses all frames in a shot, while scene-level summarization uses only shot-level key frames.

4.3 Experimental Results

In the experiment, 10 videos from RUSHES video database [51-52] are used for test. The program is coded in MATLAB. The test videos are encoded in MPEG-4 format, which can be read through MATLAB multimedia interface.

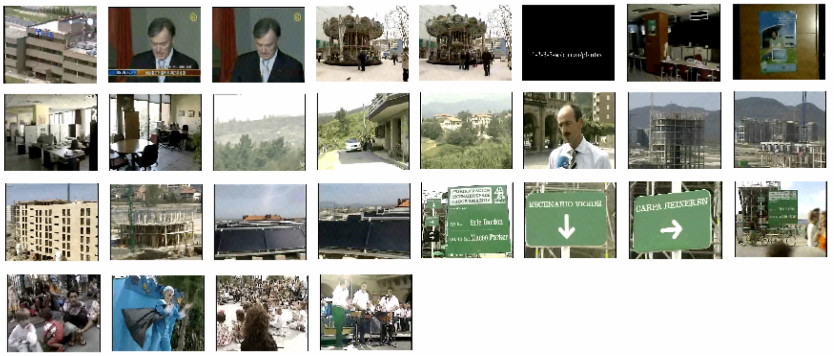


Fig. 11 Detected shot-level key frames

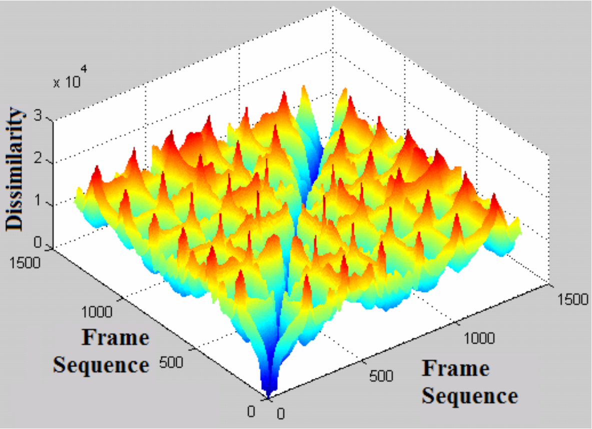


Fig. 12 Intra-shot dissimilarity matrix

In all experiments, the video frames are resized to 50% to save computation time. In the test, the video stream is first input through multimedia codec interface, and reference frames are selected from the sequence at a regular interval of one reference frame per 200 frames. For a 30-minutes video sequence, there are about 250 reference frames to be selected. Then the LSA approach is applied to these

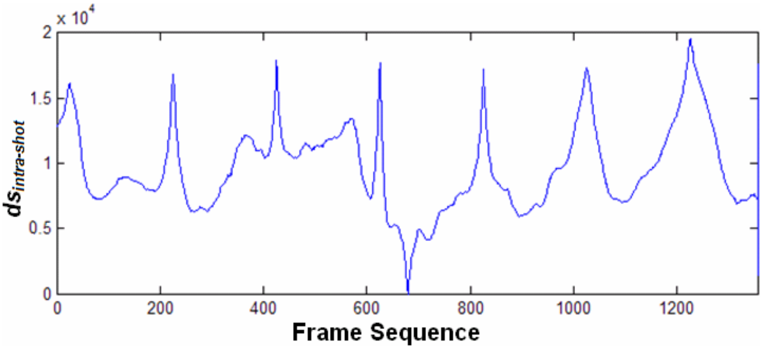
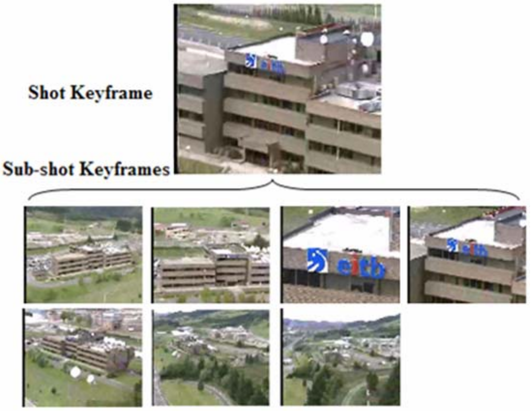


Fig. 13 Intra-shot dissimilarity from the shot-level key frame

Fig. 14 Sub-shot representative frames



reference frames to obtain the projection matrix W_{LSA}. With the projection matrix, all frames can be projected into the LSA subspace, and the dissimilarity between any two frames can be defined as their Euclidean distance (Mahalanobis distance) in the Eigen subspace.

Figure 9 has given the projection results of one test video in the Eigen space, and Figure 10 is its dissimilarity distribution. With the computed dissimilarity, the video temporal structure can be easily obtained. Figure 11 gives the detected representative frames of one test video, which shows that in comparison with the results in Figure 8 in section 3, it is clearly shown that the use of LSA and its Eigen feature can detect more key frames with less redundancy than the previous approach using colour and motion.

After the top-level structure is detected, we can further detect the sub-shot video structure. In this step, the dissimilarity of all frames in the shot are computed, and KMC approach is applied to find out most representative features are considered as sub-shot representative frames.

Figure 12 gives the intra-shot frame dissimilarity matrix of the first shot in a test video measured in the Eigen space, and Figure 13 shows an example of the measured intra-shot dissimilarity by the LSA approach. The 1st shot in this test video has 1356 frames and 7 sub-shots in total. Figure 14 lists the detected hierarchical sub-shot representative frames of this shot.

4.4 Summarize a Real Movie

Since a video summarization can provide viewer with a brief but impressive outline of the movie, most recent work on movie abstraction focuses on the generation of a short synopsis of a long feature film. In this experiment, the above LSA-based approach is applied to a well-known test movie, *Friends*, to extract its shot-level and scene-level frames for summarization. Figure 15 shows some frames in the movie. We can see there are frequent scene transitions that may make up the video story.

Fig. 15 Frames in Movie --- *Friends* Episode

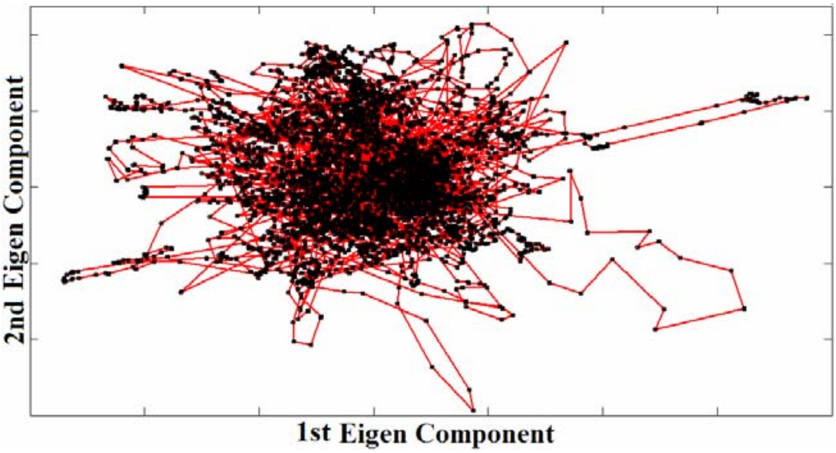


Fig. 16 The projection of frames in Eigen subspace

In the experiment, we first select reference frames by one per 200 frames. For a 20 minutes story, we can have about 200 reference frames. As described in section 4.2, the LSA approach is applied to extract the projection matrix W_{LSA}. After W_{LSA} is obtained, all frames can be projected into the Eigen space. Figure 16 is the projection result of the first 8000 frames in the movie. With this projection result, the distance between any two frames can be easily measured.

A useful dissimilarity measure is the dissimilarity between two neighboured frames, which may represent the scene transition process. Figure 17 gives the measured neighbourhood dissimilarity distribution along frame sequence. With this dissimilarity measure, shot transitions and key frames can be detected, as

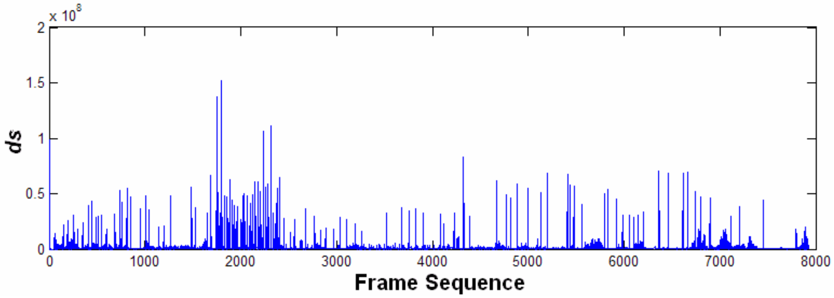


Fig. 17 Dissimilarity between neighbored frames in Eigen space

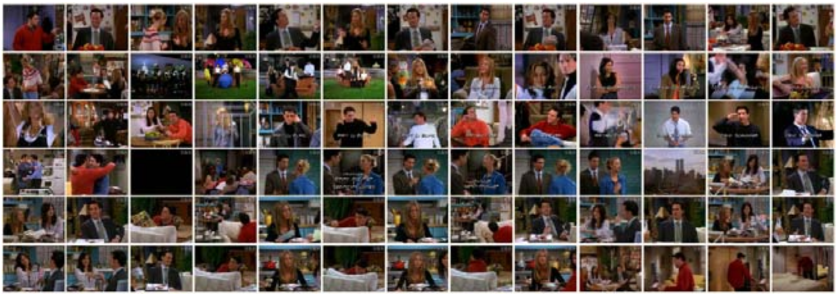


Fig. 18 Extracted shot-level key frames

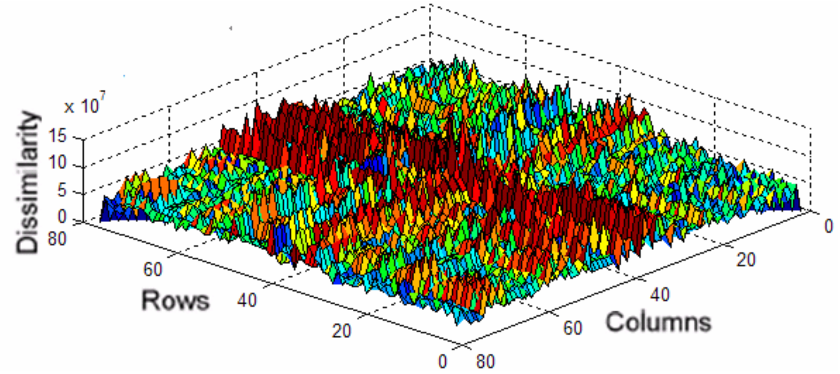


Fig. 19 Dissimilarity Matrix between shot-level key frames

shown in Figure 18. In total, about 100 key frames are found from the first 5 minutes of the movie.

As it can be seen from the detected key frames in Figure 18, shot transitions happen frequently in this movie, and video summarization may need a further abstraction to find those most representative scene frames.

In order to achieve this purpose, the dissimilarity of all detected key frames can be computed by measuring their distance in the reference Eigen space projection. Figure 19 gives the calculated dissimilarity matrix. Similarly as it is described in section 4.2, k-means clustering is applied to the dissimilarity matrix to find out which is the most representative scene-level key frames. Figure 20 is the result from KM clustering, where about 20 scene-level key frames are selected from about 100 shot-level key frames. It is shown that these scene-level key frames have less redundancy or scene similarity than shot-level key frames shown in Figure 18.

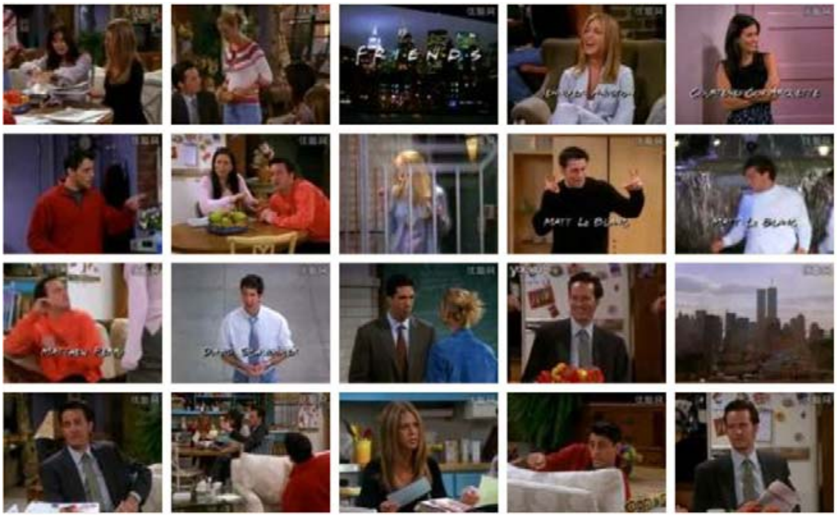


Fig. 20 Summarized scene-level representative frames

5 Conclusion

With the proliferation of digital videos, as a useful video abstraction tool for fast content browsing, skimming, transmission, and retrieval of massive video database, video summarization and skimming has become an indispensable tool of any practical video content management systems. This chapter presented state-of-the-art techniques for video abstraction and provided a review on recent advances on video summarization and skimming.

In addition, two practical video summarization schemes are presented and tested with academic-oriented videos and general-purpose movies. It is our belief that, with the technical maturity of scene classification, content understanding and video abstraction, an automatic content analysis system that facilitates navigation, browsing, and search of desired movie content will be arriving in the near future with a promising commercial market in web multimedia, mobile multimedia, interactive TV, and emerging 3D TV.

Abbreviations

LSA	---	Latent Semantic Analysis
SVD	---	Singular Value Decomposition
KMC	---	k-means Clustering
ROI	---	Region of Interest
DCT	---	Discrete Cosine Transform
SIFT	---	Scale Invariant Feature Transform
EM	---	Expectation Maximization

References

[1] Hanjalic, A.: Extracting moods from pictures and sounds: Towards truly personalized TV. IEEE Signal Processing Mag. 23(2), 90–100 (2006)

[2] Grgic, M., Grgic, S., Ghanbari, M.: A New Approach for Retrieval of Natural Images. Journal of Electrical Engineering 52(5-6), 117–124 (2001)

[3] Li, Y., Lee, S., Yeh, C., Kuo, C.-C.: Techniques for Movie Content Analysis and Skimming. IEEE Signal Processing Magazine, 76 (March 2006)

[4] VideoCollge, Microsoft Research Asia, `http://research.microsoft.com/en-us/projects/VideoCollage`

[5] VideoSue, IBM Watson Research Center, `http://www.research.ibm.com/MediaStar/VideoSue.html`

[6] Taniguchi, Y.: An intuitive and efficient access interface to real-time incoming video based on automatic indexing. In: Proc. ACM Multimedia 1995, November 1995, pp. 25–33 (1995)

[7] Mills, M.: A magnifier tool for video data. In: Proc. ACM Human Computer Interface, May 1992, pp. 93–98 (1992)

[8] Zhang, H.J., Wu, J., Zhong, D., Smoliar, S.: An integrated system for content-based video retrieval and browsing. Pattern Recognit. 30(4), 643–658 (1997)

[9] Zhuang, Y., Rui, Y., Huang, T., Mehrotra, S.: Adaptive key frame extraction using unsupervised clustering. In: Proc. ICIP 1998, October 1998, pp. 866–870 (1998)

[10] Ju, S.X., Black, M.J., Minneman, S., Kimber, D.: Summarization of video-taped presentations: Automatic analysis of motion and gestures. IEEE Trans. Circuits Syst. Video Technol. 8(5), 686–696 (1998)

[11] Toklu, C., Liou, S.P.: Automatic keyframe selection for content-based video indexing and access. In: Proc. SPIE, January 2000, vol. 3972, pp. 554–563 (2000)

[12] Iran, M., Anandan, P.: Video indexing based on mosaic representation. IEEE Trans. Pattern Anal. Machine Intell. 86(5), 905–921 (1998)

[13] Vasconcelos, N., Lippman, A.: A spatiotemporal motion model for video Summarization. In: Proc. IEEE Computer Soc. Conf. Computer Vision Pattern Recognition, June 1998, pp. 361–366 (1998)

[14] Taniguchi, Y., Akutsu, A., Tonomura, Y.: Panorama Excerpts: Extracting and packing panoramas for video browsing. In: Proc. ACM Multimedia 1997, November 1997, pp. 427–436 (1997)

[15] Doulamis, A.D., Doulamis, N.D., Kollias, S.D.: Non-sequential video content representation using temporal variation of feature vectors. IEEE Trans. Consumer Electron. 46(3), 758–768 (2000)

[16] Uchihashi, S., Foote, J., Girgensohn, A., Boreczky, J.: Video manga: Generating semantically meaningful video summaries. In: Proc. ACM Multimedia 1999, October 1999, pp. 383–392 (1999)

[17] Girgensohn, A., Boreczky, J.: Time-constrained keyframe selection technique. In: Proc. ICMCS 1999, June 1999, pp. 756–761 (1999)

[18] Yeung, M.M., Yeo, B.L.: Video visualization for compact presentation and fast browsing of pictorial content. IEEE Trans. Circuits Syst. Video Technol. 7(5), 771–785 (1997)

[19] Ratakonda, K., Sezan, M.I., Crinon, R.: Hierarchical video Summarization. In: Proc. SPIE, January 1999, vol. 3653, pp. 1531–1541 (1999)

[20] Doulamis, A.D., Doulamis, N.D., Kollias, S.D.: A fuzzy video content representation for video Summarization and content-based retrieval. Signal Process. 80(6), 1049–1067 (2000)

[21] Sun, X.D., Kankanhalli, M.S.: Video Summarization using R-sequences. Real-Time Imaging 6(6), 449–459 (2000)

[22] Huang, Q., Lou, Z., Rosenberg, A., Gibbon, D., Shahraray, B.: Automated generation of news content hierarchy by integrating audio, video, and text information. In: Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, March 1999, vol. 6, pp. 3025–3028 (1999)

[23] Xiong, Z., Radhakrishnan, R., Divakaran, A.: Generation of sports high-lights using motion activity in combination with a common audio feature extraction framework. In: Proc. IEEE Int. Conf. Image Processing, September 2003, vol. 1, pp. I-5–I-8 (2003)

[24] Omoigui, N., He, L., Gupta, A., Grudin, J., Sanocki, E.: Time-compression: System concerns, usage and benefits. In: Proc. ACM Conf. Computer Human Interaction, May 1999, pp. 136–143 (1999)

[25] Amir, A., Ponceleon, D., Blanchard, B., Petkovic, D., Srinivasan, S., Cohen, G.: Using audio time scale modification for video browsing. In: Proc. 33rd Hawaii Int. Conf. System Sciences, January 2000, vol. 3, pp. 3046–3055 (2000)

[26] Nam, J., Tewfik, A.H.: Video abstract of video. In: Proc. IEEE 3rd Workshop Multimedia Signal Processing, September 1999, pp. 117–122 (1999)

[27] Hanjalic, A., Zhang, H.J.: An integrated scheme for automated video abstraction based on unsupervised cluster-validity analysis. IEEE Trans. Circuits Syst. Video Technol. 9(8), 1280–1289 (1999)

[28] Zhu, X., Fan, J., Elmagarmid, A.K., Aref, W.G.: Hierarchical video Summarization for medical data. In: Proc. SPIE, January 2002, vol. 4674, pp. 395–406 (2002)

[29] Pfeiffer, S., Lienhart, R., Fischer, S., Effelsberg, W.: Abstracting digital movies automatically. J. Visual Commun. Image Represent. 7(4), 345–353 (1996)

[30] Russell, D.D.: A design pattern-based video Summarization technique: moving from low-level signals to high-level structure. In: Proc. 33rd Hawaii Int. Conf. System Sciences, January 2000, vol. 3, pp. 1137–1141 (2000)

[31] Salembier, P., Smith, J.: MPEG-7 multimedia description schemes. IEEE Trans. Circuits Syst. Video Technol. 11(6), 748–759 (2001)

[32] Koprinska, I., Carrato, S.: Temporal video segmentation: A survey. Signal Processing: Image Comm. 16, 477–500 (2001)

[33] Cotsaces, C., Nikolaidis, N., Pitas, I.: Video Shot Detection and Condensed Representation. IEEE Signal Processing Magazine, 28 (March 2006)

[34] Lelescu, D.: Statistical sequential analysis for real-time video scene change detection on compressed multimedia bitstream. IEEE Trans. Multimedia 5(1), 106–117 (2003)

[35] Hanjalic, A.: Shot-boundary detection: Unraveled and resolved? IEEE Trans. Circuits Syst. Video Technol. 12(2), 90–105 (2002)

[36] Zhang, H., Kankanhalli, S.S.A.: Automatic partitioning of full-motion video. ACM Multimedia Syst. 1(1), 10–28 (1993)

[37] Zabih, R., Miller, J., Mai, K.: A feature-based algorithm for detecting and classification production effects. ACM Multimedia Syst. 7(1), 119–128 (1999)

[38] Yu, J., Srinath, M.D.: An efficient method for scene cut detection. Pattern Recognit. Lett. 22(13), 1379–1391 (2001)

[39] Lienhart, R.: Reliable dissolve detection. In: Proc. SPIE, January 2001, vol. 4315, pp. 219–230 (2001)

[40] Boccignone, G., Chianese, A., Moscato, V., Picariello, A.: Foveated shot detection for video segmentation. IEEE Trans. Circuits Syst. Video Technol. 15(3), 365–377 (2005)

[41] Cernekova, Z., Kotropoulos, C., Pitas, I.: Video shot segmentation using singular value decomposition. In: Proc. 2003 IEEE Int. Conf. Multimedia and Expo, Baltimore, Maryland, July 2003, vol. 2, pp. 301–302 (2003)

[42] Zhang, W., Ye, Q., Xing, L., Huang, Q., Gao, W.: Usupervised sports video scene clustering and its application to story units detection. In: Proc. SPIE – VCIP (2005)

[43] Cohen, A., Bjornsson, C., Temple, S., Banker, G., Roysam, B.: Automatic summarization of changes in biological image sequences using algorithmic information theory. IEEE Trans. Pattern Analysis & Machine Intelligence 29 (2008)

[44] Lowe, D.: Distinctive image features from scale-invariant keypoints. Int'l J. of Computer Vision 60(2), 91–110 (2004)

[45] Ngo, C.W., Pong, T.C., Zhang, H.J.: Motion-Based Video Representation for Scene Change Detection. Int. Journal of Computer Vision (2002)

[46] Jiang, M., Crookes, D.: Approach to Automatic Video Object Motion Segmentation. Electronics Letters 43(18), 968 (2007)

[47] Delac, K., Grgic, M., Grgic, S.: Independent Comparative Study of PCA, ICA, and LDA on the FERET Data Set. International Journal of Imaging Systems and Technology 15(5), 252–260 (2005)

[48] Gong, Y.H., Liu, X.: Video Summarization Using Singular Value Decomposition. In: Int. Conf. CVPR (2000)

[49] Slaney, M., Ponceleon, D.: Hierarchical segmentation using latent semantic indexing in scale space. In: 2001 IEEE International Conf. Acoustics, Speech & Signal Processing, May 2001, vol. 3, p. 1437 (2001)

[50] Souvannavong, F., Merialdo, B., Huet, B.: Latent semantic indexing for semantic content detection of video shots. In: 2004 IEEE International Conf. on Multimedia & Expo, June 2004, vol. 3, p. 1783 (2004)

[51] European Research Project on multimedia search and retrieval of Rushes data, `http://www.rushes-project.eu`

[52] Schreer, O., Ardeo, L., Sotiriou, D., Sadka, A., Izquierdo, E.: User Requirements for Multimedia Indexing and Retrieval of Unedited Audio-Visual Footage – RUSHES. In: Ninth International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2008), May 7-9, 2008, p. 76 (2008)

New Video Abstractions for Summarisation and Browsing

Janko Ćalić

Abstract. In present-day technological developments of multimedia systems, the most challenging conundrum facing the signal processing community is how to facilitate effortless access to a rich profusion of of audio-visual data. In order to enable the intuitive access to large video archives, the areas of video summarisation and abstraction have become momentous. The main challenge of the systems for video summarisation and browsing is to achieve a good balance between removal of redundant sections of video and representative coverage of the video summary. The methods represented in this chapter make a shift towards more user-centred summarisation and browsing of large video collections by augmenting interaction rather than learning the way users create related semantics. A comprehensive survey of the state-of-the-art methods in the area of video summarisation and browsing is presented, followed by a detailed analysis of a novel approach to temporal video representation.

In order to create an effortless and intuitive access to the overwhelming extent of visual information, we propose exploitation of universally familiar abstractions such as linking ranking with spatial linear ordering and presented size, projecting highly dimensional data onto the 2D/3D visualisation space as well as using narrative structure of sequential visual art. To achieve this goal, a set of sophisticated yet robust machine learning and knowledge discovery methods are employed. By combining efficient multimedia signal processing and the computational intelligence, coupled with the user-centric interface design, the presented visualisation systems augments the vast abstract space of visual similarity, thus enabling responsive environment for intuitive experience of large-scale video databases.

1 Introduction

The recent developments in multimedia and communications technology have brought a sea change to our everyday experiences - the way we communicate, plan,

Janko Ćalić
I-Lab, University of Surrey, Guildford GU2 7XH, UK
e-mail: j.calic@surrey.ac.uk

M. Grgic et al. (Eds.): Rec. Advan. in Mult. Sig. Process. and Commun., SCI 231, pp. 51–71.
springerlink.com

watch telly and even play with friends. In the spring of 2008 BBC's video on demand iPlayer service had a monthly audience of 1.4 million and was dealing with over 700,000 daily viewing requests [31]. In addition, the projections based up on a consumer survey [4] show that we could expect over 2.2 TB of new content in an average home in 2013 including backups. By the end of 2008, there were 200,000 videos uploaded to YouTube every day. This immense data that we consume and store is becoming insurmountable, and thus the demand for effective yet intuitive multimedia management system is rapidly growing.

Having experienced this omnipresent proliferation of digital video and associated web-based services and communities, the user-centric video applications are of paramount importance. In order to facilitate effortless experience of such a vast visual information, video summarisation and browsing have re-entered the limelight in the multimedia technology world. This revival of video summarisation comes after a failure of the semantic-based approaches to video content analysis, especially knowing the striking production rate of video data and the need for fast and efficient video processing algorithms.

In order to enable the intuitive access to large video archives, the main challenge of systems for video summarisation and browsing is to achieve a good balance between removal of redundant sections of video and representative coverage of the video summary. In order to achieve this balance, the system needs to understand both user and application context, meaningfully analyse the content of the videos involved and represent the abstractions to the end user. This chapter focuses on these three themes in the following sections, while proposing a unified approach to efficient user-centric video summarisation and browsing.

2 Video Representation

In this section, we address the problem of *computational* video representation , i.e. how to abstract the audio-visual experience of the user by means of computational models. This is clearly a difficult task that has to involve both appropriate computation and processing as well as the way in which a user experiences targeted media. However, this is not a common approach to video analysis, where the focus is on the way information is extracted from the digital media, whether it makes sense to the user or not. The information flow in a content-based video analysis system has an important step between the set of automatically extracted low-level features and the user - video representation.

The foundational work that has formulated the problem of computational video representation was presented by Davis [10, 11] and Davenport et al. [9]. In [10] multi-layered, iconic annotations of video content called MediaStreams is developed as a visual language and a stream-based representation of video data, with special attention to the issue of creating a global, reusable video archive. Being radically oriented towards a *cinematic* perspective of video representation, the work presented in [9] sets the scene to a novel approach to the content-based video analysis based upon a shot, an irreducible constituent of video sequences. But this was where research community stopped following this paradigm and got attracted to extraction and analysis of low-level features, ignoring the fact that these features would make little or no sense to the end user.

And it wasn't until the definition of Computational Media Aesthetics (CMA) in a number of publications by Dorai and Venkatesh [16, 17] that the user centred representation re-emerged within the video processing community. The main idea behind CMA is to have a focus on domain distinctiveness, the elements of a given domain that shape its borders and define its essence (in film, for example, shot, scene, setting, composition, or protagonist), particularly the expressive techniques used by a domains' content creators [1]. This is clearly a diametrically opposite point of view to the common perception that the video should be indexed by the terms for which automatic detectors can be realised [34]. Nevertheless, these two different approaches are bound to merge in order to achieve the goal of *semantic* analysis of video media.

In the field of multimedia signal processing there has been a plethora of interesting research work presented recently that focuses on the problem of *semantic gap* between low-level information extracted from the media and the user's need to meaningfully interact with it on a higher level. However, the majority of ideas follow a paradigm of finding a direct mapping from low-level features to high-level semantic concepts. Not only does this approach requires extremely complex and unstable computation and processing [5], but it appears to be unfeasible unless it targets a specific and contextually narrow domain [7, 15, 23]. Little has been done to design a system capable of creating appropriate representations of video media on various levels of complexity and thus improve adaptability and reliability of multimedia analysis. As given in [2, 26], the multimedia database stands as a central point of the modern creativity and thus the challenge to effortlessly interact with the large digital media collections is our prime goal.

In order to create an effortless and intuitive interaction with the overwhelming extent of information embedded in video archives, we propose a novel approach to representation of temporal video dimension. The idea is to exploit the universally familiar narrative structure of comics to generate easily-readable visual summaries. Being defined as "spatially juxtaposed images in deliberate sequence intended to convey information" [28], comics are the most prevalent medium that expresses meaning through a sequence of spatially structured images. Exploiting this concept, the proposed system follows the narrative structure of comics, linking the temporal flow of video sequence with the spatial position of panels in a comic strip. This approach differentiates our work from more typical reverse-storyboarding [14] or video summarisation approaches.

3 Video Analysis for Summarisation and Browsing

Having established an appropriate representation of video data, the analysis stage processes the video in order to extract features relevant to the representation model. In video summarisation and browsing, the most common representation comprises a set of most relevant frames from the video sequence, called *key frames*. There are numerous approaches to the extraction of key frames, from shot boundary based methods to machine learning and here we highlight the most relevant approaches.

Zhuang et al. [39] proposed an unsupervised clustering method based on HSV colour features, where the frame closest to the cluster centre is chosen as the key frame representative for a given video shot. Utilising cluster-validity analysis, Hanjalic and Zhang [20] remove the visual content redundancy among video frames using an unsupervised procedure. An interesting approach introduced by DeMenthon et al. [12] represents the video sequence as a curve in a high dimensional space, and the summary is represented by the set of salient points on that curve. Recently, Wah et al. [8] exploited a normalised cut algorithm to globally and optimally partition the graph representation into video clusters and describe the evolution and perceptual importance of a video segment.

As given in Section 5, in order to uncover the underpinning structure of the key-frame data in an unsupervised manner, our work exploits K-way spectral clustering method [29] in perceptual grouping of extracted key-frames using locally scaled affinity matrix [40].

4 Visualising Video Abstractions

A common approach to visualising video abstractions is to render the key frame thumbnails on a 2D display. However, this approach struggles to achieve highly condensed representations, due to its frequent repetitiveness of the material and lack of the notion of temporal dimension in video. In order to tackle this, there have been attempts to utilise the form of comics as a medium for visual summarisation of videos. In [37] a layout algorithm that optimises the ratio of white space left and approximation error of the frame importance function is proposed. Following a similar approach, the work presented in [19] introduces a number of heuristic rules to optimise the layout algorithm. However, due to an inherently difficult optimisation, these attempts failed to develop a feasible layout algorithm. In Section 5.3.3 a fast and robust optimisation method that addresses all these issues is presented.

5 Efficient Comic-Based Summarisation

The work presented here makes a shift towards more user centred summarisation and browsing of large video collections by augmenting interaction rather than learning the way users create related semantics. A number of novel approaches are introduced to the algorithm pipeline, improving the processing efficiency and quality of layout optimisation. In terms of efficiency, this approach brings real-time capability to video summarisation by introducing a solution based on dynamic programming (DP) [6] and showing that the adopted sub-optimal approach achieves nearly optimal layout results. Not only does it improve the processing time of the summarisation task, but it enables new capabilities of visualisation for large-scale video archives, such as runtime interaction, scalability, and relevance feedback. In addition, the presented algorithm applies a new approach to the estimation of key-frame sizes in the final layout by exploiting a spectral clustering methodology coupled with a specific cost function that balances between good content representability and discovery of unanticipated content. In addition, a robust unsupervised estimation of number of clusters is introduced. The evaluation results compared to existing methods of video summarisation [19] [37] showed substantial improvements in terms of algorithm efficiency, quality of optimisation, and possibility of swiftly generating much larger summaries.

5.1 Key-Frame Extraction

In order to generate the visual summary, a set of the most representative frames is generated from the analysed video sequence. Initially, video data is subsampled in both space and time to achieve real-time processing capability. Spatial complexity reduction is achieved by representing a 8×8 block with its average pixel value, generating a low-resolution representation of video frames known as the *DC sequence*. By doing this, the decoding process is minimized since the DC sequence can be efficiently extracted from an MPEG compressed video stream [38]. In the temporal dimension, key-frame candidates are determined either by uniform sampling every n^{th} frame or after a cumulative pixelwise prediction error between two adjacent candidate frames reaches a predefined threshold. The latter approach distorts the time in a non-linear fashion and thus loses the notion of real motion required by the camera work classification module. Therefore, a temporal decimation with the constant factor of $n = 5$ is applied.

Having generated the low complexity data representation with dimensions $W \times H$, a dense optical flow $\overrightarrow{F}(x,y)$ of the DC sequence is estimated efficiently using the Lucas-Kanade image registration technique [25]. In order to apply model fitting of optical flow data to *a priori* generated camera work models (i.e. *zoom*, *tilt* and *pan*), specific transformations are applied to the optical flow $F^i(x,y)$ for each frame i, as given in Eq. (1)-(4).

$$\Phi_z^i(x,y) = sgn(x - \tfrac{W}{2})F_x^i(x,y) + sgn(y - \tfrac{H}{2})F_y^i(x,y) \tag{1}$$

$$M_z^i(x,y) = \Phi_z^i(x,y) \cdot \omega(x,y) \tag{2}$$

$$M_p^i(x,y) = F_x^i(x,y) \cdot \omega(x,y) \tag{3}$$

$$M_t^i(x,y) = F_y^i(x,y) \cdot \omega(x,y) \tag{4}$$

Weighting coefficients $\omega(x,y)$ favour influence of the optical flow in image boundary regions in order to detect camera work rather than a moving object, typically positioned in the centre of the frame. As shown in Eq. (5), the weighting coefficients are calculated as an inverted ecliptic Gaussian aligned to the frame centre, with spatial variances determined empirically as $\sigma_x = 0.4 \cdot W$ and $\sigma_y = 0.4 \cdot H$.

$$\omega(x,y) = 1 - e^{-(\frac{(x-W/2)^2}{\sigma_x} + \frac{(y-H/2)^2}{\sigma_y})} \tag{5}$$

The measure of optical flow data fitness for a given camera work model is calculated as a normalised sum of $M_{cw}^{i}(x,y)$ for each type of camera work (cw): zoom (z), pan (p) and tilt (t), as given in Eq. (6). If the absolute value of fitness function becomes larger than the empirically predefined threshold Th_{cw}, the frame i is labelled with one of the six camera work categories, as given in Table 1.

Table 1 Camera work categories and corresponding error threshold values

	zoom		**pan**		**tilt**	
	in	out	left	right	up	down
Th_{cw}	< -1.2	> 1.2	< -0.7	> 0.7	< -0.8	> 0.8

$$\Psi_{cw}^{i} = \frac{1}{wh} \sum_{x=1}^{W} \sum_{y=1}^{H} M_{cw}^{i}(x,y), \text{ where } cw \in \{z,p,t\} \tag{6}$$

Finally, the binary labels of camera work classes are denoised using morphological operators retaining the persistent areas with camera motion while removing short or intermittent global motion artefacts.

Once the shot regions are labelled with appropriate camera work, only the regions with a static camera (i.e. no camera work labelled) are taken into account in selection of the most representative key-frame candidates. This approach was adopted after consulting the views of video production professionals as well as inspection of manually labelled ground truth. The conclusions were that: i) since the cameraman tends to focus on the main object of interest using a static camera, the high-level information will be conveyed by the key-frame in regions with no camera work labels, ii) chances to have artefacts like motion and out-of-focus blur are minimised in those regions.

Subsequently, frames closest to the centre of mass of the frame candidates' representation in a multidimensional feature space are specifically ranked to generate the list of region representatives. The algorithm for key-frame selection is as follows:

1. Select $N_{list} \geq N_{kf}$ candidates from static regions
2. Calculate feature matrixes for all candidates
3. Loop through all candidates
    a. Rank them by L_2 distance to all unrepresented frames of the analysed shot in ascending order
    b. Select the first candidate and label its neighbouring frames as represented
    c. Select the last candidate and label its neighbouring frames as represented
4. Export N_{kf} selected key-frames as a prioritised list

The feature vector used to represent key-frame candidates is a $18 \times 3 \times 3$ HSV colour histogram, extracted from the DC sequence representation for reasons of algorithm efficiency. As an output, the algorithm returns a sorted list of N_{kf} frames and the first frame in the list is used as the key-frame in the final video summary. In addition to the single key-frame representation, this algorithm generates a video skim for each shot in the video sequence. Depending on application type, length of the skim can be either predefined ($N_{kf} = const.$) or adaptive, driven by the number of static camera regions and maximum distance allowed during the ranking process. By alternately selecting the first and the last frame from the ranked list, a balance between the best representability and discovery of unanticipated content is achieved.

5.2 *Estimation of Frame Sizes*

Our aim is to generate an intuitive and easily-readable video summary by conveying the significance of a shot from analysed video sequences via the size of its key-frame representation. Any cost function that evaluates the significance is highly dependent upon the application. In our case, the objective is to create a summary of archived video footage for production professionals. Therefore, the summary should clearly present visual content that is dominant throughout the analysed section of the video, as well as to highlight some cutaways and unanticipated content, essential for the creative process of production.

More generally speaking, being essentially a problem of high-level understanding of any type of analysed content, the summarisation task requires a balance between the process that duly favours dominant information and the discovery of the content that is poorly, if at all, represented by the summary. Keeping this balance is important especially in case of visual summarisation, where introduction of unanticipated visual stimuli can dramatically change the conveyed meaning of represented content. In a series of experiments conducted to indicate the usefulness and effectiveness of film editing [22], Russian filmmaker Lev Kuleshov (circa 1918) demonstrated that juxtaposing an identical shot with different appendixes induces completely different meaning of the shot in audiences. In other words, the conveyed meaning is created by relation and variance between representing elements of visual content. This idea of emphasizing difference, complexity, and non-self-identity rather than favouring commonality and simplicity and seeking unifying principles is well established in linguistics and philosophy of meaning through theory of *deconstruction* , forged by French philosopher Jacques Derrida in the 1960s [13].

In the case of video summarisation, the estimation of frame importance (in our case frame size) in the final video summary layout is dependant upon the underlying structure of available content. Thus, the algorithm needs to uncover the inherent structure of the dataset and by following the discovered relations evaluate the frame importance. By balancing the two opposing representability criteria, the overall experience of visual summary and the meaning conveyed will be significantly improved.

In order to generate the cost function that represents the desired frame size in the final layout: $C(i), i = 1, \ldots, N$ where $C(i) \in [0,1]$ and N is the number of extracted key-frames for a given sequence, all key-frames are initially grouped into perceptually similar clusters . The feature vector of the i^{th} frame $x_i, i = 1, \ldots, N$ used in the process of frame grouping is a $18 \times 3 \times 3$ HSV colour histogram appended with the pixel values of the DC sequence frame representation in order to maintain essential spatial information.

Large archives of raw video footage comprise of mainly repetitive video content inseparable from a random number of visual outliers such as establishing shots and cutaways. Centeroid-based methods like K-means fail to achieve acceptable results since the number of existing clusters has to be defined *a-priori* and these algorithms break down in presence of non-linear cluster shapes [21].

Being capable of analysing inherent characteristics of the data and coping very well with high non-linearity of clusters, a *spectral clustering* approach was adopted as a method for robust frame grouping. The locally scaled affinity matrix $\mathbb{W}_{loc}^{N\times N}$, introduced by Lihi and Perona [40], is calculated as:

$$\mathbb{W}_{loc}(i,j) = e^{-\frac{|x_i - x_j|}{2\cdot\sigma_i\cdot\sigma_j}} \tag{7}$$

Each element of the data set (i.e. a key-frame) has been assigned a local scale σ_i, calculated as median of κ neighbouring distances of element i. The selection of parameter value κ is independent of the scaling parameter σ and for high-dimensional data authors in [40] recommend that $\kappa = 7$.

The major drawback of K-way spectral clustering is that the number of clusters has to be known a-priori. There have been a few algorithms proposed that estimate the number of groups by analysing eigenvalues of the affinity matrix. By analysing the ideal case of cluster separation, Ng et.al. in [29] show that the eigenvalue of the Laplacian matrix $L = D - \mathbb{W}$ with the highest intensity (in the ideal case it is 1) is repeated exactly k times, where k is a number of well separated clusters in the data. However, in the presence of noise, when clusters are not clearly separated, the eigenvalues deviate from the extreme values of 1 and 0. Thus, counting the eigenvalues that are close to 1 becomes unreliable. Based on the same idea, Polito and Perona in [32] detect a location of a drop in the magnitude of the eigenvalues in order to estimate k, but the algorithm still lacks the robustness that is required in our case.

Therefore, a novel algorithm to robustly estimate the number of clusters in the data is proposed. It follows the idea that if the clusters are well separated, there will be two groups of eigenvalues: one converging towards 1 (high values) and another towards 0 (low values). In the real case, convergence to those extreme values will deteriorate, but there will be two opposite tendencies and thus two groups in the eigenvalue set. In order to reliably separate these two groups, we have applied K-means clustering on sorted eigenvalues, where $K = 2$ and initial locations of cluster centers are set to 1 for high-value cluster and 0 to low-value cluster. After clustering, the size of a high-value cluster gives a reliable estimate of the number of clusters k in analysed dataset, as depicted in Fig. 1. This approach is similar to the automatic thresholding procedure introduced by Ridler and Calvard [33] designed to optimize the conversion of a bimodal multiple gray level picture to a binary picture. Since the bimodal tendency of the eigenvalues has been proven by Ng et.al in [29], this algorithm robustly estimates the split of the eigenvalues in an optimal fashion, regardless of the continuous nature of values in a real noisy affinity matrix (see Fig. 1).

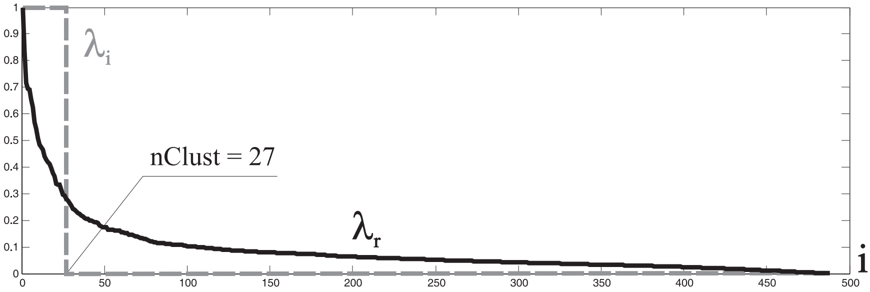


Fig. 1 Sorted eigenvalues of affinity matrix with estimated number of data clusters *nClust* in the ideal case (λ_i) and a real case (λ_r). By clustering eigenvalues in two groups, the number of eigenvalues with value 1 in the ideal case can be estimated.

Following the approach presented by Ng. et. al in [29], a Laplacian matrix $L = D - \mathbb{W}_{loc}$ is initially generated with $\widehat{\mathbb{W}}_{loc}(i,i) = 0$, where D is the degree matrix. After solving the eigen-system for all eigenvectors eV of L, the number of clusters k is estimated following the aforementioned algorithm. The first k eigenvectors $eV(i), i = 1,...,k$ form a matrix $X_{N\times k}(i,j)$. By treating each column of the row-normalised $\widehat{X}$ as a point in $\mathbb{R}^k$, N vectors are clustered into k groups using the K-means algorithm. The original point i is assigned to cluster j if the vector $\widehat{X}(i)$ was assigned to the cluster j.

To represent the dominant content in the selected section of video, the maximum cost function $C(i) = h_{max}$ is assigned to the key-frame closest to the centre of the corresponding cluster. If $d(i)$ is the i^{th} frame's distance to the central frame and σ_i is the variance of the cluster, the cost function is calculated as follows:

$$C(i) = \alpha \cdot (1 - e^{-\frac{d(i)^2}{2\sigma_i^2}}) \cdot h_{max} \tag{8}$$

Normalising $C(i)$ to the maximum row height h_{max}, scales it to the interval of frame sizes used to approximate the cost function. The parameter α controls the balance between the importance of the cluster centre and its outliers, and it is set empirically to 0.7 (see Fig. 2). As a result, cluster outliers (i.e. cutaways, establishing shots, etc.) are presented as more important and attract more attention of the user than key-frames concentrated around the cluster centre. This grouping around the cluster centres is due to common repetitions of similar content in raw video rushes, often adjacent in time. To avoid the repetition of content in the final summary, a set of similar frames is represented by a larger representative, while the others are assigned a lower cost function value.

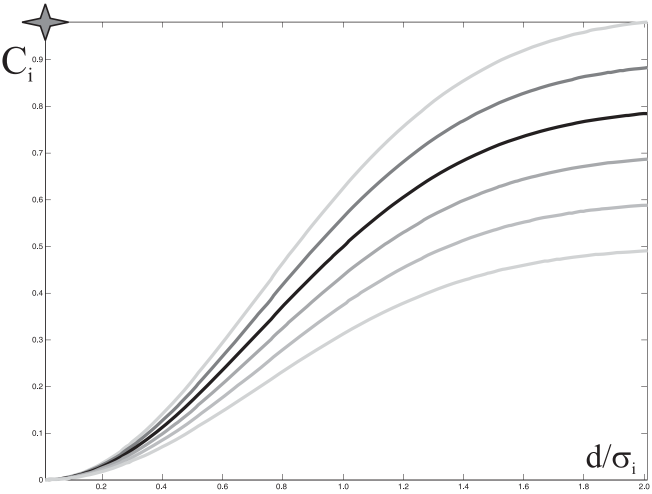


Fig. 2 Cost function dependency on distance from the cluster centre for values of parameter $\alpha \in [0.5, 1.0]$

5.3 *Summary Layout*

Given the requirement that aspect ratio of key-frames in the final layout has to be the same as aspect ratio of the source video frames, the number of possible spatial combinations of frame layouts will be restricted and the frame size ratios have to be rational numbers (e.g. 1:2, 1:3, 2:3). In addition, following the model of a typical comic strip narrative form, a constraint of spatial layout dependance on time flow is introduced. In our case, the time flow of video sequence is reflected by ordering the frames in left-to-right and top-to-bottom fashion. Excluding this rule would impede the browsing process.

Two page layout algorithms are presented. The first algorithm searches for all possible combinations of page layout and finds an optimal solution for a given cost function. However, processing time requirements make this algorithm unfeasible if the number of frames to be laid out on a single page exceeds a certain threshold. Therefore, a novel sub-optimal algorithm is introduced. It utilises dynamic programming (DP) to find the best solution in very short time. Results presented in Section 5.4 show that the error introduced by the sub-optimal model can be disregarded. Firstly, an algorithm that generates panel templates following the narrative structure of comics is presented, followed by detailed descriptions of layout algorithms.

5.3.1 Panel Generator

Following the definition of the art of comics as a sequential art [18] where space does the same as time does for film [27], this work intuitively transforms the temporal dimension of videos into spatial dimension of the final summary by following the well-known rules of comics' narrative structure.

The *panel* is a basic spatial unit of comics as a medium and it distinguishes an ordered pictorial sequence conveying information from a random set of images laid out on a page, i.e. it enables closure . Closure is a phenomenon of observing the parts and perceiving the whole. Therefore, in order to achieve an intuitive perception of the comic-like video summary as a whole, panels in the summary layout need to follow basic rules of comics' narrative structure (e.g. time flows from left to right, and from top to bottom).

Therefore, a specific algorithm that generates a set of available panel templates is developed. It creates templates as vectors of integers x_i of normalised image sizes ordered in time. Panel templates are grouped by panel heights, since all panels in a row need to have the same height. The algorithm generates all possible panel vectors $x_i, \forall h \in \{1,...,h_{max}\} \wedge w \in \{1,...,h\}$ and checks if they fit the following requirements:

1. $h \cdot w = \sum_{\forall i} x_i^2$
2. The panel cannot be divided vertically in two

The final output is a set of available panel templates for given panel heights, stored as an XML file. Examples of panel templates, for panel heights 1-4, are depicted in Fig. 3. Panelling module loads required panel templates as well as the cost function and key-frames from the database and produces a final page layout, as presented in Section 5.3.3.

5.3.2 Optimal Solution Using Full Search

In addition to the requirements for a page layout, the optimal layout solution needs to fit exactly into a predefined page width with a fixed number of images per page. This requirement enables objective comparison of layout algorithms, since the DP solution generates layout with adaptive page width and number of frames per page.

As a result of these requirements, for a given maximal row height h_{max}, a set of available panel templates is generated as described before. For a given page height h, page width w and number of images per page N, distribution of frame sizes depends on the cost function $C(i), i = 1 \ldots N$. An algorithm for calculation of the cost function is described in Section 5.2.

The main task is to find a frame layout that optimally follows the values of the cost function only using available panel templates. Each panel template generates a vector of frame sizes, that approximates the cost function values of corresponding frames. Precision of this approximation depends upon the maximum size of a frame, defined by the maximum height of the panel h_{max} which gives granularity of the solution. For a given h_{max}, a set of panel templates is generated (see Fig.3), assigning a vector of frame sizes to each template.

The page panelling algorithm is divided into two stages: i) distribution of row heights, and ii) distribution of panels for each row. Since the second stage always finds an optimal solution, the final page layout is determined by finding a minimum approximation error for a given set of row height distributions.

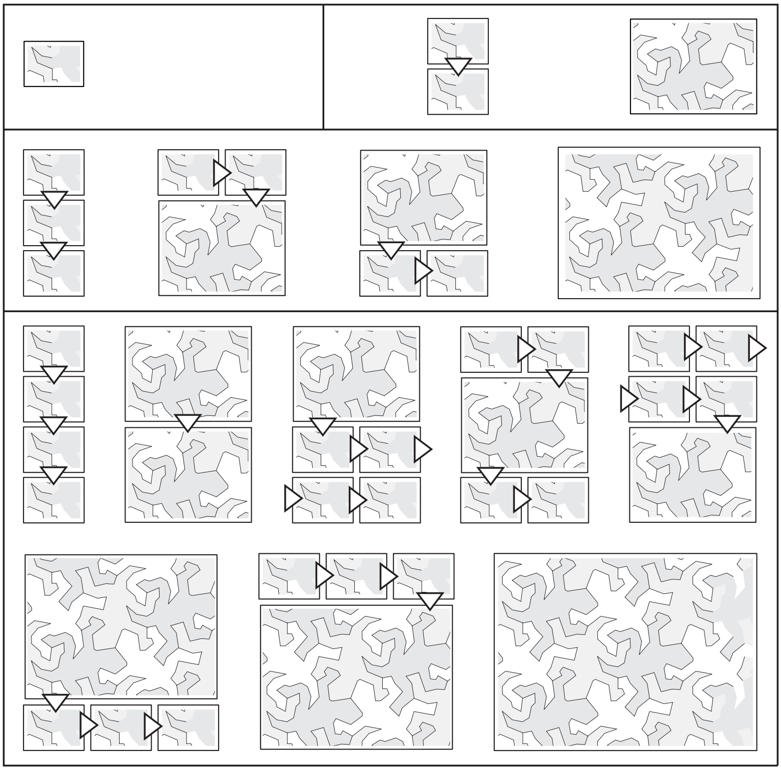


Fig. 3 Panel templates for panel heights 1 to 4. Arrows show the temporal sequence of images for each template, adopted from the narrative structure in comics.

In both parts of the algorithm, the search space is generated by the partitioning of an integer (h or w) into its summands. Since the order of the summands is relevant, it is the case of *composition* of an integer n into all possible k parts, in the form [3]:

$$n = r_1 + r_2 + \ldots + r_k, \qquad r_i \geq 0, i = 1, \ldots, k \tag{9}$$

Due to a large number of possible compositions (see Eq. (10)), an efficient iterative algorithm described in [30] is used to generate all possible solutions.

$$N_{compositions} = \binom{n+k-1}{n} \tag{10}$$

In order to find an optimal composition of page height h into k rows with heights $h(i), i = 1, \ldots, k$, for every possible $k \in [h/h_{max}, h]$, a number of frames per row $\eta(i),, i = 1, \ldots, k$ is calculated to satisfy the condition of even spread of the cost function throughout the rows:

$$\forall i, \quad \sum_{j=1}^{\eta(i)} C(j) = \frac{1}{k} \sum_{l=1}^{N} C(l) \tag{11}$$

For each distribution of rows $\eta(i), i = 1,\ldots,k$ and a given page width w, each row is laid out to minimise the difference between the achieved vector of frame sizes and the corresponding part of the cost function $C(i)$. For each composition of $\eta(i)$ a set of possible combinations of panel templates is generated. The vector of template widths used to compose a row has to fit the given composition, as well as the total number of used frames has to be $\eta(i)$. For all layouts that fulfil these conditions, the one that generates a vector of frame sizes with minimal approximation error to the corresponding part of the cost function is used to generate the row layout. The final result is the complete page layout $\Theta(i)$ with the minimal overall approximation error Δ, where Δ is calculated as given in Eq. (12).

$$\Delta = \sum_{\forall i} C(i) - \Theta(i) \tag{12}$$

5.3.3 Sub-Optimal Solution Using Dynamic Programming

There have been numerous attempts to solve the problem of discrete optimisation for spatio-temporal resources. In our case, we need to optimally utilise the available two-dimensional space given required sizes of images. However, unlike many well studied problems like stock cutting or bin packing [24] [35], there is a non-linear transformation layer of panel templates between the error function and available resources. In addition, the majority of proposed algorithms are based on heuristics and do not offer an optimal solution.

Therefore, we propose a sub-optimal solution using dynamic programming and we will show that the deviation of achieved results from the optimal solution can be practically disregarded. Dynamic programming finds an optimal solution to an optimisation problem $\min\ \varepsilon(x_1,x_2,...,x_n)$ when not all variables in the evaluation function are interrelated simultaneously:

$$\varepsilon = \varepsilon_1(x_1,x_2) + \varepsilon_2(x_2,x_3) + ... + \varepsilon_{n-1}(x_{n-1},x_n) \tag{13}$$

In this case, solution to the problem can be found as an iterative optimisation defined in Eq. (14) and Eq. (15), with initialisation $f_0(x_1) = 0$.

$$\min\ \varepsilon(x_1,x_2,...,x_n) = \min f_{n-1}(x_n) \tag{14}$$

$$f_{j-1}(x_j) = \min[f_{j-2}(x_{j-1}) + \varepsilon_{j-1}(x_{j-1},x_j)] \tag{15}$$

The adopted model claims that optimisation of the overall page layout error, given in Eq. (12), is equivalent to optimisation of the sum of independent error functions of two adjacent panels x_{j-1} and x_j, where:

$$\varepsilon_{j-1}(x_{j-1},x_j) = \sum_{i\in\{x_{j-1}\cup x_j\}} (C(i)-\Theta(i))^2 \tag{16}$$

Although the dependency between non-adjacent panels is precisely and uniquely defined through the hierarchy of the DP solution tree, strictly speaking the claim about the independence of sums from Eq. (13) is incorrect. The reason for that is a limiting factor that each row layout has to fit to required page width w, and therefore, width of the last panel in a row is directly dependent upon the sum of widths of previously used panels. If the task would have been to lay out a single row until we run out of frames, regardless of its final width, the proposed solution would be optimal. Nevertheless, by introducing specific corrections to the error function $\varepsilon_{j-1}(x_{j-1},x_j)$ the sub-optimal solution often achieves optimal results.

The proposed sub-optimal panelling algorithm comprises following procedural steps:

1. Load all available panel templates x_i
2. For each pair of adjacent panels

   a. If panel heights are not equal, penalise
   b. Determine corresponding cost function values $C(i)$
   c. Form the error function table $\varepsilon_{j-1}(x_{j-1},x_j)$ as given in Eq. (16)
   d. Find optimal $f_{j-1}(x_j)$ and save it

3. If all branches reached row width w, roll back through optimal $f_{j-1}(x_j)$ and save the row solution
4. If page height reached, display the page. Else, go to the beginning

Formulation of the error function table $\varepsilon_{j-1}(x_{j-1},x_j)$ in a specific case when panel reaches the page width w, the following corrections are introduced:

- if current width $w_{curr} > w$, penalise all but empty panels
- if current width $w_{curr} = w$, return standard error function, but set it to 0 if the panel is empty
- if current width $w_{curr} < w$, empty frames are penalised and error function is recalculated for the row resized to fit required width w, as given in Eq. (17).

$$\varepsilon_{j-1}(x_{j-1},x_j) = \sum_i (C(i) - \frac{w_{curr}}{w}\cdot\Theta(i))^2 \tag{17}$$

In this context, penalising means assigning the biggest possible error value to $\varepsilon_{j-1}(x_{j-1},x_j)$ and w is the required page width. Typically, normalised dimensions of the page, its width w and height h, are determined from the cost function and two values set by the user: expected number of frames per page $\mathscr{N}$ and page aspect ratio $\mathscr{R}$, as given in Eq. (18).

$$w = \sqrt{\frac{1}{\mathscr{R}}\sum_{i=1}^{\mathscr{N}} C(i)^2}, \quad h = \mathscr{R}\cdot w \tag{18}$$

This procedure generates a set of sequential displays without any screen size limitation. In other words, this algorithm targets application where the video summary is being displayed on a computer screen or is being printed as a page in video archive catalogue. In case of the small screen devices, such as mobile phones or PDA's, this approach is not feasible. The following section introduces an adaptation of the video summarisation algorithm to small screen displays.

5.4 Results

The experiments were conducted on the TRECVID 2006 evaluation content, provided by NIST as the benchmarking material for evaluation of video retrieval systems. In order to evaluate the results of the DP sub-optimal panelling algorithm, results are compared against the optimal solution, as described in Section 5.3. Results in Table 2 show the dependency of approximation error defined in Eq. (19) for two main algorithm parameters: maximum row height h_{max} and number of frames on a page $\mathscr{N}$.

$$\Delta = \frac{1}{\mathscr{N} \cdot h_{max}} \sqrt{\sum_{i=1}^{\mathscr{N}} (C(i) - \Theta(i))^2} \tag{19}$$

Table 2 Approximation error Δ as a function of maximum row height h_{max} and number of frames on a page $\mathscr{N}$, expressed in [%]

$h_{max} \backslash \mathscr{N}$	40	80	120	160	200	240
1	6.40	3.92	3.42	2.81	2.58	2.34
2	2.16	1.83	1.65	1.61	1.39	1.46
3	2.24	2.02	1.81	1.53	1.32	1.43
4	2.67	2.17	1.68	1.65	1.31	1.28

As expected, error generally drops as both h_{max} and $\mathscr{N}$ rise. By having more choice of combinations for panel templates with bigger h_{max}, the cost function can be approximated more accurately. In addition, the effect of higher approximation error has less impact as number of frames per page $\mathscr{N}$ rises. As we described in Section 5.3, the reason behind this phenomenon is the finite page width, that results in sub-optimal solution of the DP algorithm. On the other hand, the approximation error rises with h_{max} for lower values of $\mathscr{N}$, due to a strong boundary effect of our sub-optimal solution for small values of W.

The first three columns of Table 3 show the approximation error of the optimal method, while the other three columns show absolute difference between errors of the optimal and sub-optimal solutions. Due to the high complexity of the optimal algorithm, only page layouts with up to 120 frames per page have been calculated. The overall error due to the sub-optimal model is on average smaller than 0.5%

Table 3 Approximation error Δ using optimal algorithm for given h_{max} and $\mathcal{N}$, expressed in [%]

	$\Delta_{optimal}$			$\lvert\Delta_{DP} - \Delta_{optimal}\rvert$		
$h_{max} \setminus \mathcal{N}$	40	80	120	40	80	120
1	6.40	3.92	3.42	0.00	0.00	0.00
2	1.87	1.57	1.45	0.29	0.26	0.20
3	2.05	1.34	1.81	0.19	0.68	0.00
4	2.21	1.62	1.60	0.39	0.55	0.08

of the value of cost function. Therefore, the error can be disregarded and this result shows that the much faster sub-optimal solution achieves practically the same results with the optimal method. The optimal algorithm lays out 120 frames on a page in approximately 30 minutes, while the sub-optimal algorithm does it in a fraction of a second (see Table 4).

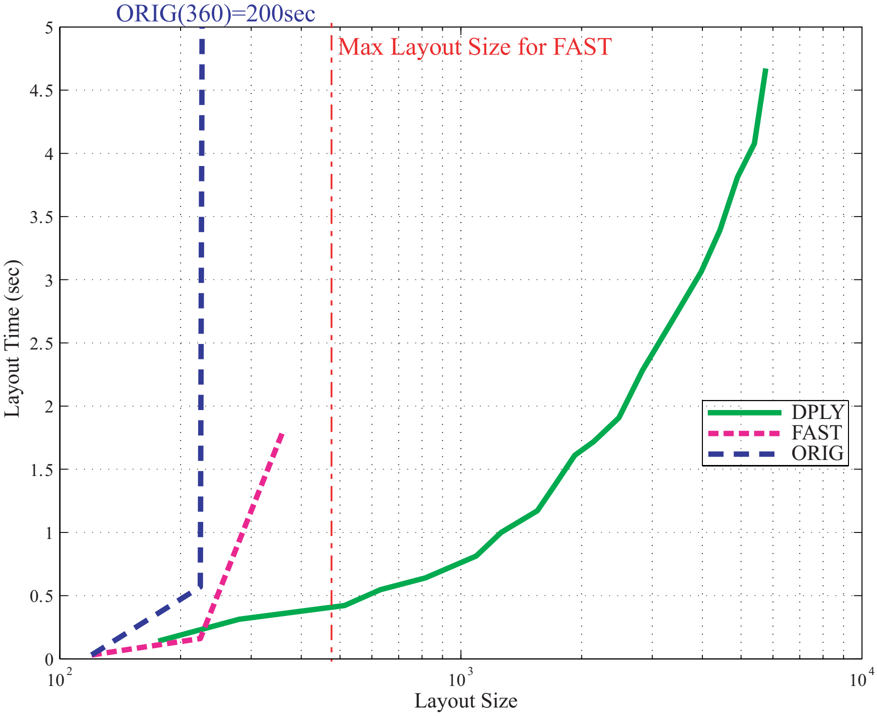


Fig. 4 Comparison of the layout algorithm speed for methods presented in [36] [ORIG], [19] [FAST] to our method [DPLY]. Linear complexity of the proposed layout algorithm is observable.

The page layout optimisation algorithm is an *NP* hard problem. Therefore, the approach presented in [36], as well as our optimal solution, regardless the speedup achieved by various heuristics [19], is not feasible for larger layouts. In [19], the authors limit the size of the final layout to 484 (22×22). The layout times for the sequence TRECVIDnews.mpg of the algorithms presented in [36] (T_{ORIG}) and [19] (T_{FAST}), compared to the proposed method (T_{DPLY}) are depicted in Fig. 4 and numerically given in Table 4.

Table 4 Comparison of layout algorithm speeds, depending upon number of frames on a page $\mathcal{N}$, page width W and height H.

$\mathcal{N}$	W	H	$W \cdot H$	T_{ORIG}	T_{FAST}	T_{DPLY}
25	12	10	120	0.03	0.03	0.127
75	16	14	224	0.57	0.16	0.241
125	20	18	360	200	1.8	0.382
150	19	27	513	×	×	0.547
1000	42	59	2478	×	×	1.907
2400	64	90	5760	×	×	4.672

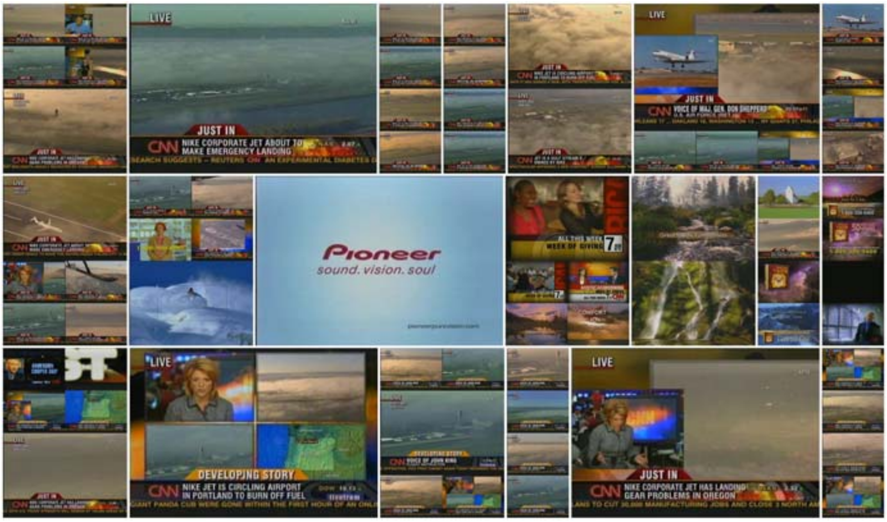

LIVE
JUST IN
CNN NIKE CORPORATE JET ABOUT TO MAKE EMERGENCY LANDING
JUST IN
CNN VOICE OF MAJ. GEN. DON SHEPPERD
Pioneer
sound. vision. soul
LIVE
DEVELOPING STORY
CNN NIKE JET IS CIRCLING AIRPORT IN PORTLAND TO BURN OFF FUEL
LIVE
JUST IN
CNN NIKE CORPORATE JET HAS LANDING GEAR PROBLEMS IN OREGON

Fig. 5 News sequence from the TRECVID 2006 search corpus, summarised using layout parameters $\mathcal{N} = 70$ and $H/W = 3/5$. Repetitive content is always presented by the smallest frames in the layout. On the other hand, outliers are presented as big (e.g. a commercial break within a newscast, row 2 frame 11) which is very helpful for the user to swiftly uncover the structure of the presented sequence.

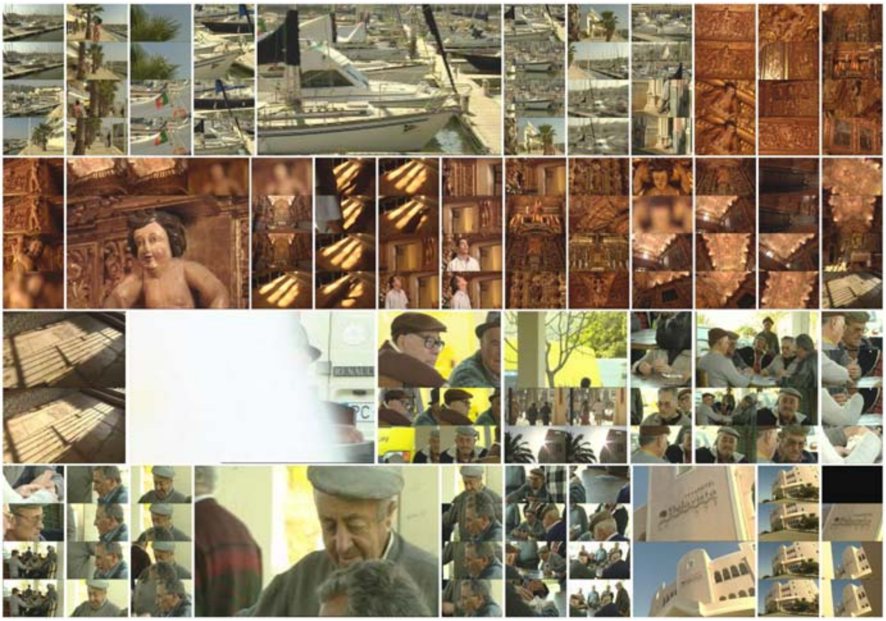


Fig. 6 A sequence from the TRECVID 2006 rushes corpus, summarised using layout parameters $\mathscr{N} = 150$ and $H/W = 1$. Since there is a lot of repetition of the content, this type of data fully exploits functionality of the presented system: the largest frames represent the most frequent content and in some cases extreme outliers (e.g. a capture error due to an obstacle in row 3, frame 3); middle sized frames represent similar, but a bit different content to the group represented by the largest frames; the smallest frames are simple repetitions of the the content represented by the largest frames.

Examples of three contrasting content types, news broadcast and rushes, from the TRECVID corpus, as well as an edited TV show, are presented in Fig. 5, Fig. 6 and Fig. 7 respectively. The news sequence is summarised using layout parameters $\mathscr{N} = 70$ and $H/W = 3/5$. It can be observed that the repetitive content is always presented by the smallest frames in the layout. On the other hand, outliers are presented as big (e.g. a commercial break within a newscast, row 2 frame 11) which is very helpful for the user to swiftly uncover the structure of the presented sequence. Finally, a sequence from the TRECVID 2006 rushes corpus is summarised using layout parameters $\mathscr{N} = 150$ and $H/W = 1$. Since there is a lot of repetition of the content, this type of data fully exploits the functionality of the presented system. The largest frames represent the most frequent content and in some cases extreme outliers (e.g. a capture error due to an obstacle in row 3, frame 3); middle sized frames represent slightly different content to the the largest frames, while the smallest frames are simple repetitions of the the content represented by the largest frames.

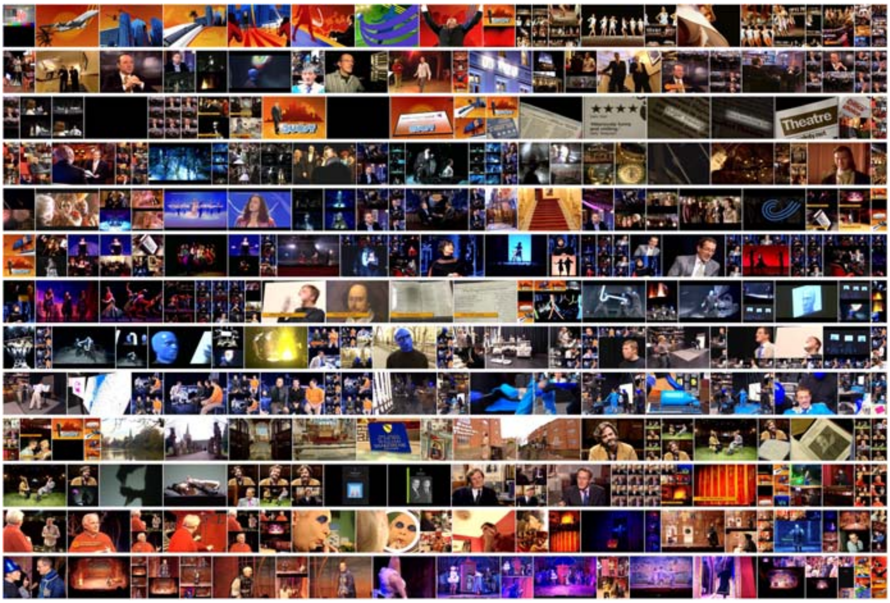


Fig. 7 An edited material comprising 662 shots, summarised using proposed method. At one key frame per shot, this summary represents the whole 1:40h long show and conveys the temporal structure at the same time.

References

[1] Adams, B.: Where does computational media aesthetics fit? IEEE Multimedia Magazine, spec. ed. Computational Media Aesthetics (2003)

[2] Anderson, S.: Select and combine: The rise of database narratives. Res Magazine 7(1), 52–53 (2004)

[3] Andrews, G.E.: The theory of partitions. In: Encyclopedia of mathematics and its applications, vol. 2. Addison-Wesley, Reading (1976)

[4] Associates, C.: Consumer survey on digital storage in consumer electronics 2008 (2007), `http://www.researchandmarkets.com/reports/648703`

[5] Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D.M., Jordan, M.I.: Matching words and pictures. J. Mach. Learn. Res. 3, 1107–1135 (2003)

[6] Bellman, R.E., Dreyfus, S.E.: Applied Dynamic Programming. Princeton University Press, Princeton (1962)

[7] Bertini, M., Del Bimbo, A., Nunziati, W.: Highlights modeling and detection in sports videos. Pattern Analysis and Applications (2005)

[8] Chong-Wah, N., Yu-Fei, M., Hong-Jiang, Z.: Video summarization and scene detection by graph modeling. IEEE Transactions on Circuits and Systems for Video Technology 15(2), 296–305 (2005)

[9] Davenport, G., Smith, T.A., Pincever, N.: Cinematic primitives for multimedia. IEEE Comput. Graph. Appl. 11(4), 67–74 (1991), `http://dx.doi.org/10.1109/38.126883`

[10] Davis, M.: Media streams: representing video for retrieval and repurposing. In: MULTIMEDIA 1994: Proceedings of the second ACM international conference on Multimedia, pp. 478–479. ACM Press, New York (1994), http://doi.acm.org/10.1145/192593.197412

[11] Davis, M.: Media streams: representing video for retrieval and repurposing. Ph.D. thesis, Cambridge, MA, USA (1995)

[12] DeMenthon, D., Kobla, V., Doermann, D.: Video summarization by curve simplification. In: MULTIMEDIA 1998: Proceedings of the sixth ACM international conference on Multimedia, pp. 211–218. ACM Press, New York (1998)

[13] Derrida, J.: Of grammatology. Johns Hopkins University Press, Baltimore (1997)

[14] Dony, R., Mateer, J., Robinson, J.: Techniques for automated reverse storyboarding. Vision, Image and Signal Processing, IEE Proceedings 152(4), 425–436 (2005)

[15] Dorado, A., Calic, J., Izquierdo, E.: A rule-based video annotation system. IEEE Transactions on Circuits and Systems for Video Technology 14(5), 622–633 (2004)

[16] Dorai, C., Venkatesh, S.: Media computing: computational media aesthetics. In: The Kluwer international series in video computing. Kluwer Academic Publishers, Boston (2002)

[17] Dorai, C., Venkatesh, S.: Bridging the semantic gap with computational media aesthetics. IEEE Multimedia 10, 15–17 (2003)

[18] Eisner, W.: Comics and sequential art. Poorhouse, Tamarac, Fla., Great Britain (2001)

[19] Girgensohn, A.: A fast layout algorithm for visual video summaries. In: Proceedings. 2003 International Conference on Multimedia and Expo, 2003. ICME 2003, vol. 2, pp. 77–80 (2003)

[20] Hanjalic, A., HongJiang, Z.: An integrated scheme for automated video abstraction based on unsupervised cluster-validity analysis. IEEE Transactions on Circuits and Systems for Video Technology 9(8), 1280–1289 (1999)

[21] Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. ACM Computing Surveys 31(3), 264–323 (1999)

[22] Kuleshov, L., Levaco, R.: Kuleshov on film: writings by Lev Kuleshov. University of California Press, Berkeley (1974)

[23] Leonardi, R., Migliorati, P., Prandini, M.: Semantic indexing of soccer audio-visual sequences: a multimodal approach based on controlled Markov chains. IEEE Transactions on Circuits and Systems for Video Technology 14(5), 634 (2004)

[24] Lodi, A., Martello, S., Monaci, M.: Two-dimensional packing problems: A survey. European Journal of Operational Research 141(2), 241–252 (2002)

[25] Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI 1981), pp. 674–679 (1981)

[26] Manovich, L.: The language of new media. Leonardo. MIT Press, Cambridge (2001)

[27] McCloud, S.: Understanding Comics. Tundra Publishing Ltd, Northhampton (1993)

[28] Mccloud, S.: Understanding Comics. HarperPerennial, New York (1994)

[29] Ng, A.Y., Jordan, M.I., Weiss, Y.: On spectral clustering: Analysis and an algorithm, pp. 849–856 (2002)

[30] Nijenhuis, A., Wilf, H.S.: Combinatorial algorithms: for computers and calculators, 2nd edn. Computer science and applied mathematics. Academic Press, New York (1978)

[31] Ofcom: The communications market 2008 (2008), http://www.ofcom.org.uk/research/cm/cmr08/

[32] Polito, M., Perona, P.: Grouping and dimensionality reduction by locally linear embedding, pp. 1255–1262 (2002)

[33] Ridler, T., Calvard, S.: Picture thresholding using an iterative selection method. IEEE SMC 8(8), 629–632 (1978)

[34] Snoek, C., Worring, M.: Multimodal video indexing: A review of the state-of-the-art. Multimedia Tools and Applications 25(1), 5–35 (2005)

[35] Sweeney, P.E., Paternoster, E.R.: Cutting and packing problems: a categorized, application-oriented research bibliography. J. Operational Research Society 43(7), 691–706 (1992)

[36] Uchihashi, S., Foote, J.: Summarizing video using a shot importance measure and a frame-packing algorithm. In: Proceedings. 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 1999, vol. 6, pp. 3041–3044 (1999)

[37] Uchihashi, S., Foote, J., Girgensohn, A., Boreczky, J.: Video manga: generating semantically meaningful video summaries. In: MULTIMEDIA 1999: Proceedings of the seventh ACM international conference on Multimedia, pp. 383–392. ACM Press, New York (1999)

[38] Yeo, B.L., Liu, B.: Rapid scene analysis on compressed video. IEEE Transactions on Circuits and Systems for Video Technology 5(6), 533–544 (1995)

[39] Yueting, Z., Yong, R., Huang, T.S., Mehrotra, S.: Adaptive key frame extraction using unsupervised clustering. In: Proceedings. 1998 International Conference on Image Processing, ICIP 1998, vol. 1, pp. 866–870 (1998)

[40] Zelnik-Manor, L., Perona, P.: Self-tuning spectral clustering, pp. 1601–1608 (2005)

Multi-dimensional BPTs for Content Retrieval

Shirin Ghanbari, John C. Woods, and Simon M. Lucas

Abstract. This chapter documents the use of region image based analysis within binary partition trees to produce meaningful segmentations suitable for MPEG-7 description. An image is pre-segmented into a large number of homogenous regions which are subsequently merged according to their similarity with the process documented using a binary partition tree. The trees are derived using multi-dimensional descriptors such as colour, and texture. The correlations between these domains are studied leading to a tree where objects are constrained to individual branches rather than being fragmented. The tree is then pruned to retain meaningful areas leading to high density semantic image segments that are highly valuable in other applications such as in content retrieval systems. Object nodes generated from the BPT are indexed and matched through a combination of descriptors. Results justify the use of segmentation within retrieval systems. Based on MPEG-7 colour descriptors of colour, texture and edge histograms, a semi-automatic segment based image retrieval is studied.

1 Introduction

When observing a scene, the human visual system segments the view into a series of discrete objects. This process is so efficient that a person is not usually confused, and observes a series of familiar objects. The precise mechanism the brain is employing remains a mystery but it is clear that the boundary condition makes a large contribution to cognition. To attempt this, images must be first segmented. Image segmentation refers to the process of partitioning a digital image into multiple regions (sets of pixels) that represents an image into something that is more meaningful and easier to analyze. Each of the pixels in a particular region is similar (homogenous) with respect to some characteristic or computed property, such as colour, intensity, or texture. In other-words, adjacent regions are significantly different with respect to the same characteristic(s).

A proposed methodology is in the segmentation of images into 'identifiable' regions, using multiple descriptors to be represented using Binary Partition Trees (BPT) [1]. Essentially the BPT allows a database representation of images, documenting the merging order of homogeneous regions for manual or automatic

Shirin Ghanbari, John C. Woods, and Simon M. Lucas
School of Computer Science and Electronic Engineering, University of Essex, UK
e-mail: sghanb@essex.ac.uk

M. Grgic et al. (Eds.): Rec. Advan. in Mult. Sig. Process. and Commun., SCI 231, pp. 73–90.
springerlink.com

retrieval. Leaf nodes represent the initial image partition, with the remaining nodes created by merging two child nodes to form a parent node according to their similarity. With this structure the tree can represent a set of regions at different scales of resolution. The generation of BPT can be considered in different forms. This chapter provides a multi-dimensional approach where multi descriptors are combined to define nodes or regions of the BPT.

The proposed multi-dimension BPT generates high density partial objects with clearly defined boundaries using a combination of colour and spatial frequency. These complete objects are now ready for further processing for a range of applications. In particular, content based retrieval systems can truly benefit from segmentation and can be a key pre-requisite.

In this chapter the effectiveness of multiple descriptors in describing content; be it entire images or smaller regions is presented along with the justification of how segmentation can considerably enhance content retrieval systems in the form of a semi-automatic tool.

2 Image Segmentation

The history of segmentation of digital images can be traced back 40 years. Since then, this field has evolved very quickly and has undergone great change [2]. Image segmentation is often described as the process that sub-divides or partitions images into non-overlapping regions or areas that are homogenous with respect to some characteristic (colour or texture) [3]. The level to which the subdivision is carried depends on the problem being solved.

It can be used for analysis of the image and/or for further processing of the image. Some of the practical applications of image segmentation include machine vision, medical imaging, face recognition and even traffic control systems. One of the common problems encountered in image segmentation is choosing a suitable approach for isolating different objects from their background. Generally, difficulty arises when we encounter occlusion, uneven brightness and when images are overwhelmed with irrelevant detail.

Image segmentation can be approached from different philosophical perspectives; boundary-based and region-based methods. The first approach is based on discontinuity and partitions an image by detecting isolated points, lines and edges according to abrupt changes of local properties. Usual techniques are edge detectors and energy minimization approaches (i.e. snake model). The algorithms from the second approach exploit the homogeneity of spatially dense information (e.g. colour, texture properties, etc.) to produce the segmented image, and includes thresholding, clustering, region-growing [4].

Thresholding is computationally cheap and fast (especially if repeated on similar images) and is the oldest segmentation method and is still widely used in simple applications. However, there are no guarantees of object coherency, for example there may be the presence of extraneous pixels. For simple noise-free data, detection of edges usually results in straightforward region boundary delineation. However, edge detection on noisy, complex image data often produces missing edges and extra edges that cause the detected boundaries to not necessarily form a set of close connected curves that surround connected regions. Region growing approaches to segmentation are preferred here because region growing exploits spatial information and guarantees the formation of closed, connected regions.

In region based methodologies two techniques can be employed either a top-down or bottom-up approach; split-and-merge and region growing. Typical split-and-merge techniques [5] consist of two basic steps and are based on a quad tree data representation. First, the whole image is considered as one region. If this region does not satisfy a homogeneity criterion the region is split into four quadrants (sub-regions) and each quadrant is tested in the same way; this process is recursively repeated until every square region created in this way contains homogeneous pixels. In the second step, all adjacent regions with similar attributes may be merged following other (or the same) criteria. The criterion of homogeneity is generally based on the analysis of the chromatic characteristics of the region.

In the bottom-up approach, region growing, based on the growth of a region, starts from an initial, incomplete segmentation and attempts to aggregate the yet unlabelled pixels to one of the given regions. The initial regions are usually called seed regions or seeds. Adjacent pixels are iteratively examined and added to the region if they are sufficiently similar to that region. Otherwise if a pixel is deemed as too dissimilar to the current region, it is used to start a new region. Several advantages have been noted; firstly borders of regions are found to be thin since pixels are only added to the exterior of a region and connected. Also, it is said to be stable under noise, as a properly defined region will never contain too much of the background. Interestingly for my research is that membership of a region can be based upon multiple criteria; allowing for the use of multiple image properties at the same time (such as colour intensity and gradient level).

Despite the numerous segmentation techniques, image segmentation is still a subject of on-going investigations and it has not been conclusively stated that the segmentation problem has been solved because of the application's diversity. This is merely because one 'persons' perception of a segment/object may not be that of others. Hence, allowing the presence of different resolutions of segments can considerably aid this situation. A region based methodology, Binary Partition Trees, provides such means and has shown to output 'useful' segmentations by partitioning images into homogenous regions.

3 Creation of Binary Partition Tree

The Binary Partition Tree should be created in such a way that the most "interesting" or "useful" regions are represented [1]. Essentially it requires the labelling of nodes and its computation consists generally of two stages. Firstly, rather than working with millions of pixels, a pre-segmentation stage is required resulting in an initial partition of the image in the form of small homogenous regions. These numerous regions make browsing of individual leaf nodes extremely difficult and regions often have no semantic meaning. However, this can be rectified in the second stage, region merging, where candidate region segments from the first stage are merged continuously until a single region (following a homogeneity criterion) is obtained.

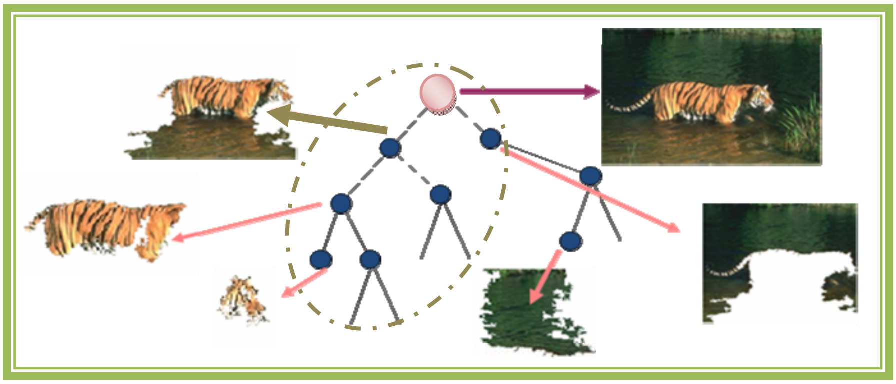
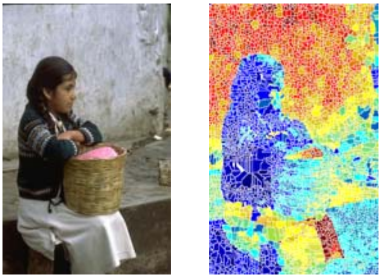


Fig. 1 Binary Partition Tree

3.1 Pre-segmentation

To obtain initial regions (seeds), images are pre-segmented or partitioned generally using the Watershed [6] transform. The transform is applied to a gradient magnitude image, resulting in watershed lines that divide the image plane into regions associated with regional minima; or partitions of an image into non-overlapping regions.

Fig. 2 Watershed Transform



Although, the Watershed transform is an unsupervised method it is highly sensitive to gradient noise resulting in over-segmentation; even small noise can make a homogenous region 'break' into lots small watersheds. However, compared to other pre-segmentation techniques (K-means clustering and mean shift algorithm) its regions preserve detailed object boundaries, which is our main objective in the generation of a BPT. Figure 2 illustrates the labelling of the Watershed transform (computed using Matlab), and clearly highlights the formation of numerous small regions that have no semantic meaning, other than homogeneity of a metric such as colour. Each region is subsequently assigned to a unique label and mapped onto a Region Adjacency Graph (RAG).

The RAG encodes neighbourhood information; information regarding the order in which regions can be merged (figure 3). In essence a RAG is a graph which is constituted by a set of nodes representing regions of the space and a set of links connecting two spatially neighbouring nodes. Computation of the RAG using the labels of the regions makes subsequent merging stages possible and provides a complete description of all the neighbourhoods [7]. Consequently, the RAG is constructed based on an initial partition, where each region of the partition image is associated to the nodes of the graph. Here two nodes are connected if their associated regions are neighbours in the partition image.

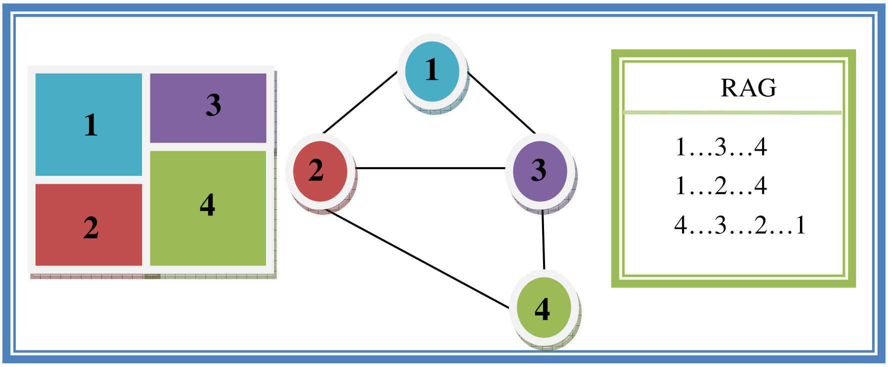


Fig. 3 Region Adjacency Graph

3.2 *Region Merging*

Merging is required to eliminate very small regions into larger with possibly semantic meaning. This stage requires e examination of statistics of adjacent regions. If similar (according to a criterion) they may be invited to be merged. A merging approach is specified by three notions [1]:

- The region model defines how regions are represented and contains region statistics from which the similarities between regions are calculated.
- The merging order specifies the order in which the regions are processed and is defined by a similarity measure between regions.
- The merging criterion decides if the merging has to be done or not.

An example of region merging is shown in figure 4. The original partition involves four regions. The regions are indicated by a number. The algorithm merges the four regions in three steps. In the first step, the pair of most similar regions, 1 and 3, are merged to create region 5. Then, region 5 is merged with region 2 to create region 6. Finally, region 6 is merged with region 4 generating region 7, which corresponds to the entire image. As can be seen, the merging is performed iteratively between pairs of regions. The resulting tree is thereby binary.

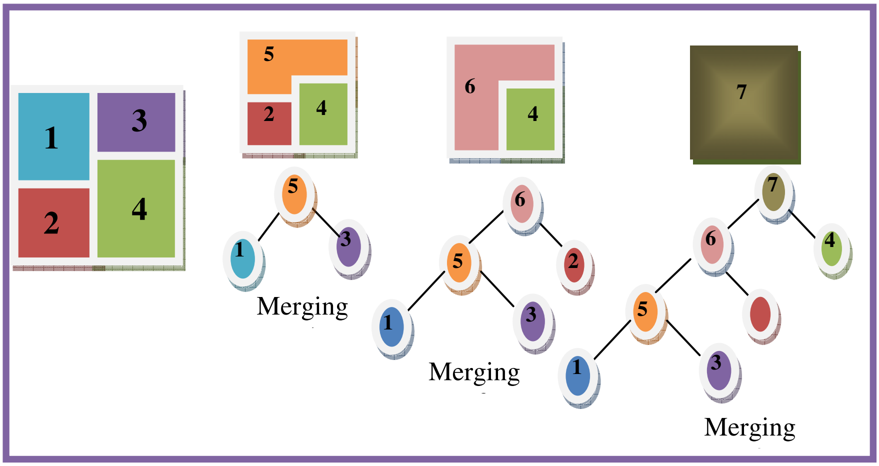


Fig. 4 BPT with region merging algorithm

The result of a region merging process depends on the order in which regions are merged. The merging criterion simply states that the merging continues until a termination criterion is reached; for example target number of regions. At each step the algorithm looks for the pair of most similar regions. In describing regions, descriptors from the MPEG-7 standard [8] are recommended.

Formally referenced as the 'Multimedia Content Description Interface', MPEG-7 unlike its predecessors focuses on description of multimedia content. Its ultimate goal is allowing interoperable searching, indexing, filtering and access of audio-visual (AV) content, by allowing interoperability among devices and applications with AV content descriptions [8]. MPEG-7 Visual Descriptors describe AV media content on the basis of visual information. For images the content may be described by the shape of objects, texture, and colour or simply by object size.

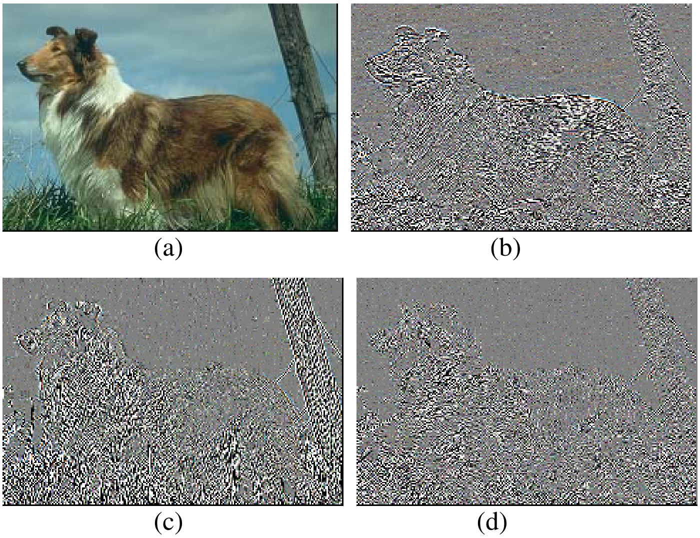
3.2.1 Colour Model

Colour is the most commonly employed low-level descriptor in BPTs [1], both for its simplicity and results. It is clear that the human eye responds more quickly to what is happening in a scene if it is in colour. Colour is helpful in making many objects ‘stand out’. Therefore, appropriate use of colour can significantly extend the capabilities of a vision system. What needs to be investigated is in finding the colour space which is most suitable for this particular segmentation technique. Some are hardware-oriented (e.g., RGB, and CMY colour space), as they were defined by taking into account properties of the devices used to reproduce colours. Others are user-inspired (e.g., L*u*v*, L*a*b*) as they were defined to quantify colour differences as perceived by humans. With this in mind the CIE $L*a*b*$ colour space is implemented accordingly.

3.2.2 Texture Model

Segmentation results can be improved through the inclusion of other descriptors such as texture. Texture features represent visual patterns with homogeneous properties in an image. Textual properties of the image can be extracted using statistical features, spatial-frequency models, stochastic models, and so forth. However, the most commonly used feature for texture analysis in the wavelet domain is the energy of its sub-band coefficients.

Wavelets are a mathematical tool for hierarchically decomposing functions in the frequency domain whilst at the same time preserving the spatial domain locality. Using a four band wavelet, an image pyramid can be produced representing entropy levels for each frequency band. The transform is successively applied along the rows and columns of the image; resulting in the decomposition into four distinct sub-bands. The image can now be considered as its constituent parts: the lowest frequency, the Low Low (LL) band and the higher-frequency sub-bands, that is Low High (LH), High Low (HL) and High High (HH). It is within the higher spatial frequency sub-bands in the horizontal, vertical and diagonal directions that the spatial information used in this work is extracted, as it is here that finer detail can be found.



(a) (b)

(c) (d)

Fig. 5 Wavelet Transform

Figure 5, illustrates the four bands where clearly in the LL band the majority of the image content has been persevered. The finer detail manifests in the other remaining bands. The high frequency bands are shown in gray-scale and have been magnified by a factor of 16 for visualization purposes. From the sub-bands we can locate areas having significant levels of texture, for example the dog's fur (Figure 5b), the foliage and the tree in the background (Figure 5c).

3.2.3 Colour-Texture Model

It is often difficult to obtain satisfactory results when only using a single descriptor for segmentation of complex images such as outdoor and natural images, which involves additional difficulties due to effects such as shading, highlights, non-uniform illumination or texture. For example, imagine a photograph of a tiger in the jungle; there is a wide variation in the levels of illumination due to shadowing and the animal may be indistinguishable from the background due to camouflage. With this in mind we propose a model that combines colour and texture information.

We construct a colour based binary tree directly from the LL band. The algorithm begins by generating a Watershed based pre-segmentation of the LL band which becomes the bottom of the tree forming the leaf nodes. The merging order between two adjacent regions is then calculated in two steps. Firstly, the mean colour distance, C_{ij}, between two adjacent region nodes i and j is evaluated as in [9], for the LL band.

> Let R_i denote the ith region in the tree. The region model $M(R_i)$ is defined as the average colour of the region in the CIE $L*a*b*$ colour space. The CIE space is an arbitrary selection; similar results are reported for other representations. The merging order (cost) C for a pair of adjacent regions R_i and R_j is defined as:
>
> $$C(R_i.R_j) = N(R_i\left|M(R_i) - M(R_i \cup R_j)\right|_2 + N(R_j\left|M(R_j) - M(R_i \cup R_j)\right|_2 \qquad (1)$$
>
> where $N(R_i)$ is the area of region R_i and $|.|_2$ denotes the L2-norm.

Then the regions texture energies are considered, by examining the region's corresponding LH, and HL bands energies with the objective of grouping regions with similar energy. It should be noted that the HH band can be avoided here due to their low energy and their sensitivity to noise. The total energy in region i is defined as:

$$E_i = e_{LHi} + e_{HLi} \qquad (2)$$

where e is the energy contained in one of the wavelet sub-bands. From this the merging order D, for the pair of adjacent regions i and j is defined as:

$$D_{ij} = C_{ij} \times \left|E_i - E_j\right| \qquad (3)$$

where $\left|E_i - E_j\right|$ is the L1 norm texture differences between adjacent regions. Equations 2 and 3 are slight variations of that in [10] and [11] but has the advantage of being threshold free.

3.3 Analysis

When observing a scene, the human visual system segments the view into a series of discrete objects, and places emphasis on certain objects. However, recognized by many ([12], [13], and [14]) the ability to quantify this in a general way is an ill posed problem. In the absence of a specific application requirement, we expect the segmentation of an image to agree with that performed by own vision system. In this work we evaluate our methodology using an empiral discrepancy approach, comparing results with the ground truth obtained from object boundary. The best match (image segment) is the single node within the tree which best approximates the content inside the boundary. The segmentation is then quantified according to the percentage of pixels inside and outside the mask of the ground truth. Here, we term these measures as intrinsic I and extrinsic E values; representing under and over segmentation respectively [11]:

$$I = 100 * \text{size}(\text{intersection}(g,s)) / \text{size}(g) \tag{4}$$

$$E = 100 * \text{size}(\text{intersect}(g,s)) / \text{size}(s) \tag{5}$$

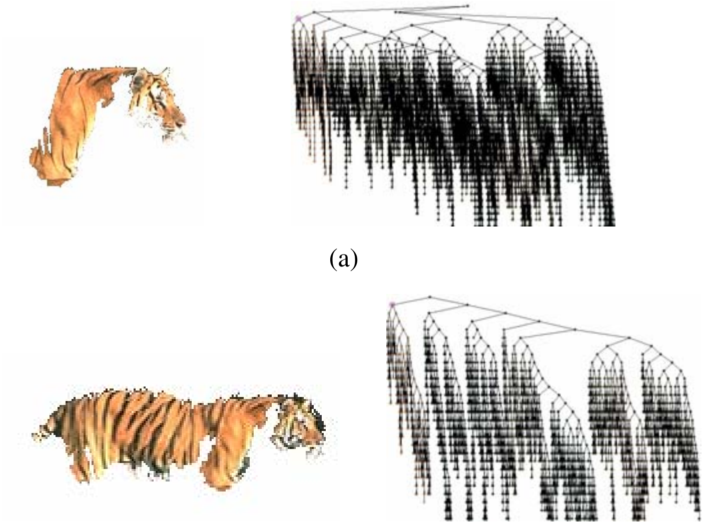
where g is set of ground truth pixels, and s is the set of selected pixels at this node.

Consider figure 6, from the Berkeley Segmentation Dataset (BSD) [15], one would be inclined to separate the animal (tiger) from its background. Figure 6a shows the BPT produced using a colour only model [9]. The BPT is clearly very dense and difficult for a user to browse, where there are even branches crossing over one another. When browsing this tree no single node is evident which represents the tiger, it is fragmented and requires more than one node/branch to represent it.

Figure 6b shows the application of the colour-texture methodology. Note that the BPT is much less dense and more amenable to browsing. A single branch/node can be found which in this case produces a more holistic segmentation of the tiger. The segmentation produced is robust and this type of behaviour is observed to be general. A good segmentation will be one with a high intrinsic score accompanied by a low extrinsic score; the ground-truth is ‘coloured-in’ with very little outside the boundary. When searching for an optimum node, if a choice is made too low in the tree a low intrinsic value will result and therefore a fragmented segmentation. If a choice is made too high in the tree the intrinsic value will be high, but so will the extrinsic. A compromise between these two extremes can be quantitatively derived.

It is widely accepted [16] that image segmentation is complicated by a number of factors including occlusion, background noise and changes in illumination. The ability of the two methodologies, colour and colour-texture, to cope with these impediments is examined in the following experiments: table 3 shows the images and their associated segmentations using the two methods, and a summary of the statistics is presented in tables 1 and 2 for images from the BSD test set.

In image 157055, we present a colour-texture segment that clearly defines a female figure wearing textured clothing. Although, image number 86016 contains only a single colour (green and distinctly different from the background) the colour-texture has matched the performance of the colour only. Images 109053 and 134035 present rather complicated natural images, as the target objects (animals)


(a)

(b)

Fig. 6 BPT Comparison (a) Colour based BPT, (b) Colour-Texture based BPT

are camouflaged and have lighting variations. The colour method is unable to segment these types of objects, but higher density segmentations with improved intrinsic values are returned using the texture-colour method. Even though the image's background and foreground are highly textured the colour-texture method has been able to separate these regions. The Black Panther in image no. 304034 should be suitable for colour only segmentation but lighting variation has prevented the recovery of the complete object. Closer inspection shows the colour method has failed to include the man's clothing (189080). Clearly, image 108005 is camouflaged but the colour-texture has provided a 'better' segmentation without including too much of its background (reinforced by table 1).

Whilst a general trend for improvement using the colour-texture approach is apparent, there are a few anomalies which invite comment. In image no. 42049 the colour method has produced a higher intrinsic value than the colour-texture method, but has done so at the expense of including more than double the extrinsic content. In image no. 69015 the koala has been effectively segmented into a meaningful object by the colour-texture approach but carries the penalty of a higher intrinsic value. When locating the optimum node/branch for the segmentation, perfect intrinsic and extrinsic scores are seldom reported: too high in the tree and the extrinsic value will be high and likewise too low in the tree and the intrinsic value will be low.

If it is assumed the optimum node has been found it becomes necessary to analyze the amount of intrinsic and extrinsic content to decide which has performed the best; colour or colour-texture. Qualitatively this is a difficult thing to do, so the

Table 1 Intrinsic and extrinsic values of representative segmentations from the Berkeley BSD test set

Image No.	Colour (Intrinsic %)	Colour-Texture (Intrinsic %)	Colour (Extrinsic %)	Colour-Texture (Extrinsic %)
42049	99.03	88.52	58.72	15.94
69015	74.23	80.5	2.91	4.15
78004	57.42	75.48	0.33	0.61
85048	68.52	75.26	1.30	7.42
86016	98.08	97.98	1.70	5.14
109053	52.55	55.82	13.29	3.85
134035	38.11	78.76	12.64	11.15
189080	86.49	94.37	0.55	2.71
291000	47.51	80.41	0.21	7.02
302008	49.94	58.09	0.24	2.32
304034	79.33	89.93	4.02	10.22
376043	48.06	55.67	0.31	1.09
Average	66.61	78.81	8.02	7.03

results are combined into a single metric of segmentation quality: the bi-trinsic quality measure:

$$\text{Bi-trinsic} = \text{Intrinsic x } (100 - \text{Extrinsic})/100 \qquad (6)$$

This single value is computed for all test cases and shown in table 2, illustrating significant overall gain for the proposed colour-texture approach.

Table 2 Segmentation Quality

Image No.	Colour(Bi-trinsic %)	Colour-Texture(Bi-trinsic %)
42049	40.88	74.40
69015	72.70	77.16
78004	57.23	75.02
85048	67.63	69.68
86016	96.41	92.94
109053	45.57	53.67
134035	33.29	69.98
189080	86.01	91.81
291000	47.41	74.77
302008	49.82	56.74
304034	76.14	80.74
376043	47.91	55.06
Average	60.08	72.66

Looking at this table one can conclusively see the link between the bi-trinsic values and segmentation quality. For example, the bi-trinsic values for images 109053 and 376043 are quite low and have resulted in correspondingly ‘fragmented’ images (table 3). As seen previously for, image no. 42049, the colour method has performed well, but due to its high extrinsic value it has caused its bi-trinsic value to be considerably lower than the colour-texture method. The koala segmented by the colour-texture method has a higher extrinsic value, but due to its superior intrinsic value, its bi-trinsic value is still higher.

Table 3 Segmentations generated from colour and colour-texture based BPTs

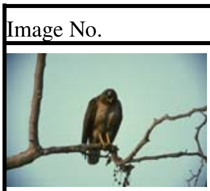
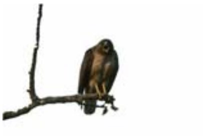
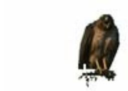
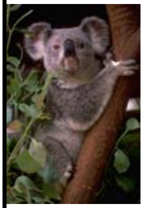
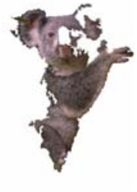
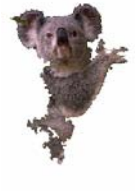
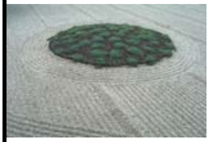
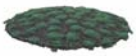
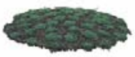
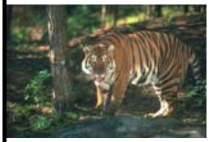
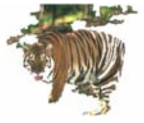
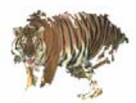
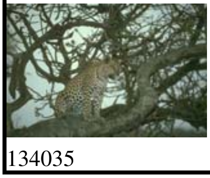
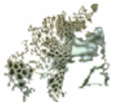
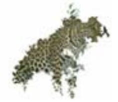
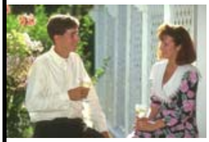
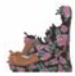
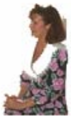
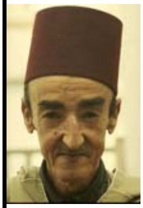
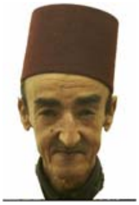
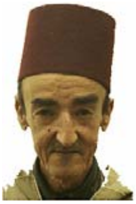
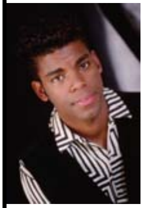
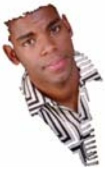
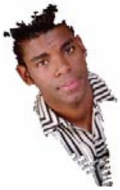
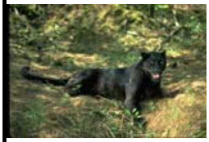
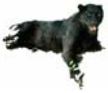
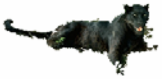
Image No.	Colour	Colour-Texture
 42049		
 69015		
 86016		
 108005		
 134035		

Table 3 (*continued*)

 157055		
 189080		
 302008		
 304034		

This work has shown that the combination of low-level descriptors can significantly aid in computer vision computation in particular within segmentation. Each descriptor provides their own unique characteristic aiding to represent objects that have semantic meaning. Returning such object segments with high density can be highly valuable for a range of applications including content-based image retrieval systems.

4 Case Study: Content Retrieval

Even with the amount of activity achieved for content retrieval systems [17] there still remain problems unsolved. In particular, the so-called semantic gap; "… the lack of coincidence between the information that one can extract from visual data and the interpretation that the same data have for a user in a given situation" [18]. In such tasks man has traditionally outperformed machines, whereby "while text is man's creation, images are a replica of what man has seen since birth" [19]. This notion with the theory that the human vision system has evolved over the years, naturally makes the interpretation of what one sees hard to characterize, and even harder to teach a machine.

Providing users with informed options and assisting in the preparation of the query dramatically increases the chances of the system returning relevant results. Segmentation can be classified as a pre-requisite for CBIR systems. BPT's provides the basis for such an opportunity; where as described users have the opportunity of interacting with nodes within the BPT to select regions of the image that most suit their query. For example a user may have a picture of a tiger but may only wish to retrieve foliage; they can do so by selecting one of the nodes which may represent the background image. From the query node a search or retrieval procedure is instantiated returning the best match representing the root node of the BPT. In this manner the user can refine an inquiry; if the first rank ordered candidate set is unacceptable a further node can be selected to refine the search.

4.1 Node Descriptors

In representing or describing nodes within the BPT, a combination of MPEG-7 descriptors can be employed. The histogram is the most commonly used structure in representing the global composition of an image. It is invariant to image translations and transformations, making it ideal for indexing and retrieval. For this reason a combination of histograms are implemented. Firstly, the commonly used colour histogram is deployed through 8 bins (composed of four green, 2 red and 2 blue) which are quantized and normalized accordingly.

Other than colour and texture, edges constitute as an important feature in representing content for both machine and man (were the human eye is normally sensitive to edge features). In MPEG-7 edges can be represented through an Edge Histogram Descriptor (EHD) ([8] and [20]). The histogram represents the frequency and the directionality of the brightness changes within images. This is said to be a unique feature that cannot be duplicated by either colour histograms or texture features [20] and hence can complete our node description phase.

According to the MPEG-7 standard [8], the EHD represents the distribution of five types of edges; vertical, horizontal, 45-degree diagonal, 135-degree diagonal and non-directional edges. These edges are located locally within an image or sub-image. In our implementation each node represents a segment, where it is then sub-divided into 4x4 sub-images. And the local edge distribution for each sub-image may be represented by a histogram. This results in a total of 80 (5x16) histogram bins, which are quantized and normalized. For each image-block we may extract edge information; existence of edges and their type. The predominant edge is determined and the histogram value of the corresponding edge is incremented by one.

Since the EHD describes the distribution of non-directional, as well as non-edges and four directional edges, the extraction scheme is block-based. In this case the sub-image is further divided into non-overlapping square image-blocks (where its size depends on the resolution of the image) and the coefficients for the edge categories are applied accordingly. [20] extends EHD by including global and semi-global edge histograms. However, in our node description, we also consider global information (representing edge distribution for the whole image space - node) within the EHD.

4.2 Similarity Match

For retrieval, firstly, to manage complexity and storage, the BPT of images both for the query and the database can be simplified using region evolution [9]. The criterion used can simply be the number of requested nodes. This ability to simplify the BPT provides a hierarchy of search planes that provide a coarse to fine granularity of search.

After pruning of nodes (figure 7), comes the process of matching nodes between the database images and our query node from the query image. Since the tree nodes are based on colour and texture it is logical to incorporate their statistics in the matching process.

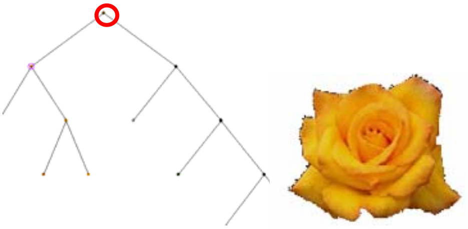


Fig. 7 Semi-automatic tool; user selects query node from a pruned BPT

For each node comparison their absolute colour mean, texture variance and EHD differences are computed. For the similarity matching of the colour histogram, the distance $M(Q,X)$ between histograms Q (representing the query node) and X (candidate nodes) is defined as:

$$M(Q,X) = \sum_{i=0}^{7} \left| Q[i] - X[i] \right| \tag{7}$$

Secondly, for texture, the wavelet transform is applied to each node and their energies computed. The differences between the energies of the LH and HL band for two nodes are summed accordingly:

$$N(Q,X) = \left| e_{LH_Q} - e_{LH_B} \right| + \left| e_{HL_Q} - e_{HL_B} \right| \tag{8}$$

Thirdly, for the similarity matching of EHD the distance of two histograms Q and X representing local and global EHD is defined as [19]:

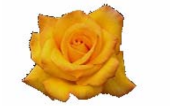
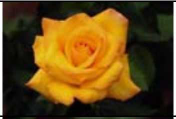
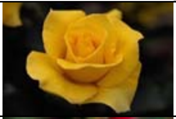
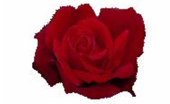
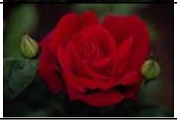
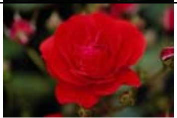
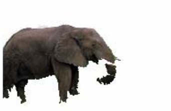
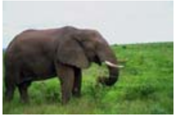
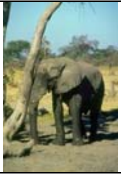
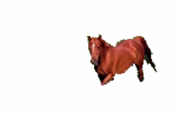
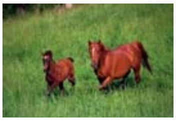
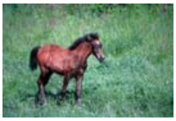
$$P(Q,X) = \sum_{i=0}^{79} \left| Local\_Q[i] - Local\_X[i] \right| + 5 \times \sum_{i=0}^{4} \left| Global\_Q[i] - Global\_X[i] \right| \tag{9}$$

For each node descriptor (colour, texture, and edge) their n most similar nodes and their corresponding root image are recorded. However, if more than one node in the same tree is found only one is recorded (the least distance). Finally the common images among all the descriptors are found as the most likely image. These common images can further be rank ordered based upon their edge histogram.

4.3 Results and Analysis

To demonstrate the semi-automatic system, a sub-set of the images in [21] were used as query images. Table 4 illustrates retrieval results from selected query nodes. In this table the node image segments are shown in the first column and their root images are shown in the second column. The most similar retrieved image after the query image is represented at the third column for each trial.

Table 4 Retrieval results

Query Node	Query Image	Most Similar Image
		
		
		
		

The results confirm the accuracy of node-based image retrieval. The multiple descriptors have each provided their unique characteristics in describing individual nodes. The node-based retrieval system has the advantage that an object within an image is not masked by the background image. For example, although in the second and third row, both elephant and horse objects are camouflaged in a predominately green background, with the use of node descriptors they are easily discriminated from each other.

In this system, accurate segmentation also plays an important role, which is realized as a combination of colour and texture in the derivation of a BPT. This is because the fidelity of this system very much depends on the meaningful objects at the nodes of the BPT, which is best realized by a good segmentation.

5 Summary

This work represents the novel application of the Binary Partition Tree, which has been computed within the wavelet domain where spatial frequency or texture has been used in conjunction with colour to produce a threshold free binary partition tree for segmentation. Results are compared against colour only based segmentations and quantified against ground truths according to intrinsic and extrinsic measures and shown to be superior in test cases. The application of a single quality metric incorporating the intrinsic and extrinsic values reinforces these observations.

Our algorithm has facilitated subjectively best derived segmentations which are meaningful, smoother and more amenable to browsing by constraining salient objects to individual branches within the BPT. This work presents how high density image segments can enhance the performance of image processing applications such as in content based retrieval system.

As an example application, this work presented a semi-automatic tool for content retrieval based on the interactive browsing of a binary partition tree. Object nodes generated from the BPT are indexed and matched through a combination of descriptors. Significantly, users' had more freedom in their choice of query and results illustrated that semantic segments can even enhance retrieval results for natural images.

References

[1] Salembier, P., Garrido, L.: Binary Partition Tree as an Efficient Representation for Image Processing, Segmentation, and Information Retrieval. IEEE Transactions on Image Processing 9(4), 561–576 (2000)

[2] Zhang, Y.: An Overview of Image and Vide Segmentation in the Last 40 Years, ch. 1

[3] Haralick, R.M., Shapiro, L.G.: Image segmentation techniques. In: CVGIP, vol. 29, pp. 100–132 (1985)

[4] Zouagui, T., et al.: Image segmentation functional model. In: Pattern Recognition, pp. 1785–1795 (2004)

[5] Freixenet, J., Muñoz, X., Raba, D., Martí, J., Cuff, X.: Yet Another Survey on Image Segmentation: Region and Boundary Information Integration. Springer, Heidelberg (2002)

[6] Vincent, L., Soille, P.: Watersheds in digital spaces: An efficient algorithm based on immersion simulations. IEEE Transactions on Pattern Analysis and Machine Intelligence 13(6), 583–598 (1991)

[7] Andrade, E.L., Khan, E., Woods, J.C., Ghanbari, M.: Segmentation and Tracking Using Region Adjacency Graphs, Picture Trees and Prior Information

[8] Manjunath, B.S., Salembier, P., Sikora, T.: Introduction to MPEG-7: Multimedia Content Description Interface. Wiley, Chichester (2002)

[9] Lu, H., Woods, J.C., Ghanbari, M.: Binary Partition Tree for Semantic Object Extraction and Image Segmentation. IEEE Transactions on Circuits and Systems for Video Technology 17(3), 378–383 (2007)

[10] Ghanbari, S., Woods, J.C., Rabiee, H.R., Lucas, S.M.: Wavelet Domain Binary Partition Trees for Semantic Object Extraction. Electronics Letters 43(22) (October 2007)

[11] Ghanbari, S., Woods, J.C., Rabiee, H.R., Lucas, S.M.: Wavelet Domain Binary Partition Trees for Image Segmentation. In: CBMI (June 2008)

[12] Zhang, Y.J.: A Survey on evaluation methods for image segmentation. In: Pattern Recognition, vol. 29, pp. 1335–1346 (1996)

[13] McCane, B.: On the Evaluation of Image Segmentation algorithms. In: Proceedings of Digital Image Computing: Techniques and Applications, pp. 455–460. Massey University, Albany Campus, Auckland (1997)

[14] Ge, F., Wang, S., Liu, T.: Image-Segmentation Evaluation from the Perspective of Salient Object Extraction. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, June 2006, vol. 1, pp. 1146–1163 (2006)

[15] Martin, D., Fowlkes, C., Tal, D., Malik, J.: A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics. In: Proc. 8th Int'l Conf. Computer Vision, July 2001, vol. 2, pp. 416–423 (2001)

[16] Xia, Y., Feng, D., Zhao, R.: Optimal Selection of Image Segmentation Algorithms Based on Performance Prediction. In: Pan-Sydney Area Workshop on Visual Information Processing (VIP 2003), vol. 36, pp. 105–108. Conferences in Research and Practice in Information Technology, Sydney (2004)

[17] Rui, Y., Huang, T.S., Chang, S.: Image Retrieval: Current Techniques, Promising Directions, and Open Issues. Journal of Visual Communication and Image Representation 10, 39–62 (1999)

[18] Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-Based Image Retrieval at the End of the Early Years. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(12), 1349–1380 (2000)

[19] Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Survey 40(2), Article 5 (April 2008)

[20] Won, C.S., Park, D.K., Park, S.: Efficient Use of MPEG-7 Edge Histogram Descriptor. ETRI Journal 24(1) (February 2002)

[21] Corel Data Set, http://www.corel.com

Logarithmic r-θ Map for Hybrid Optical Neural Network Filter for Object Recognition within Cluttered Scenes

Ioannis Kypraios, Rupert C.D. Young, and Chris R. Chatwin

Abstract. Space-variant imaging sensors can be designed to exhibit in-plane rotation and scale invariance to image data. We combine the complex logarithmic r- θ mapping of a space-variant imaging sensor with the hybrid optical neural network filter to achieve, with a single pass over the input data, simultaneous invariance to: out-of-plane rotation; in-plane rotation; scale; projection and shift invariance. The resulting filter we call a complex logarithmic r- θ mapping for the hybrid optical neural network filter. We include in the L-HONN filter's design a window based unit for registering the translation invariance of the input objects, initially lost by applying the logarithmic mapping. We test and record the results of the L-HONN filter for single and multiple input objects of the same class within cluttered still images and video frame sequences.

1 Introduction

In digital image processing [1] the image data is acquired from a camera or any relevant signal input source and stored in the form of a raster image array. Translating the image data from Cartesian to Polar co-ordinates results to an r-θ mapping, which is rotation invariant about its centre. Hence, an in-plane rotation of an object in the image space will produce a shift of the image in the r-θ space. By applying the log along r on the r- θ mapping, we produce scale and projection invariance, too. In effect, when sampling the original image data from the Cartesian space by the proper sensor geometry in the r- θ space, its logmap cells increase logarithmically from the centre to the edges of the mapping. This structure resembles the mammalian retina [2-6], where the resolution of the resulting mapping decreases gradually from the centre to the outside edges of the fovea. Many researchers in their recent work on pattern recognition have emphasized the need to design and build systems which imitate the visual system of the human brain for engineering applications. Such systems can be used for modeling the brain functions and therefore help us in better understanding as well as offer superior solutions in machine vision problems [7-21, 22, 23, 24-29]. This transformation of

Ioannis Kypraios, Rupert C.D. Young, and Chris R. Chatwin
Laser and Photonic Systems Research Group, Department of Engineering and Design
University of Sussex, Falmer, Brighton BN1 9QT, UK
e-mail: i.kypraios@sussex.ac.uk

M. Grgic et al. (Eds.): Rec. Advan. in Mult. Sig. Process. and Commun., SCI 231, pp. 91–120.
springerlink.com

Cartesian to Polar mapping has many machine vision and pattern recognition applications. Such applications include autonomous robot navigation [30-33], efficient methods for tomographic reconstruction [34, 35], stepped-frequency inverse synthetic aperture radar (ISAR) [36, 37] image reconstruction and several others [38, 39].

Here, we describe in details the design and implementation of the complex logarithmic r- θ mapping [40, 41, 2, 30, 31, 42-44] for the hybrid optical neural network (L-HONN) filter. The overall filter with a single pass over the input data is to simultaneous exhibit out-of-plane rotation, in-plane rotation, scale and projection invariance and resistance to clutter. However, the shift invariance obtained by applying the log r- θ mapping on the input images is lost and a window-based system is used to restore shift-invariance to the L-HONN filter.

An augmented multi-layer perceptron for in-plane rotation and scale invariant hand-written numeral recognition has been presented previously in the literature by S. Kageyu., N. Ohnishi et al. [43]. Their system has used a three stage approach to the problem composed of complex logarithmic r- θ mapping, Fourier transformation and a multi-layer neural network architecture. H. Y. Kwon and B. C. Kim [42] have proposed a two stage system composed of complex logarithmic r- θ mapping and an augmented second order neural network to achieve scale and in-plane rotation. Their system is based on the observation that the scaling or the in-plane rotation of the original image data are shown on the logmap as positional translation. In effect, the transformations are interpreted as horizontal and vertical shifts on the logmap, which can cause wrap-around translation of the image data. The second order neural network (SONN) has been proven previously to be translation invariant. Thus, Kwon et al. have augmented the SONN by integrating the condition for the weight connections to be wrap-around translation invariant and deriving a new updated learning rule of each unit of the network. In our system, the translation invariance is assured in the optical correlator part of the L-HONN filter. The optical correlator block, besides super-fast performance, maintains the shift invariance of the recognised object. M. Fukumi [26] has pointed out that one of the main drawbacks of existing systems that utilise high-order neural networks has been the combinatorial explosion of units. The concept in psychology of mental rotation [29] is integrated in the design and implementation in his rotation-invariant neural pattern recognition system, which estimates the pattern's rotation angle. It consists of a feature extraction stage and two main artificial neural networks (NNETs). The first network is based on the feed-forward architecture with the back-propagation learning algorithm. The second network is split into two parts, the pattern recognition part and the angle estimation part. Briefly, the input test pattern from the retina is processed and edge features are properly selected. If the recognition result of the tested pattern is acceptable from the first network, the recognition process through the system is stopped. If the results from the first network are unclear, the input features are passed to the second network, which tries to correctly recognise the pattern and stop the process. If still the results are unclear, then based on its angle estimation, the pattern is rotated and the process starts again from the first network. Satisfactory recognition accuracy and angle estimation was exhibited by the system for binary patterns and 256-colour grey-scale coin images. However, though the results from their test applications recorded were promising, in its essence our system is different. Fukumi's system assumes a passive retina which acquires the input pattern and passes it to the feature extractor. The space-variant sensor, integrated in the design of our logarithmic r-θ map for the hybrid optical neural network filter, actively transforms the input image data to the in-plane and scale invariant logmap data. In effect, much effort spent in Fukumi's system for selecting the edge features and dividing the second NNET to two parts for separating the pattern recognition task and angle estimation task is not a requirement in the L-HONN filter. Also, the patterns used in Fukumi's system have been pre-processed and are assumed to be in a plain background. Our filter inputs directly the image data, even complicated cluttered input scenes, without the use of a separate feature extraction stage and the in-plane rotation angle extraction of the input pattern is inherent in the logmapping transformation. Nevertheless, the notion of the mental rotation [29] and its integration to future systems offers new interesting possibilities.

Sect. 2 describes the sensor's properties, its geometry, the logmap of an x-y raster input image and the sensor's parameters affecting its resolution. Sect. 3 gives details of the window based sub-image acquisition unit. Sect. 4 presents the design and implementation of the L-HONN filter. Sect. 5 presents the simulation results of the filter for single and multiple input objects of the same class within cluttered still images and video frame sequences. Sect. 6 describes an Internet multimedia application of the L-HONN filter and Sect. 7 concludes.

2 Space-Variant Image Processing

We have applied similar notation rules throughout this paper to keep consistency between the different mathematical symbols of artificial neural networks and optical correlators and to unify their representation. We denote the variable names and functions by non-italic letters, the names of the vectors by italic lower case letters and the matrices by italic upper case. The frequency domain vectors, matrices, variable names and functions are represented by bold letters and the space domain vectors, matrices, variables and functions by plain letters.

The motivation to use a space-variant imaging sensor [38, 39] arises from the need to increase even more the distortion invariant abilities and producing a real-time performance in our final object recognition system. It is shown mathematically the space-variant sensor geometry to be scale, rotation and projection invariant about its centre or focus of expansion. The main advantage of the sensor is a wide visual field whilst maintaining high resolution and data reduction, using a minimum amount of memory and computational resources. The structure of the sensor is based on a Weiman [2-6] polar exponential grid and consists of concentric exponentially spaced rings of pixels, which increase in size from the centre to the edge.

The sensor allows image data on the circular region of the Cartesian x-y space to be mapped into a rectangular region of the Polar image space r-θ. The sensor's geometry maps concentric circles in the Cartesian space into vertical lines in the polar space and radial lines in the Cartesian space into horizontal lines in the Polar space. This transformation offers scale and rotation invariance about its centre.

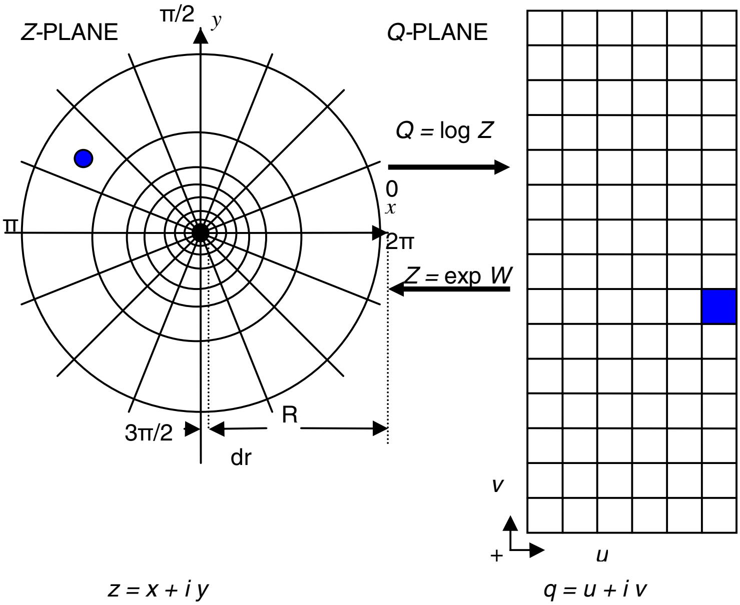


Fig. 1 It shows the complex logarithmic mapping. Here, we have marked one randomly selected cell with a colour-filled circle on the space-variant sensor and we have marked the corresponding logmap pixel with a colour-filled square. The marked cell is located at (5,6) *z*-plane co-ordinates, i.e. on the fifth ring and the sixth radial line. When translated to the *q*-plane of the logmap, the corresponding marked pixel is located at the sixth column and the seventh row respectively. Both, the sensor and the logmap starts from co-ordinates (0,0) (adapted by [38]).

By modifying the r- θ mapping to log r- θ mapping or logmap, now offers scale, rotation and projection invariance about its focus of expansion (centre). The space-variant sensor we use performs a complex logarithmic mapping of the image data from the circular retinal region into a rectangular region. Fig. 1 shows the complex logarithmic mapping performed by the sensor geometry. The vertical lines in the *q*-plane map the concentric circles in the *z*-plane and the horizontal lines in the *q*-plane map the radial lines in the z-plane. The complex logarithmic mapping [2, 30, 32] can be written as:

$$q = \log z \tag{1}$$

or applying the complex form mathematical notation:

$$z = x + i\,y \tag{2}$$

where

$$r = \sqrt{x^2 + y^2} \tag{3}$$

and

$$\theta = \arctan\left(\frac{y}{x}\right) \tag{4}$$

Thus,

$$q \;=\; \log r + i\theta \tag{5}$$

$$= u + iv \tag{6}$$

$$u = \log r \tag{7}$$

$$v = \theta \tag{8}$$

Hence, an image in the z-plane with co-ordinates x and y is mapped to the q-plane with co-ordinates u and v. The mapped image from the Cartesian space (z-plane) in to the Polar space (*q*-plane) is referred to as the log-polar mapping or the logmap.

The following invariances kept by the logmap which are attractive for image processing applications, in addition to the wide field of view, reduced pixel count and a certain highly focused area inherited in the space variant sensor geometry. First, rotating the image by an angle α results in a vertical shift in the mapped image by log of the rotation angle [38]. It can be shown that:

$$z = r\,e^{i(\theta+\alpha)} \tag{9}$$

$$\begin{aligned} q &= \log r + i\theta + i\alpha \\ &= u + iv + i\alpha \end{aligned} \tag{10}$$

Second, the log-polar mapping offers scale invariance. Scaling the image by the factor β has resulted in a horizontal shift in the mapped image by log of the scaling factor [38]. It can be shown that:

$$z = r\beta e^{i\theta} \tag{11}$$

$$\begin{aligned} q &= \log r + \log\beta + i\theta \\ &= u + \log\beta + iv \end{aligned} \tag{12}$$

Third, the logmapping offers projection invariance, too. If we move towards a fixed point in the *z*-plane then the mapped image will shift horizontally in the *q*-plane. The size and shape of the original *z*-plane image will stay the same. This is equivalent to progressively scaling the *z*-plane image, but with a different change in the perspective ratio. We should emphasize here, that the log-polar mapping is not shift invariant, so the properties described above hold only if they are with respect to the origin of the Cartesian image space.

Transforming the image from the log-polar co-ordinates to the Cartesian co-ordinates is called the inverse mapping of the image [38]. We have:

$$z = e^{q} \tag{13}$$

and, from eqn. (6) we have:

$$q = u + iv \tag{14}$$

then, we can get:

$$z = e^{u}\left[\cos(v) + i\sin(v)\right] \tag{15}$$

Let

$$x = e^{u}\cos(v) \tag{16}$$

and

$$y = e^{u}\sin(v) \tag{17}$$

From eqn. (3) and eqns. (5) and (6) we can find:

$$\begin{aligned} r &= \sqrt{x^{2} + y^{2}} \\ &= e^{u} \end{aligned} \tag{18}$$

$$\Rightarrow u = \log r \tag{19}$$

Also, we can find:

$$\begin{aligned} \tan\theta &= \frac{e^{u}\sin(v)}{e^{u}\cos(v)} \quad \therefore \\ \theta &= v \end{aligned} \tag{20}$$

In effect, from eqn. (18) and eqn. (19) we find that the rows of the logmap (q-plane) represent the spokes of the inverse mapping (z-plane). Next, we study briefly some aspects of the sensor's geometry, so to be able to understand the design considerations in relation to the hybrid optical neural network filter.

2.1 Polar-Exponential Sensor Geometry

We define to an equal number of pixels in each of the rings of the inverse mapping. There is an increase in size of these pixels inside the concentric exponential circles from the centre to the edge of the mapping. This leads to a considerable reduction in the number of pixels required. The pixels of the Cartesian space image are forced to be averaged together at the edges of the mapping, to produce a blurred peripheral view, but are not averaged together at the centre, creating a highly focused area at the centre to retain all useful information in the image. As mentioned before, the geometry of the space-variant sensor to achieve a wide visual field whilst maintaining high resolution needs to be based on the log-polar mapping. The first sensor in literature to perform the log-mapping was suggested by Weiman [2-6]. A suitable mapping is to have the sensor's angularity proportional to cell size. This has resulted to a sensor array where the sensor cells are roughly square and increase exponentially in size with distance from the centre of the mapping. Our implemented sensor's geometry consists of an array of u concentric, exponentially spaced rings of n pixels. Each sensor cell corresponds to one logmap square pixel, or equivalently the indices of the logmap, which is a rectangular grid of square cells; these are used to index the spokes and rings in the sensor array. In effect, the polar ($r\angle\theta$) co-ordinates of the sensor array are mapped to the Cartesian (u, v) co-ordinates of the logmap.

2.2 Logmap of an x-y Raster Image

There are two basic mapping methods for performing the logmap of an x-y raster image. Fig. 2 shows an example of a logmap of an x-y raster test image of the True-class object, a Jaguar S-type, inserted in a car park scene [45]. Here, the offset distance dr from the sensor's centre to the first ring rmin is set to zero. As we will see in Sect. 2.3, dr defines the blind spot area of the sensor, which in turns it affects the overall resolution of the mapping. For the first mapping method, weighted methods are employed [38]. Hence, the intensity of each image pixel is assigned directly to the proportion of the corresponding sensor cells that it covers. Then, we divide the total intensity of the pixel by the total weight of the pixels within the range of its sensor cell to compute the logmap pixel. If each and every pixel percentage included inside the sensor cell is calculated together, then it forms the ideal sensor geometry. If each image pixel contributes only once to each sensor cell, then it forms the many-to-one [38], as it is called, mapping sensor geometry. By using look-up tables [38] to perform the latter sensor geometry we can speed-up the required calculations. The second mapping method is based on interpolation techniques. Now, each sensor cell samples by interpolating the intensities of all the image pixels it covers.

To avoid the problem of oversampling pixels at the centre of the mapping and undersampling pixels at the periphery of the mapping, which results in an overall broken logmap, not fully scale and rotation invariant, the central part of the mapping itself can be removed and vary accordingly the sensor's parameters (see Sect. 2.3) so that each sensor pixel only contributes to one logmap pixel, or map back into the logmap separately the data at the centre. A more ideal mapping [38] is produced by subdividing each pixel into smaller sub-pixels and assigning a proportion of the intensity of each pixel to the sensor cell that each sub-pixel falls within, e.g. 16 smaller sub-pixels, each with $\frac{1}{16}\mathrm{th}$ of the intensity of the image pixel. Of course this extra sub-division of each sensor cell prior to assigning the intensities of the image pixel increases the overall computation time of the logmapping, which, however, can be

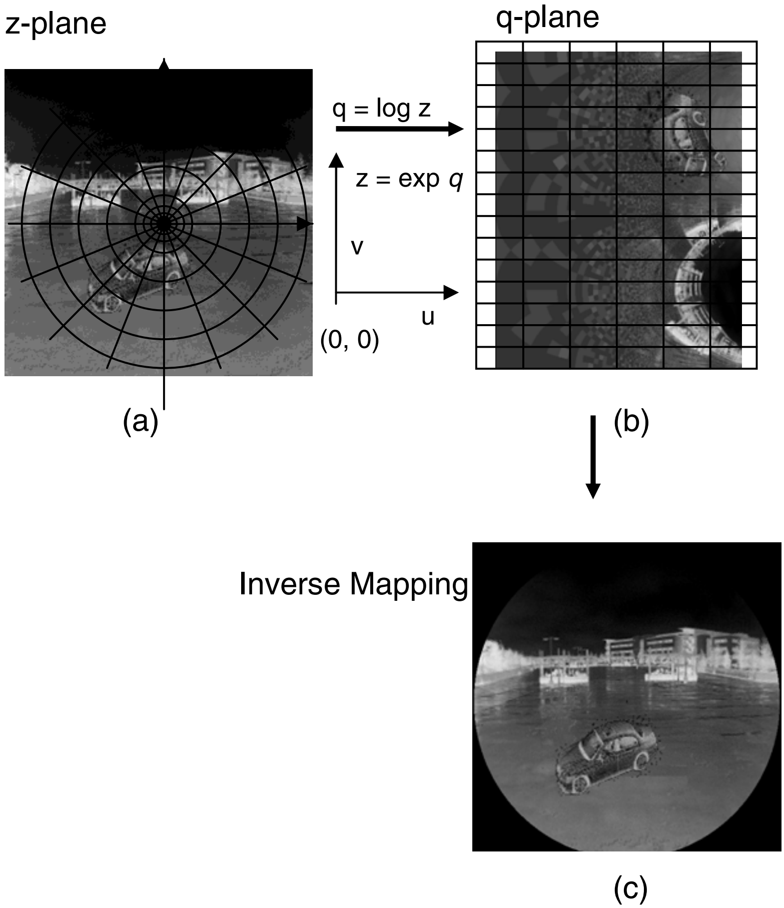


Fig. 2 An example is shown of the complex logarithmic x-y mapping of a raster test image of the True-class object, a Jaguar S-type car, inserted in a car park scene. (a) The space-variant sensor samples the polar $r\angle\theta$ co-ordinates in the z-plane of the image data and translates them to the complex logarithmic (u,v) co-ordinates of the q-plane (the origin (0, 0) of the axes (u, v) of the q-plane is shown on (b)). (c) It shows the resulted inverse mapping of the logmap.

reduced with look-up tables, so only the overlapping pixels are subdivided. The coarseness of the image decreases as the number of sub-pixels increases, since more of the overlapping pixels now are used to contribute to each logmap pixel. By increasing the number of sub-pixels to more than sixteen, we achieve a more accurate percentage of each pixel, but it is at the expense of a heavy computational load.

For our results we applied an interpolation method [38] rather than sub-dividing the overlapping pixels of the image. It is noticed that if the image is large and the mapping contains a large number of rings and spokes, so that each image pixel contributes to several logmap pixels, then it is preferable to sample the image pixels for every logmap pixel, instead of assigning each image pixel to several logmap pixels. There are several sampling methods existing in the literature, all of them suggested in an effort to achieve the best resolution of the logmap without significant loss of data. For our purpose we used the nearest neighbour interpolation [38] method, rather than other methods such as the bilinear interpolation and variable width mask interpolation. Briefly, the nearest neighbour mapping [38] results in a blurred peripheral view, where the sensor cells are large, leading to a pure sampling of the image pixels. For our simulations tests (see Sect.5) we assume the True-class object to be centered within the cluttered input still image or video scene. In effect, the produced blurred peripheral view by the mapping does not significantly affect our tests results. The bilinear interpolation [38] method undersamples the image data at the periphery of the mapping, where there are many image pixels per sensor cell, but it oversamples the image pixels at the centre of the mapping where there are only a few corresponding pixels per sensor cell. The resulting logmap is blurred at the edges and at the centre of the mapping. The variable width interpolation mask [38] produces a finer logmap of the image data, since the size of the masks follow the exponential increase of the sensor's concentric circles. However, due to the Cartesian-rectangular-nature of the masks, it does not fully conform to the ideal sensor cell geometry. Hence, the resulting logmap is blurred at the edges, where several pixels are averaged together by the masks at the periphery, but highly focused at the centre, where there is a one-to-one mapping between the mask size and the image pixels.

2.3 Space-Variant Imaging Sensor's Parameters and Resolution

All the sensors are characterised using their parameters $\left(\mathrm{n},\, \mathrm{R},\, \mathrm{r}_{\min},\, \mathrm{dr}\right)$ values [38, 39]. By varying the values of the sensor parameters we affect its resolution. First, we define the grain of the sensor array as the number of the sensor cells in each of the rings of the array, but the number of the sensors cells is equal to the number of the spokes in the sensor array. Thus, n spokes cover a range of 2π radians, and so the angular resolution of each sensor cell from the centre of the array is $\frac{2\pi}{\mathrm{n}}$. It is found experimentally that by decreasing the grain of the sensor array n [38], the resolution at the edges of the logmap becomes more blurred, but there is sill enough information retained at the centre to recognize the image. Second, we assume that the values of the granularity, the inner ring and the offset of the blind spot stays constant and we vary only the field width. Then, we can experiment on the effect of varying the value of R [38]. It is found that the value of the field width affects the size of the logmap as expected. By inverse mapping it is discovered that the actual size of the logmap determines the available resolution at the centre of the mapping. For smaller size logmap images the inverse mapping is less focused at the centre. Thus, if we are interested in maintaining most of the original image data at the centre of expansion, it is apparent we need to perform the sampling for the logmap from a large field width. However, this results to a high compression ratio of the logmap from the original image, but it is at the expense of the computation time required and available memory. The innermost ring, rmin [38], and the offset, dr [38], parameters are directly related with the size of the blind spot and its effect on the overall sensor's resolution. So, the third parameter to be discussed is the value of the innermost ring. Assuming we keep constant the values of the granularity, the field width and the offset and we only vary the value of rmin, this fixes the extent of the logmap, so the larger the inner most ring value the smaller the size of the logmap. Additionally, the increase of rmin increases the compression ratio of the logmap, but it removes the central area from the original image. If we set fixed values all the parameters of the sensor, but vary only the granularity g of the sensor array, then it is found the increase of the grain n (i.e it can be shown mathematically from the grain and granularity equation, that, then the granularity g is decreased, and vice-versa by decreasing the value for the grain n we increase the value for granularity g of the sensor array) can produce smoother images (from the inverse mapping) even when we increase the value of rmin and without any significant increase in the storage requirements of the logmap. If the image data at the centre is needed, then researchers have suggested the blind spot is filled by rectangular pixels or an $r\text{-}\theta$ mapping can be used. Last, by changing the value of the offset parameter, the resolution at the periphery of the image is affected. In effect, by increasing the offset dr, we produce finer resolution at the edges, but we increase the blind spot area at the centre, which reduces the information available in that area. By tuning the values of both parameters dr and rmin, we can produce a smaller size logmap and increase the peripheral view, but lose the information in the central area.

In our implementation of the L-HONN filter, and for the purposes of our conducted simulation results with the used train and test sets of still images as well as video sequences, we set the parameters of the sensor to be the granularity, n equal to the dimension (x- or y-) of the square size input image (size of x- equals size of y-). In effect, there is one ring covering each pixel column and row. The width of the field, R is set to be equal to half the dimension of the square image or frame, so the centre of the logmap expansion is the centre of the input image (or frame). The inner most ring, rmin is set to 1 to cover all the area at the centre of the input image or frame, since it is assumed the target object to be positioned at the centre of the processed by the filter input image or frame. Offset dr is set to zero, since we wish to maintain the data at the centre of the input image (or frame).

3 Window Based Sub-image Acquisition Unit

All the input images and frames (test set or training set) are passed through the window unit [46-50] and then each sub-image is logmapped. The window unit restores the shift invariance of the L-HONN filter, which otherwise it is lost since the rotation, scale and projection invariances of logmapping only hold if they are taken from the centre of expansion. If the True-class object is translated off the centre of the image before logmapping, then the logmap produced will not be shifted accordingly, so it will not contain the same image data. Fig. 3 shows the block-diagram of the window-based unit system we used for the implementation

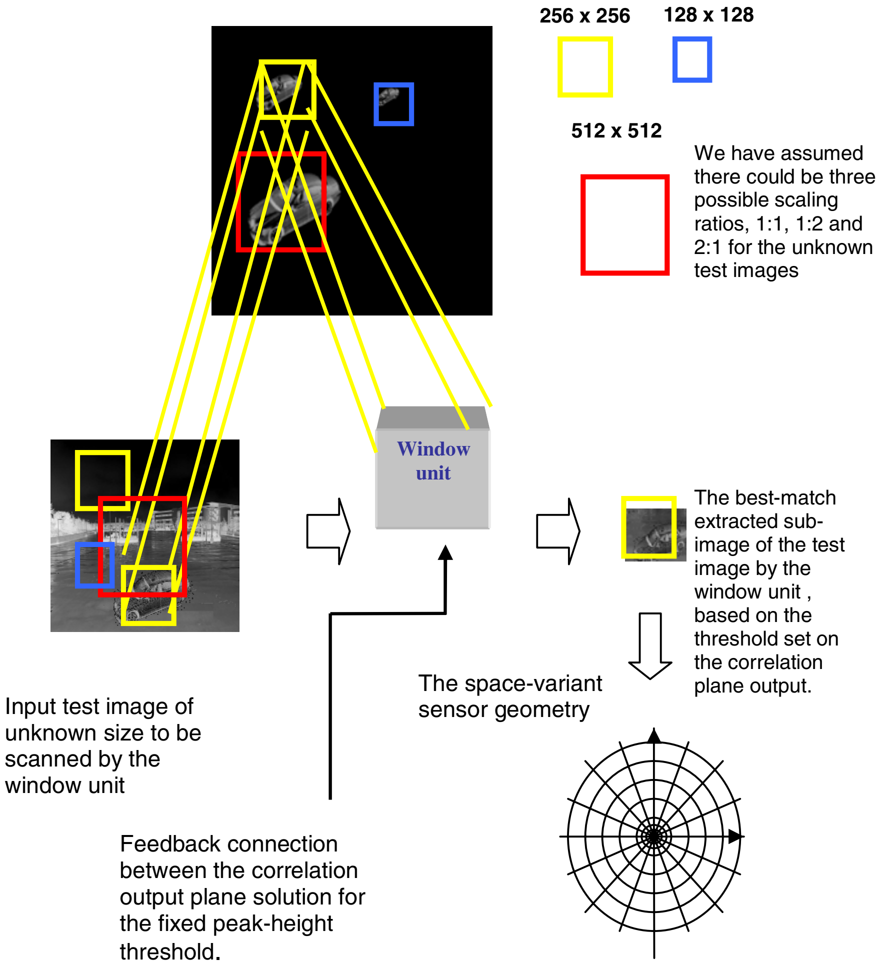

256 x 256
128 x 128
512 x 512
We have assumed there could be three possible scaling ratios, 1:1, 1:2 and 2:1 for the unknown test images
Window unit
The best-match extracted sub-image of the test image by the window unit , based on the threshold set on the correlation plane output.
The space-variant sensor geometry
Input test image of unknown size to be scanned by the window unit
Feedback connection between the correlation output plane solution for the fixed peak-height threshold.

Fig. 3 It shows the window-based unit. The test input image is passed through the window-based unit before being sampled by the space-variant sensor. We have assumed there are three possible scaling ratios, 1:1, 1:2 and 2:1with which the test image can be input to the unit. Then, the window-based unit scans the test image for the different scaling ratios. Here, a simple *ad-hoc* heuristic for updating each of the window's steps was devised to reduce the number of sub-mages produced by the system. There is a feedback connection between the correlation output plane of the filter and the window-based unit to locate the best-match sub-image extracted from the input image.

of the L-HONN filter. A. Pantle and R. Sekuler [27], in their work on mechanisms for detecting different sizes of the same object in human vision, have suggested that scale invariance is achieved by appropriately tuning visual processing channels to different scales of the input pattern. Their concept was implemented in a neural network architecture for achieving scale-invariant feature extraction of the input data by T. Nagano and M. Ishikawa [51].

The window unit we implemented subdivides the original image into smaller size images, before applying the logmapping on top of each of these sub-images. Many similar units have been presented in the literature as part of face or object recognition systems. Rowley et al. [52] presented a neural network-based face detector. In their work, a small window passes over all the locations of the input image and examines each sub-image for the presence or not of faces. In brief, their system consists of two stages. The first stage applies a set of neural network based filters to the extracted sub-mage from the window unit. Two main preprocessing steps are used before applying the set of neural networks to it. First, a linear function, which approximates the overall brightness of each part of the window, is fit to the intensity values in an oval region inside the window. By subtracting out the function from each window, the lightning conditions are unified amongst the several sub-images and any extreme lighting conditions are compensated. Then, a histogram equalisation step follows to correct for the gain variations of the different cameras and to improve contrast across the window. Next, the preprocessed sub-image is passed through the neural network sets. Each neural network consists of different types of hidden units acting as receptive fields for detecting features such as pairs of eyes, noses or mouths. The second stage of the system consists of methods to eliminate the false detections from the set of the NNETs; two separate strategies are presented. The first approach merges the overlapping detections from a single network, whereas the second approach arbitrates among multiple networks by ANDing, ORing or voting amongst the outputs of the NNETs. Drucker et al. [53] and Sung et al. [54] suggested algorithms for a “bootstrap” method that reduces the size of the training set needed by training iteratively a neural network with the examples on which the previous networks had failed and by selectively adding images to the training set as training progresses. Several other systems have been suggested recently. Osuna et al. [55] combined his window based face detector with a “support vector machine”. Moghoddam et al. [56] developed a face detector that uses the window unit together with a two-component distance metric and combines the two distances to produce a multimodal Gaussian distribution for all the face images. Colmenarez et al. [57] experimented with a window based face detector employing statistical methods to build probabilistic models of the sets of faces and non-faces. Each input window in his system compared the inputs with two probabilistic categories. We have selectively adapted some of the techniques used in window unit based systems for face detection or recognition to our object recognition filter.

It should be emphasized that our implementation of the window unit is aiming to solve the shift invariance lost in the overall log r-θ hybrid optical neural network filter and not to experiment with different settings of window based systems. As pointed out by Rowley et al. [52] the amount of the filter invariance determines the number of scales and positions to which the window unit must be applied on the image. Umezaki [58] has presented a license plate detector, which combines a window unit with a neural network. The neural network is made invariant to translations of about 25 percent of the size of the license plate to reduce the number of windows processed. Similarly, Rowley et al. adapted Umezaki's idea of translation invariance to decrease the number of windows processed in total for each input image in their face detector. In effect, the window unit becomes even more attractive for use with our hybrid optical neural network filter. The logarithmic r-θ mapping of the processed window offers scale and in-plane rotation invariance. Thus, the number of windows processed by our L-HONN filter decreases due to the scale invariance and, also, the number of training images is significantly less (see Sect. 5 for simulation results) than the window based systems mentioned above due to the in-plane rotation invariance.

Each sub-image is logmapped and processed by the filter. The highest correlation peak should be produced by the sub-image which includes the True-class object closest to the centre of expansion of the window. By providing a priori the maximum and minimum size, in pixel count, of the wanted object, we have worked out a simple heuristic to improve the speed of the window unit system and reduce even more the number of sub-images being processed by the filter (see Sect. 5). We have assumed there are three possible scaling ratios, 1:1, 1:2 and 2:1with which the test image can be input to the unit. Then, the window-based unit scans the test image for the different scaling ratios. We set the size of the window such as to convert the maximum size of the object and the step size along x- and y- axes in even or uneven size steps (depending on the size of the True-class object) equal to half the size of the object in both dimensions (equal steps, if we assume the object to have roughly square size):

$$WU \ \left(m,n\right) \tag{21}$$

where, WU is the window unit with co-ordinates (*m, n*) on the image.

$$WU \ \left(m',n'\right) = WU \ \left(m+\Delta m, n+\Delta n\right) \tag{22}$$

When the window unit WU is updated on the image, it moves to the new position with co-ordinates $\left(m',n'\right)$. The new co-ordinates came from the previous position co-ordinates $\left(m',n'\right)$ incremented by $\left(\Delta m, \Delta n\right)$:

$$step = \left(\Delta m, \Delta n\right) \tag{23}$$

and

$$\text{object size } = \left[2\cdot\Delta m, 2\cdot\Delta n\right] = \left[m,\, n\right] \tag{24}$$

In effect, the step size of the window unit is equal to (Δ*m,* Δ*n*), but the object size is given by [2·Δ*m*, 2·Δ*n*] in distance units measured on the image. If we assume that the size of the object in pixels is given by [*m*, *n*], then the step size along the x- and y- axes in pixels is written, from eqns. (23) and (24), as:

$$\Delta m = \frac{m}{2} \text{ and } \Delta n = \frac{n}{2} \tag{25}$$

These settings allow the True-class object to be centred with respect to the window unit each time. The window unit step size has been represented with respect to the unit's centre of axes. Observe carefully on Fig. 3 that there is a feedback connection between the correlation output plane of the filter and the window-based unit to locate the best-match sub-image extracted from the input image.

4 Design and Implementation of the Logarithmic r-θ Map for the Hybrid Optical Neural Network Filter

Fig. 4 shows the block diagram of the logarithmic r-θ map for the hybrid optical neural network filter. As for all the hybrid optical neural network-type filters, the input images or frames are processed through the NNET block [59-61] (see Fig. 5 for the NNET detailed block diagram) and multiplied by the corresponding dot product of the input and layer weights of the NNET. Now, in the L-HONN filter, the extracted logmapped images are passed through the NNET. From the newly transformed images of the NNET the composite image of the filter is built. We have chosen to constrain the correlation peak heights of the filter, similarly to the constrained-HONN (C-HONN) filter [59]. The L-HONN filter is described as follows:

$$L-HONN = \sum_{i=1\cdots N}^{N} \mathrm{a}_{\mathrm{i}} \cdot S_i \left[Q(u,v) \right] \tag{26}$$

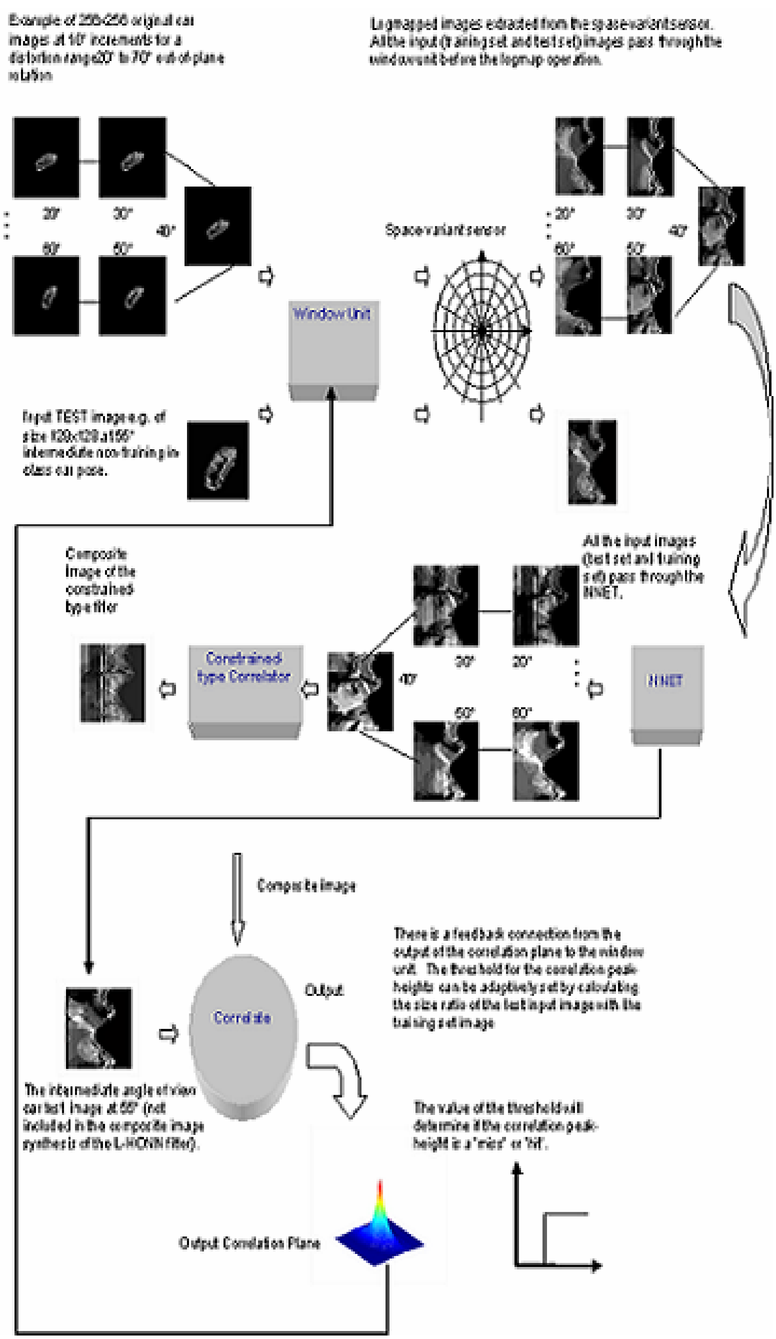
or eqn. (26) is simplified and written as:

$$\begin{aligned} L-HONN &= \sum_{i=1\cdots N}^{N} \mathrm{a}_{\mathrm{i}} \cdot S_i (u,v) \\ &= \mathrm{a}_1 \cdot (\Gamma_1 \cdot Q_1(u,v)) + \mathrm{a}_2 \cdot (\Gamma_2 \cdot Q_2(u,v)) + \cdots \\ &\quad + \mathrm{a}_{\mathrm{N}} \cdot (\Gamma_N \cdot Q_N(u,v)) \end{aligned} \tag{27}$$

or in frequency domain it is written as:

$$\boldsymbol{L-HONN} = \sum_{\boldsymbol{i}=\mathbf{1}\cdots \boldsymbol{N}}^{\boldsymbol{N}} \mathbf{a}_{\mathbf{i}} \cdot \mathbf{S}_{\mathbf{i}} (\mathbf{u},\mathbf{v}) \tag{28}$$

This is the filter's transfer function. The L-HONN filter is composed of a non-linear space domain superposition of the logmapped training set images (or frames). As for all the hybrid optical neural network-type filters, the multiplying coefficient now becomes a non-linear function of the input weights and the layer weights, rather than a simple linear multiplying constant $\mathrm{a}_{\mathrm{i}=1\ldots\mathrm{N}}$, as used in a constrained linear combinatorial-type filter synthesis procedure. Thus, the non-linear L-HONN filter is inherently shift invariant and it may be employed in an



Example of 256x256 original car images at 10° increments for a distortion range 20° to 70° out-of-plane rotation

Logmapped images extracted from the space variant sensor. All the input (training set and test set) images pass through the window unit before the logmap operation.

Input TEST image e.g. of size 128x128 at 55° intermediate non-training in-class car pose.

All the input images (test set and training set) pass through the NNET.

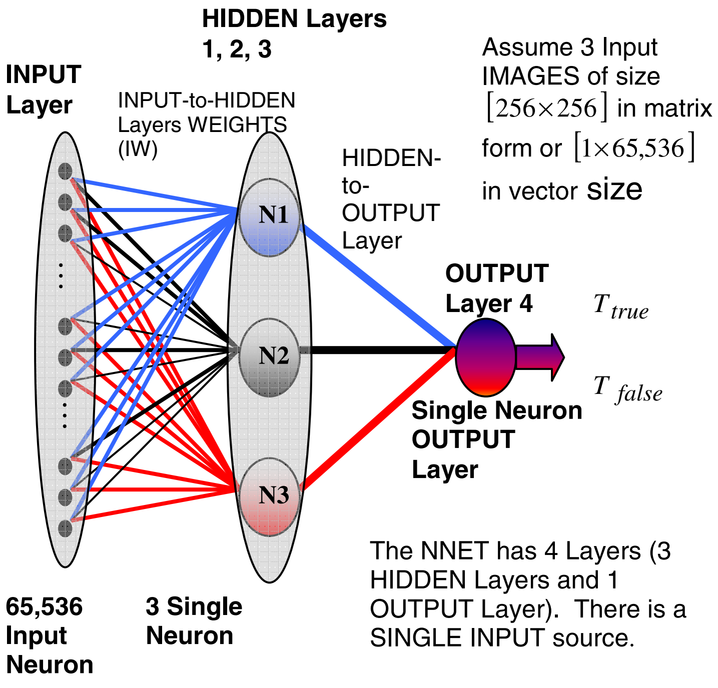
Composite Image of the constrained type filter

There is a feedback connection from the output of the correlation plane to the window unit. The threshold for the correlation peak-heights can be adaptively set by calculating the size ratio of the test input image with the training set image

The intermediate angle of view car test image at 55° (not included in the composite image synthesis of the L-HONN filter).

The value of the threshold will determine if the correlation peak-height is a 'miss' or 'hit'.

Fig. 4 Complex logarithmic r-θ mapping for the hybrid optical neural network (L – HONN) filter design


HIDDEN Layers 1, 2, 3
INPUT Layer
INPUT-to-HIDDEN Layers WEIGHTS (IW)
Assume 3 Input IMAGES of size [256×256] in matrix form or [1×65,536] in vector size
HIDDEN-to-OUTPUT Layer
N1
N2
N3
OUTPUT Layer 4
T true
T false
Single Neuron OUTPUT Layer
The NNET has 4 Layers (3 HIDDEN Layers and 1 OUTPUT Layer). There is a SINGLE INPUT source.
65,536 Input Neuron
3 Single Neuron

N1 is the Hidden Layer 1 Neuron learned the first Image,
N2 is the Hidden Layer 2 Neuron learned the second Image,
N3 is the Hidden Layer 3 Neuron learned the third Image.

Fig. 5 It shows the architecture of the NNET block of the L-HONN filter

optical correlator as would a linear superposition constrained-type filter, such as the SDF-type [62-64] filters. It may be used as a space domain function in a joint transform correlator architecture or be Fourier transformed and used as Fourier domain filter in a 4-f Vander Lugt [65] type optical correlator. Here, the elements of matrix $\Gamma_{i=1\cdots N}$ (or written in vector form as $\gamma_{i=1\cdots N}$), with N the total number of training set images are the dot products of the layer and input weights of the NNET. We then, as for all the HONN-type filters, calculate the dot product of the matrix elements of $\Gamma_{i=1\cdots N}$ with the corresponding logmapped training image matrix elements of $Q_{i=1\ldots N}$ (element wise multiplication of the two matrices). The logmapped image was transformed by the space-variant sensor from the original training image $x_{i=1\cdots N}$ and after, forming its polar form $(r\angle\theta)_{i=1\ldots N}$, mapped to the z-plane. Assume we have N logmapped training images of size $[\mathrm{u}\times\mathrm{v}]$ and we represent each training image with

$Q_{i=1\ldots N}(u,v)$, the weights from node ι to node κ with $w_{\iota\kappa}^{Q_i}$ and the layer weights with $l_{\iota\kappa}^{Q_i}$. It is noted that ι is the symbol used for the neuron node and $i=1,\ldots,N$ is the index used for the logmapped training set images. Then:

$$\Gamma_{i=1\cdot\cdot N}=W^{Q_i}\cdot L^{Q_i}$$

$$=\begin{bmatrix} w_{11}^{Q_i}l_{11}^{Q_i}+w_{12}^{Q_i}l_{21}^{Q_i}+\cdots+w_{1v}^{Q_i}l_{v1}^{Q_i}\cdots w_{11}^{Q_i}l_{1\omega}^{Q_i}+w_{12}^{Q_i}l_{2\omega}^{Q_i}+\cdots+w_{1v}^{Q_i}l_{v\omega}^{Q_i} \\ w_{21}^{Q_i}l_{11}^{Q_i}+w_{22}^{Q_i}l_{21}^{Q_i}+\cdots+w_{2v}^{Q_i}l_{v1}^{Q_i}\cdots w_{21}^{Q_i}l_{1\omega}^{Q_i}+w_{22}^{Q_i}l_{2\omega}^{Q_i}+\cdots+w_{2v}^{Q_i}l_{v\omega}^{Q_i} \\ \vdots \\ w_{u1}^{Q_i}l_{11}^{Q_i}+w_{u2}^{Q_i}l_{21}^{Q_i}+\cdots+w_{uv}^{Q_i}l_{v1}^{Q_i}\cdots w_{u1}^{Q_i}l_{1\omega}^{Q_i}+w_{u2}^{Q_i}l_{2u}^{Q_i}+\cdots+w_{uv}^{Q_i}l_{vu}^{Q_i} \end{bmatrix} \quad (29)$$

$$S_{i=1\cdots N}=\Gamma_{i=1\cdots N}\cdot Q_{i=1\cdots N}(u,v) \quad (30)$$

where W^{Q_i} and L^{Q_i} are the matrices of the input and layer weights. $w_{uv}^{Q_i}$ are the input and layer weights from the input neuron of the input vector element at row u and column v to the associated hidden layer for the training image $Q_{i=1\cdots N}(u,v)$. $l_{uv}^{Q_i}$ are the input and layer weights from the hidden neuron of the layer vector element at row u and column v to the associated output neuron. It is emphasised that above in the text and in eqn. (29) 'Q_i' and, 'uv' and '$\iota\kappa$' are used as upper script and lower script notations and not as power or logarithmic base indices. In our case, ω = 1 since the output layer has only one neuron. $S_{i=1\cdots N}(u,v)$ is the transformed image calculated from the dot product of the matrix elements of $\Gamma_{i=1\cdots N}$ with the corresponding training image matrix elements of $Q_{i=1\cdots N}(u,v)$.

5 Simulation Results

It was confirmed experimentally that, as for all the HONN-type filters [59-61], by choosing different values of the target classification levels for the True-class $\mathrm{T}_{\mathrm{true}}$ and False-class $\mathrm{T}_{\mathrm{false}}$ objects i.e. the output layer's neuron's target output for the True-class object and the False-class object, respectively, of the NNET (see Fig. 5), the L-HONN filter's behaviour can be varied to suit different application requirements. Hence, by increasing the absolute distance of the target classification levels between the True-class objects and the False-class objects, $\Delta T = \left| T_{true} - T_{false} \right|$ the L-HONN filter exhibited generally sharp peaks and good clutter suppression but is more sensitive to intra-class distortions i.e. it behaves more like a high-pass biased filter, whereas by decreasing the absolute distance ΔT the L-HONN filter exhibited relatively good intra-class distortion invariance but producing broad correlation peaks i.e. it behaves more like a minimum variance synthetic discriminant function (MVSDF) filter [66]. For our application purposes it was found to be adequate to use for both the still images data sets and the video frame sequences of the single object True-class and of the multiple objects of same True-class, $T_{true} = +40$ for the True-class object's target classification level and $T_{false} = -40$ for the False-class object's target classification level. Similarly, we could have used different T_{true} and T_{false} values for the video frame sequences than the values we used for the still images.

First, we test the L-HONN filter with still images and its tolerance to background clutter in the input scene by the insertion of training and non-training single and multiple of the same class object images into different car park scenes. Both training set and test set images were used in grey-scale bitmap format. As for all the HONN-type filters, all the training set and test images prior being processed by the NNET are concatenated row-by-row into a vector of size $u \times v$. The training set consisted of at least two images of the Jaguar S-type for a distortion range over 20° to 70° degrees at 10° increments. We added at least one of the three used in total background images of typical empty car park scenes (different than the actual background scenes of the test images) in the training set of the NNET to fall inside the False-class. For increasing the discrimination ability between the different object classes of the L-HONN filter when the filter it is tested with the cluttered images we can add in the training set an image of the Police patrol car model Mazda Efini RX-7 False-class object. The training set images were of size 256×256. All the False-class object images were constrained to zero peak-height constraint and all the True-class object images to unit peak-height constraint in the synthesis of the L-HONN filters' composite image (at the correlator-type block, see Fig. 4).

Fig. 6 (a) and Fig. 7 (a) shows the test input still images used for our conducted simulations. It consists of one image of the Jaguar True-class object at 55°, inserted in a car park scene, in-plane rotated to 90° and shifted off the centre of the image, scaled down from the original training set images to 128×128 and a second image of the Jaguar True-class object at 40° inserted in a different car park scene, in-plane rotated to 90° and shifted off the centre of the image (at a different shifted position than the first image), which was scaled down from the original training set images, of size 256×256, to 96×96. During the process of inserting the object in to the car park scenes some Gaussian noise is added, too. Fig 6 (b) and Fig. 7 (b) show the isometric correlation planes of the first test set image and of the second test set image, respectively. Fig. 6 (a) and Fig. 7 (a) show the window unit locked on top of the part of the images the L-HONN filter gave the highest and

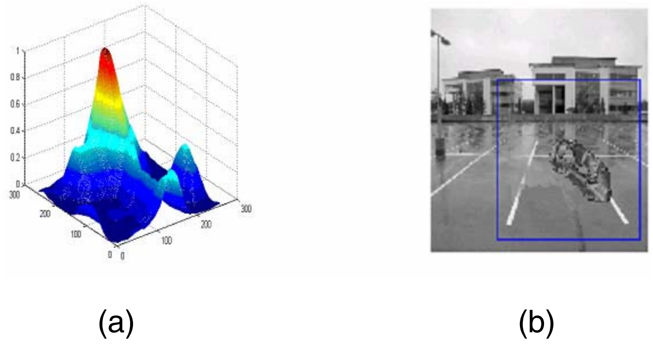


Fig. 6 Simulation results for the test set image of the Jaguar true-class object at 55˚ inserted in a car park scene, in-plane rotated to 90˚ and shifted off the centre of the image, scaled down from the original training set images to 128 x 128 with added noise, too: (a) the isometric correlation plane output and (b) the actual test image with the window unit on top of the recognised area.

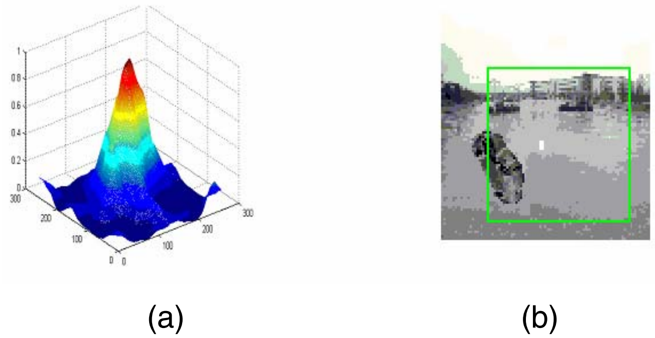


Fig. 7 Simulation results for the second test set image of the Jaguar true-class object at 40˚ inserted in a different car park scene, in-plane rotated to 90˚ and shifted off the centre of the image to a different position than the first test image, scaled down from the original training set images of size 256 x 256 to 96 x 96 with added noise, too: (a) the isometric correlation plane output and (b) the actual test image with the window unit on top of the recognised area.

sharpest correlation plane peak. For Fig. 6 (a) the window unit size, WU was 84×84 pixels and the step size was 2×2 pixels for the x-y axis. The total number of sub-images produced was 484. For Fig. 7 (a) the window unit size, WU was 60×60 pixels and the step size was 3×3 pixels for the x-y axis. The total number of sub-images was 144. Both isometric plots are normalised to the maximum correlation peak-height value of the correlation planes of the entire test sub-images produced by the window unit for each test image. The rest of the sub-images, outside the recognised area of the True-class object, were suppressed sufficiently by the L-HONN filter.

The design of the L-HONN filter's window unit can accommodate the recognition of multiple input objects of the same class within cluttered still images and video frame sequences. Fig. 8 shows the sub-images created by the window unit of size WU equal to 45×45 pixels. The step size was set to 20×20 pixels. The input image is scaled down from the original training set images size of 256×256 to 96×96. Then, the total number of sub-images created was 9. We have modified slightly the initial design of the window unit, and, more specifically, the feedback connection between the output correlation plane (see Fig. 4) and the window unit. Now, to suit the design of the window unit multiple input objects of the same class the feedback connection is used to check whether or not the maximum output correlation plane peak intensity for each created sub-image is greater than the average peak intensity value of all the sub-images and not, as before, to signal the detection of the maximum correlation peak height for locking the mask on top of the sub-image that produced it.

Fig. 8 shows one of the input still images we used for testing the L-HONN filter's performance of recognizing multiple objects of the same class within a cluttered scene. It consists of two Jaguar True-class objects out-of-plane rotated at 40°, inserted in a car park scene, one is at the centre of the image and the second Jaguar object is shifted off the centre of the image. The image is scaled down from the original training set images size of 256×256 to 96×96. As before, during the process of inserting the object in to the car park scenes some Gaussian noise is added, too. Fig. 9 shows the window unit mask locked on top of the sub-images that gave correlation peak intensity values greater than the average correlation peak intensity of all the sub-images and not, as for the single True-class object case we described before, that gave the highest and sharpest correlation plane peak intensity value. Also, it shows the isometric correlation planes for those sub-images. The window unit size, WU was 30×30 pixels. For setting the step size of the window unit we implemented the simple heuristic rule we described in Sect. 3. Hence, we set the step size to be 20×20, approximately equal to half of the True-class object size. Then, the total number of sub-images created was 16. As before, all the isometric planes are normalized to the maximum correlation peak-height value of the correlation planes of the entire test sub-images produced by the window unit for each test image.

From the recorded results for the first series of tests with still images, it is obvious that the L-HONN filter was able to detect and classify correctly the Jaguar S-type car for both cases of a single object and of multiple objects within an unknown car park scene and suppress the background clutter. From Fig. 7 and Fig. 8, the filter was able to detect and classify correctly the Jaguar S-type car within the unknown car park scene and suppress the background clutter for both test images. Hence, from the shown results, the filter exhibited in-plane rotation invariance and scale invariance by recognising the 90° in-plane rotated True-class object within the car park scene, which was scaled to different sizes each time from the 256×256 size training set images. The True-class Jaguar S-type car object was shifted off the centre for both test set images and different out-of plane rotation angles were used; at 55° for the first test image and at 40° for the second test image. The L-HONN filter successfully recognised the object for both cases by

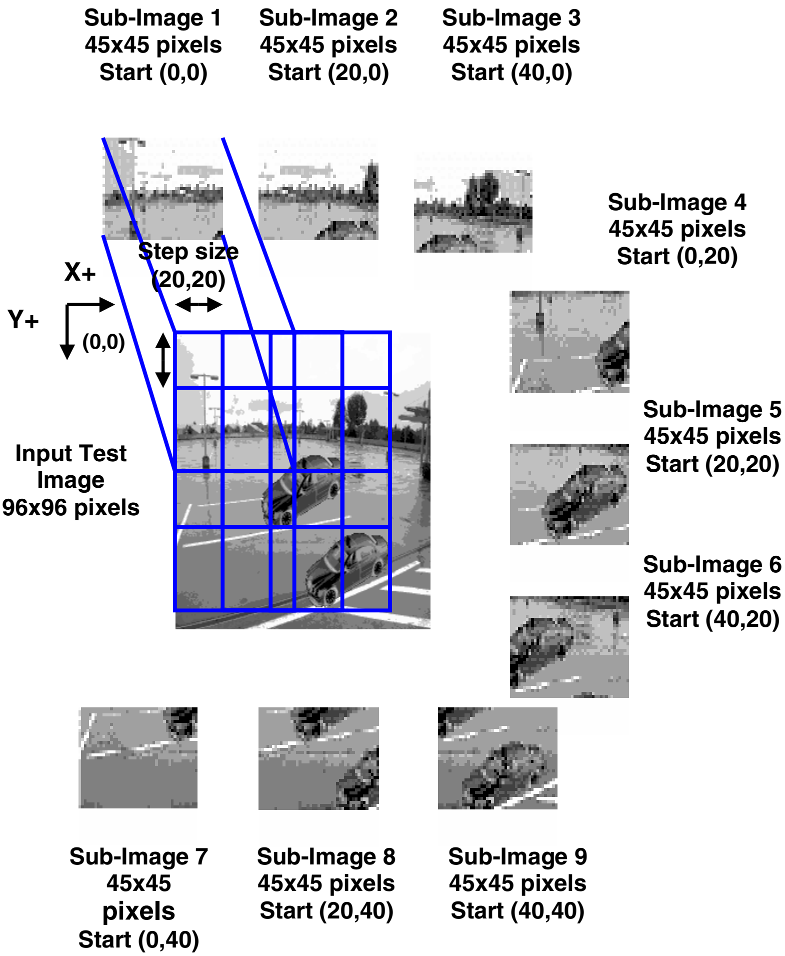


Fig. 8 We test the L-HONN filter's performance for multiple objects of the same True-class recognition. It shows the sub-images created by the window unit of size WU to be equal to 45×45 pixels and the step size to be equal approximately half the size of the Jaguar S-type car object i.e. 20×20 pixels. The test input image is scaled down to 96×96 from the original input image size of 256×256. In effect, there are 9 sub-images created.

exhibiting shift and out-of-plane rotation of the True-class object invariance. The unknown (not previously included in the training set) car park scene used in the second test set image was more difficult to suppress by the L-HONN filter, due to the several False-class objects interfering in the correct recognition of the Jaguar. The car park scene used in the first test image consisted of fewer non-uniform areas of False-class objects, making the True-class object recognition an easier task for the filter. It was found that by varying the step size of the window unit to 1×1 pixels, the overall performance of the L-HONN filter was improved due to the

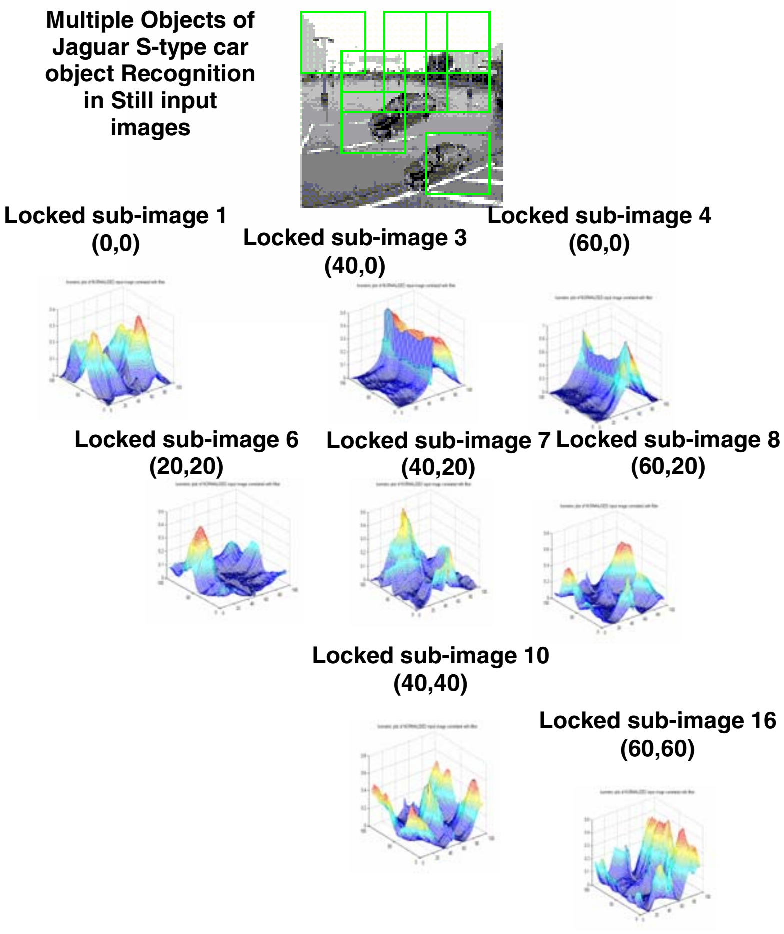


Fig. 9 It shows the window unit mask locked on top of the sub-images that gave correlation peak intensity values greater than the average correlation peak intensity of all the sub-images. Also, it shows the isometric correlation planes for those sub-images. The window unit size, WU was 30×30 pixels. We set the step size to be 20×20, approximately equal to half of the True-class object size. Then, the total number of sub-images created was 16.

sensitivity of the logmap in translation of the object being 1 or 2 pixels off the point of expansion.

Similarly, from Fig. 9, the L-HONN filter was able to detect and classify correctly both Jaguar S-type car True-class objects within the unknown car park scene and suppress the background clutter. Thus, the filter exhibited out-of-plane rotation invariance by recognising both Jaguar S-type objects out-of-plane rotated at 40° (not included in the training set), and scale invariance by recognising the Jaguar S-type objects within the car park scene, which was scaled from the original size of 256×256 to 96×96. Also, the window unit's modified design confirmed that it enable the recognition at different off the centre positions of the Jaguar S-type objects. Here, it worth mentioning that though there were more than one masks, including couple of False-class masks drawn on top of background regions, locked on top of the area of each of the Jaguar S-type cars, it is obvious from the density and the two formed separate clusters of locked masks that two different True-class objects, as expected, were recognized. That is due to the window unit's modification of the feedback connection to the output correlation plane of the filter, the mask is locked on top of the sub-image that give a correlation peak intensity value greater than a pre-specified threshold value rather than the one that gives the maximum correlation peak intensity value (as for Fig. 7 and Fig. 8). It was found that by increasing the threshold value and by reducing the step size as well as the WU size the filter was able to give less False-class locked masks (around the background regions).

The second series of tests we conducted was with video frame sequences of multiple objects of the same True-class within cluttered scenes. Fig. 10 shows four of the original video frames used for our simulations. On the figure we have positioned each one of the frames next to each other successively in time. The frame rate for the video sequence was 25 frames per second (fps). For the purposes of this chapter we are showing indicatively four of the frames of the video sequence used. Both training set and test set video sequences were used in grey-scale bitmap format. All the input video frames prior being processed by the NNET are concatenated row-by-row into a vector of size $u \times v$. The training set consisted of at least two still frames (images) of the Jaguar S-type for a distortion range over 20° to 70° degrees at 10° increments. We added at least one background frame of the original video sequence we used for the test. However, as we saw from Fig. 7 and Fig. 8 we could have used different than the tested video sequence background frames, since the L-HONN filter was proven to be able to recognize the True-class

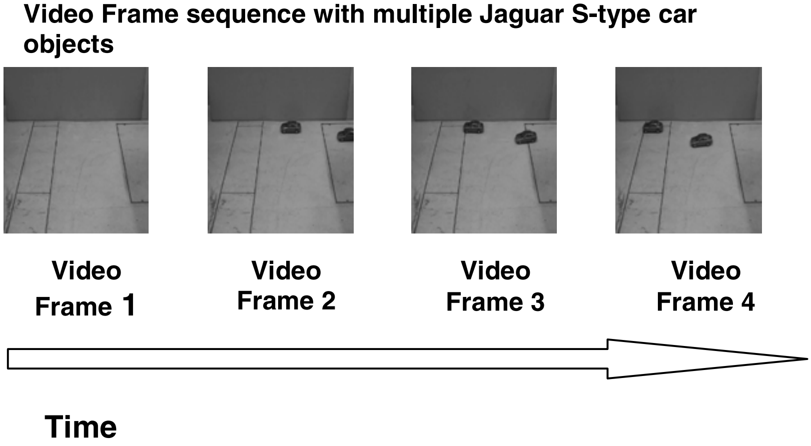

Video Frame sequence with multiple Jaguar S-type car objects
Video Frame 1
Video Frame 2
Video Frame 3
Video Frame 4
Time

Fig. 10 It shows four frames of the original video sequence used for the second series of tests with multiple objects of the same True-class. The frame rate for the video sequence was 25 fps. Each frame is positioned on the figure next to each other successively in time.

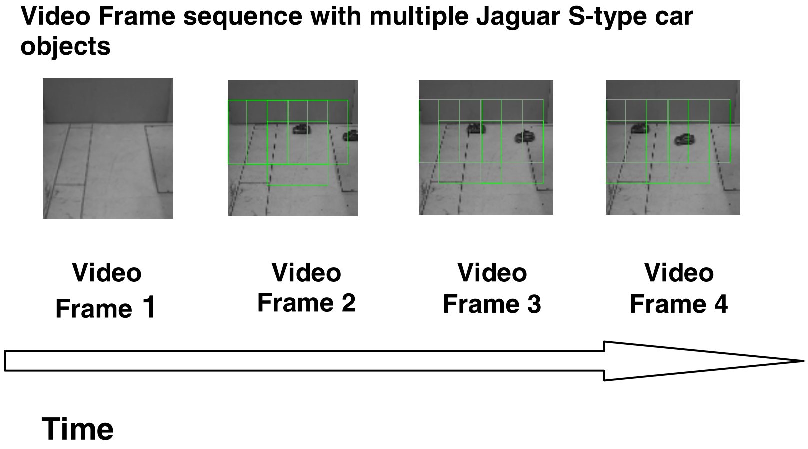

Video Frame sequence with multiple Jaguar S-type car objects
Video Frame 1
Video Frame 2
Video Frame 3
Video Frame 4
Time

Fig. 11 It shows the window unit mask locked on top of the True-class objects which are tracked throughout the four shown frames. WU was 45×45 pixels. We set the step size to be 15×15. Then, the total number of sub-frames (sub-images) created was 16. The filter has correctly recognized and tracked the Jaguar S-type car objects for the shown frames and suppress the background clutter. We can improve the performance of the L-HONN filter and reduce the False-class locked masks by increasing the threshold value, and by reducing the step size and the WU size of the filter.

within unknown cluttered scenes (i.e. cluttered scenes, e.g. of car parks, not included in the training set). As before, for increasing the discrimination ability between the different object classes of the L-HONN filter when the filter it is tested with the cluttered video frames, thus producing sharper correlation peaks, we can add in the training set a still frame (image) of the Police patrol car model Mazda Efini RX-7 False-class object. The training set frames were resized to 256×256. All the False-class object frames were used in the training set were constrained to zero peak-height constraint and all the True-class object images to unit peak-height constraint in the synthesis of the L-HONN filters' composite image.

Fig. 11 shows the window unit mask locked on top of the True-class objects which are tracked throughout the four shown frames. Now, the window unit size, WU was 45×45 pixels. We set the step size to be 15×15. Then, the total number of sub-frames (sub-images) created was 16. For both of the test series each sub-image or sub-frame of the window unit was processed by a Dual Core CPU at 2.4 GHz with 3.50GB RAM in few a msec for all the conducted experiments from the L-HONN filter.

From Fig. 11 we see that the filter has correctly recognized and tracked the Jaguar S-type car objects for the shown frames and suppress the background clutter. Again, the L-HONN filter has exhibited out-of-plane rotation invariance. Here, it worth noticing that during the video sequence as the car's wheel steers to follow its route the out-of-plane rotation angle changes, too. In effect, during the complete video sequence the out-of-plane rotation angle of one of the cars changed from approximately 0° to 15°. The filter was still able to correctly recognize and track that car. It exhibited scale invariance by recognizing the Jaguar S-type objects within the background scene, which was scaled (resized) from the original size of 256×256 to 96×96. As for the multiple objects of the True-class test of still images, we can improve the performance of the L-HONN filter and reduce the False-class locked masks by increasing the threshold value, and by reducing the step size and the WU size of the filter. Nevertheless, for both test series the L-HONN filter suppressed successfully the added Gaussian noise and the unknown background scene to give good detection peaks at the output correlation plane for recognising the True-class object, as shown by the isometric correlation planes.

6 Internet Multimedia Applications

Searching for web pages for retrieving certain information of our interest is one of the most common tasks performed on the web. However, as the size and popularity of Internet grows in combination with the lack of structure style and adequate web search technologies for locating any type of multimedia content on the web pages turns web search into a difficult task for being performed satisfactorily. The continuously growing demands of the users for accessing new types of content, such as audio, video, graphics and images resulted in introducing hypermedia embedded in the web pages. Nowadays, most of the web pages are a mixture of text, images, audio and video content. Currently, most multimedia search engines use text searches for locating any multimedia content. The difficulty in implementing fast and accurate multimedia content web search engines is directly related to indexing the multimedia content of the web pages.

The simultaneous properties of input object shift, in-plane rotation, out-of-plane rotation, scale and projection invariances can reduce the number of stored images of the same object and, consequently, reduce the time needed for an Internet image-to-image search engine (content-based search engine for still images or stored video frames) to search the complete dataset of matched images. Hence, the integration of the L-HONN filter with the current content-based Internet search engines can prove advantageous [67-69]. Already, in literature have been described several Internet image-to-image search engines (in contrast to text-to-image that current popular Internet search engines perform). The PicToSeek [70] image-content web search engine exploits the use of web-robots to discover and collect the web image documents. ImageRover [71], as in the PicToSeek system, collects the images by the use of web-robots. The ImageScape [72] image-to-image Internet search engine allows queries based on keywords, semantic icons and user-drawn sketches. WebSeer [73] image content-based web search engine uses text input for forming the query for an image. VisualSeek [74] enables querying by image regions and image spatial layout. It combines both, feature-based image indexing with spatial query methods. Each image is decomposed into regions which have feature properties, such as colour, and spatial properties such as size, location and relationships to other regions. Moreover, by augmenting the number of neurons in the output layer of the NNET block it is suggested that more than one different class objects can be recognized by the same L-HONN filter. Thus, the training times are reduced instead of training several filters for the different object classes. Finally, the good performance in recognizing the True-class object and suppressing the clutter, even for unknown scenes, exhibited by the L-HONN filter (see Sect. 5.1) is further advantageous in being integrated with the Internet content-based search engines (for still images or stored video frames) where the searched object can be inserted in different background scenes.

7 Conclusion

We have been motivated to combine a space-variant imaging sensor with the HONN filter to increase its distortion invariant abilities and produce a real-time performance in our final object recognition system. The polar exponential geometry of the sensor offers in-plane rotation, scale and projection invariances to the combined filter with respect to its centre expansion. Moreover, the sensor parameter values can be tuned to fit the application's requirements in resolution, compression rate, and the available amount of memory and computational resources.

L-HONN filter was proven above from the conducted simulations with both still images and video frames to exhibit invariance to in-plane rotation and scaling inherited by the logmapping. The overall design of the L-HONN filter provided invariance to out-of-plane rotation and shift, and noise and clutter tolerance. We have demonstrated the dual importance of the window-based system. First, it maintains the otherwise lost translation invariance off the centre of the logmap expansion by sub-diving the original image or video frame into smaller size images, before applying the logmapping on top of each of these sub-images (or sub-frames). Second, the window unit design can be slightly modified to allow multiple objects of the same True-class within the input image or video frame to be recognised. It is emphasized that a single L-HONN filter and a single pass of the input data through the filter achieves the mentioned invariances and tolerance to noise and clutter without the need of several filters separately being trained for achieving each time any of them or several input data passes. There is no need for a separate pre-processing stage of the input video frames for background segmentation, but the filter is able to successfully recognize and track the True-class objects and suppress even unknown (not included in the training set) background scenes.

Acknowledgments. The authors would like to thank their colleagues in the Laser and Photonics Systems Research Group and everyone else that contribute towards the completion of this chapter.

List of Abbreviations

hybrid optical neural network	HONN
logarithmic r-θ mapping for the hybrid optical neural network	L-HONN
inverse synthetic aperture radar	ISAR
artificial neural network	NNET
constrained-HONN	C-HONN
synthetic discriminant function	SDF
minimum variance SDF	MVSDF

References

[1] Gonzalez, R.C., Woods, R.E.: Digital image processing. Addison-Wesley Publ. Co., Reading (1993)

[2] Weiman, C.F.R., Chaikin, G.: Logarithmic spiral grids for image processing and display. Comp. Graph and Imag Process 11, 197–226 (1979)

[3] Weiman, C.F.R.: 3-D sensing with polar exponential sensor arrays. In: Digit and Opt. Shap Represent and Pattern Recognit. Proc. SPIE Conf. on Pattern Recognit. and Signal Process, vol. 938, pp. 78–87 (1988)

[4] Weiman, C.F.R.: Exponential sensor array geometry and simulation. In: Digit and Opt. Shap Represent and Pattern Recognit. Proc. SPIE Conf. on Pattern Recognit. and Signal Proc., vol. 938, pp. 129–137 (1988)

[5] Weiman, C.F.R., Juday, R.: Tracking algorithms using log-polar mapped image coordinates. In: Proc. SPIE Conf. Intell Robots and Comp. Vis. VIII, vol. 1192, pp. 843–853 (1989)

[6] Weiman, C.F.R.: Video compression via log-polar mapping. In: Real-Time Imag Process II, SPIE Symp. on OE/Aerosp Sens., vol. 1295, pp. 266–277 (1990)

[7] Grossberg, S.: Adaptive pattern classification and universal recoding, II: feedback, expectation, olfaction, and illusions. Biol. Cybern 23, 187 (1976)

[8] Grossberg, S.: Studies of mind and brain: neural principles of learning, perception, development, cognition, and motor control. Reidel, Boston (1982)

[9] Grossberg, S.: Competitive learning: from interactive activation to adaptive resonance. Cogn. Sci. 11 (1987)

[10] Grossberg, S.: Neural Networks and natural intelligence. MIT Bradford Press, Cambridge (1988)

[11] Carpenter, G.A., Grossberg, S.: A massively parallel architecture for a self-organizing neural pattern recognition machine. Comp. Vis. Graph Imag Process 37 (1987)

[12] Carpenter, G.A., Grossberg, S.: ART 2: Stable self-organization of pattern recognition codes for analog input patterns. In: IEEE Proc. 1st Int. Conf. on Neural Networks, San Diego (1987)

[13] Carpenter, G.A., Grossberg, S.: Invariant pattern recognition and recall by an attentive self-organizing ART architecture in an non-stationary world. In: IEEE Proc. 1st Int. Conf. on Neural Networks, San Diego (1987)

[14] Carpenter, G.A., Grossberg, S.: ART 2: Self-organization of stable category recognition codes for analogue input patterns. Appl. Opt. 26, 4919–4930 (1987)

[15] Carpenter, G.A., Grossberg, S.: The ART of adaptive pattern recognition. IEEE Comp. 21 (1988)

[16] Carpenter, G.A., Grossberg, S.: ART3: Hierarchical search using chemical transmitters in self-organising pattern-recognition architectures. Neural Networks 3, 129–152 (1990)

[17] Carpenter, G.A., Grossberg, S., Reynolds, J.H.: ARTMAP: Supervised real-time learning and classification of nonstationary data by a self-organizing neural network, vol. 4, pp. 565–588 (1991)

[18] Carpenter, G.A., Grossberg, S.: Pattern recognition by self-organising neural networks. MIT Press, Cambridge (1991)

[19] Carpenter, G.A., Grossberg, S., Rosen, D.B.: Fuzzy Art: an adaptive resonance algorithm for rapid, stable classification of analog patterns. In: Proc. Int. Conf. Neural Networks II, pp. 411–420 (1991)

[20] Carpenter, G.A., Grossberg, S., Rosen, D.B.: Fuzzy Art: fast stable learning and categorisation of analog patterns by an adaptive resonance system. Neural Networks 4, 759–771 (1991)

[21] Carpenter, G.A., Grossberg, S., Markuzov, N., Reynolds, J.H.: Fuzzy ARTMAP: a neural network architecture for incremental supervised learning of analogue multidimensional maps. IEEE Trans. on Neural Networks 3, 698–713 (1991)

[22] Fukushima, K., Miyake, S., Ito, T.: Neocognitron: a neural network model for a mechanism of visual pattern recognition. IEEE Trans. Syst. Man Cybern. 13, 826–834 (1983)

[23] Fukushima, K.: Analysis of the process of visual recognition by the Neocognitron. Neural Networks 2, 413–420 (1989)

[24] Fukumi, M., Omatu, S.: A new back-propagation algorithm with coupled neuron. IEEE Trans. Neural Networks 2, 535–538 (1991)

[25] Fukumi, M., Omatu, S., Takeda, F., Kosada, T.: Rotation-invariant neural pattern recognition system with application to coin recognition. IEEE Trans. Neural Networks 2, 272–279 (1992)

[26] Fukumi, M., Omatu, S., Nishikawa, Y.: Rotation-invariant neural pattern recognition system estimating a rotation angle. IEEE Trans. Neural Networks 8, 568–581 (1997)

[27] Pantle, A., Sekuler, R.: Size-detecting mechanisms in human vision. Sci. 162, 1146–1148 (1968)

[28] Shepard, R.N., Metzler, J.: Mental rotation of three-dimensional objects. Sci. 171, 701–703 (1971)

[29] Takano, Y.: Perception of rotated forms: a theory of information types. Cogn. Psychol. 121, 1–59 (1989)

[30] Sandini, G., Dario, P.: Active vision based on space-variant sensing. In: 5th Int. Symp. Robot. Res., pp. 75–83 (1989)

[31] Tistarelli, M., Sandini, G.: On the advantages of polar and log-polar mapping for direct estimation of time-to-impact from optical flow. IEEE Trans. Pattern Anal. Mach. Intell. 15, 401–410 (1993)

[32] Schwartz, E.L., Greve, D., Bonmasser, G.: Space-variant active vision: definition, overview and examples. Neural Networks 8, 1297–1308 (1995)

[33] Bederson, B., Wallace, R., Schwartz, E.L.: A miniaturised space-variant active vision system: Cortex-I. Mach. Vis. Appl. 8, 101–109 (1995)

[34] Lewitt, B.: Reconstruction algorithms: transform methods. Proc. IEEE 71, 390–408 (1983)

[35] Easton Jr., R.L., Barrett, H.H.: Tomographic transformations in optical signal processing. In: Horner, J.L. (ed.) Opt. Signal Process, pp. 335–386. Acad. Press, N Y (1987)

[36] Unzueta, M., Flores, B.C., Vargas, R.A.: Two-dimensional polar format algorithm for high-quality radar image formation. SPIE Conf. Radar Process, Techol. Appl. 2845, 151–162 (1996)

[37] Choi, H., Krone, A.W., Munson Jr., D.C.: ISAR imaging of approaching targets. IEEE Trans. Imag. Process. (1997)

[38] Ho, C.G.: Comparison of interpolation techniques for polar to rectangular co-ordinate transformation with application to real-time image processing. DPhil thesis, Sch. of Eng., Univ. of Sussex, U.K (2000)

[39] Ho, C.G., Young, R.C.D., Chatwin, C.R.: Sensor geometry and sampling methods for space-variant image processing. Pattern Anal. Appl. 5, 369–384 (2002)

[40] Casasent, D., Psaltis, D.: Position, rotation, and scale invariant optical correlation. Appl. Opt. 7, 1795–1799 (1976)

[41] Mersereau, K., Morris, G.: Scale, rotation, and shift invariant image recognition. Appl. Opt. 25, 2338–2342 (1986)

[42] Kwon, H.Y., Kim, B.C., Cho, D.S., Huang, H.Y.: Scale and rotation invariant pattern recognition using complex-log mapping and augmented second order neural network. Elect. Lett. 29, 620–621 (1993)

[43] Kageyu, S., Ohnishi, N., Sugie, N.: Augmented multi-layer perceptron for rotation-and-scale invariant hand-written numeral recognition. Int. Jt. Conf. on Neural Networks 1, 54–59 (1991)

[44] Messner, R.A., Szu, H.H.: An image processing architecture for real time generation of scale and rotation invariant patterns. CV GIP 31, 50–66 (1985)

[45] Car parks image database (October 2004), `http://www.locations-uk.com/CarParks.html`

[46] Hsu, Y.N., Arsenault, H.H., April, G.: Rotation-invariant digital pattern recognition using circular harmonic expansion. Appl. Opt. 21, 4012 (1982)

[47] Hsu, Y.N., Arsenault, H.H., April, G.: Optical pattern recognition using circular harmonic expansion. Appl. Opt. 21, 4016 (1982)

[48] Hsu, Y.N., Arsenault, H.H., April, G.: Pattern discrimination by multiple circular harmonic components. Appl. Opt. 23, 841 (1984)

[49] Jensen, A.S., Lindvold, L., Rasmussen, E.: Transformation of image positions, rotations, and sizes into shift parameters. Appl. Opt. 26, 1775–1781 (1987)

[50] Bryngdahl, O.: Geometrical transformations in optics. J. of Opt. Soc. of Am. 64, 1092 (1974)

[51] Nagano, T., Ishikawa, M.: A neural network for size-invariant feature extraction. In: Proc. Int. Conf. Neural Networks, vol. 1, p. 103 (1990)

[52] Rowley, H.A., Baluja, S., Kanade, T.: Neural network-based face detection. IEEE Tran. on Pattern Anal. Mach. Intell. 20, 23–38 (1998)

[53] Drucker, H., Schapire, R., Simard, P.: Boosting performance in neural networks. Int. J. Pattern Recognit. and Artif. Intell. 7, 705–719 (1993)

[54] Sung, K.K.: Learning and example selection for object and pattern detection. PhD thesis MIT AI Lab, AI Tech. Rep. 1572 (1996)

[55] Osuna, E., Freund, R., Girosi, F.: Training support vector machines: an application to face detection. Comp. Vis. and Pattern Recognit., 130–136 (1997)

[56] Moghaddam, B., Pentland, A.: Probabilistic visual learning for object detection. In: Proc. 5th Int. Conf. Comp. Vis., pp. 786–793 (1995)

[57] Colmenarez, A.J., Huang, T.S.: Face detection with information-based maximum discrimination. Comp. Vis. and Pattern Recognit., 782–787 (1997)

[58] Umezaki, T.: Personal Communication (1995)

[59] Kypraios, I., Young, R.C.D., Birch, P., Chatwin, C.R.: Object recognition within cluttered scenes employing a hybrid optical neural network (HONN) filter. Opt. Eng. Special Issue on Trends in Pattern Recognit. 43, 1839–1850 (2004)

[60] Kypraios, I., Young, R.C.D., Chatwin, C.R.: Performance assessment of unconstrained hybrid optical neural network (U-HONN) filter for object recognition tasks in clutter. In: Proc. SPIE Opt. Pattern Recognit. XV, vol. 5437, pp. 51–62 (2004)

[61] Kypraios, I., Lei, P.W., Birch, P., Young, R.C.D., Chatwin, C.R.: Appl. Opt. 47, 3378–3389 (2008)

[62] Caulfield, H.J., Maloney, W.: Improved discrimination in optical character recognition. Appl. Opt. 8, 2354–2356 (1969)

[63] Hester, C.F., Casasent, D.: Multivariant technique for multiclass pattern recognition. Appl. Opt. 19, 1758–1761 (1980)

[64] Bahri, Z., Kumar, B.V.K.: Generalized synthetic discriminant functions. Appl. Opt. 5, 562–571 (1988)

[65] Vander Lugt, A.: Signal detection by complex spatial filtering. IEEE Trans. on Inf. Theory 10, 139–145 (1964)

[66] Kumar, B.V.K.: Minimum variance synthetic discriminant functions. J. Opt. Soc. Am. 3, 1579–1584 (1986)

[67] Kypraios, I.: Introduction: Background theory on web search technologies. In: Integrating invariant object recognition tools into an Internet search engine, MSc in Modern Digital Communication Systems Thesis, University of Sussex, U.K (2005)

[68] Kypraios, I.: Invariant Object Recognition Tools. In: Integrating invariant object recognition tools into an Internet search engine, MSc in Modern Digital Communication Systems Thesis, University of Sussex, U.K (2005)

[69] Kypraios, I.: Visual Internet search engine (VISE). In: Integrating invariant object recognition tools into an Internet search engine, MSc in Modern Digital Communication Systems Thesis, University of Sussex, U.K (2005)

[70] Najork, M., Heydon, A.: High performance web crawling. COMPAQ Sys. Res. Cent. SRC Res. Rep. 173, `http://www.research.compaq.com/SRC/` (accessed September 1, 2005)

[71] Sclaroff, S., Taycher, L., La Cascia, M.: ImageRover: a content-based image browser for the world wide web. In: Proc. IEEE Workshop on Content-Based Access of Image and Video Libr. (1997)

[72] Gevers, T., Smeulders, A.: The PicToSeek www image search system. In: Proc. IEEE Int. Conf. on Multimed Comput. and Sys., Florence, Italy, pp. 264–269 (1999)

[73] Swain, M.J., Frankel, C., Athitsos, V.: WebSeer: an Image search engine for the world wide web. In: CVPR 1997 (1997)

[74] Smith, J.R., Chang, S.F.: VisualSeek: a fully automated content-based image query system. In: ACM Multimedia 1996, Boston, MA (1996)

2D-3D Pose Invariant Face Recognition System for Multimedia Applications

Antonio Rama, Francesc Tarrés, and Jürgen Rurainsky

Abstract. Automatic Face recognition of people is a challenging problem which has received much attention during the recent years due to its potential multimedia applications in different fields such as 3D videoconference, security applications or video indexing. However, there is no technique that provides a robust solution to all situations and different applications, yet. Face recognition includes a set of challenges like expression variations, occlusions of facial parts, similar identities, resolution of the acquired images, aging of the subjects and many others. Among all these challenges, most of the face recognition techniques have evolved in order to overcome two main problems: illumination and pose variation. Either of these influences can cause serious performance degradation in a 2D face recognition system.Some of the new face recognition strategies tend to overcome both research topics from a 3D perspective. The 3D data points corresponding to the surface of the face may be acquired using different alternatives: A multi camera system (stereoscopy), structured light, range cameras or 3D laser and scanner devices The main advantage of using 3D data is that geometry information does not depend on pose and illumination and therefore the representation of the object does not change with these parameters, making the whole system more robust. However, the main drawback of the majority of 3D face recognition approaches is that they need all the elements of the system to be well calibrated and synchronized to acquire accurate 3D data (texture and depth maps). Moreover, most of them also require the cooperation or collaboration of the subject during a certain period of time. All these requirements can be available during the training stage of many applications. When enrolling a new person in the database, it could be performed off-line, with the help o human interaction and with the cooperation of the subject to be enrolled. On the contrary, the previous conditions are not always available during the test stage. The recognition will be in most of the cases in a semi-controlled or uncontrolled scenario, where the only input of the system will probably consists of a 2D intensity image acquired from a single camera. This

Antonio Rama and Francesc Tarrés
Department of Signal Theory and Communications
Universitat Politècnica de Catalunya (UPC)
e-mail: {tonirama,tarres}@gps.tsc.upc.edu

Jürgen Rurainsky
Image Processing Department
Fraunhofer Institute for Telecommunications
Heinrich-Hertz-Institut (HHI)
e-mail: rurainsky@hhi.fraunhofer.de

M. Grgic et al. (Eds.): Rec. Advan. in Mult. Sig. Process. and Commun., SCI 231, pp. 121–144.
springerlink.com

leads to a new paradigm where 2D-3D mixed face recognition approaches are used. The idea behind this kind of approaches is that these take profit of the 3D data during the training stage but then they can use either 3D data (when available) or 2D data during the recognition stage. Belonging to this category, some of 2D statistical approaches like Eigenfaces of Fisherfaces have been extended to fit in this new paradigm leading to the Partial Principal Component Analysis (P^2CA) approach. This algorithm intends to cope with big pose variations (±90◦) by using 180◦ cylindrical texture maps for training the system but then only images acquired from a single, normal camera are used for the recognition. These training images provide pose information from different views (2.5D data). Nevertheless they can also be extended to a complete 3D multimodal system where depth and texture information is used. This chapter is structured as follows: First, a brief overview of the state-of-the-art in face recognition is introduced. The most relevant methods are grouped by multimedia scenarios and concrete applications. Afterwards, novel 2D-3D mixed face recognition approaches will be introduced.

1 Introduction: Face Recognition in Multimedia Applications

One of the reasons face recognition has attracted so much research attention and development over the past 30 years is its great potential in numerous multimedia commercial applications. Zhao and Chellappa [Zhao06] grouped face recognition technology into five different categories of multimedia applications and described their advantages and disadvantages. Another example of the evolution of face processing research is the continuously proposal of different benchmarks and performance evaluation initiatives [FRVT02, FRGC05], indicating that the field is far from maturity. In this subsection, face processing approaches are briefly described or mentioned depending on how suitable they are for the addressed multimedia application scenario. A total of five different scenarios have been proposed: Access Control Points, Intelligent Human Computer Interfaces, Virtual Collaboration, Video Indexing and Retrieval and Video Surveillance.

1.1 Control Access Application Scenario

Nowadays, there are multiple access points where the verification or identification of the person is required. Most of them use a card that should be passed through a magnetic card reader. The main problem is that the card is not associated with a person. If the card is lost or stolen, anyone that gets it, will access the restricted area without any further problem. Thus, it is becoming more common the use of biometric technology together with the card to improve the security of the system. Access control by face recognition has several advantages in comparison with other biometrics systems: First, there are no requirements for expensive or specialized equipment since the system consists on a simple video camera and a personal computer. Secondly, the system is passive and the complete cooperation of the user is not needed. Any user walking or staying in front of the camera is processed by the recognition system. Generally speaking, control access applications are not only entering a physical place but also a virtual one. For instance, imagine a communication system or any other multimedia application where the person is recognized and some settings are loaded depending on the privileges of this person.

The main characteristics of this kind of scenario are that conditions are controlled and the subject to be recognized may cooperate in acquiring an almost frontal image without any kind of occlusions and with a neutral facial expression. Furthermore, the illumination should not vary significantly since the access points are in most of the cases indoors localizations. Hence, this scenario implies a low-medium challenge for the current face detection and recognition technologies. Skin detectors [McKenna98, Albiol01, Jones02, Vezhnevets03, Chenaoua06, Aureli07, Pierrard07], template-based face detectors [Craw92, Yuille92, Sinha94, Kwon94, Samal95, Miao99] or the well known state-of-the-art Adaboost face detector [Viola01, Lienhart02] may achieve a very good performance in this kind of applications. For the recognition stage, 2D statistical approaches like Principal Component Analysis (PCA) or Eigenfaces [Turk91], Fisherfaces [Belhumeur97], Independent Component Analysis [Barlett02], Kernel PCA [Kim02], or 2D PCA [Yang04] may be enough for solving the problem with a high performance.

1.2 Intelligent Human Computer Interface (IHCI)

In Intelligent HCI applications a key goal is usually to give machines some perceptual abilities that allow them to interact naturally with people: To recognize people and remember their preferences and peculiarities, to know what they are looking at, and to interpret their gestures and facial expression changes. This could be very useful to facilitate typical tasks with machines especially for illiterate, blind or dumb people. An example of these tasks could be getting money from a cash machine. The IHCI may be able to recognize and interpret the movements of the lips or gestures of the hands as different commands. This can be extended to other environments such as smart rooms where a gesture may indicate to switch on the light projector or to pass the slide of a presentation.

The a priori requirements can vary from controlled conditions (office desk) to semi-controlled conditions (outdoor cash machine). Thus, the methods proposed have to be more robust towards some challenges such as small (mid-) pose and illumination variations, or even they have to consider the possibility of some occlusions like wearing sun glasses or any other clothing accessories. Nevertheless, since the goal is the interaction with the machine it is also supposed that the user will collaborate with the system; thus, an almost frontal image of the face with an acceptable resolution may be acquired.

For this reason, the previous commented face detection and recognition methods proposed in the "access control scenario" may also be used. However, since some small pose and illumination variations should be taken into account, face detection approaches based on neural networks [Propp92, Rowley98] and variations of the Adaboost method [Lienhart02] may produce better results. The recognition stage requires also pose robust algorithms; thus, variations of the previous statistical approaches like Self-Eigenfaces [Torres00], Eigenfeatures [Pentland94, Emidio00] can be applied. Additionally, 2D-3D mixed schemes like the Partial PCA or the Partial LDA, that will be described in detail in the next section, show their robustness towards pose variations and can be more efficient for this kind of IHCI applications. Finally, Active Appearance Models (AAM) [Cootes01, Batur05] and Elastic Graph Matching (EGM) [Lades93, Wiskott99] are also more suitable in this IHCI multimedia scenario since there are supposed to be very robust for even big pose variations.

1.3 Virtual Collaboration

Intelligent Human Computer Interface is only a small, but important, part from bigger applications for a virtual collaboration scenario. Virtual Collaboration encloses different areas or possibilities like smart rooms, e-learning, or entertainment. In these cases, face detection and recognition should be applied under semi-controlled conditions. Here there are a lot of possibilities or requirements since this kind of applications could go from a conventional 2D video conference with only one single camera available, to a complete 3D tele-inmersive system with multiple cameras acquiring different images simultaneously. Nevertheless, almost all virtual collaboration scenarios are indoors so that the illumination conditions should be relative well controlled to avoid problems of shadows. Furthermore, occlusions may also not be one of the biggest challenges in this kind of applications since the person is supposed to collaborate with the environment. However, pose and facial expression changes may occur continuously.

Thus, one main multimedia application of the Virtual Collaboration scenario represents a meeting between people in different places. An example is represented in Figure 1. A team of four individuals from a university (Terminal 1) were teamed up with a fifth individual, who is at a remote office (Terminal 2), through the Virtual Collaboration Desk for a discussion on producing a promotional video for the university.

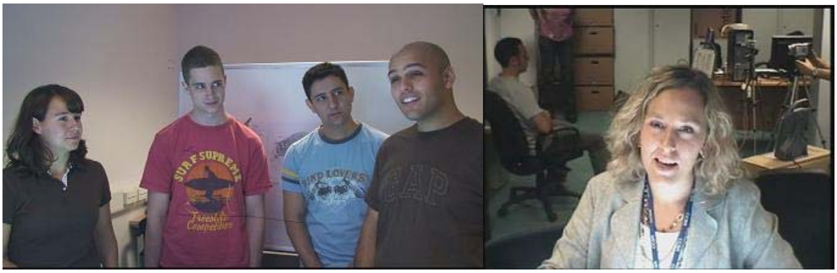


Fig. 1 Sample screenshots from a Virtual Collaboration recordings at terminal T1 (left) and terminal T2 (right)

One important point is to know which person is talking for producing, for instance, automatic reports or summaries of the meeting. Thus, face recognition may be fused with other information such as audio to detect who is talking at each moment. As an example, the audio of the sequence can be recorded not only using a conventional microphone but also a microphone array. This is useful for audio source localization which can be combined with face detection in order to identify which person is talking. Another possibility could be to fuse speaker recognition with face recognition. A visual example of output of such automatic e-meeting reports can be illustrated in Figure 2.

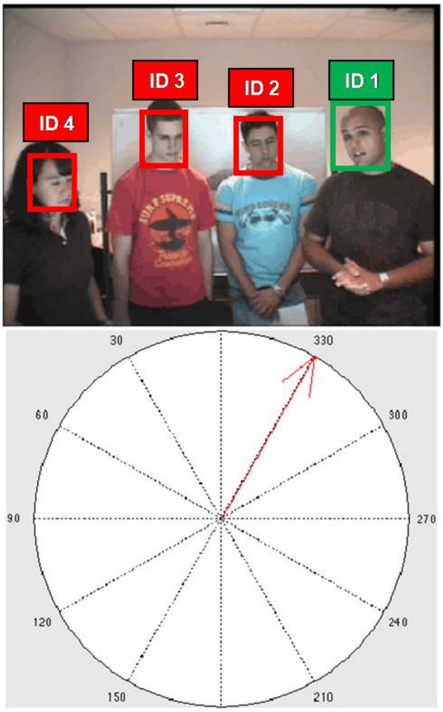


Fig. 2 E-meeting multimedia application for automatic minutes extraction

Thus, in this kind of multimedia applications face detection methods that may cope with different view angles should be developed. Examples are neural network approaches or techniques that use deformable templates. In the case of the identification stage, 3D face recognition methods [Cartoux89, Lee90, Beymer95, Beumier01, Georghiades01, Blanz03, Bowyer04, Papatheodorou04, Tsalakanidou04, Bronstein05 , Chang05, Lu05, Feng06, Lu06, Onofrio06, Samani06, Antini06, Mahoo07, Yueming07, Sen07, Kakadiaris07, Faltemier08] can improve face recognition results in virtual collaboration applications due to the possibility of extracting 3D data in a more or less accurate way since multiple cameras can be integrated and the view area of the application may be spatially limited.

1.4 Video Indexing and Retrieval

Nowadays, the amount of created digital data each year exceeds the 250 Exabytes, and more or less the 90% of this corresponds to audiovisual data. The storage of this data is a quite small problem compared with the management and the intelligent reuse of this data. Therefore, video indexing and retrieval has become a main issue in the video processing field. It has been demonstrated that faces are an important clue for searching visual data (images or videos) in large databases. Thus, automatic labeling of persons in visual databases is an important application scenario for face processing. Regarding personal multimedia applications, nowadays people use digital cameras and usually they shot hundreds of pictures during the holidays. All these pictures are stored in folders and sometimes the quantity of them is so huge that these are never again viewed. This occurs basically due to the time needed for searching to some specific picture in those folders. Thus, face recognition may be a very useful tool for making intelligent searches like "I want to seek pictures where person A and person B appear". This automatic photo album indexing is one of the main issues in face processing and particularly in face recognition. In this case, a 100% of recognition accuracy is not needed since it can be improved by the user by means of a verification process.

Particularly, video indexing and retrieval of faces depends a lot on the features of the data that wants to be catalogued. For this reason, faces can vary from controlled conditions to uncontrolled conditions depending on the nature of the video. Nevertheless, it is not necessary that the recognition rate is above the 100% since errors can be solved manually in most of the cases. Additionally, the detection and recognition stages should not be performed in real time since all the video indexing process may be done off-line. Consequently, the methods proposed in this application area have to be more robust towards more challenges such as pose and illumination variations, or some occlusions independently of the computational cost.

Face detection methods robust to illumination variations and occlusions such as a techniques using illumination robust features like Local Binary Patterns (LBP) [Heusch06, Ahonen06] or Gabor filters can be developed; or algorithms based on topology verification like the component-based face detector proposed by Goldmann et al. [Goldmann07]. In the case of face recognition for video indexing applications methods based on multiple still images [Arandjelovic05] or video-based approaches [Li01, Zhou03] would be a good selection.

1.5 Surveillance Application

Video Surveillance is one of the most common application scenarios in face processing. One or various cameras placed in some special positions in order to detect and recognize suspicious or dangerous persons without the cooperation of those. This kind of scenarios represents a big challenge in face processing since the conditions may range from semi-controlled to uncontrolled. Faces might be hidden either intentionally (with sunglasses or a mask), or unintentionally (crowded places, scarves). Furthermore, the scenario permits the presence of multiple frontal and non-frontal faces in a crowded background, with some inclination due to the angle of the camera, with changes in illumination, and another important challenge: The low resolution of the images (faces). Nevertheless, in some small or medium places the cameras could be installed in strategic places where almost frontal faces can be acquired with relatively controlled conditions. An example of this kind of VS applications is illustrated in next Figure.

Fig. 3 Sample screenshots from the Video Surveillance recordings using (*a*) Standard Definition camera in the frontal desktop and (*b*) High Definition camera pointing the main entrance

If a more complex scenario is defined, such as VS for huge places like airports or train stations, then the results produced by face recognition systems may not be very reliable. Only recent infrared-based face recognition approaches [Cutler96, Wilder96, Socolinsky04, Jinwoo06] can represent an alternative to cope with all the challenges mentioned for this application scenario.

The most important multimedia application scenarios for face processing have been presented at this point of the chapter together with some references to the more representative methods of the literature that work better in such scenarios or challenges. Now an alternative method for face recognition is described. Concretely, this method is an extension of conventional statistical approaches that tries to add a new concept: Training the system with 3D data, but then performing the test stage using 3D or 2D data depending on the main features of the scenario, i.e. if it is possible or not to acquire the 3D data during the test stage.

The rest of the chapter is structured as follows: Section 2 describes the novel concept of 2D-3D face recognition schemes. Section 3 explains some improvements in terms of accuracy and computational cost together with a novel method for the automatic creation and alignment of the 3D training using a multi-camera system. Finally, section 4 presents some results and also how the method may perform in the scenarios explained in Section 1. Finally, conclusions and future work are briefly commented.

2 Mixed 2D-3D Face Recognition Schemes

Recently some of the new face recognition strategies tend to overcome the different challenges from a 3D perspective. The 3D data points corresponding to the surface of the face may be acquired using different alternatives: a multi camera system (stereoscopy), structured light, range cameras or 3D laser and scanner devices. The main advantage of using 3D data is that depth information does not depend on pose and illumination and therefore the representation of the object do not change with these parameters, making the whole system more robust. However, the main drawback of the majority of 3D face recognition approaches is that they need all the elements of the system to be well calibrated and synchronized to acquire accurate 3D data (texture and depth maps). Moreover, most of them also require the cooperation or collaboration of the subject making them not useful for uncontrolled or semi-controlled scenarios where the only input of the algorithms will be a 2D intensity image acquired from a single camera. For this reason, the main objective of this work is to intend to ask the following question: "*It is possible to develop a face recognition framework that takes advantage of 3D data during the training stage of the system, but then, use either 2D or 3D data in the test stage depending on the possibilities of the scenario?*" Thus, this work is focused on the development of a 2D-3D face recognition framework. This framework would provide the recognition system a great flexibility so that it could be adapted to the application scenario, i.e. the recognition system will use only the information available in each situation.

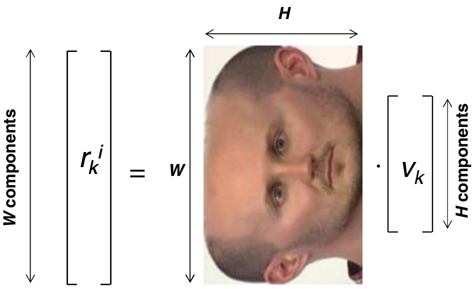
2.1 Overview of Partial Principal Component Analysis (P^{2}CA) Approach

The objective of P^{2}CA is to implement a mixed 2D-2.5D method (2D-3D when using texture and depth maps [Onofrio06]), where either 2D (pictures or video frames) or 2.5D data (180º texture images in cylindrical coordinates) can be used in the recognition stage. However, the method requires a cylindrical representation of the 2.5D face data for the training stage. Like in the majority of face recognition methods, in P^{2}CA the dimensionality of the face images is reduced through the projection into a set of *M* optimal vectors which composed the so called *feature space* or *face space*. The vectors representing the i^{th} individual are obtained as:

$$\mathbf{r}_k^{\,i} = \mathbf{A}_i^{\,T} \cdot \mathbf{v}_k \qquad k = 1,..,M, \tag{1}$$

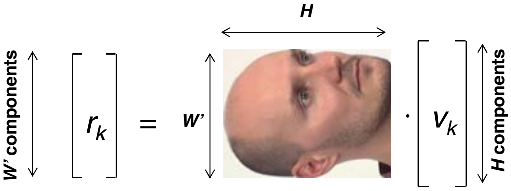
where $\boldsymbol{A}_i^T$ is the transposed if the HxW image representing individual *i*, and $\mathbf{v}_\text{k}$ are the *M* optimal projection vectors that maximize the energy of the projected vectors $\mathbf{r}_\text{k}$ averaged over the entire database. These vectors could be interpreted as unique signatures that identify each person. The projection described in equation (1) is depicted in Figure 4. The (training) texture map of the subject *i* is represented by the *M* vectors $\mathbf{r}_\text{k}^\text{i}$. Each vector $\mathbf{r}_\text{k}^\text{i}$ has *W* elements, where *W* is the width of the matrix $\boldsymbol{A}_i$.

Fig. 4 Description of a texture map (A_i^T) by means of projection vectors using P^{2}CA (training stage)



The main advantage of this representation scheme is that it can also be used when only partial information of the individual is available. Consider, for instance, the situation depicted in Figure 5, where it is supposed that only one 2D picture of the individual is available. In this case, the *M* vectors $\boldsymbol{r}_k$ representing the 2D picture, have a reduced dimension *W'*. However, it is expected that these *W'* components will be highly correlated with a section of *W'* components in the complete vectors $\boldsymbol{r}_k^i$ computed during the training stage.

Fig. 5 Projection of a "partial" 2D image through the vector set v_k (recognition stage)

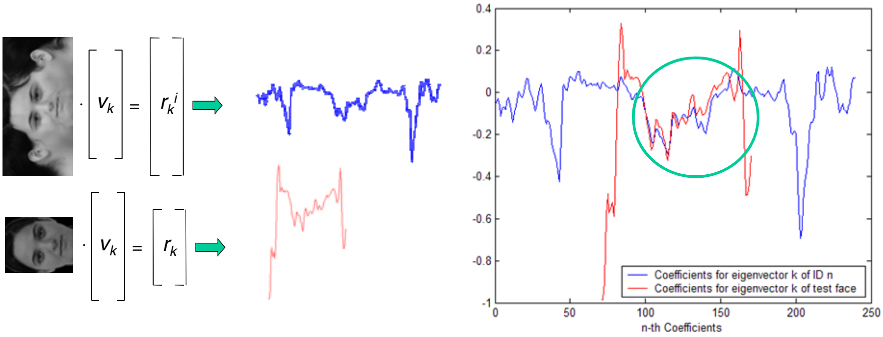


Therefore, the measure proposed below has been used to identify the partial available information (*W'* components) through the vectors $\boldsymbol{r}_k^i$ [Rama05]:

$$\min_{(i,j)}\left\{\sum_{k=1}^{M}\sum_{l=1}^{W'}\left(r_k(l)-r_k^{\,i}(l+j)\right)^2\right\} \qquad (2)$$
$$i=1,..,L; \quad j=0,..,W-W' \qquad ,$$

with L being the total number of gallery images (subjects) of the database.

In other words, the training texture maps coded 180° angle information of the person in the database. So during the test stage, any pose view image acquired from a single camera can be used to perform the recognition. It is expected that this 2D partial information after the projection in the face space is highly correlated with some part of the higher dimensionality vector of the texture map as shown in Figure 6.



a) b)

Fig. 6 a) Projection of the training texture map and a test image using only one eigenvector (face space is one dimensional). b) High correlation area of the test feature vector and the training feature vector

The Square Differences (SD) similarity measure of equation (2) has been proposed due to its relatively low computational cost. However, other measures have been used with better results for pattern matching like Normalized Cross Correlation (NCC) but at the expense of a higher computational cost:

$$\max_{(i,j)}\left\{\frac{\sum_{k=1}^{M}\sum_{l=1}^{W'}\left(r_k(l)\cdot r_k^{\,i}(l+j)\right)}{\sqrt{\sum_{k=1}^{M}\sum_{l=1}^{W'}r_k(l)^2\cdot\sum_{k=1}^{M}\sum_{l=1}^{W'}r_k^{\,i}(l+j)^2}}\right\} \tag{3}$$

$$i=1,..,L;\quad j=0,..,W-W'$$

In section 3, the computation of P^{2}CA in the frequency domain is formulated allowing us the introduction of the NCC for identification efficiently. The other limitation of the method is that the training image (Figure 4) should be well aligned when computing the face space [Rama05, Rama06]. If this is not the case, the correlation between test and the gallery images will be reduced after projecting them into the face space leading to a reduction of the face recognition rate. This is especially important for the relevant facial features since these are the key points used for the creation of the training texture maps. A process for a more accurate alignment of these texture maps is presented.

2.2 Computation of the Face Space Using P^{2}CA

The set of vectors which maximize the projection of Eq. (1) may be obtained as the solution to the following optimization problem: Find v_k for k=1,..,M such that $\xi=\sum_k\sum_n (r_k^n)^T\cdot r_k^n$ is maximum, where r_k^n is defined as the projection of image n through the vector v_k and n accounts for the number of images in the training set. The function to be maximized may be expressed as:

$$\xi=\sum_k\sum_n\left(A_n^T v_k\right)^T\cdot\left(A_n^T v_k\right)=\sum_k v_k^T\left(\sum_n A_n\cdot A_n^T\right)\cdot v_k\ , \tag{4}$$

which states that the optimum projection vectors may be obtained as the eigenvectors associated to the M largest eigenvalues of the *m*x*m* positive definite matrix C_s

$$C_s=\sum_n A_n\cdot A_n^T \tag{5}$$

This vector set will be used for feature extraction and recognition from partial information:

$$\{\mathbf{v}_1,\ldots,\underline{\mathbf{v}}_M\}$$

The procedure for feature extraction from an intensity image A consists in projecting the *transposed* image through every eigenvector:

$$\mathbf{r}_k=A^T\cdot\mathbf{v}_k\qquad k=1,\ldots,M$$

Therefore, a total of *M* feature vectors are available, with *n* (width) components each, for the image. The image has been compressed to a total of *n*x*M* scalars with *M* always being smaller than *m*.

When a complete image sample *A* (*mxn*) is available, the recognition stage is straightforward. First, the projected vectors of the sample image are computed using the previous equation and then, the best match is found as the individual *i* whose representing vectors minimize the Euclidean distance:

$$\min_{i}\left\{\xi_k = \sum_{k=1}^{M}\sum_{l=1}^{n}\left(r_k(l) - r_k{}^{i}(l)\right)^2\right\} \qquad i = 1,..,L\,, \tag{6}$$

where L represents the number of individuals in the database.

The most outstanding point of this procedure is that the image projected in the *n*-dimensional space does not need to have dimension *m*x*n* during the recognition stage so that partial information can be used. It is possible to use a reduced *p*x*n* (*p*<*m*) image which is projected to a smaller subspace.

If only partial information is used, a classification method is needed to compare the partial projection with the data in the whole space. In this case, it is not possible to use nearest neighbour classifier like in conventional PCA and correlation of partial difference methods like the criteria defined in (2) or (3) have to be applied.

The procedure is quite different from conventional PCA. Certainly, in PCA a scalar number is obtained when the vector image is projected to one eigenvector, whereas in P^{2}CA, an *n*-dimensional vector (**r**$_k$) is obtained, when the image (in matrix form) is projected to an eigenvector. It can seem that the P^{2}CA approach demands more computational cost because it uses vectors instead of numbers to represent the projections. However, the number of eigenvectors {**v**$_k$} needed in P^{2}CA for an accurate representation is much lower than in PCA.

It should be mentioned that the mathematical theory behind this approach is similar to one recent method which has extended the conventional PCA method [Turk91] from 1D to 2D; this technique was called 2DPCA [Yang04].

3 Improvements in P^{2}CA Approach

3.1 P^{2}CA in the Frequency Domain

In [Lewis95] the author presented an efficient way of computing NCC (equation (3)) in the frequency domain. Based on this work a new formulation of P^{2}CA is presented. The numerator of the NCC in equation (3) can be expressed again as:

$$c(i,j) = \sum_{k=1}^{M}\sum_{l=\frac{-(W'-1)}{2}}^{\frac{(W'-1)}{2}}\left(r_k(l)\cdot r_k{}^{i}(l+j)\right) \qquad i = 1,..,L; \quad j = -\frac{W-W'}{2},...,\frac{W-W'}{2} \tag{7}$$

For convenience we have accepted a change in the vector index *l*, choosing the zero coordinate in the center. In this case, we have to correlate vectors of L components with vectors of *W'* components in *½(W-W')* lags. This condition can be implemented very efficiently in the frequency domain. So taking the Discrete Fourier Transform (DFT) of the inner sum of the previous equation:

$$S(u) = R_k(u) \cdot R{^i}_k{}^*(u) \tag{8}$$

The correlation between $\boldsymbol{r}_k$ and $\boldsymbol{r}_k^i$ has a total of *W+W'-1* lags from which only *W-W'-1* samples are interesting for the computation. Thus, we have to avoid that spectral overlapping occurs when transforming these *W-W'-1* central. For this reason it is necessary to compute the *W*-points-DFT of $\boldsymbol{r}_k^i$ and the *W*-points-DFT of $\boldsymbol{r}^{zp}{}_k$ which consists on the zero padded version of $\boldsymbol{r}_k$. Now taking equation (1), $R_k(u)$ can be expressed as follows:

$$R_k(u) = DFT\left[\mathbf{r}_k^i\right] = DFT\left[\mathbf{A}_i^T \cdot \mathbf{v}_k\right] = DFT2D\left[\mathbf{A}_i^T\right] \cdot \mathbf{v}_k', \tag{9}$$

where $\mathbf{v}_k$' are the eigenvectors that minimize the energy projection of a given training set after applying two dimensional DFT. The demonstration of this statement have already been proved for the 1D case in [Savvides04] but the same parallelism can be followed if the images ($\boldsymbol{A}_i$) are treated like matrices.

The test stage of the P^{2}CA in frequency domain will be summarized as follows:

- Given an image $\boldsymbol{A}_{test}$, normalize this image (same as P^{2}CA in spatial domain [Rama05]).
- Extend $\boldsymbol{A}_{test}$ with columns of zeros until we have $\boldsymbol{A}^{zp}{}_{test}$ with *W* columns (same width as the training texture maps).
- Compute the DFT-2D of $\boldsymbol{A}^{zp}{}_{test}$ and project the result into the face space $\mathbf{v}_k$' obtained during the training stage.
- Obtain the product between the frequency domain test coeffients and the frequency domain weights of each identity (equation (8)).
- Compute the IDFT of (8) for the *M* different coefficients vectors where only the *W-W'-1* central samples should be considered because these are not affected by aliasing.
- Identify the identity of the database that gets a max value of the sum of the *M* IDFT vectors.

3.2 Automatic Creation of the Texture Maps and a Local Alignment Method

3.2.1 Introduction

The goal is it to create a cylindrical projection of a face, like the one shown in Figure 7 with a desired view angle range of ±90° using multi view images. Such image can be used as texture map for 3D models or pose invariant face recognition tasks. The creation of texture maps from one or multiple views are described in many publications. The common idea is to project the captured images onto a cylinder using a more or less detailed surface representation or to find registration data, which related one captured view to another. Blending rules at the overlapping areas define the quality of the synthetic image. The approach implemented here is based on a simple approximation of a human head in contrast to a detailed surface representation used for similar results. Therefore a detailed reconstruction of the face is not required. Registration techniques require high frequency parts, which maybe not available for the complete face or hair. Marker points are not very likely, because of the more complicated capture and removal process. On the other hand, ghost edges are very likely, if the cylindrical approximation is not placed at the right position or the system calibration is not adequate. Another important issue in the creation of this kind of images is the correct aligned of the images for the posterior face space creation.

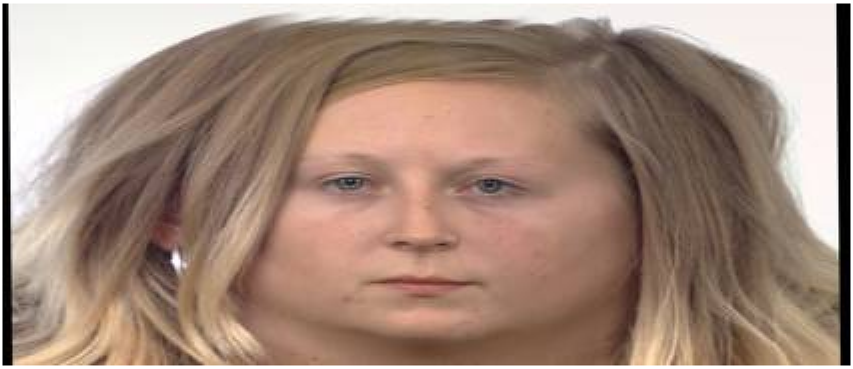


Fig. 7 Texture map created from an image set of nine images captured with a multi-view camera setup

3.2.2 Image Stitching

Image stitching algorithms blend images in a seamless manner, taking care to deal with potential problems such as blurring or ghosting caused by parallax and scene movement as well as varying image exposures. Our goal is the creation of cylindrical face projections from several different views. Thus, the idea is to use a cylinder as approximation of a human head. The combination of several views captured with a system as shown in Figure 8 to a synthetic image as shown in Figure 7 requires information about the surface of the captured object. Such information can range from an exact 3D model over some points in 3D space to an approximation of the basic shape. If such surface information is not available, they can be extracted from the captured view images with a wide variety of methods. Triangulation only requires corresponding point pairs in two views. Surface point reconstruction for a couple of corresponding point pairs allows the definition of an approximation of the desired surface up to a detailed representation. If the captured images show a unique silhouette, a method known as shape from silhouette can be used to reconstruct a volume model. This method is limited to convex surfaces. Other methods use registration information, created from gradients or marker positions.

A cylindrical approximation is adapted to the captured head and positioned in the middle of the assumed head axis. Head dimensions are defined in the book of Farkas [Farkas95] with average values of 151mm for the width of a head and 197mm for the depth of a head. Similar numbers can also be found in the report of

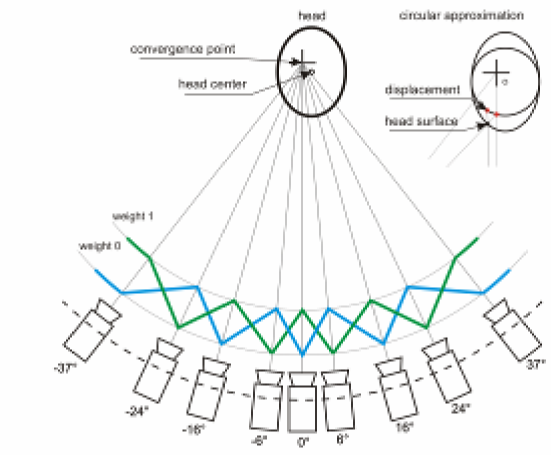


Fig. 8 Capture system for training data acquisition consists of nine cameras at different horizontal angles. (left) face surface point displacement due to the circular approximation. (middle) Blending rule for the combining of the input images

Young [Young93]. Considering a head with hair and two ears, a circular assumption is one possible approximation to the ellipse shaped head therefore suitable for the given problem. Besides the head size (radius for the cylinder), which can be taken from the mentioned statistical publications, there are other remaining unknown positions. Both offsets (x- and z-axis) have to be estimated.

Therefore the following equation describes the problem.

$$r^2 = (x - o_x)^2 + (z - o_z)^2 \tag{10}$$

This non-linear equation can be solved by Levenberg-Marquadt approach as described in the book of Scales [Scales85]. We have used the 3D reconstructed locations of both eye pupil centers as well as left and right eye corner of both eyes as fitting locations. The radius was fixed to a slightly bigger value than given for the head depth by the statistical publications, because of the requirements for the resulting texture map (±90°). The determined cylinder offset parameter for the x- and z-axis are used to place the cylinder at the face surface with the highest impact for the alignment, like the eyes and mouth. Therefore, the edges of these features are aligned and the displacement error for other features are not visible, e.g. nose tip. The fitted cylinder and a face surface model are shown in Figure 9. Therefore, some face surface points are reconstructed using the calibration data and perspective projection. Due to the fact, that several views provide 2D locations of the same feature point, a multi view approach for the reconstruction is used for more reliable 3D feature point locations. The handling of outliers, which can be a result of the point correspondences or calibration data, is crucial at this step. The average of permuted reconstructed 3D point locations using two views as well as a closed-form solution for all views has the drawback of moving the solution to one or more outliers. We have used a combination of the permutation and closed-form solution, which takes advantage of the back projection error in all considered views and is therefore an iterative solution. The consideration of the minimum back projection error as measurement for the 3D point localization accuracy leads to reliable surface points and identifies possible problems with 2D point locations. The evaluation result can not only be used for more exact 3D point locations by excluding outliers, but also to define weights according to the quality of back projection and therefore to define the influence of each view to the to be created texture map.

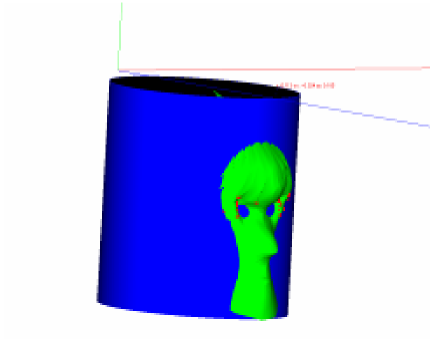


Fig. 9 Interpolated face surface model with a cylinder fitted at both eyes

The composition is mainly described by the projection of each view onto the object surface and in our case a cylindrical approximation. The perspective warping of images onto a cylindrical surface is described in the technical report of Szeliski [Szeliski06]. Afterwards a projection will take place, in order to create the texture map. Sampling the cylinder surface with the desired resolution of the virtual view is used for this approach. This method allows a set of DOFs, like the horizontal and vertical resolution as well as defining a specific region of interest. A linear blending rule, like the one shown in Figure 8 is used to incorporate adjacent views. For each horizontal position along the circumference left and right views are selected according to the camera rotation around the y-axis. The angle differences between the view vector and the selected left and right views are converted to weights. This method is constrained on the assumption, that the views are now rotating around the head and cylinder center and not longer around the convergence point of all views. The 3D reconstructed face features show a very small offset along the x axis and a strong offset along the z axis, so that middle view still refers to the most frontal face view. Using a linear interpolation between adjacent views leads to a weight of 0.5 in the middle of these both views.

The result is a texture map shown in Figure 7, where nine views are incorporated to one cylindrical projected face image showing ±90°. At the top and the bottom of the created image some ghost parts are visible. These wrong stitched parts are caused by the cylindrical assumption and the difference to the real object surface. A more detailed 3D object would decrease such miss alignments of the regions below the chin.

3.2.3 Image Alignment

The importance of local aligned face features for recognition is described by Tsapatsoulis *et al.* [Tsapatsoulis98]. There, the local alignment is described as resizing in contrast to a complete affine transformation. In most face recognition systems, the eye centers are used for the alignment of the images, whereas other face features have usually not received attention although it has been demonstrated that they are as important as eyes centers for performing recognition [Kouzani99]. In this work, a two step alignment approach is proposed. First, a global alignment is applied to the training texture maps, which is based on 2D affine transformation. The parameters for the transformation are determined by using a limited number of manually selected face feature locations. The average feature point location is used as reference and all transformations are calculated with respect to these data.

In order to achieve a better alignment result, the regions of the selected face features are additionally aligned locally. The generic triangle mesh of Figure 10 is adapted and placed at the global transformed face feature locations and the associated texture information for each triangle is extracted. The face feature locations are transformed to the desired position and thus affine warping is performed locally for each triangle.

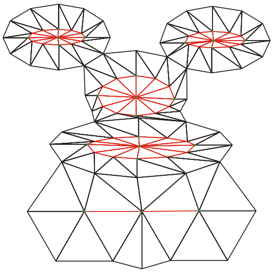


Fig. 10 Adapted triangle mesh for the local alignment of face features including eyes, nose, mouth and chin

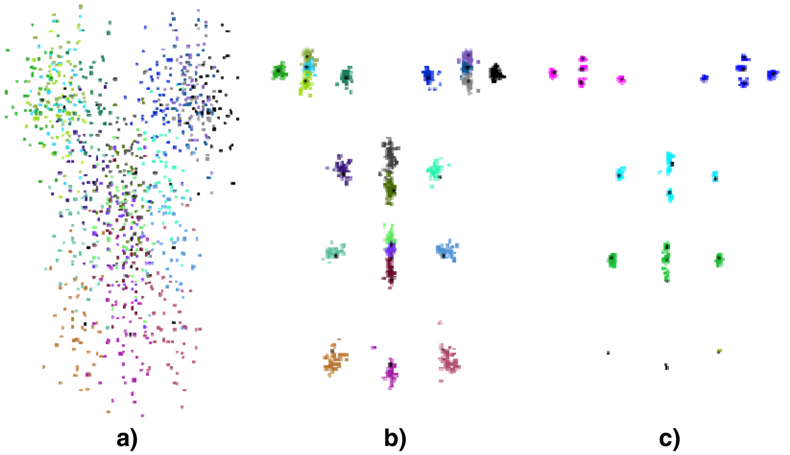


Fig. 11 Facial Feature localizations of the facial images a) before any transformation, b) after global transformation, and c) after local transformation using the wired mesh

Using this approach leads to a better alignment of the local features as depicted in Figure 11. Figure 11a) shows the point clouds for the selected feature points on the texture maps (only frontal features). Figure 11b) represents the same points after applying the 2D affine global transformation to each image. And finally, Figure 11c) illustrates the consistency of all the facial feature points after the local affine transformation. Since alignment of the feature points is done jointly for each feature, the feature points shown in Figure 11c) do not match in one center. A more visual comparison between just global or global plus local aligned texture maps is given in Figure 12.

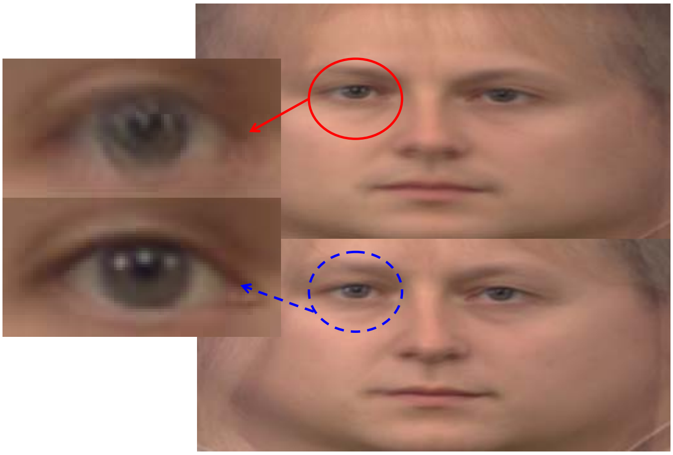


Fig. 12 Average image. Top: after global alignment. Bottom: after global plus local alignment.

4 Experimental Results

4.1 Dataset and Experiment Description

Two different databases are used for the experimental results. The first one is the HHI face database which is composed of 10 different subjects. The HHI database contains one 180° texture map for each subject that have been aligned using only global alignment; and another texture map using the global and local alignment process described in Section 3.2. These texture maps are used as the training and gallery sets. The test set for recognition is composed of a total of 9 views with different head poses (0°, ±6°, ±16°, ±25° and ±37°) for each subject (90 test images) acquired in a second session.

The second database is the UPC database [UPC-FaceDatabase]. This database includes a test set of 30 persons with 9 pictures per person which correspond to different pose views (0°, ±30°, ±45°, ±60° and ±90°). Furthermore, a total of 30 different 180° texture maps have been created by morphing only five views [Rama05], these texture maps are aligned and used as the training and gallery ensemble.

The first database is used for testing the improvement of the recognition rate when using the alignment images with the proposed approach of section 3.2, whereas the second one is used for testing the computational time when performing the P^{2}CA approach in the frequency domain. Finally, both databases are used for comparing the results using the SD and the NCC measures. A final experiment is focused on a comparative between the Partial Principal Component Analysis technique and the conventional *Eigenfaces* approach using the UPC-Face Database.

4.2 Experimental Results

Table 1 summarizes the recognition accuracy for the two correlation methods proposed in equations (2) and (3). From the results it is clear that using NCC improves the recognition rate results since this measure is more robust towards slight changes in illumination. For the UPC database the improvement is more visible since 10 out of the 30 identities have been enrolled on the database in a different session with slightly different illumination conditions.

Table 1. Recognition Accuracy

Dataset	**SD**	**NCC**
HHI	94.44%	96.66%
UPC	81.85%	89.63%

Table 2 presents the results for face recognition when using the alignment process described above. Although only 10 different persons are used in the experiments, results show that the locally aligned images present a slight improvement in the recognition rate. The improvement of this rate has been obtained for the 0° and ±6° views since these enclose all the face features used for the alignment.

Table 2. Face Recognition Results using NCC

Dataset	**FR (global alignment)**	**FR (global + local alignment)**
HHI	93.33%	96.66%

The computational time is analyzed when using the P^{2}CA approach in the spatial and in the frequency domain. Simulations have been run in MATLAB using a 2.0 GHz μP with 1GB of RAM. Table 3 shows the computational time for P^{2}CA in Spatial and in Frequency Domain depending on the total number of eigenvectors used for computing the face space. This time comprises the matching of one image to the 30 enrolled persons. Results illustrate the importance of performing P^{2}CA in FD for higher dimensions since the reduction in computational time is between 70 and 150.

Table 3. Computational Time for NCC measure

dim	P^2CA in SD	P^2CA in FD	factor
1	0.14 sec	0.011 sec	12
20	0.96 sec	0.042 sec	22
60	8.15 sec	0.114 sec	71
122	32.9 sec	0.222 sec	150

In the final experiment, we verify the robustness of P^2CA in front of the conventional 2D strategies. Thus, Eigenfaces (PCA), and 2DPCA have been implemented. For the training of the conventional 2D strategies, 5 different face views for each subject have been used as training and gallery data (the same 5 images used for the creation of the texture maps as explained in [Rama05]).

Table 4. Recognition Accuracy of the different algorithms

Method \ Exp.	**Neutral illumination**	**3 illuminations**
PCA / Eigenfaces	72.22%	60.45%
2DPCA	75%	61.24%
P^2CA	89.63%%	72.9%

The results presented in Table 4 show that the novel 2D-3D mixed scheme (P^2CA) outperforms its respective two dimensional approaches (PCA and 2DPCA) when varying pose.

5 Conclusions

5.1 Performance of the System in the Different Application Scenarios

Although the experiments have been carried out on a small database, the idea is to present a framework for the extension of any statistical approach such as LDA, ICA, Kernel-PCA to this novel 2D-3D mixed framework. Thus, the performance of the Partial PCA may decrease for some scenarios such as Video Surveillance since PCA is sensitive to illumination changes. However, if there were enough training samples, a Partial LDA or even a Partial ICA could be implemented to make the face recognition system more robust towards these illumination changes.

Partial PCA is supposed to get good results for Access Control, HCI or VC scenarios. In those scenarios, the main challenge is pose variation and as already demonstrated in the previous section P^2CA is developed considering this challenge. In these 3 scenarios the illumination may not vary considerably and the resolution of the images is medium high. For the Video Surveillance it may depend a lot on the type of room or place of the scenario. In a bank-hall like the one illustrated in Figure 3, P^2CA may perform acceptable since the face can be acquired in semi-controlled conditions. Nevertheless, the accuracy may decrease considerably in VS for airports or train stations where the resolution of the face is smaller, the inclination of the cameras or, for all of them, illumination and occlusions. In the case of video and image indexing and retrieval, the illumination conditions may vary a lot. Nevertheless, the main problem is to acquire for all the persons of the photo album or video database different views for the creation of the 3D training data. In this case, manual interaction for the creation of the texture and depth maps may be needed. This could be one of the main drawbacks of the 2D-3D mixed face recognition schemes for the image and video indexing scenario.

5.2 Future Work

Face processing is a very hot topic for several multimedia applications like video surveillance, virtual collaboration, HCI, video indexing or control access points. In each multimedia application there are different challenges that should be overcome, but pose is one common problem in all of them. Recently, a new trend of 3D face recognition approaches showed an increase in the recognition rate under the presence of big pose and illumination variations if 3D data is available. Nevertheless, cost of the set-up, acquisition time and cooperation of the subjects are still some of the requirements for obtaining accurate 3D data that may not be available during the recognition stage. Thus, we have presented here a possible alternative following a mixed 2D-3D face recognition philosophy, i.e. the system is trained with 3D data but it can use either 2D or 3D data in the test stage. However, this philosophy may be extended also to other face recognition statistical approaches like LDA or ICA which have shown a higher robustness in the presence of illumination variations. Additionally, we have presented an automatic approach for the creation of aligned virtual view images using nine different views. These aligned virtual view images are used as training data for the P^2CA technique. The virtual view image is created by using a cylindrical approximation for the real object surface.

Furthermore, two improvements for the P^2CA approach have been proposed: First, a local alignment method of the training images. Second, a reformulation of the complete approach in the frequency domain is proposed. Both improvements lead to an increase of the recognition accuracy and a reduction of the computational time making it suitable for face recognition multimedia scenarios such as security control access point, Human Computer Interface or Virtual collaboration applications where the processing should be on real time.

Acknowledgement

The work presented was developed within VISNET II, a European Network Of Excellence funded under the EC IST FP6 programme.

References

[Ahonen06] Ahonen, T., Hadid, A., Pietikainen, M.: Face Description with Local Binary Patterns: Application to Face Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(12), 2037–2041 (2006)

[Albiol01] Albiol, A., Torres, L.: Ed. Delp. Optimum Color Spaces for Skin Detection. In: IEEE International Conference on Image Processing, Thessaloniki, Greece, October 7-10 (2001)

[Antini06] Antini, G., Berretti, S., Del Bimbo, A., Pala, P.: 3D Face Identification Based on Arrangement of Salient Wrinkles. In: IEEE International Conference & Multimedia Expo, Toronto, Canada, July 9-12 (2006)

[Arandjelovic05] Arandjelovic, O., Shakhnarovich, G., Fischer, J., Cipolla, R., Darrell, T.: Face recognition with Image Sets using Manifold Density Divergence. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, San Diego, USA, vol. 1 (2005)

[Aureli07] Soria-Frisch, A., Verschae, R., Olano, A.: Fuzzy Fusion for Skin Detection. Fuzzy Sets and Systems 158(3), 325–336 (2007)

[Barlett02] Barlett, M.S., Movellan, J.R., Sejnowski, T.J.: Face Recognition by Independent Component Analysis. IEEE Trans. on Neural Networks 13(6) (November 2002)

[Batur05] Batur, A.U., Hayes, M.H.: Adaptive active appearance models. IEEE Transactions on Image Processing 14(11), 1707–1721 (2005)

[Belhumeur97] Belhumeur, R.N., Hespanha, J.P., Kriegman, D.J.: Eigenfaces vs Fisherfaces: Recognition Using Class Specific Linear Projection. IEEE Trans. Pattern Analysis and Machine Intelligence 19(7) (July 1997)

[Beumier01] Beumier, C., Acheroy, M.: Face verification from 3D and grey level clues. Pattern Recognition Letters 22, 1321–1329 (2001)

[Beymer95] Beymer, D., Poggio, T.: Face Recognition from One model View. In: Proc. Fifth Int'l Conf. Computer Vision (1995)

[Blanz03] Blanz, V., Vetter, T.: Face Recognition based on fitting 3D morphable model. IEEE Trans. Pattern Analysis and Machine Intelligence 25(9), 1063–1074 (2003)

[Bowyer04] Bowyer, K., Chang, K., Flynn, P.: A Survey of Approaches to 3D and Multi-Modal 3D+2D Face Recognition. In: IEEE Intl.Conf. on Pattern Recognition (2004)

[Bronstein05] Bronstein, A.M., Bronstein, M.M., Kimmel, R.: Three-dimensional face recognition. International Journal of Computer Vision 64(1), 5–30 (2005)

[Cartoux89] Cartoux, J.Y., Lapreste, J.T., Richetin, M.: Face authentification or recognition by profile extraction from range images. In: Workshop on Interpretation of 3D Scenes, pp. 194–199 (1989)

[Chang05] Chang, K.I., Bowyer, K.W., Flynn, P.J.: An Evaluation of Multimodal 2D+3D Face Biometrics. IEEE Trans. on Pattern Analysis and Machine Intelligence 27, 619–624 (2005)

[Chenaoua06] Chenaoua, K., Bouridane, A.: Skin Detection using a Markov Random Field and a New Color Space. In: IEEE International Conference on Image Processing, October 8-11, pp. 2673–2676 (2006)

[Cootes01] Cootes, T., Edwards, G., Taylor, C.: Active Appearance Models. IEEE Transactions on Pattern Analysis and Machine Intelligence 23, 681–685 (2001)

[Craw92] Craw, I., Tock, D., Bennett, A.: Finding Face Features. In: Proc. Second European Conf. Computer Vision, pp. 92–96 (1992)

[Cutler96] Cutler, R.: Face Recognition Using Infrared Images and Eigenfaces., Technical Report. UMI Order Number: CSC 989, University of Maryland at College Park (1996)

[Emidio00] Emidio, T., Schmidt, R., Marcondes, R.: A Framework in Face Recognition from Video Sequences using GWN and Eigenfeature Selection. In: I WAICV - Workshop on Artificial Intelligence and Computer Vision, São Paulo, Brazil (2000)

[Faltemier08] Faltemier, T.C., Bowyer, K.W., Flynn, P.J.: A Region Ensemble for 3-D Face Recognition. IEEE Transactions on Information Forensics and Security 3(1), 62–67 (2008)

[Feng06] Feng, S., Krim, H., Gu, I., Viberg, M.: 3D Face Recognition Using Affine Integral Invariants. In: IEEE International Conference on Acoustics, Speech and Signal Processing, Toulouse, May 14-19 (2006)

[FRGC05] Phillips, P.J., Flynn, P.J., Scruggs, T., Bowyer, K.W., Chang, J., Hoffman, K., Marques, J., Min, J., Worek, W.: Overview of the Face Recognition Grand Challenge. In: IEEE Conf. Computer Vision and Pattern Recognition(CVPR) (2005)

[FRVT02] Phillips, P.J., Grother, P., Micheals, R., Blackburn, D., Tabassi, E., Bone, J.: Face Recognition Vendor Test 2002: Evaluation Report. In: NISTIR 6965, National Institute of Standards and Technology (2003), `http://www.frvt.org`

[Georghiades01] Georghiades, A.S., Belhumeur, P.N., Kriegman, D.J.: From Few to Many: Illumination Cone Models for Face Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(6), 643–660 (2001)

[Goldmann07] Goldmann, L., Mnich, U.J., Sikora, T.: Components and their topology for robust face detection in the presence of partial occlusions. IEEE Transactions on Information Forensics and Security 2 (September 2007)

[Heusch06] Heusch, G., Rodriguez, Y., Marcel, S.: Local binary patterns as an image preprocessing for face authentication. In: 7th International Conference on Automatic Face and Gesture Recognition, 2006. FGR 2006, April 10-12, 6 p (2006)

[Jinwoo06] Kang, J., Borkar, A., Yeung, A., Nong, N., Smith, M., Hayes, M.: Short Wavelength Infrared Face Recognition for Personalization. In: IEEE International Conference on Image Processing, October 8-11, pp. 2757–2760 (2006)

[Jones02] Jones, M.J., Rehg, J.M.: Statistical Color Models with Application to Skin Detection. Int. Journal of Computer Vision 46(1), 81–96 (2002)

[Kakadiaris07] Kakadiaris, I.A., Passalis, G., Toderici, G., Murtuza, M.N., Lu, Y., Karampatziakis, N., Theoharis, T.: Three-Dimensional Face Recognition in the Presence of Facial Expressions: An Annotated Deformable Model Approach. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(4), 640–649 (2007)

[Kim02] Kim, K.I., Jung, K., Kim, H.J.: Face Recognition Using Kernel Principal Component Analysis. IEEE Signal Processing Letters 9(2) (February 2002)

[Kouzani99] Kouzani, A., Sammut, K.: Quadtree principal component analysis and is application to facial expression classification. In: IEEE Int. Conf. on Systems, Man, and Cybernetics (October 1999)

[Kwon94] Kwon, Y.H., da Vitoria Lobo, N.: Face detection using templates. In: Proc. of IAPR International Conference on Pattern Recognition, pp. 764–767 (1994)

[Lades93] Lades, M., Vorbruggen, J.C., Buhmann, J., Lange, J., von der Malsburg, C., Wurtz, R.P., Konen, W.: Distortion Invariant Object Recognition in the Dynamic Link Architecture. IEEE Transactions on Computers archive 42(3), 300–311 (1993)

[Lee90] Lee, J.C., Milios, E.: Matching range images of human faces. In: Proc. IEEE International Conference on Computer Vision, pp. 722–726 (1990)

[Lewis95] Lewis, J.P.: Fast Normalized Cross-Correlation. Vision Interface (1995)

[Li01]Li, Y., Gong, S., Liddell, H.: Constructing facial identity surfaces in a nonlinear discriminating space. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition (2001)

[Lu05] Lu, X., Jain, A.K.: Integrating Range and Texture Information for 3D Face Recognition. In: Proc. IEEE WACV, Breckenridge, Colorado (2005)

[Lu06] Lu, X., Jain, A.K.: Matching 2.5D face scans to 3D models. IEEE Transactions Pattern Analysis and Machine Intelligence 28, 31–43 (2006)

[Mahoo07] Mahoo, M.H., Abdel-Mottaleb, M.: 3D Face Recognition Based on 3D Ridge Lines in Range Data. In: IEEE International Conference on Image Processing, 2007. ICIP 2007, September 16-October 19, vol. 1, pp. I-137–I-140 (2007)

[McKenna98] McKenna, S., Gong, S., Raja, Y.: Modelling Facial Colour and Identity with Gaussian Mixtures. Pattern Recognition 31(12), 1883–1892 (1998)

[Miao99] Miao, J., Yin, B., Wang, K., Shen, L., Chen, X.: A hierarchical multiscale and multiangle system for human face detection in a complex background using gravity-center template. Pattern Recognition 7(32), 1237–1248 (1999)

[Onofrio06] Onofrio, D., Rama, A., Tarres, F., Tubaro, S.: P2CA: How Much Information is needed. In: IEEE International Conference on Image Processing, Atlanta, USA (October 2006)

[Papatheodorou04] Papatheodorou, T., Rueckert, D.: Evaluation of Automatic 4D Face Recognition Using Surface and Texture Registration, fgr. In: Sixth IEEE International Conference on Automatic Face and Gesture Recognition, p. 321 (2004)

[Pentland94] Pentland, A., Moghaddam, B., Starner, T.: View based and modular eigenspaces for face recognition. In: IEEE Conf. on Computer Vision and Pattern Recognition (1994)

[Pierrard07] Pierrard, J.-S., Vetter, T.: Skin Detail Analysis for Face Recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR 2007, June 17-22, pp. 1–8 (2007)

[Pretzel07] Pretzel, Lotz. Research project: Face recognition as a search tool. Technical report, Bundeskriminalamt Wiesbaden (2007)

[Propp92] Propp, M., Samal, A.: Artificial neural network architecture for human face detection. Intell. Eng. Systems Artificial Neural Networks 2, 535–540 (1992)

[Rama05] Rama, A., Tarrés, F.: P2CA: A new face recognition scheme combining 2D and 3D information. In: IEEE International Conference on Image Processing, Genova, Italy, September 11-14 (2005)

[Rama06] Rama, A., Tarres, F.: Partial PCA Vs Partial LDA. In: IEEE International Conference on Multimedia and Expo, Toronto, Canada, July 9-12 (2006)

[Rowley98] Rowley, H.A., Baluja, S., Kanade, T.: Neural network-based face detection. IEEE Trans. Pattern Anal. Mach. Intell. 20, 23–38 (1998)

[Samal95] Samal, A., Iyengar, P.A.: Human face detection using silhouettes. International Journal of Pattern Recognition and Artificial Intelligence 9(6), 845–867 (1995)

[Samani06] Samani, A., Winkler, J., Niranjan, M.: Automatic Face Recognition Using Stereo Images. In: IEEE International Conference on Acoustics, Speech and Signal Processing, Toulouse, May 14-19 (2006)

[Savvides04] Savvides, M., Kumar, B.V., Khosla, P.K.: Eigenphases vs. Eigenfaces. In: Int. Conf. on Pattern Recognition, Washington DC (2004)

[Sen07] Wang, S., Wang, Y., Jin, M., Gu, X.D., Samaras, D.: Conformal Geometry and Its Applications on 3D Shape Matching, Recognition, and Stitching. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(7), 1209–1220 (2007)

[Sinha94] Sinha, P.: Object recognition via image invariants: A case study. Investigative Ophthalmology and Visual Science 35(4), 1735–1740 (1994)

[Socolinsky04] Socolinsky, D.A., Selinger, A.: Thermal face recognition in an operational scenario. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004, June 27-July 2, vol. 2, pp. II-1012–II-1019 (2004)

[Szeliski06] Szeliski, R.: Image alignment and stitching: a tutorial. Found. Trends. Comput. Graph. Vis. 2(1), 1–104 (2006)

[Torres00] Torres, L., Lorente, L., Vila, J.: Automatic Face Recognition of Video Sequences Using Self-eigenfaces. In: International Symposium on Image/video Communication over Fixed and Mobile Networks, Rabat, Morocco, April 17-20 (2000)

[Tsalakanidou04] Tsalakanidou, F., Malassiotis, S., Strintzis, M.: Integration of 2D and 3D Images for enhanced face authentication. In: Sixth International Conference on Automated Face and Gesture Recognition, May 2004, pp. 266–271 (2004)

[Tsapatsoulis 98] Tsapatsoulis, N., Doulamis, N., Doulamis, A., Kollias, S.: Face extraction from non-uniform background and recognition in compressed domain. In: IEEE ICASSP, Seattle, WA, USA (May 1998)

[Turk91] Turk, M.A., Pentland, A.P.: Face recognition using eigenfaces. In: Proceedings of the IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, Maui, Hawaii, pp. 586–591 (1991)

[UPC-FaceDatabase] UPC Face Database in, `http://gps-tsc.upc.es/GTAV`

[Vezhnevets03] Vezhnevets, V., Sazonov, V., Andreeva, A.: A survey on pixel-based skin color detection techniques. In: Proc. Graphicon.(2003)

[Viola01] Viola, P., Jones, M.: Rapid Object Detection using a Boosted Cascade of Simple Features. Computer Vision and Pattern Recognition (2001)

[Wilder96] Wilder, J., Phillips, P.J., Jiang, C., Wiener, S.: Comparison of visible and infrared imagery for face recognition. In: Proceedings of the 2nd international Conference on Automatic Face and Gesture Recognition (FG 1996), October 14-16, p. 182. IEEE Computer Society, Washington (1996)

[Wiskott99] Wiskott, L., Fellous, J.M., Kruger, N., von der Malsburg, C.: Face recognition by elastic bunch graph matching. IEEE Trans. Pattern Analysis and Machine Intelligence 19(7), 775–779 (1999) (Revised version)

[Yang04] Yang, J., Zhang, D., Frangi, A.F., Yang, J.: Two-Dimensional PCA: A New Approach to Appearance-based Face Representation and Recognition. IEEE Trans. on Pattern Analysis and Machine. Intel. (January 2004)

[Young93] Young, J.W.: Head and Face Anthropometry of Adult U.S. Civilians. Technical Report ADA268661, Civil Aeromedical Institute, Federal Aviation Administration (1993)

[Yueming07] Wang, Y., Pan, G., Wu, Z.: 3D Face Recognition in the Presence of Expression: A Guidance-based Constraint Deformation Approach. In: IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR 2007, June 17-22, pp. 1–7 (2007)

[Yuille92] Yuille, A.L., Hallinan, P.W., Cohen, D.S.: Feature extraction from faces using deformable templates. International Journal of Computer Vision 8(2), 99–111 (1992)

[Zhao06] Zhao, W., Chellapa, R.: Face Processing: Advanced modeling and methods. Academic Press, London (2006)

[Zhou03] Zhou, S., Krueger, V., Chellappa, R.: Probabilistic recognition of human faces from video. Computer Vision and Image Undestanding 91, 214–245 (2003)

Interfaces That Should Feel Right: Natural Interaction with Multimedia Information

Radu-Daniel Vatavu

Abstract. Multimedia content, organization and availability influences the way we work, learn or spend our free time even more than we realize it: we share photos, play movies or listen to FM internet radio stations as ubiquitous activities. Multimedia information has literally invaded our lives: more and more resourceful computing equipments; continuously-growing internet on-line activity; inexpensive cameras and recording equipment at everyone's reach. With all these advances, interaction with multimedia information is still one step behind being restricted to the use of mouse and keyboard in the detriment of a more fluent and natural interaction. If we take a look at how our interactions look like in the real world we easily identify gestures as our every day means in order to interact with real objects, convey information or express emotions. Quite opposite, when it comes to computers and handling digital information we need to go back to the mouse and keyboard pair. The chapter addresses the problem of intuitive and natural interfaces for interacting with multimedia information. Gesture acquisition and recognition has been around for a while and many advances have been reported in what is today a very challenging research domain. Aside from successfully capturing and recognizing meaningful gesture commands, the real challenge starts when constructing gesture dictionaries, designing interaction metaphors and techniques that would feel "right" for a given application.

1 Introduction

Interaction techniques and interfaces have definitely evolved in the last decades with many concepts being introduced while technologies became more and more available and affordable. And while they did so, multimedia content evolved many times

Radu-Daniel Vatavu
Research Center in Computer Science, University "Stefan cel Mare" of Suceava, 13 University Street, 720229 Suceava, Romania
e-mail: vatavu@eed.usv.ro

M. Grgic et al. (Eds.): Rec. Advan. in Mult. Sig. Process. and Commun., SCI 231, pp. 145–170.
springerlink.com

more. The dimension, content, availability and structure of multimedia demand appropriate interfaces for visualization, handling and manipulation, storing and consuming for which WIMP style interactions provide less and less support. Although multimedia has literally invaded our lives (photo sharing, video webcasts, on-line activity, virtual worlds emerging) we still interact via mice and keyboards. This very much contrasts with the interactions we perform in our every day lives where we gesture in order to manipulate objects, express emotions, convey meaning and make us understood or simply to enhance and add new dimensions to our speech.

This chapter discusses gestures as an alternative form of input for the current mouse and keyboard. The main reason for using human gestures as the interface is naturalness, intuitiveness and familiarity that can be immediately gained. Gestures play an important part in our lives including art, science, music, dance, allowing us to work, communicate, express feelings, enhance and accompany speech. Using gestures is something we have been training for all our lives, still in the process of learning, and make use of it according to our personalities, jobs, social situations and events, most of the time without even realizing it. The naturalness and familiarity of gesturing are revealed even more by the fact that blind people gesture when they speak just as much as sighted individuals do, even when they know their listener is also blind [23].

Designing gestural interfaces has been introduced a while ago and evolving ever since into what is today a very challenging research domain [49]. As every new technology, it shouldn't be used just because it is new: the right applications that would indeed benefit of this kind of interaction must be sought. GUIs and WIMP are fine where they are productive and give good results however there exist applications (some just emerge) where gestures find themselves belonging in the interface. Large screens and information visualization, touch-sensitive tabletops and interactive surfaces, virtual reality simulators or computer games are only a few examples.

The chapter includes an overview on definitions, taxonomies, technology and techniques used to capture and recognize human gestures. A case study is presented for a simple application of a virtual piano playing for which gesture-based interaction feels *more real* or feels *right* in opposition to mouse clicks and key taps. Different technologies (sensors and computer vision) were purposely chosen in order to achieve this goal. Practical details of the implementation are discussed so that they may serve practitioners faced with similar challenges of building interfaces that *should feel right.*

2 An Overview on Human Gestures

Gestures express ideas, feelings and intentions, sometimes replacing words and enhancing speech. They convey information and are accompanied by content and semantics. Various psycholinguistic studies have been conducted in what concerns the understanding of gesture communication and they provide an excellent starting material for gesture studying and understanding [6, 9, 27, 41]. The vocabulary of gestures that people use can be at once informative, entertaining but also dangerous.

Gestures may be instructive such as the signs made by an airport officer guiding planes or those performed by lecturers when sustaining their presentations; gestures may be warm in the form of giving a hug or a confident hand-shake; they may even be menacing: for example imagine two drivers that 'accidentally' met on a freeway. We start our discussion on gestures by following a few definitions of how gestures are perceived by several interested communities after which we proceed to investigating properties and taxonomies.

2.1 Definitions: Trying to Put Meaning on the Word "Gesture"

Gestures may be bluntly looked upon as physical movements of hands, arms, face and body with the intent of conveying information and meaning.

Actual definition of gestures varies in accordance with the research community that studies them: sociologists, biologists, linguists or computer scientists look at gestures from different perspectives. For example, from a biological and sociological point of view, gestures are loosely defined and thus researchers are free to visualize and classify them as they see fit. Biologists define gestures generally, as in Nespoulous et al. [45]:

> *the notion of gesture is to embrace all kinds of instances where an individual engages in movements whose communicative intent is paramount, manifest, and openly acknowledged*

this being caused by the

> *lack of consensus (which) seems primarily to originate from intrinsic ambiguities related to the problem of defining, in the entire spectrum of motor activity, those activities which properly can be defined as gestures.*

Computer scientists need more thorough definitions of what a gesture is, and more important, of how it can be understood and represented as form of input for human-computer interaction. The definition of articulated gesture as in Kurtenbach et al. [35] is more appropriate in this case as it allows isolating those interactions for which gestures are articulated and recognized:

> *A gesture is a motion of the body that contains information. Waving goodbye is a gesture. Pressing a key on a keyboard is not a gesture because the motion of a finger on its way to hitting a key is neither observed nor significant. All that matters is which key was pressed.*

Distinction must be made between *gesture* and *posture*. There is the tendency to capture the dynamic part in gesture while to consider posture as being static [43]. Consulting a few dictionaries [22, 42], we end up with the following definitions for posture and gesture[1]:

[1] Only fragments cited.

> *Posture* (noun): 1. a. A position of the body or of body parts: a sitting posture. b. An attitude; a pose: assumed a posture of angry defiance.
>
> *Posture* (noun): 1. The position or bearing of the body whether characteristic or assumed for a special purpose: erect posture. 2. A conscious mental or outward behavioral attitude.
>
> *Posture synonyms*: attitude, carriage, pose, stance. These nouns denote a position of the body and limbs: erect posture; an attitude of prayer; dignified carriage; a defiant pose; an athlete's alert stance.
>
> *Gesture* (noun): 1. A motion of the limbs or body made to express or help express thought or to emphasize speech. 2. The act of moving the limbs or body as an expression of thought or emphasis. (verb intr.) To make gestures (verb tr.) To show, express or direct by gestures.

We are thus further considering posture as describing the position of body or of body parts, e.g. holding the hand in the victory sign for a certain amount of time is considered to be a posture. Advancing further this definition, LaViola [37] sees postures as *static movements* that can be either simple or complex. A simple posture will have all the fingers either extended or flexed but not in between, e.g. fist, thumbs up, index pointing, victory sign are simple postures. A complex posture would allow for fingers to be bent at different angles other than 0 or 90 degrees as in the ok sign, pinch or various sign language postures. A gesture is defined as a *dynamic movement* such as waving good bye or describing the shape of a circle in mid-air. According to LaViola [37], dynamic movements are also simple and complex. Simple movements are represented by either a posture held still while changing the position and orientation of the hand or keeping the hand still by moving fingers, i.e. changing postures. A complex movement will include changes in posture, position and orientation of the hand.

Similar concepts for postures and gestures for the purpose of human-computer interaction are introduced by Vatavu and Pentiuc [56]. By considering the amounts of static and dynamic information that are needed in order to sufficiently and completely describe a gesture command, the authors identify four distinct types of gestures that are referenced as simple static, static generalized (or complex), simple dynamic and dynamic generalized as Figure 1 illustrates. The four types range from simple static postures to sequences of postures, motion trajectories and sequences

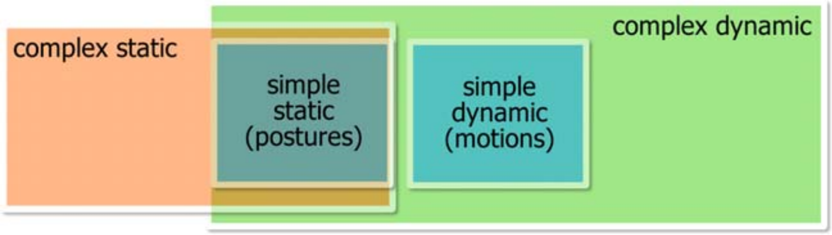


Fig. 1 Gesture types function of the amount of posture and motion information included in their structure

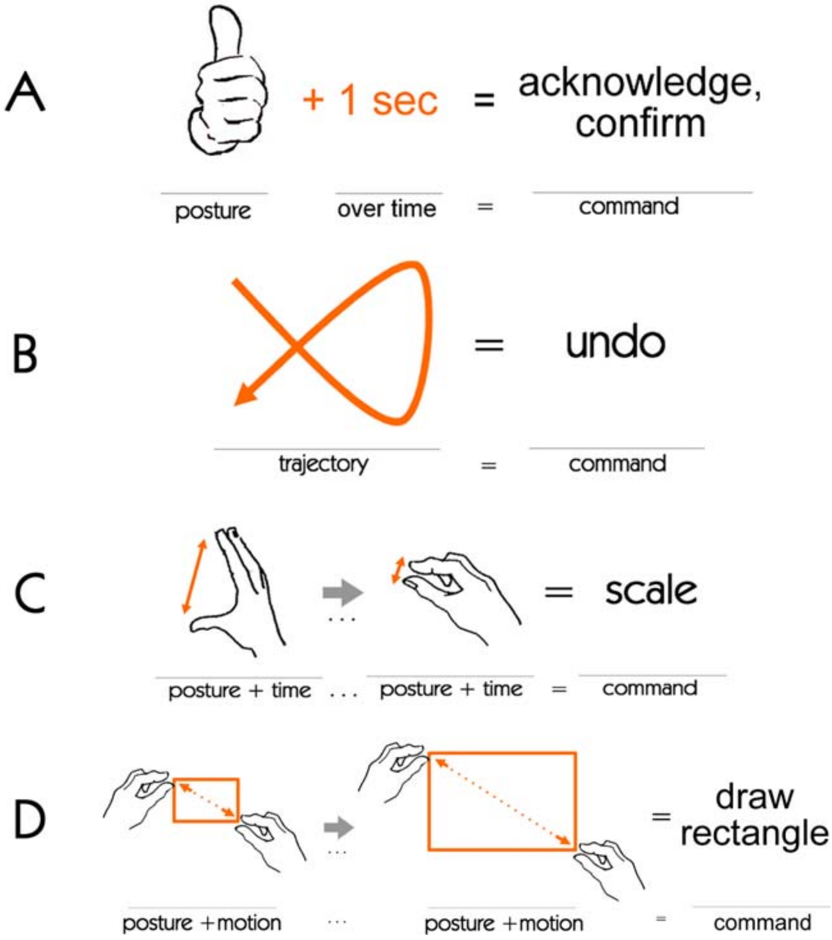


Fig. 2 Examples of simple, complex, static and dynamic gestures: (A) simple static gesture: the "thumbs up" posture held for a period of say 1 second may be associated with user acknowledging in response of an application confirmation inquiry; (B) simple dynamic gesture: the gesture represented by an X-cross may be associated with performing an "undo" operation; (C) complex static gesture: the scaling gesture executed with one hand indicates a change in size or equivalently, a zoom operation that would be proportional to the distance between the index and thumb finger; (D) generalized dynamic gesture: specifying a rectangle shape may be performed by two hands that control the distance between two opposite corners.

of consecutive or parallel motions. Figure 2 presents a few examples of gestures in accordance to Vatavu and Pentiuc [56] definitions.

We will consider posture as being a set of measurements $\{m_0, m_1, ...m_n\}$ that describe the position of body or body parts at one instant moment in time while gesture may be represented as a function of time with values into the Cartesian product of a coordinates space and the set of all postures P: $gesture : R \rightarrow R^d \times P$ where d represents the dimensionality of the space (2D, 3D, ...).

2.2 Properties That Give Gestures Away

Gestures posses a variety of distinct characteristics that make them distinguishable among other types of activities that relate to the human body [29] such as practical actions or postural adjustments:

- Gestures may be looked upon as excursions [50] from a rest position always returning to a rest state after execution.
- They posses a peak structure with the center associated to the actual meaning of gesture.
- Gestures are well bounded: the action phrases which are perceived and identified as gestures have clear onsets and offsets.
- Gestures are symmetric: it is remarkable difficult to spot the differences of someone caught gesturing on a film that runs backwards and forwards.

Similarly to speech, gestures serve a variety of functions. They convey information to listeners [28]; facilitate some aspects of memory [5, 34]; facilitate the smoothness of interactions and increase linking between interaction partners [11]; communicate attitudes and emotions both voluntarily and involuntarily [19]; can provide insight into a speaker's mental representations [41].

2.3 Gestures Taxonomies

Several taxonomies and classification criteria have been proposed for human gestures. We therefore start by citing and providing a small overview on commonly accepted classifications for human gestures as they may be encountered in the literature.

Cadoz [6] identifies three main categories by taking into account the functionality aspect of gestures that manipulate, sense and communicate with the environment (Figure 3 illustrates a few examples):

- *Ergotic* gestures are associated with the idea of work and ability to model and manipulate the environment. The ergotic gesture acts directly on the environment by altering its form and properties, e.g. hand made pottery, knitting or sculpting.
- *Epistemic* gestures offer information that reveals the environment through perception of temperature, pressure, surface quality for a given object, shape, orientation, weight, and so on. The environment gets revealed through tactile experience or haptic exploration.
- *Semiotic* gestures produce meaningful informational messages for the environment and come as a result of commonly shared cultural experience. The intent is to convey information.

Semiotic gestures are further classified by McNeill [41] according to their role in communication:

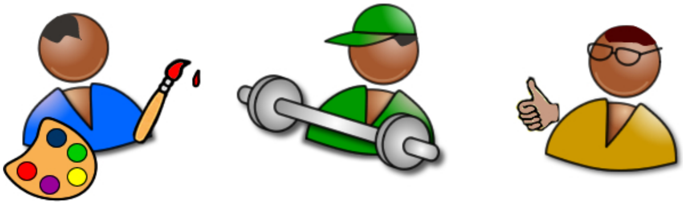


Fig. 3 Ergotic (painting, manipulating the environment), epistemic (heavy lifting, sensing the environment) and semiotic (giving the ok sign) gestures

- *Iconic* gestures describe an actual concrete object or event. They are closely related to the semantic content of speech, illustrating what is being said, e.g. when using hands to describe a physical item in order to show how big or small it is.
- *Metaphoric* gestures are similar to iconics but referring to abstract objects or events, depicting a general abstract idea.
- *Deictic* or pointing gestures.
- *Beat-like* which are gestures that accentuate the meaning of a word or a phrase, e.g. rhythmic beating of a finger or hand.

Kendon [27] describes a gesture continuum that goes from gesticulation up to sign languages, as follows:

- *Gesticulation* or spontaneous movements of hands and arms that take place during speech and always accompany speech.
- *Language-like* gestures that represent gesticulation actually integrated into speech that replaces a word or a phrase. An example would be the following phrase: *I enjoyed eating the grapes but the cake that came after was [gesture]* where [gesture] integrates grammatically inside the phrase.
- *Pantomime* are gestures that depict objects, events or actions that may or not be accompanied by speech.
- *Emblems* or familiar gestures, e.g. the V sign for victory or thumbs up for ok.
- *Sign languages* are sets of gestures and postures that define linguistic communication systems such as ASL, the American Sign Language.

Starting from gesticulation to sign languages, the association with speech gets more and more reduced, spontaneity decreases and social regulation increases. The first category, gesticulation, as being spontaneous and associated with speech, represents approximate 90% of the total amount of gesturing people perform. Familiar gestures are culture dependent. Very few gestures are universally understood and interpreted with the same signification. What is perfectly acceptable in one culture may prove be rude, inappropriate or even obscene in others. For example, nodding head up and down to intend and say *Yes* actually means *No* in Bulgaria and Greece; an "okay" gesture common in Western Europe is insulting in Turkey or Greece; passing an item to someone with one hand is considered to be very rude in Japan, etc.

Nespoulous et al. [45] consider a more detailed approach and propose a 3-level classification. With respect to the universality of gestures, the following types are identified:

- *Arbitrary* or uncommon gestures that need to be learned.
- *Mimetic* gestures which are usually encountered within a culture.
- *Deictic* gestures similar to the classification of [41].

Cassell [9] considers two categories of gestures: *autonomous* (or independent) and *natural* (or spontaneous). Autonomous gestures are not necessarily associated with verbal communication, possess fixed spatial-temporal properties and are speaker independent. Natural gestures are usually associated with speech in a conscientiously or unconscientiously manner. They are speaker dependent and much influenced by educational and cultural factors as well as by the actual situation at the moment they are produced. They were further classified by McNeill [41] into iconic, metaphoric, deictic and beat-like as discussed above.

The most numerous category of gestures is represented by hand actions due to the the ability of the human hand to acquire a large number of discernible configurations (the sign languages being a good example in this case). Much of the hand actions are related to manipulating the physical world hence a further classification of such ergotic hand gestures may be considered in accordance with physical characteristics: object type, change effectuated, hands involved, direct or indirect manipulation [43].

3 Catching Gestures

3.1 Existing Technology for Gesture Acquisition

The first step prior to use gestures as input when interacting with a computer system is represented by data acquisition. Technologies for the acquisition of human gestures have been very rapidly proliferating and a great variety of trackers, pointing or whole hand or body devices are available today commercially. One main property of all these input devices is the number of DOFs (Degrees of Freedom) they posses. Data may be collected in several ways [37] by using:

- Capture devices that are worn by users and which may provide a fine level of gesture representation (such as small variations when bending fingers as given in the output of sensor gloves). Users are however required to wear additional equipment which may feel cumbersome and disturbing, burdening the actual interaction.
- Video cameras, one or multiple, that capture a sequence of images and allow for detection of a 2D or 3D gesture. The main advantage is the feel of natural interaction but the capturing accuracy and frequency are lower than in the previous case.
- Hybrid approaches that combine the above technologies.

Karam and Schraefel [25] consider *perceptual* and *non-perceptual* input. Non-perceptual input includes devices or equipments that need actual physical contact in order to reply spatial or temporal information. Non-perceptual input includes: mice and pens, touch, electronic sensing (gloves, wearable devices) and audio input. On the other hand, perceptual input enables gestures to be recognized without requiring any physical contact via any input device or physical objects, allowing the user to communicate gestures without having to wear, hold or make any sort of physical contact. Computer vision is an example of technology that allows perceptual interaction.

Taking into account the type of events the capturing devices are able to generate, one may distinguish between:

- *Discrete* input devices that generate one event at a time according to the user's need (e.g., events are fired when the user presses a button). Examples include the traditional keyboard or the pinch glove. Considering for example the case of Fakespace Labs Pinch Gloves[2], users will pinch two or more fingers for the device to signal an event. The gloves detect when two or more fingers are in contact and, after contact verification, a signal is fired. The time that elapsed between two consecutive gestures is also recorded. A variety of different actions starting from the pinch gesture can be included into applications (a pinching gesture may be used to grab a virtual object; a finger snap between the middle finger and thumb can be used to initiate a given action, etc).
- *Continuous* input devices that generate a stream of events. Data gloves or 2D/3D trackers report periodically the measurements being performed usually at high frequencies.
- *Hybrid* devices that combine both discrete and continuous events.

With regards to tracking devices, one can discriminate between several technologies: magnetic, mechanical, acoustic, inertial, vision/video camera based or hybrid.

Magnetic trackers (such as the Ascension's Flock of Birds[3]) use the low-frequency magnetic field emitted by a transmitter for the receiver sensor to determine its position and orientation with respect to the magnetic source. The main disadvantage is the distortion of the magnetic field that metal or conductive metals will produce as well as the interference with nearby monitors. Bolt's system Put-That-There [4], the first system that implemented posture recognition, made use of such a magnetic-based space-sensing cube device attached to the user's wrist in order to capture position and orientation parameters. The system was designed to allow users to control simple shapes on a large display inside the Media Room by using pointing gestures (deictic) and voice commands.

Mechanical trackers (such as the BOOM Tracker from Fakespace Labs) have the advantage of accuracy and low latency. They may however be big and cumbersome with reduced mobility. Acoustic trackers (the Fly Mouse from Logitech[4]) prove to be relatively inexpensive, light weighted and with no interference with metals but

[2] http://www.fakespacelabs.com, last visited Jan. 2009

[3] http://www.ascension-tech.com, last visited Jan. 2009

[4] http://www.logitech.com, last visited Jan. 2009

they exhibit line-of-sight issues as well as sensitiveness to noise. They use high-frequency sound emitted from a source placed on the object to be tracked while microphones placed in the environment pick up the signals from the source in order to determine its position and orientation. Inertial trackers (IS300 from Intersense[5]) use a variety of inertial measurement devices such as accelerometers or gyroscopes and have the advantage of speed, accuracy, long working range, non interference with metals as well as no need for a transmitter. The sensors prove however to be bigger than the magnetic-based ones and they are likely subject of error accumulation. Hybrid trackers present the advantage of combining multiple technologies for improving on accuracy and reducing latency.

With respect to the main forms of feedback, one can classify devices into:

- *Ground referenced* such as Sensable Phantom devices[6]
- *Body referenced* Cyber Grasp from Immersion[7] which is a light weighted force-reflecting exoskeleton that fits over a Cyber Glove data glove wired version and adds resistive force feedback to each finger. Grasp forces are produced by a network of tendons routed to the fingertips via the exoskeleton
- *Tactile* such as Cyber Touch from Immersion, a tactile feedback option for Immersion's wired Cyber Glove instrumented glove. It disposes of small vibrotactile stimulators on each finger and the palm of the Cyber Glove system.

An example of a gesture recognition system using a data glove is Charade [2] which allows a speaker controlling presentation with free-hand gestures while still using gestures for communicating with the audience. A VPL DataGlove measures the bending of each finger and the 3D orientation of the hand: when the pointing direction of the hand enters an active zone, a cursor appears on the screen and follows the hand. A gesture is detected when the user's hand is pointing in the active area while gesture segmentation is performed using start and end positions defined by wrist orientation and finger positions.

3.1.1 Gesture-Based Interaction for Gaming

The gaming industry is a great target for gesture-based interfaces that allow players to immerse even more into the game environments. Gesture-based game controlling was introduced by Playstation 2 in the form of EyeToy[8]. EyeToy makes use of a video camera that gets mounted on the top of the TV screen facing the player that sits in the range of a couple of meters away. The players are brought into the game and allowed to interact with various game elements by using legs, arms, head or the whole body. The processing technique is based on motion, color detection and sound. Limitations of EyeToy are those common to computer vision applications: the video camera needs to be used in a well-lit room and the player must be in view in a given distance range. In order to help letting the player know when there is

[5] http://www.isense.com, last visited Jan. 2009

[6] http://www.sensable.com, last visited Jan. 2009

[7] http://www.immersion.com, last visited Jan. 2009

[8] http://www.eyetoy.com, last visited Jan. 2009

not enough light, a red LED on the front of the camera will flash indicating too dark working conditions. Also, the gestures are limited to basic hit-like motions area-wise around the player's body.

Nintendo released at the end of 2006 the Wiimote[9] controller for the Wii console. By using its motion sensing capability implemented using accelerometer and optical sensors, the controller allows players to interact with game elements by moving, shaking or pointing. The Wiimote is multi-functional and adapts to all kinds of games and scenarios. It may serve as the racket during a tennis match that players swing with their arms or as the steering wheel during a car race.

3.1.2 Multi-Touch, Tabletops and Interactive Surfaces

Touching is intuitive and presents several advantages such as direct manipulation and haptic feedback the user immediately receives.

Modern touchscreen interfaces usually employ one of three technologies: resistive, surface wave, or capacitive. Han [20] introduced a simple and inexpensive technique for enabling multi-touch sensing at high resolution for interactive surfaces based on a technology called frustrated total internal reflection. Dietz and Leigh [12] describe a technique for creating a touch-sensitive device that allows multiple users to interact simultaneously. Their surface generates modulated electric fields at each location that are capacitively coupled to receivers installed in the work environment.

Multi-touch may be implemented using computer vision as well. Wilson's system Touchlight [62] uses image processing techniques in order to combine video frames acquired from two cameras. The IR video cameras are placed behind a semi-transparent plane facing the user together with a projector and mirror system. By combining the distortion-corrected information from the two video sources, detection of objects that touch or are in a short distance of the surface plane is achieved. Surface computing is an intuitive way to interact with digital content [63, 64, 68]. Current applications include browsing photographs, playing videos, listening to music, viewing map locations or ordering menus[10]. Interesting interactions are achieved by connecting to external devices such as digital cameras or bluetooth mobile phones in order to exchange photographs [65].

3.2 Video-Based Systems

Camera-based or vision trackers make use of video information and video-based processing to achieve face, hand, fingers, arm or whole body tracking. The main advantage that comes with vision gesture acquisition and which provides the comfortable feeling of natural interaction is the fact that the technology is non intrusive and does not require users to wear additional equipments or devices. Users may interact freely with the system with no need for wearing or interacting via an additional device that may distract, restrict or burden the natural movement (e.g. wires more

[9] http://www.nintendo.com/wii, last visited Jan. 2009

[10] http://www.surface.com

or less heavy attached to gloves or hand-wearable trackers, glove sizes that may be a bit small or large). Of equally importance, vision-based solutions are relatively inexpensive compared to trackers that exhibit a price range from several hundreds to tens of thousands dollars.

3.2.1 Skin Color for Hands Detection

Color is an important feature that has been intensively used for hands detection. Instead of wearing distinctly colored gloves or rings [13, 40] or using LEDs or any other marker systems [53], skin color detection has proved to be an important intermediate step for face and hands tracking algorithms [1, 36, 58]. It has been observed on large image datasets [7, 8, 24] that skin color clusters in predefined intervals in several color spaces. Based on this comfortable property, a few approaches have been proposed: histograms of skin probability at different resolutions [24], single or mixtures of Gaussians modeling [7], elliptical [39], adaptive methods or various curved and linear polygonal segments for skin cluster delimitation. A common conclusion is that skin color indeed clusters under known limits in several color spaces.

There are several advantages when using skin color detection: invariance to orientation and size of the object being tracked (e.g. hands) and low computation time. Skin processing also comes with a few drawbacks such as the dependence on illumination: skin appears different under different illuminations hence skin color depends on the scene context; skin color varies from person to person (Caucasian, African, Asian, etc.); many real life objects present the same color as skin, i.e. the color of skin (or its reflectance) is not unique and does not pertain to skin only.

A very simple and low cost procedure is to filter the current video frame using simple static thresholds in a given color space, for example, the intervals $\left[h_{low}, h_{high}\right]$ and $\left[s_{low}, s_{high}\right]$ for the hue and saturation components in the HSV color space.

3.2.2 Challenges in Video Acquisition

There are also drawbacks when it comes to processing video information for detecting and recognizing gestures, many of which are commonly encountered when it comes to computer vision applications (see Figure 4):

- There is the dependency on the environment that translates into several issues: time varying lighting conditions; video cameras settings; users skin color; background in motion. All these issues burden the task of detectors which must prove to be stable, robust and continuously adapting to the environment.
- A calibration stage before using the system is required, usually related to color or point of view correction.
- There are sometimes constraints on the gesture dictionary with regards to the postures the system is able to successfully recognize. For example, fingers may occlude themselves while in movement in a single camera view; hands may partially or totally occlude fingers or they can be as well occluded by other objects from the scene.

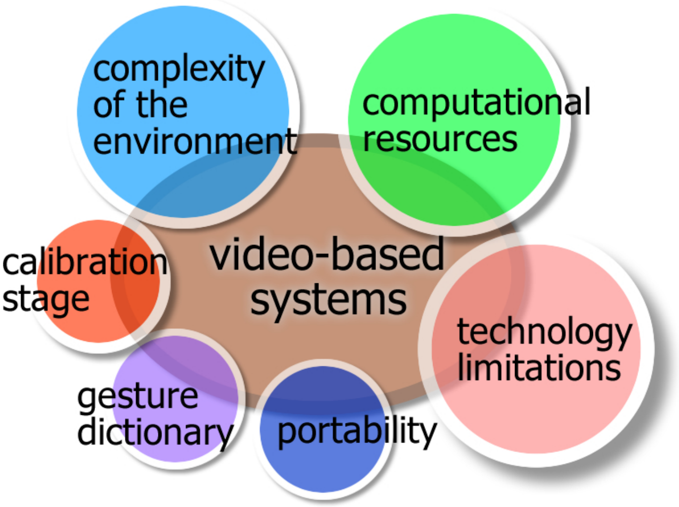


Fig. 4 Challenges encountered in video-based systems

- Video processing is CPU intensive in the conditions where real-time processing is a must for natural interaction. There is a great demanding of processing power especially if multiple cameras are involved. Latest research on GPGPU (General Purpose computing on the Graphics Processor Unit) demonstrated the use of the processing power of the GPU in order to run and execute highly-intensive CPU demanding computer vision algorithms [17]. OpenNVIDIA [18] is an open source library that implements vision algorithms on computer graphics hardware using OpenGL and Cg, available at[11].
- Technology and processing requirements limitations: video resolution may not be sufficient for detecting high fidelity finger movements; 25 fps of ubiquitous video capture devices may not be enough to capture quick hand movements (hand is quicker than the eye and or than the capture device in this case).
- Portability is an issue for most vision systems that require still placements of the video cameras. Nonetheless, mobile solutions exist and they consider wearable video computing (laptops with built-in cameras, head mounted displays and cameras [30]) but they come with all the problems related to wearable equipments.

Special considerations for using video-based acquisition solutions relate to:

- the number of cameras to be used (whether if 2D or 3D gestures are targeted or depth or stereo information is needed).
- cameras placement in space should not affect the visibility of body parts that are performing the gesture.

Commonly employed techniques include motion detection and motion flow following, color detection (and particularly skin color detection as a preprocessing

[11] http://openvidia.sourceforge.net, last visited Jan. 2009

stage for hands or face detection), feature based detectors (e.g. Haar-like features), non standard trackers (e.g. flock of birds) all combined with pattern recognition methods and techniques.

4 Gesture Recognition: Detecting Similarities

The section gives an overview on the current state-of-the-art in gesture recognition including detection, tracking, pose and trajectory recognition techniques. The focus is mainly oriented on gestures performed with one or two hands in a video-based acquisition scenario although references to other body gestures or capture devices are equally mentioned. [37, 46, 54, 60] provide excellent reviews on gesture recognition techniques. Also, good introductory courses on sketching exist such as [38].

4.1 Hands Tracking

There are two main streams of research encountered in the literature with regards to hand tracking: model-based and view-based approaches [46]. Model-based approaches make use of articulated 3D models of the hand that get projected onto the image plane. An error function is computed between the parameters of the model and various image features and the model parameters are adapted in correspondence with the the minimization of a certain cost functional. The view-based approaches use sets of features which are associated with a certain hand pose and common classifiers are trained from a previously collected database of feature samples. The set of features are searched for within the image, usually at multiple scales, looking for a high classifier output. The approach proves useful when the number of hand poses and the feature set is small.

Kolsch and Turk introduce a fast hand tracking technique that uses a flock of KLT features together with a learned foreground color distribution [33]. The KLT features, named after Kanande, Lucas and Tomasi [51], are based on the observation that a steep brightness gradient along at least two directions makes for a promising feature candidate to be tracked over time. A flock of features contains multiple instances of KLT features whose locations are updated independently from one video frame to another: they move to a new location for which the highest match correlation is found. The features don't follow a uniform direction, some might be lost while others may venture far from the flock. The concept of a flock inforces additional conditions with regards to features distribution: no two features must be closer to each other than a threshold distance and no feature must be farther away from the feature median than a second threshold distance. The tracker is fast, robust against background noise and has the ability to track objects that undergo rapid changes in orientation or deformations. An implementation is available under the HandVu library [31].

Haar-like features as introduced by Viola and Jones [59] for rapid object detection have been used for tracking and detection of several hand postures [32]. The Haar-like features are defined for gray-levels images as differences between the pixel intensities of various rectangle areas. A learning process will find the features that perform best on the training set and combine them into more powerful classifiers using Ada boosting methods. To achieve fast-real time detection, the classifiers are arranged in a cascade structure for which the first levels will discard most of the wrong candidates. The method is fast and combined with image pyramids structures assures size invariant object detection as well.

When only motion is important, the hands postures information may be discarded which takes the pressure off the tracking algorithms. Simple hand detection may be done using color tracking such as following colored gloves [13] or bright LEDs. Atlas Gloves[12] is a physical interface for controlling 3D mapping applications. The user interface is a pair of gloves that have LEDs attached. The LEDs are used to track intuitive hand gestures like grabbing, pulling, reaching and rotating. A video camera translates each LED-enabled gesture into a set of possible actions: pan, zoom, rotate and tilt.

Although tracking techniques are useful as they offer the possibility of continuously knowing the position, orientation or other parameters of the object of interest, they may not always be the best approach for acquiring gestures. For example, Wilson [66] presents a very intuitive method of interacting by means of gesture using the pinch posture or the TAFFI interface (Thumb and Fore Finger). A camera placed on the top of the computer screen detects the pinch gestures, i.e. when the thumb and the fore fingers of the user's hand touch. Tracking algorithms are avoided by only paying attention to the moment when the actual gesture takes place, i.e. when the user performs the pinch posture. Detection of the pinch posture is achieved using blob analysis by identifying the blob that fits the specific constraints of the region between the thumb and the fore finger. A good contrast between the user's hand and the background (represented by the keyboard in the majority of cases) is mandatory. Also the user is required to maintain the pinch posture as accurately as possible in the horizontal plane so that a blob could form between the two fingers.

Active shape models or smart snakes have been designed for exactly locating a feature in an image when given an initial guess [10, 21]. A contour, which is roughly the shape of the feature to be located is placed on the image at the feature approximate location. The contour is attracted to edges that are located near and its parameters change deforming it so that in the end it will converge to the actual shape of the feature. The process is iterative, moving and deforming the contour. The technique can be extended to tracking features across video frames: the position and shape of the feature in one frame is used as an approximation for the next frame.

4.2 Pose Recognition

Recognizing postures is important as it gives a great variety of information: touching the index finger with the thumb in the hand pinch posture has been exploited for select, drag or zoom operations; the index finger pointed means select; face emotions may be used as context. We explore template matching, PCA (principal component analysis) and neural networks as techniques that have been used before for posture recognition.

[12] http://atlasgloves.org, last visited Jan. 2009

4.2.1 Template Matching

Template matching is one of the simplest techniques that have been used for recognizing hand postures [52, 60, 69]. Template matching determines if a new posture can be classified to a number of class templates that have been previously computed from training sets. The classification part is implemented by computing a distance measure to each template. Commonly encountered measures are the sum of absolute differences or city-block distance and the sum of squares. Freeman and Weissman [16] demonstrate a TV-set control application using the template matching technique. The method may be efficient for a relative limited set of postures and it is simple to implement. It doesn't work well however for large sets of postures due to overlapping templates [60].

4.2.2 Principal Component Analysis

Principal component analysis (PCA) is a statistical technique for reducing the dimensionality of a data set while retaining as much of the variation in the data as possible. New variables are derived (called principal components), in decreasing order of importance, that are linear combinations of the original variables. Eigenvectors and eigenvalues are computed for the original data set and the eigenvector with the highest eigenvalue holds the highest variance. Similarly put, PCA produces an orthogonal coordinate system in which the axes are ordered in terms of the amount of variance in the original data for which the corresponding principal components account [61].

4.2.3 Neural Networks

Neural networks have been intensively used for gesture recognition [14, 15]. Common examples include multi-layered perceptrons, radial basis functions or Kohonen networks. Neural networks are previously learnt using a training set consisting of both positive and negative samples. For feed-forward networks having differentiable activation functions there exists a powerful and computationally efficient training method called error back-propagation for finding the derivatives of an error function with respect to the weights and biases of the network [3]. Glove-TalkII [14] translates hand gestures to speech using an adaptive interface. Hand gestures are mapped to 10 control parameters of a speech synthesizer. This makes the hand act as an artificial vocal tract for producing speech. Vatavu et al. [55] use a multi-layered perceptron in order to recognize a set of hand postures for the purpose of interacting with virtual objects.

4.3 Trajectory Matching: Recognizing Motions

Gestures do not consist in posture information only but they have motion trajectories associated as well. For example, describing the shape of a circle in mid-air using the hand may represent a gesture command for the creation of a circle or sphere-like object in a virtual environment. Trajectory recognition is an important problem and it relates to other application areas such as handwriting or signature recognition or general shape recognition. The difficulty of shape recognition in general, be it gesture, handwriting or signature is represented by the variability that is present within the data: users execute the same shape/gesture differently with each execution. Figure 5 illustrates an example of variability present within execution. Providing robust algorithms that would take into account these variations whilst providing good accuracy is a considerable challenge.

Gesture trajectories, as acquired using a given capture technology, are represented by an array of 2D or 3D points $\{p_1, p_2, \ldots p_n\}$. The data may be processed directly for recognition or other features may be extracted. Feature extraction and analysis usually consists in processing the low-level information from the raw data in order to produce higher-level information to help the further representation and recognition levels.

Rubine's GRANDMA system [48] performs recognition of 2D strokes using a set of 13 features such as: sine and cosine of the initial angle of the gesture; the length and angle of the diagonal of the bounding box; distance between first and last point; cosine and sine of the angle between the first and the last point; the total gesture length; total angle traversed; sum of the absolute value at each point; sum of the squared angles; maximum speed (squared) of the gesture; time duration of the gesture. Classification is done using a linear discriminator and reported recognition rates are above 98%.

Wobbrock et al. [67] introduce and make available to the HCI community an implementation of a simple yet robust algorithm for gesture trajectory matching. They entitle their classifier a "$1 recognizer" as it is easy, cheap and may be implemented

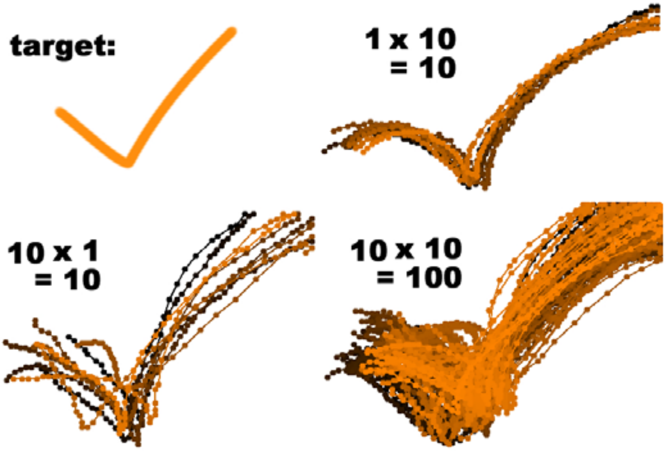


Fig. 5 Variability in execution. Top left: the *check* gesture. Top right: 10 executions performed by the same user. Bottom left: 10 executions from 10 users (one execution per user). Bottom right: 100 executions from 10 users.

in about 100 lines of code. The accuracy attained over a large set gesture samples was over 97% with only 1 loaded template and 99% accuracy with 3+ loaded templates.

Vatavu et al. [57] describe a method for recognizing gesture motions based on elastic deformable shapes and curvature templates. Gestures are modeled using spline curve representations that are further enhanced with elastic properties: the entire gesture or any of its parts may stretch or bend. The energy required to transform a gesture into a given template gives an estimation of the similarity between the two.

Simple trajectory matching of mouse gestures is supported under the Mozilla Firefox browser, Mozilla Thunderbird email client and Chatzilla. The gestures are captured by holding down a mouse button (usually the right one) and moving the mouse in a certain pre-defined way (simple horizontal, diagonal or vertical strokes) in order to form a gesture after which the mouse button is released. Mozilla gestures are combinations of URLD strokes (up/right/left/down) and 012 (left/middle/right mouse buttons). Mouse gestures operations include: working with history (backward/forward); working with windows and tabs (new window, minimize, next tab, new tab, previous tab, close); scroll (up, down) and zoom; working with images (double size, half size, hide image) and working with links (link in new tab); miscellaneous (view source, add bookmark). The source code is also available[13].

5 Case Study: Playing the Piano

We continue by discussing a practical case study of a gesture-based interface designed in order to allow users to play a virtual piano. We discuss the technical details of our implementation from a practitioner's point of view that needs to start development of similar natural user interfaces. The piano application was chosen as it provides several challenges with regards to technology and implementation. There are many piano simulators and applications that allow piano playing (a simple Google search revealed many downloadable links) however playing is achieved either via the keyboard by key association or via mouse clicks. The mouse and keyboard don't provide a natural interaction mechanism while the application is clearly suited for a gesture interface.

5.1 Building the Gestures Set

The requirements of the interface are to allow users to scroll and hover their virtual hand over the piano keys and press or hit a specific key. There is thus the need of a hand tracking mechanism combined with a *touch* detection when the index finger is above the corresponding piano key for a note to be produced. We simplify the interface to one-hand piano playing and one-finger key touching (the index finger) for discussion simplification although extension would be straightforward. Figure 6 illustrates the gestures set: horizontal hand motions are needed

[13] http://optimoz.mozdev.org/gestures/, last visited Jan. 2009

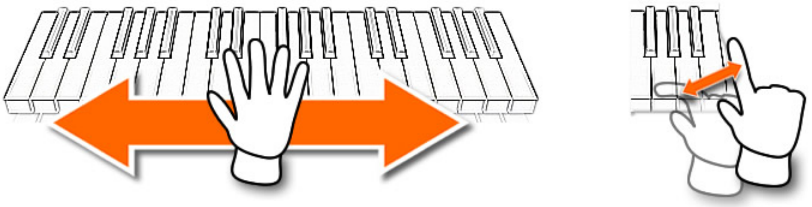


Fig. 6 Operations and associated gestures: travel to key by hovering above the piano (left); hitting a piano key by flexing the finger (right)

for traveling to piano keys while flexing the index finger will be used for hitting a given key.

5.2 *Gestures Acquisition*

A video camera was used in order to track the user hand and a 5DT glove[14] employed in order to detect the index finger flexing corresponding to hitting a piano key. The video camera is placed on the computer screen monitoring the desk surface and the user's hand. The scenario is similar to standard desktop computer usage: the user is seated in front of the computer screen while the hands rest comfortably on the surface of the desk. A snapshot of the scenario is illustrated in Figure 7.

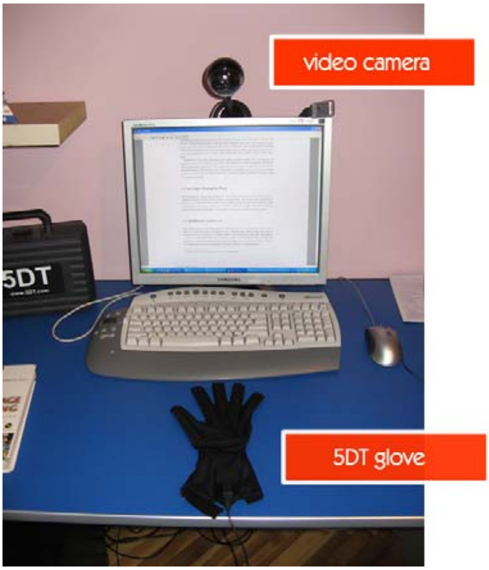


Fig. 7 Gesture acquisition scenario: video camera points down toward the desk tracking the user's hand; a 5DT sensor glove detects finger flexion

[14] http://www.5dt.com/products/pdataglove5u.html, last visited Jan. 2009

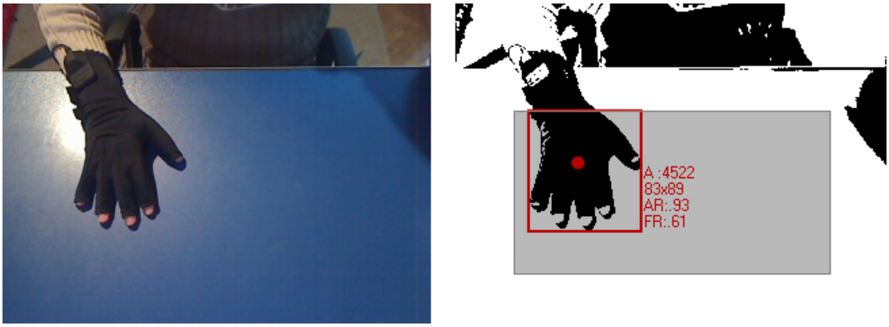


Fig. 8 Desktop view with user's hand wearing the 5DT glove. Left: webcam view, Right: hand identified after low-pass filtering and blobs detection.

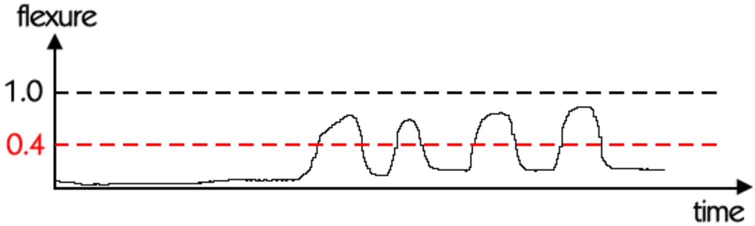


Fig. 9 Normalized output of the sensor associated to the index finger: 4 consecutive flexions are shown. A threshold of 0.4 was sufficient in order to detect key touches.

The video camera works at 25fps with a resolution of 320 x 240 pixels. Tracking the user hand is achieved by segmenting the black moving objects (as the 5DT glove the user is wearing comes in a black color) using a simple low-pass filter after which groups of connected pixels are detected as blobs (binary large objects) [47]. The segmentation process is simple and easy to implement due to the fact that the glove is of a homogeneous black color which can be easily and accurately detected on the surface of a desk (provided that the desk color is not black or dark). Figure 8 illustrates the segmentation result. The center of the detected hand blob is further used as (x,y) input for sliding across the keys of the virtual piano.

We further use the information provided by the sensor glove in order to detect hitting keys. The glove provides 5 sensors which measure the flexion of each finger in the normalized scale $[0..1]$ where 0 codes the finger being relaxed while 1 a finger completely flexed. The sensors are read with a frequency of 66 measurements per second. We are only interested in the index finger hitting a key or equivalently when it flexes just enough to hit a key. We experimentally chosen the threshold for the finger at 0.4 which is about in the middle of values domain with the meaning that any value greater than the threshold would mean the finger was flexed to hit a key. Figure 9 presents variation of the sensor normalized output over time.

We describe below the pseudo code for the hand-based interface trying to keep it simple and intuitive. Dummy procedure names with intuitive meaning are referenced for the purpose of shortening the code. This is an example of a simple controlled scenario which provides for quick implementation and fast results.

```
procedure HANDS-BASED-PIANO-PLAYING()
{
   while(running)
   {
      // read and process video frame
      image <- GET-NEXT-VIDEO-FRAME()
      // keep only 'black' pixels <= 50
      FILTER-BLACK-PIXELS(image)
      blobs <- GROUP-PIXELS(image)
      // hand is the biggest connected group of pixels
      hand <- BIGGEST-BLOB(blobs)

      // give visual feedback of a virtual
      // hand over the piano
      UPDATE-ICON-ON-SCREEN(hand.X, hand.Y)

      // read and process glove sensors
      flexion <- READ-GLOVE-MEASUREMENTS()
      if (flexion >= 0.4)
        PERFORM-CLICK(hand.X, hand.Y)
   }
}
```

The virtual piano is an example of application that can benefit of gesture interaction. Sliding the hand over the desk while a visual feedback of a virtual hand is provided on the screen together with finger flexing for simulating a key touch increases the experience users perceive. The mouse and keyboard, clearly unappropriated for such a task where replaced with free hand interaction. The approach taken in this study case is not by all means unique nor the solution bond to the particular technology that was used. The sensor glove was used together with the video camera in order to demonstrate a hybrid technology working together. Each sensor took separately could not have met the requirements: the glove does not come with a position reporting mechanism nor is the video camera powerful enough in order to spot slight finger flexions.

6 Conclusions

Gesture-based interfaces are definitely evolving and we will encounter more and more of them in our everyday lives. Be it in the form of touch screens, presence sensors that trigger actions (lights on/off on hallways, water control in bathrooms washbasins) or complex vision systems that understand free hand gestures in mid-air, gestures will play an important part in future human-computer interfaces.

There is a catch in all this: the fact of gesture acquisition and recognition being a new and impressive technology that attracts the attention should not make them all-purpose interfaces. Instead, the right applications or tasks that would indeed benefit of such powerful interactions must be investigated. Gestures are natural to perform in the real-world as a mean of interaction and conveying information. At the same time, gestures can be described as imprecise and not self revealing. For example, a touchscreen that accepts gesture motions as input but doesn't present any form of assistance of how to do this is as misleading as a blank sheet of paper: what gestures to use? where and how should one begin to interact? This discoverability issue was far less important in GUIs where buttons, icons and menus give you an overview of the possible actions. Ergonomics plays an important part as well as interfaces for general purpose and wide public applications should be designed so that they can be used by people of different ages or different motor skills. Muscles fatigue in time if the design of the gesture set is inappropriate so comfortable interactions should be a priority.

When designing gesture-based interfaces, several guidelines should be considered: give permanent feedback to users to improve confidence and system responsiveness; provide hints on how the technology was implemented (video, sensors, audio), take into account and let users know limitations of the chosen technology - just as right expectation; combine gestures with existing GUIs or traditional commands and actions - use gestures as shortcuts for advanced users that want to perform tasks faster; construct gesture dictionaries that are appropriate (not demanding special physical condition, not offending to any culture, meaningful and easy to understand, memorize and recall) - take users' feedback on gesture association with given tasks. It is also important that interfaces should be fun to use as this gives a new dimension to the system, draws users attention and in the end make users forgiveness and more willing to tolerate recognition errors.

References

[1] Angelopoulou, E.: Understanding the color of human skin. In: Proceedings of the SPIE Conference on Human Vision and Electronic Imaging VI, pp. 243–251. SPIE Press (2001)

[2] Baudel, T., Beaudouin-Lafon, M.: Charade: remote control of objects using free-hand gestures. Communications of the ACM 36(7), 28–35 (1993)

[3] Bishop, C.M.: Neural Networks for Pattern Recognition. Oxford University Press, Oxford (1995)

[4] Bolt, R.A.: Put-that-there: Voice and gesture at the graphics interface. In: Proceedings of the 7th annual conference on Computer graphics and interactive techniques (SIGGRAPH 1980), Seattle, Washington, United States, pp. 262–270. ACM Press, New York (1980)

[5] Butterworth, B., Beattie, G.: Gesture and silence as indicators of planning in speech. In: Campbell, R., Smith, P. (eds.) Recent advances in psychology of language: Formal and experimental approaches, pp. 347–360. Plenum, New York (1978)

[6] Cadoz, C.: Le geste canal de communication homme-machine. La communication instrumentale. Technique et Science Informatique 13(1), 31–61 (1994)

[7] Caetano, T.S., Olabarriaga, S.D., Barone, D.A.C.: Performance Evaluation of Single and Multiple-Gaussian Models for Skin Color Modelling. In: Proceedings of the 15th Brazilian Symposium on Computer Graphics and Image Processing, pp. 275–282 (2002)

[8] Caetano, T.S., Olabarriaga, S.D., Barone, D.A.C.: Do mixture models in chromaticity space improve skin detection? Pattern Recognition 36(12), 3019–3021 (2003)

[9] Cassell, J.: A Framework for gesture generation and interpretation. In: Cipolla, R., Pentland, A. (eds.) Computer Vision in Human-Machine Interaction, pp. 191–215. Cambridge University Press, New York (1996)

[10] Chang, J.S., Kim, E.Y., Jung, K., Kim, H.J.: Real Time Hand Tracking Based on Active Contour Model. In: Gervasi, O., Gavrilova, M.L., Kumar, V., Laganá, A., Lee, H.P., Mun, Y., Taniar, D., Tan, C.J.K. (eds.) ICCSA 2005. LNCS, vol. 3483, pp. 999–1006. Springer, Heidelberg (2005)

[11] Chartrand, T.L., Bargh, J.A.: The chameleon effect: The perception-behavior link and social interaction. Journal of Personality and Social Psychology 76(6), 893–910 (1999)

[12] Dietz, P., Leigh, D.: DiamondTouch: a multi-user touch technology. In: Proceedings of the 14th annual ACM symposium on User interface software and technology (UIST 2001), Orlando, Florida, United States, pp. 219–226. ACM Press, New York (2001)

[13] Dorner, B.: Chasing the Colour Glove: Visual Hand Tracking, Simon Fraser University (1994)

[14] Fels, S., Hinton, G.: Glove-TalkII: an adaptive gesture-to-formant interface. In: Proceedings of the SIGCHI conference on Human factors in computing systems (CHI 1995), Denver, Colorado, United States, pp. 456–463. ACM Press, New York (1995)

[15] Finlay, J., Beale, R.: Neural networks and pattern recognition in human-computer interaction. SIGCHI Bull. 25(2), 25–35 (1993)

[16] Freeman, W.T., Weissman, C.D.: Television Control by Hand Gestures. In: Proceedings of the 1st International Conference on Automatic Face and Gesture Recognition (1994)

[17] Fung, J., Mann, S.: Computer Vision Signal Processing on Graphics Processing Units. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Montreal, Quebec, Canada (2004)

[18] Fung, J., Mann, S., Aimone, C.: OpenVIDIA: Parallel GPU Computer Vision. In: Proceedings of the 13th annual ACM international conference on Multimedia, Singapore, November 2005, pp. 849–852. ACM Press, New York (2005)

[19] Graham, J.A., Argyle, M.: A cross-cultural study of the communication of extra-verbal meaning by gestures. International Journal of Psychology 10(1), 57–67 (1975)

[20] Han, J.Y.: Low-cost multi-touch sensing through frustrated total internal reflection. In: Proceedings of the 18th annual ACM symposium on User interface software and technology (UIST 2005), Seattle, Washington, United States, pp. 115–118. ACM Press, New York (2005)

[21] Heap, A.J., Samaria, F.: Real-Time Hand Tracking and Gesture Recognition Using Smart Snakes. In: Proceedings of Interface to Real and Virtual Worlds (1995)

[22] Houghton Mifflin Company: American Heritage Dictionary of the English Language, 4th edn. Houghton Mifflin Company (2000)

[23] Iverson, J.M., Goldin-Meadow, S.: Why people gesture when they speak. Nature 396(6708), 228 (1998)

[24] Jones, M.J., Rehg, J.M.: Statistical color models with application to skin detection, TR 98/11, Cambridge Research Laboratory (1998)

[25] Karam, M., Schraefel, M.C.: A Taxonomy of Gestures in Human Computer Interactions, TR ECSTR-IAM05-009, Electronics and Computer Science, University of Southampton (2005)

[26] Kendon, A.: Gesticulation and speech: two aspects of the process of utterance. In: Key, M.R. (ed.) The relation of verbal and nonverbal communication, pp. 207–227. Mouton, The Hague (1980)

[27] Kendon, A.: Current issues in the study of gesture. In: Nespoulous, J.-L., Perron, P., Lecours, A.R. (eds.) The Biological Foundation of Gestures: Motor and Semiotic Aspects, pp. 23–47. Lawrence Erlbaum Assoc., Hillsdale (1986)

[28] Kendon, A.: Do gestures communicate? A review. Research on Language and Social Interaction 27(3), 175–200 (1994)

[29] Kendon, A.: An agenda for gesture studies. Semiotic Review of Books 7(3), 8–12 (1996)

[30] Kolsch, M., Turk, M., Hollerer, T.: Vision-based interfaces for mobility. In: Proceedings of the 1st Annual Int. Conf. on Mobile and Ubiquitous Systems: Networking and Services (2004)

[31] Kolsch, M., Hollerer, T., DiVerdi, S.: HandVu: A New Machine Vision Library for Hand Tracking and Gesture Recognition, demo at ISWC/ISMAR (2004)

[32] Kolsch, M., Turk, M.: Robust Hand Detection. In: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (2004)

[33] Kolsch, M., Turk, M.: Hand tracking with Flocks of Features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2005)

[34] Krauss, R.M., Morrel-Samuels, P., Colasante, C.: Do conversational hand gestures communicate? Journal of Personality and Social Psychology 61(5), 743–754 (1991)

[35] Kurtenbach, G., Hulteen, E.: Gestures in human-computer communications. In: Laurel, B. (ed.) The Art of Human Computer Interface Design, pp. 309–317. Addison-Wesley, Reading (1990)

[36] Cho, K.-M., Jang, J.-H., Hong, K.-S.: Adaptive skin color filter. Pattern Recognition 34(5), 1067–1073 (2001)

[37] LaViola, J.: A survey of hand posture and gesture recognition techniques and technology, TR CS-99-11, Department of Computer Science, Brown University, Providence RI (1999)

[38] LaViola, J.: Sketching and gestures 101. In: SIGGRAPH 2007: ACM SIGGRAPH 2007 courses, p. 2. ACM Press, New York (2007)

[39] Lee, J.Y., Yoo, S.I.: An elliptical boundary model for skin color detection. In: Proceedings of the Int. Conf. on Imaging Science, Systems and Technology, Las Vegas, USA (2002)

[40] Zhang, L.-G., Chen, Y., Fang, G., Chen, X., Gao, W.: A vision-based sign language recognition system using tied-mixture density HMM. In: Proceedings of the 6th international conference on Multimodal interfaces (ICMI 2004), State College, PA, USA, pp. 198–204. ACM Press, New York (2004)

[41] McNeill, D.: Hand and mind: What gestures reveal about thought. University of Chicago Press, Chicago (1992)

[42] Merriam-Webster Inc.: Merriam-Webster's Medical Dictionary, Merriam-Webster Inc. (2002)

[43] Mulder, A.: Hand Gestures for HCI: Research on human movement behavior reviewed in the context of hand centered input, TR 96-1, Simon Fraser University (1986)

[44] Napier, J.R.: Hands. Princeton University Press, Princeton (1993)

[45] Nespoulous, J.-L., Perron, P., Lecours, A.R.: The Biological Foundations of Gestures: Motor and Semiotic Aspects. Lawrence Erlbaum Associates, Hillsdale (1986)

[46] Pavlovic, V., Sharma, R., Huang, T.: Visual interpretation of hand gestures for human-computer interaction: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(7), 677–695 (1997)

[47] Pratt, W.K.: Digital Image Processing, PIKS Inside, 3rd edn. John Wiley & Sons Inc., Chichester (2001)

[48] Rubine, D.: Specifing Gestures by Example. In: Proceedings of SIGGRAPH 1991, pp. 329–337. ACM Press, New York (1991)

[49] Saffer, D.: Designing Gestural Interfaces. O'Reilly Media Inc. (2009)

[50] Schegloff, E.A.: On some gestures' relation to speech. In: Atkinson, J.M., Heritage, J. (eds.) Structures of social action: Studies in conversational analysis, pp. 226–296. Cambridge University Press, Cambridge (1984)

[51] Shi, J., Tomasi, C.: Good features to track. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (1994)

[52] Sturman, D.J.: Whole-hand Input, PhD thesis, Massachusetts Institute of Technology (1992)

[53] Sturman, D.J., Zeltzer, D.: A Survey of Glove-based Input. IEEE Computer Graphics and Applications 14(1), 30–39 (1994)

[54] Tappert, C.C., Suen, C.Y., Wakahara, T.: The state-of-the-art in on-line handwriting recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 12(8), 787–808 (1990)

[55] Vatavu, R.D., Pentiuc, S.G., Chaillou, C., Grisoni, L., Degrande, S.: Visual Recognition of Hand Postures for Interacting with Virtual Environments. In: Proceedings of the 8th International Conference on Development and Application Systems, DAS 2006, Suceava, Romania, pp. 477–482 (2006)

[56] Vatavu, R.D., Pentiuc, S.G.: Multi-Level Representation of Gesture as Command for Human Computer Interaction. Computing and Informatics 27(6), 837–851 (2008)

[57] Vatavu, R.D., Grisoni, L., Pentiuc, S.G.: Gesture Recognition Based on Elastic Deformation Energies. In: Dias, M.S., Gibet, S., Wanderley, M.M., Bastos, R. (eds.) GW 2007. LNCS (LNAI), vol. 5085, pp. 1–12. Springer, Heidelberg (2009)

[58] Vezhnevets, V., Sazonov, V., Andreeva, A.: A survey on pixel based skin color detection techniques. In: Proceedings of Graphicon 2003, Moscow, Russia, pp. 85–92 (2003)

[59] Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 511–518 (2001)

[60] Watson, R.: A survey of gesture recognition techniques, TR TCD-CS-93-11, Trinity College Dublin (1993)

[61] Webb, A.: Statistical Pattern Recognition. John Wiley & Sons, Ltd., West Sussex (2002)

[62] Wilson, A.D.: TouchLight: an imaging touch screen and display for gesture-based interaction. In: Proceedings of the 6th international conference on Multimodal interfaces (ICMI 2004), State College, PA, USA, pp. 69–76. ACM Press, New York (2004)

[63] Wilson, A.D.: PlayAnywhere: A Compact Tabletop Computer Vision System. In: Proceedings of the 18th Symposium on User Interface Software and Technology (UIST 2005), Seattle, WA, USA, pp. 83–92. ACM Press, New York (2005)

[64] Wilson, A.D., Robbins, D.C.: PlayTogether: Playing Games across Multiple Interactive Tabletops. In: IUI Workshop on Tangible Play: Research and Design for Tangible and Tabletop Games (2006)

[65] Wilson, A.D., Sarin, R.: BlueTable: Connecting Wireless Mobile Devices on Interactive Surfaces Using Vision-Based Handshaking. In: Proceedings of Graphics Interface 2007, pp. 119–125. ACM Press, New York (2007)

[66] Wilson, A.D.: Robust computer vision-based detection of pinching for one and two-handed gesture input. In: Proceedings of the 19th annual ACM symposium on User interface software and technology (UIST 2006), Montreux, Switzerland, pp. 255–258. ACM Press, New York (2006)

[67] Wobbrock, J.O., Wilson, A.D., Li, Y.: Gestures without libraries, toolkits or training: a $1 recognizer for user interface prototypes. In: Proceedings of the 20th annual ACM symposium on User interface software and technology (UIST 2007), Newport, Rhode Island, USA, pp. 159–168. ACM, New York (2007)

[68] Wu, M., Balakrishnan, R.: Multi-finger and whole hand gestural interaction techniques for multi-user tabletop displays. In: Proceedings of the 16th annual ACM symposium on User interface software and technology (UIST 2003), Vancouver, Canada, pp. 193–202. ACM Press, New York (2003)

[69] Zimmerman, T.G., Lanier, J., Blanchard, C., Bryson, S., HArvill, Y.: A hand gesture interface device. In: Proceedings of the SIGCHI/GI conference on Human factors in computing systems and graphics interface (CHI 1987), Toronto, Ontario, Canada, pp. 189–192. ACM Press, New York (1987)

Super Resolution for Multimedia, Image, and Video Processing Applications

K. Malczewski and R. Stasiński

Abstract. In this chapter an overview of super-resolution (SR) methods used in processing of multidimensional signals is given. The emphasis is laid on presentation of possibly wide range of SR algorithms rather than on detailed description of few ones. The term *super resolution* is used for naming any technique that exploits the knowledge contained in several low-resolution signals to form a high-resolution one. For example, few frames from a cell-phone video containing hazy shots of some text can be transformed into its clearly readable image. There are many applications of super-resolution, it is used successfully for improving medical imaging systems, satellite imaging, in surveillance, astronomical imaging; new ideas are emerging all the time.

The chapter begins with some introductory remarks, and with a description of the general model of imaging systems (observation model). Then, an extensive overview of super-resolution techniques is given, firstly multi-frame, and then single-frame ones are described. The following section contains examples illustrating some problems and solutions encompassed by the domain of super-resolution. At the end a table summarizing properties of the most important super-resolution algorithms is provided.

1 Introduction

The statement that an image or video has *high resolution* means that its fine details are particularly well rendered. In digital technology the prime (but not unique) factor determining resolution is how many pixels constitute an image, or video frame. The pixel cost, whatever is its definition, is an ever decreasing factor, nevertheless, this does not mean that the problem of insufficient resolution of equipment at our disposal now and here will definitely disappear. In problematic situations a good idea is to apply a technique of *super-resolution* (SR). Namely, super-resolution is a common name for a group of methods of improving image or video resolution above its initial level. The term should not be confused with image *enhancement*, or *reconstruction*, in contrast to an enhanced image the SR-processed one potentially

K. Malczewski and R. Stasiński
Poznań University of Technology
Faculty of Electronics and Communications
Polanka 3, PL-60-965 Poznań, Poland
e-mail: `{kmal,rstasins}@et.put.poznan.pl`

M. Grgic et al. (Eds.): Rec. Advan. in Mult. Sig. Process. and Commun., SCI 231, pp. 171–208.
springerlink.com

contain some additional information, if compared to the unmodified source. Of course, adding details is impossible without some side information accompanying the input low-resolution image. SR techniques found applications in such various domains as standard definition TV to HDTV signal conversion, medical image processing, military applications (infrared imaging), satellite imaging, surveillance systems, and many others. The most challenging, while very promising, area of SR application is low-end multimedia equipment, like cell phones. The overview of SR techniques done in this chapter shows that the domain is matured enough to undertake any challenge, as there is a multitude of options and solutions.

Insufficient equipment resolution may happen on any complexity level – from cell-phone camera to telescope imaging system. In the latter case the reason is obvious, much more complex hence, more instructive is the analysis of the former one. Making photos and video clips isn't the main application of a cell-phone, which means that its camera should be possibly small and unobtrusive. As a consequence the imaging sensor is often very small, which combined with cheap, primitive lens results in low pixel count. Namely, even moderate number of pixels on a small sensor implies high pixel density hence, demand on lens high throughput. In more multimedia-oriented phones lens are better, nevertheless, other limitations emerge. One is individual pixel size – small pixel diameter means large sensor bandwidths, noise power may still decrease with sensor dimensions, but not as fast as signal power. Noise can be somewhat ameliorated by signal processing, but this increases computational burden of phone processors. This is linked with other important limitation, characteristic of any portable equipment in general – acquisition of each image pixel costs processor time, memory occupation, and battery charge. We cannot count here on steady improvement in this domain, as introduction of every new "trendy" cell-phone feature diminishes camera share in phone's processing, memory, and energy budgets. Finally, many cell-phone users are absolutely happy with their mediocre quality photos and video clips, which is not surprising at all when they watch them on diminutive cell-phone displays.

In the case of webcams the above budgets are unrestricted, the resolution limiting factor is here the bandwidth of slow Internet connections. Another obstacles rule construction of surveillance systems. Namely, among their most important virtues are reliability and environmental conditions resilience, which results in somewhat "retarded" but proven solutions. Additionally, if we have large, expensive, and still working flawlessly system we would think twice before replacing it by a new one, only because its higher resolution might be useful sometimes. As can be seen from the above analysis, pixel cost is by no means the sole factor for application of not-so-high resolution sensors.

Super-resolution algorithms can be divided into two groups: multiple-frame super resolution methods and single-frame super-resolution ones (single-frame image zooming, also: quasi super-resolution). Most of them address the first group. Side information for SR methods is usually extracted from few low-resolution images, in the case of single-frame SR ones these are training images. Nevertheless, the former term is also used when the only a priori knowledge concerns image blur/noise, or maybe other properties. In such cases a good question is whether these are SR methods, or simply image enhancement ones, but in a modern disguise. The single-frame high-resolution image is found in a solution space, the search is steered by the a priori knowledge. The techniques are useful when collecting multiple images of the same scene is problematic or impossible. Nevertheless, there is a price for it, as the processing often introduces artifacts. Namely, the result is in part nothing more than an educated guess.

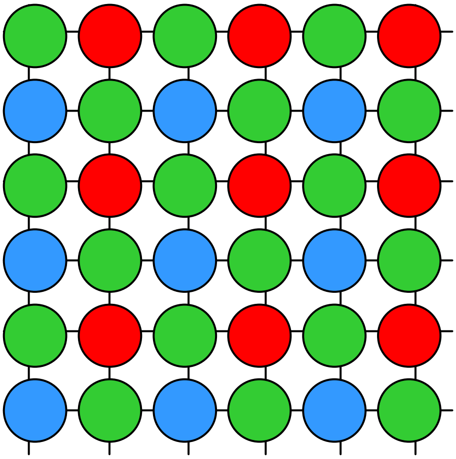


Fig. 1 Localization of RGB pixels in a Bayer sensor

The multiple-frame SR methods, often referred to simply as SR ones, are usually more reliable. To understand how these techniques work it is worth to analyze the (non-SR) method of obtaining full resolution RGB images from a Bayer sensor implemented in high-end medium format camera backs. As can be seen from Figure 1, resolution of an *N*-pixel Bayer sensor is much smaller than *N* image samples: they are *N*/2 samples in green channel, and only *N*/4 samples in channels red and blue. Moreover, red, green and blue images are collected on slightly different positions, their mutual shift is one to two pixels. This is the source of annoying and clearly visible chrominance moiré effects, and the reason why the majority of manufactures puts anti-aliasing filters on their sensors. In the case of medium format equipment such filters are usually optional. Instead of applying them it is often possible to shot (at least) 4 photographs of the same scene: one normal, one shifted by one pixel horizontally, one shifted by one pixel vertically, and one for both shifts combined. In this way we are obtaining a "true" *N*-sample image – it consists of *N*-samples in red and blue channels, and two *N*-sample copies of the green one, these can be combined for diminishing noise, rejecting "wild" pixels etc. Note that now there is no shift between color images, and that resolution of red and blue channels has been increased two times.

The super-resolution techniques try to mimic what is done to red and blue channels in medium-format backs without changing the construction of an ordinary camera. Of course, making perfectly controlled shifts without rebuilding the equipment is not feasible, random shifts are accepted instead. Then, the first step of any SR algorithm consists in measuring these shifts. Precision of the measurements is crucial for the success of the whole method, and several approaches to their accomplishment have been proposed. Once the shifts are known, we can position samples of collected low-resolution images on the plane of "analog" image projected by a lens on the camera sensor. Then, the following step is image interpolation, of course, we are not interested in full reconstruction of the "physical" image, but rather in determining its samples on a sampling grid much denser than that implied by pixel pitch of the camera sensor. A rule of thumb is that due to inevitable errors in shifts determination for obtaining twofold increase in pixel density six or more good quality low-resolution images should be collected (in contrast to four ones when shifts are exactly known, as in medium format backs). Fourfold increase requires much more than sixteen images, which makes it hardly feasible, if possible at all. Good quality means here that the low-resolution images can be effectively de-blurred, and de-noised. Fundamental limits on how much the resolution may be improved exist and under certain assumptions can be computed analytically [3].

The chapter is organized as follows: In the next section the general model of imaging systems is described, the so-called *observation model*. The overview of SR algorithms is done in the following two sections devoted to multi-frame and single-frame techniques, respectively. Multi-frame methods are divided into three groups: intuitive (inspired by observation model), back-projection, and regularized, the last group contain some particularly effective approaches (Bayesian, least-squares, POCS). Single-frame techniques form a rather mixed bag aiming not only at resolution improvement, but also at "plain" visual image enhancement, or even image mosaicing. Then follows a chapter containing results of experiments with exemplary SR reconstruction techniques, the chosen methods represent all groups of SR approaches. Finally, the conclusion is supported by a table, in which properties of the most important SR techniques are summarized.

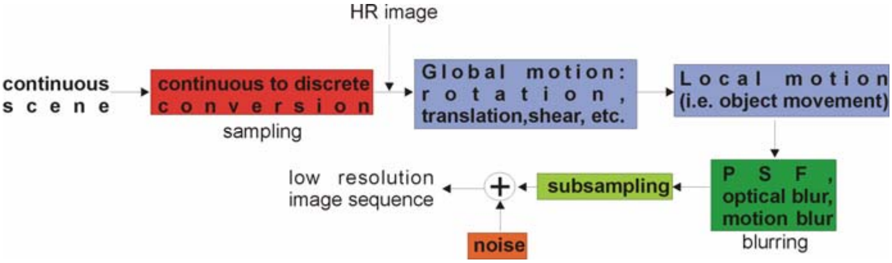

HR image
continuous scene
continuous to discrete conversion
sampling
Global motion: rotation, translation,shear, etc.
Local motion (i.e. object movement)
PSF, optical blur, motion blur
blurring
subsampling
+
noise
low resolution image sequence

Fig. 2 Observation model

2 Observation Model

Let us start with description of relations linking input low-resolution (LR) image, or images, and the output high-resolution (HR) one. Namely, in this section we are describing the *observation model*, Figure 2. Several observation models have been proposed in the literature [76], they can be divided into models for still images and for video sequences. We are proposing a simplified universal one.

It is assumed that the desired HR image is defined by a vector of image samples $x = [x_1, x_2, ..., x_N]^T$, where N is given by $N = M_1N_1 \times M_2N_2$. The x is assumed to be sampled at or above the Nyquist rate from a continuous scene supposed to be bandlimited. For the LR images, the horizontal and vertical directions are subsampled by the factors given by M_1 and M_2, respectively. The k-th LR image is represented by the vector $y = [y_{k,1}, y_{k,2}, ..., y_{k,M}]^T$ where $k = 1,2,...,p$ and $M = N_1 \times N_2$.

Our simplified observation model takes into consideration all important image degradation factors: warping, blurring, downsampling, noising and motion degradations, Figure 2. Information about these factors is crucial for the success of SR reconstruction. The model for k-th image is described by the following formula:

$$y_k = DB_kM_kx + n_k \text{ for } 1 \le k \le p \tag{1}$$

where M_k and B_k are warp and blur matrices, respectively, D is a subsampling matrix, n_k denotes a noise vector, and p is the number of LR images. The warp matrix M_k represents the motion in observed scene that takes place when acquiring k-th image, leading to global or local geometric degradations. The blur matrix B_k is describing effects of blurring by the optical system, effects of relative motion between the original scene and the imaging system when k-th image is acquired, and the point spread function (PSF) of camera sensor (note that its resolution is low – LR). In the case of SR image reconstruction, this LR sensor PSF is usually modeled as a spatial averaging operator. Finally, aliased LR images are generated by downsampling the HR one, this is done by multiplying the warped and blurred HR image by matrix D.

The above observation model may be expressed in a simpler form as follows:

$$y_k = W_kx + n_k \text{ for } 1 \le k \le p \tag{2}$$

where W_k represents all degradation operators. Our goal is to estimate HR image x from the set of LR ones given by vectors y_k.

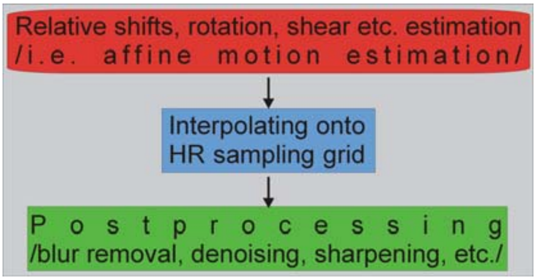

Relative shifts, rotation, shear etc. estimation /i.e. affine motion estimation/
Interpolating onto HR sampling grid
Postprocessing /blur removal, denoising, sharpening, etc./

Fig. 3 Direct inversion of the observation model – scheme for the intuitive SR image reconstruction

To counteract distortions introduced by observation model any kind of super-resolution multi-frame algorithm should perform the following three actions on low-resolution images: registration, interpolation, and restoration. Their meaning is provided in Figure 3. In this chapter the algorithms in which these steps are done in distinct stages, and only once, are named *intuitive* SR methods, their simplified flowchart is shown in Figure 3. In other techniques two, or all three actions are done simultaneously, and repeated iteratively, see section 3.

Multi-frame Super-Resolution is feasible only when a scene is (almost) static, hence, capturing of a set of Low-Resolution (LR) images is possible, when this is not a case, single-frame methods can be applied. What is particularly important LR images should be taken from various camera positions, sub-pixel shifts between LR images are crucial. Namely, if fractional shifts between LR images are not present, all images contain the same information. In such situation a better solution might be simple image interpolation for image enhancement. No special devices for camera displacement are needed here, chaotic movements of hand that holds it are sufficient.

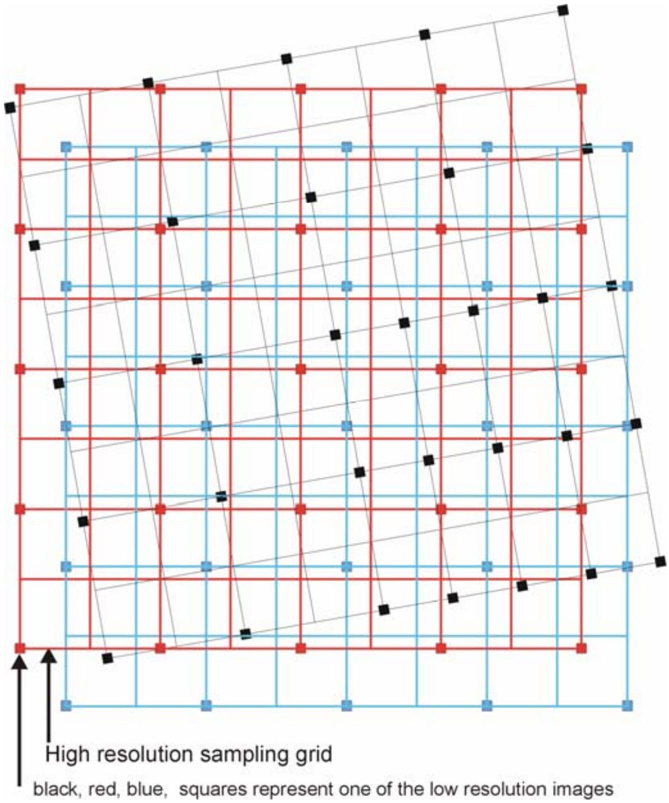

High resolution sampling grid
black, red, blue, squares represent one of the low resolution images

Fig. 4 Relative displacements between LR images and HR sampling grid

The displacement (or motion) estimation, or registration, is extensively studied in various fields of image processing [3, 8]. This is the most vulnerable part of SR algorithms, namely, even slightest estimation errors may result in visible HR image distortions and artifacts. As can be seen from Figure 4, parameters evaluated in the registration stage might be not only horizontal and vertical shifts, but also rotation angles, and possibly other ones. The set of measured parameters depends on the motion model implemented in an SR reconstruction technique.

The next step of an SR algorithm is a reconstruction of High Resolution (HR) image from samples on nonuniformly sampled grid. Non-uniformity of LR image samples positions on image plane comes from a somewhat chaotic pattern of displacements between LR images, see Figure 5.

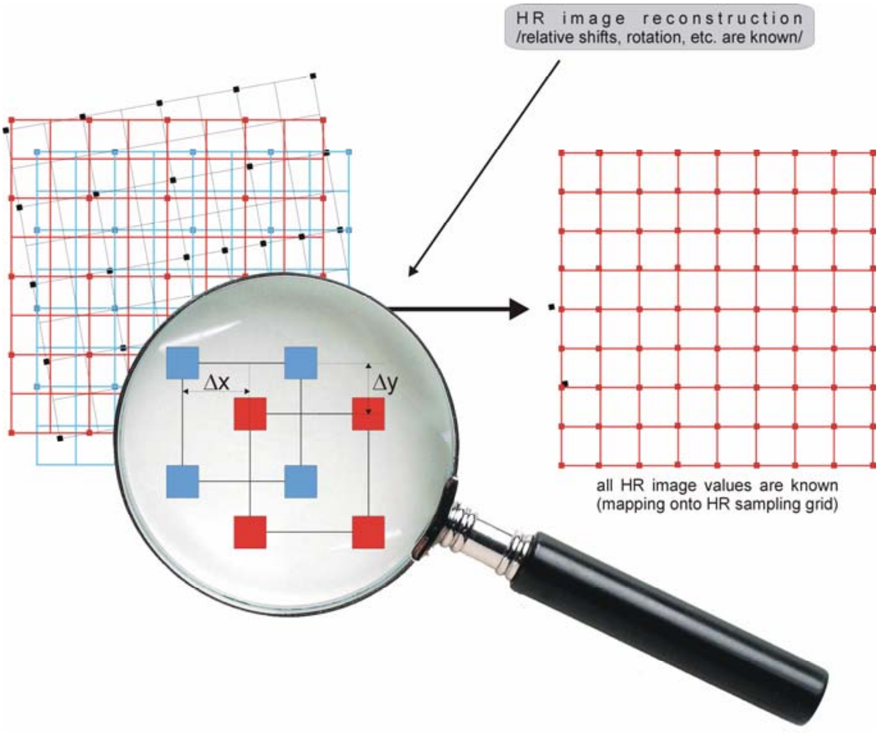

HR image reconstruction
/relative shifts, rotation, etc. are known/
Δx
Δy
all HR image values are known
(mapping onto HR sampling grid)

Fig. 5 High resolution image reconstruction from irregular set of samples

As a final touch, some postprocessing is applied to the upsampled image, called restoration. Its goal is the removal (or at least amelioration) of blur and noise. The restoration can be done either directly on image samples (spatial methods), or in the frequency domain. The tutorial paper of Borman and Stevenson [40] provides a comprehensive and complete overview of SR image reconstruction algorithms until around 1998. In our work we are basing classification of SR techniques on the approach taken to inverse the observation model from Figure 2.

3 Super-Resolution Algorithms

The history of super-resolution began with Tsai and Huang work [64]. Authors described an algorithm to register multiple frames simultaneously using nonlinear minimization technique approach in frequency domain. Their method for registering multiple aliased images was based on the fact that the original, high-resolution signal is band-limited. It was not clear, however, if such a solution was unique and if such an algorithm would not converge to a local minimum. Their algorithm implemented the same principle as the formulation in time domain given by Papoulis [3].

As it has been said in section 2, in each (multi-frame) super-resolution method the following three actions should be somehow implemented: first all LR images are registered in the same coordinate system, then a high-resolution image is reconstructed from the irregular set of samples, finally deblurring and denoising procedures are applied, see Figure 3. The Registration step could be completed either in the spatial or in the frequency domain. Unfortunately, due to basic properties of the Fourier transform, frequency domain methods are limited to global translational and rotational motion models. This can be quite limiting, as accuracy of subpixel motion estimation is one of principle requirements for successful image reconstruction. On the other hand, the aliasing is much easier to explain and understand in the frequency domain. Nevertheless, much more flexible and realistic higher order motion models, such as affine, projective can be combined with this approach. In [48] authors proposed preliminary image sequence local motion segmentation by incorporating modified region growth and optical flow approaches. Irani et al. [48] presented a technique that overcame both object transparency and the problem of occluding motions in an image sequence. They proposed to track different objects using segmentation and temporal integration. Gluckman in [26] presented a registration method that utilized gradient field distribution of images. Vandewalle [62] and Gluckman devised algorithms in which planar shift estimations followed cancellation of rotation done by a phase correlation method.

Of course, image registration can be done in the spatial domain, too [41]. Incorporation of a priori knowledge into the solution can be done using the projection onto convex sets (POCS) approach (Patti et al. [15]). POCS techniques are usually iterative (estimated reconstruction is repeatedly projected on consecutive convex sets), each set represents a constraint to the reconstructed image, which have been derived from some measurements and assumptions about a signal. Elad and Feuer [15] combined the POCS idea with the maximum likelihood (ML), or maximum a posteriori (MAP) probability approach to update the convex optimization formulation.

Mathematically, super-resolution image reconstruction is a kind of inversion problem, hence, it is usually ill-conditioned. A widely used method for stabilization of inversion problems is their regularization. Successful regularization strongly depends on the choice of proper regularization parameters. Furthermore in many practical situations image blur (mathematically – operation to be inverted) is unknown or is known only within a set of parameters. It is then essential to incorporate a priori information about the blur classification into the restoration procedure. The ill-posed problem also means that a small perturbation of input may produce a huge unexpected disturbance in the reconstructed image.

A variety of regularization techniques have been proposed, such as half-quadratic regularization (HQR) [12], directional regularization [46] and adaptive regularization [25]. Nevertheless, Tikhonov regularization (TR) [4] is still one of the most commonly used methods to solve the ill-posed problem because of easy implementation and high efficiency. Being advantageous in implementation, it results in images with deteriorated edges, possibly globally smoothed, or even with ringing artifacts.

Due to ill posed nature of SR image reconstruction, utilizing a priori knowledge is very important. Then, Capel and Zisserman [30], and Schultz et al. [64] have applied maximum a posteriori (MAP) statistical methods to build a high-resolution image.

The method which is known as the iterative back-projection [34] iteratively verifies the value of current HR image estimate by creating from it a set of low-resolution images using the imaging model. Obviously, in each iteration the HR estimate is updated according to the difference between real and simulated low-resolution images. The method has been improved by Zomet et al. [5] by replacing means by medians of errors, and then generalized in [48]. Farsiu et al. [18] proposed a more robust super-resolution approach, in which instead of more frequent L2 norm minimization the L1 was used, as a result sharper high-resolution images have been obtained. They have shown that this change improved the Zomet et al. algorithm [5], too.

For now, temporarily, let's forget about regularizations, there are much simpler, more intuitive SR solutions. Due to essence of observation model from Figure 2, section 2, the basic scheme which relates low resolution images and high resolution one is a three stage cascade structure, Figure 3. The premise for SR usage is the existence of relative displacements between LR images, intuitively, it involves estimation of relative motion. With the motion information estimated, the HR image on non-uniformly spaced sampling grid is obtained. Then, the direct or iterative reconstruction procedure follows to calculate uniformly spaced sampling points. This very obvious framework is discussed in the next subsection.

3.1 Intuitive Super-Resolution Algorithms

In SR image reconstruction, the LR images represent different views of the same scene: they are subsampled, mutually (sub-pixel) shifted so that they contain complementary information, hence, they can be merged into a single image with higher resolution. The shift estimation (registration) methods can be divided into two groups. Methods belong to the first operate in the spatial domain space. Algorithms from the second solve the registration issue in the frequency domain space.

Assuming that the relative displacement of LR images is known, samples of "continuous" image (projected by camera lens on its sensor) in nonuniformly spaced sampling points are obtained. After LR samples mapping onto the image plane computation of samples on the dense HR grid is done. It could be done directly, although more complex iterative procedures result in smaller interpolation errors.

The last step of an intuitive algorithm is addressed to the restoration problem, formally deblurring and denoising. Restoration can be performed by applying any deconvolution method that is sufficiently robust in the presence of noise. Note, however that due to unfavorable blur and noise characteristics, degradations could limit performance of this SR algorithm step.

3.1.1 Spatial Domain Intuitive SR Methods

Differences between intuitive spatial domain SR methods mainly concern the motion estimation part. Spatial domain methods generally allow for much more flexible and realistic motion models, such as homographies. Nevertheless, displacement of the whole image or only a set of selected corresponding feature vectors is usually considered [52]. Keren et al. [41] developed an iterative planar motion estimation algorithm based on Taylor series expansions. A pyramidal scheme is implemented to increase the precision for large displacements. When the relative motion is reduced to the global one, it may be estimated via the technique based on Taylor series motion estimation, Keren et al. [41]. The reference image $f_{reference}$ can be approximated by the first order Taylor series expansion of another one f_{second}:

$$f_{reference} \approx f_{\sec ond} + x_1 + \frac{\partial}{\partial x} f_{\sec ond} + y_1 + \frac{\partial}{\partial y} f_{\sec ond} \,. \tag{3}$$

By minimizing the squared difference of both sides we obtain a sub-pixel shift estimator [41].

The first order Taylor series expression contains a gradient operator. It may be implemented by appropriate use of the Gaussian gradient. It implies that both images are convolved with the Gaussian blur function. As a positive side effect, this removes higher frequencies, which are most affected by the aliasing. The resulting formula is

$$f_{reference}^{(\sigma)} \approx f_{\sec ond}^{(\sigma)} + x_1 + \frac{\partial}{\partial x} f_{\sec ond}^{(\sigma)} + y_1 + \frac{\partial}{\partial y} f_{\sec ond}^{(\sigma)}, \text{ with } f_{reference}^{(\sigma)} = f_{reference} \otimes g^{(\sigma)}$$

$$\frac{\partial}{\partial x} f_{reference}^{(\sigma)} = f_{reference} \otimes \frac{\partial}{\partial x} g^{(\sigma)}$$

$$g^{(\sigma)} = \frac{1}{2\pi\sigma^2} e^{\frac{-x^2+y^2}{2\sigma^2}} \tag{4}$$

The estimated shifts are only accurate if they are small in comparison to the extent of the Gaussian surface. To overcome this, images can be initially corrected for coarse shifts, their sizes equal to multiplicities of pixel pitch, sub-pixel shifts remain, only. This can be done in two ways:

- Estimate integer-pixel shifts as in the CPF (Cross-correlation with phase fitting) method [30].
- Estimate integer-pixel shifts repeatedly at decreasing scales. In each iteration the unity distance (current pixel pitch) is reduced 10 times.

The Taylor series expansion based motion estimation approach can be adapted to higher order motion models [53] (i.e. affine motion model, see Figure 6).

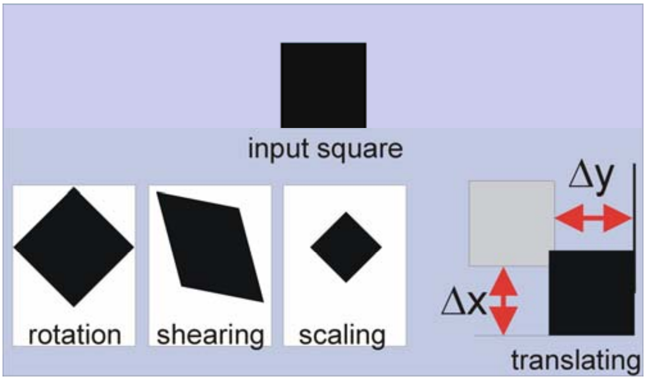


Fig. 6 Affine motion model

Another motion estimation method based on mean square error criterion applied to minimization of error between reference and arbitrarily chosen images can be found in [33].

A motion estimation error results in a degradation of the reconstructed HR image. It should be noted that in the case of total motion estimation failure a better choice is to interpolate one of low-resolution images than to apply an SR procedure with incorrect motion parameters. The next nontrivial issue is how to derive correctly information from registered images to the reconstruction of a sharp high-resolution one. Usually a nontrivial deconvolution operation is required to inverse the blur imposed by the imaging system.

Ur and Gross [29] proposed a nonuniform interpolation of an ensemble of spatially shifted LR images by implementing the Papoulis generalized multichannel sampling scheme [3]. It should be noted that the authors ignored the motion estimation problem. Komatsu et al. [44] presented an image resolution enhancement scheme by applying the Landweber algorithm [45]. Authors used the block-matching technique to measure relative shifts. While using the Landweber algorithm Shah and Zakhor were aware of the inaccuracy of the registration algorithm, and searched for a set of candidate motion estimates for each pixel instead of a single block motion vector. They used both luminance and chrominance information to estimate the motion field. Nguyen and Milanfar [56] proposed an efficient wavelet-based SR reconstruction algorithm. They exploited the “interlacing” structure of the sampling grid of combined LR images, and derived a computationally efficient wavelet interpolation for interlaced two-dimensional (2-D) data, which made real-time applications possible. However, in this approach, degradation models were rather simplistic.

Many of these techniques have been developed under assumption that the system operates in a constrained environment, for example: objects in the scene are rigid, or only simple transformations are allowed. As a consequence many of these techniques are not applicable to images containing human faces due to obvious inappropriateness of assumptions. Some of issues are [47]:

- Non-Planarity: It is usually not correct to assume that all objects are planar, as the human face is far from planar.
- Non-Rigidity: When facial expressions change, local deformations occur, hence, rigid-object models do not work.
- Occlusions: Movement of head or face will result in many partial self occlusions.
- Illumination and Reflectance Variation: Faces are subject to specular reflections that vary across the face, particularly off the cheeks and forehead.

To overcome some of these problems, in particular the face non-planarity and non-rigidity, it is possible to use optical flow techniques (see Figure 7) to define a dense flow field that describes a deformation (or other non-trivial mapping) for every pixel in the scene. By determining local flows it is possible to track motion of a complicated non-planar and non-rigid object such as a human face. The remaining two problems: occlusions and illumination variation can be addressed by applying robust estimation methods.

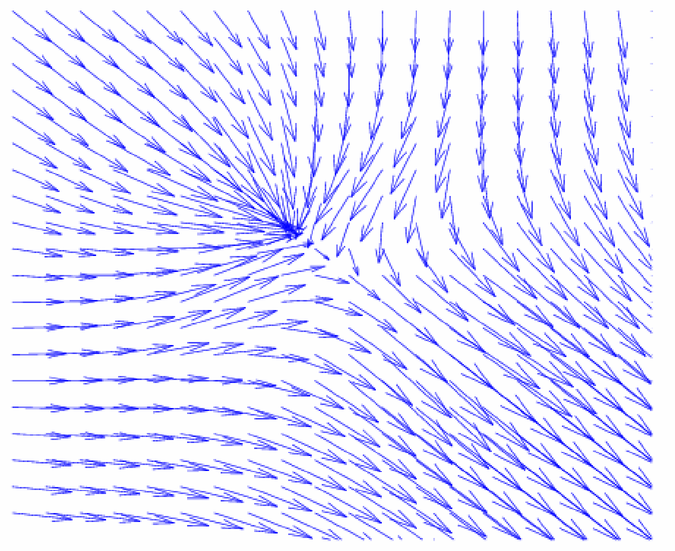


Fig. 7 The optical flow graphical interpretation. Each pixel has a velocity instead of a color, showing what direction the pixel in the underlying image is taking.

3.1.2 Frequency Domain Intuitive SR Methods

Registration can be done either in spatial or in frequency domains. The main advantage of frequency domain approach is the fact, that degradations resulting from bandwidth limitations are much easier to describe there than in the spatial domain. Unfortunately, as pointed out before, frequency domain methods are limited to global motion models. In general, they address only planar shifts, and possibly planar rotation and scaling, which transformations can be easily expressed in the Fourier domain. On the other hand, aliasing is much easier both to describe and to handle in the frequency domain than it is in space.

Let us assume that displacement between the reference image and some other one from the set of low resolution ones occurs in the image plane, and is of translational and rotational type [61], only. Additionally, static scene is considered, as in such case global motion sufficiently describes differences between LR images. Then, for such premises displacement between two images can be simply described by three parameters: horizontal and vertical translations, Δx_1, Δx_2, and planar rotation factor represented by an angle φ. In the Fourier transform domain the relation between two mutually shifted and rotated images can be expressed as follows:

$$\begin{aligned} F_{\sec ond}(\omega) &= \iint_x f_{\sec ond}(x) e^{-j2\pi\omega^T x} dx = \\ &= \iint_x f_{reference}(R(x+\Delta x)) e^{-j2\pi\omega^T x} dx = \quad , \\ &= e^{-j2\pi\omega^T \Delta x} \iint_{x'} f_{reference}(R(x')) e^{-j2\pi\omega^T x'} dx' \end{aligned} \tag{5}$$

where R means rotation operation expressed in planar coordinates:

$$R = \begin{bmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{bmatrix}, \tag{6}$$

and x and x' represent the coordinates of the reference and shifted images, respectively, $\Delta x=x'-x$. A displacement in the spatial domain does not change absolute value of a spectrum (absolute values of rotated and reference images spectra F_{second} and $F_{reference}$ do not depend on the shift Δx). Then, rotation angle φ can be computed from amplitude spectra of images:

$$\left|F_{\sec ond}(\omega)\right| = \left|F_{reference}(R\omega)\right| .$$

Rotation estimation is based on a simple premise that the un-rotated version of rotated image f_{second} has a spectrum that correlates best with that of $f_{reference}$. Unfortunately such approach needs a computation of correlation for many rotation angles, which is not computationally efficient. Transforming spectra into polar coordinates simplifies the approach in an important way. Rotation by an angle may be estimated by computing phase difference between signals. On the other hand, transformation into polar coordinates from the Cartesian ones usually involves an interpolation and is not trivial.

If the rotation has been estimated, the shift may be computed as follows:

$$\begin{aligned} F_{\sec ond}(\omega) &= \iint_x f_{\sec ond}(x) e^{-j2\pi\omega^T x} dx = \\ &= e^{-j2\pi\omega^T \Delta x} F_{reference}(\omega) \end{aligned} . \tag{7}$$

Translation values can be evaluated from $\angle \frac{F_{\sec ond}}{F_{reference}}$.

Foroosh et al. [20] proved that the power of signal obtained when phase correlating images corresponds to the polyphase transform of system impulse response. Lucchese and Cortelazzo [61] presented a rotation estimation algorithm based on the property that the magnitude of Fourier transform of an image and the reflected version of the magnitude of Fourier transform of a rotated image have a pair of orthogonal zero-crossing lines. The angle that these lines make with the axes is equal to half the rotation angle between the two images [61]. The horizontal and vertical displacements are then estimated using a standard phase correlation method. Bergen et al. developed a hierarchical approach to estimate motion in a multiresolution data structure [5].

3.2 Iterative Back-Projection Irani and Peleg's Scheme

Primarily proposed in [34], Iterative Back-Projection (IBP) is based on a similar idea to the computer-aided tomography where a 2-D object is reconstructed from its 1-D projections. The technique contains a registration stage, an iterative adjustment for displacement estimation, and a simulation of the imaging process (the blurring effect) using a point spread function, see Figure 8. The algorithm starts by guessing an initial HR image. The initial HR image can be generated e.g. by upscaling an LR one. The initial HR image is then downsampled to simulate the observed LR images which are then subtracted from the observed LR ones. If the initial HR image would be the real HR one, then the simulated and observed LR images would be identical, hence, their differences equal to zero. Therefore, the computed updates can be "back-projected" to improve the initial guess. The back-projecting process is repeated iteratively to minimize the difference between the simulated and the observed LR images, and in a consequence to produce a better HR image. Assuming that the algorithm is converging, the error energy becomes smaller. Ill-posed nature of an SR reconstruction problem cause that there is no unique solution to it. Unfortunately, the choice of back-projection filter in the IBP method is arbitrary. Moreover, compared to other approaches, such as regularization, it is more difficult here to incorporate a priori information.

Being of practical interest original Irani and Peleg's SR algorithm does not work when local motion is present. Malczewski and Stasinski generalized Irani and Peleg's Super-Resolution method [48], their extension concerned dynamic scenes, see Figure 9. The approach has been based on a premise, that LR images could be segmented into areas with coherent motion trajectory. These common motion regions were processed separately. After computing motion parameters and applying SR image restoration procedure, these image segments could be enhanced separately. Appropriately processed sub-images are combined into one

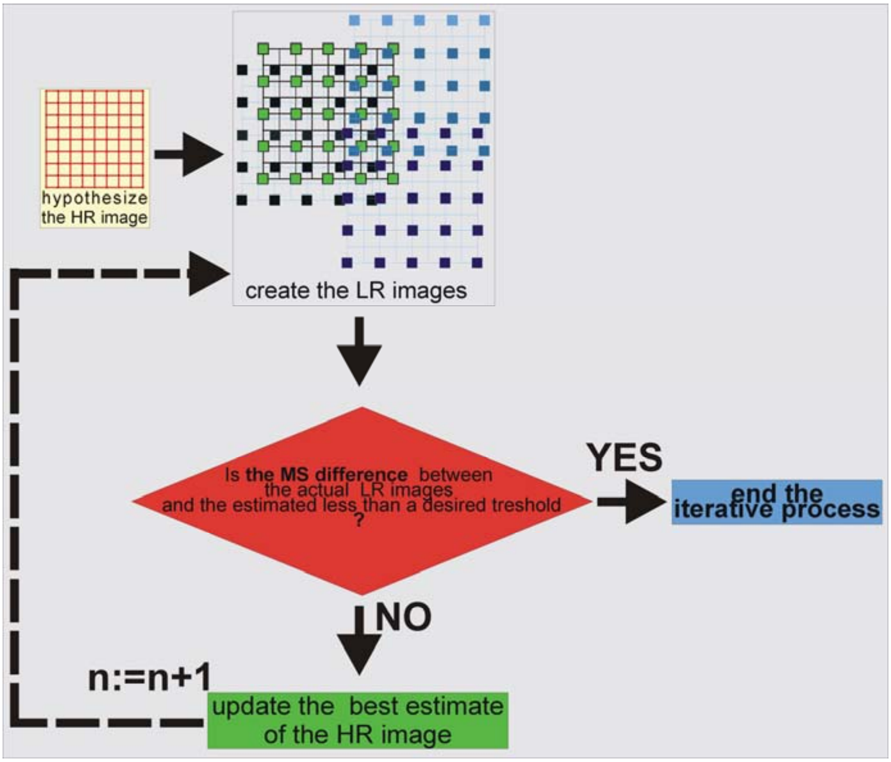

hypothesize the HR image
create the LR images
Is the MS difference between the actual LR images and the estimated less than a desired treshold ?
YES
end the iterative process
NO
n:=n+1
update the best estimate of the HR image

Fig. 8 Iterative back-projection super-resolution framework

Super-Resolution image by a modified IBP scheme, see Figure 9. The improvements resulted in lower approximation errors and higher convergence speed.

3.3 Regularized Methods

Super-resolution is often referred as the problem of estimating high resolution image details when the low resolution image sequence is given. Mathematically, such problems are usually highly ill-conditioned, what became a motivation for the use of Bayesian techniques and generic smoothness assumptions about the solution.

The regularization is often formulated in terms of a prior distribution of image values extending over the high resolution image, in which case the solution can be interpreted as a MAP (maximum a posteriori) optimization. (Prior distribution, or simply prior, is a function depicting a priori knowledge about probability distribution of estimated signal “shape”.) Baker and Kanade [3] have tried to improve the performance of super-resolution algorithms by developing domain-specific image priors, for example applicable to faces, or text, which were learned from data. The algorithm was effectively producing perceptually plausible high frequency features.

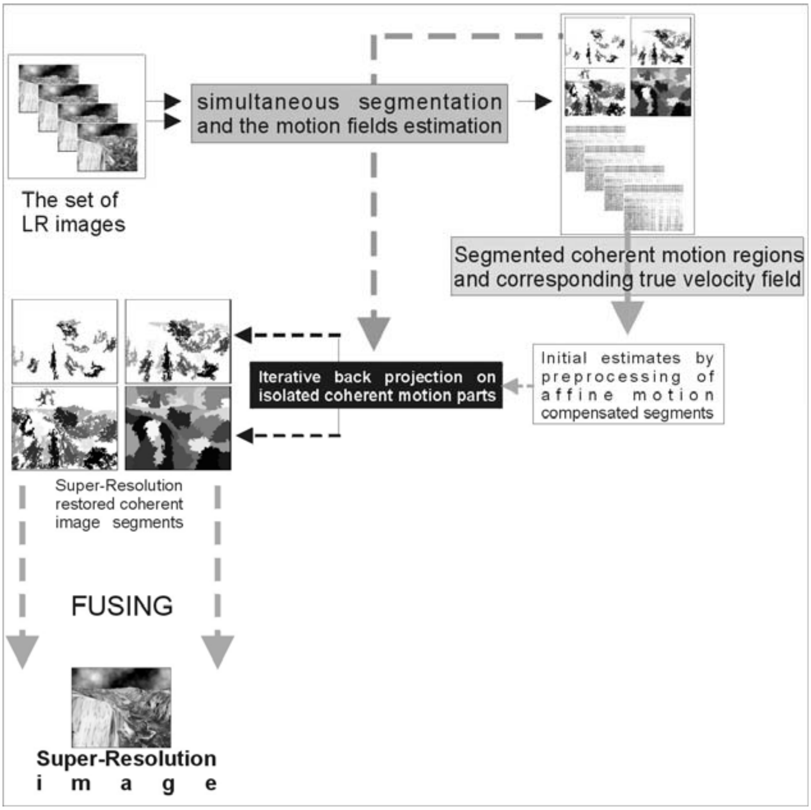

The set of LR images
simultaneous segmentation and the motion fields estimation
Segmented coherent motion regions and corresponding true velocity field
Initial estimates by preprocessing of affine motion compensated segments
Iterative back projection on isolated coherent motion parts
Super-Resolution restored coherent image segments
FUSING
Super-Resolution image

Fig. 9 Generalized iterative back-projection algorithm for super-resolution moving object extraction

In this chapter, a deterministic and stochastic regularization approaches to SR image reconstruction are presented [76]. The most frequently used are constrained least squares (CLS) and maximum a posteriori (MAP) SR image reconstruction methods.

3.3.1 Bayesian Super-Resolution

Stochastic SR image reconstruction, usually based on the Bayesian approach, provides a flexible and convenient way to model a priori knowledge concerning the solution. Bayesian estimation methods are used when the a posteriori probability density function (PDF) of the original image can be established [52, 40].

The model of generating LR images from the HR one for multi-frame super-resolution assumes existence of a known scene x (vectorized, size N×1), and vectors $\theta^{k)}$, each describing registration of the k-th LR image. These are used to generate (vectorized) M-pixel low-resolution images $y^{(k)}$ through multiplication by a system matrix $W^{(k)}$. Gaussian *i.i.d.* noise having variance $1/\beta$ (or precision β) is then added:

$$y^{(k)} = \lambda_{\alpha}^{(k)} W\left(\theta^{(k)}\right)x + \lambda_{\beta}^{(k)} + \varepsilon^{(k)} \tag{8}$$

$$\varepsilon^{(k)} \sim N\left(0, \beta^{-1} I\right)$$

Photometric parameters λ_α and λ_β provide global affine correction for the scene illumination, λ_β is simply an M×1 vector filled out with a value λ_β . Each row of $W^{(k)}$ determines a single pixel in $y^{(k)}$, the row entries are responses to low-resolution pixels, derived from (vectorized) point-spread functions (PSF) of the considered system [30, 61, 39]. The PSF is usually assumed to be an isotropic Gaussian function, though for some motion models (e.g. planar projective) this does not necessarily lead to a Gaussian shapes of responses to LR pixels. Conditional probability distribution to occur for an individual low-resolution image is then

$$p\left(y^{(k)} \mid x, \theta^{(k)}, \lambda^{(k)}\right) = \left(\frac{\beta}{2\pi}\right)^{\frac{M}{2}} \exp\left\{-\frac{\beta}{2} \| y^{(k)} - \lambda_{\alpha}^{(k)} W\left(\theta^{(k)}\right)x - \lambda_{\beta}^{(k)} \|_2^2\right\} \tag{9}$$

When the registration is not exactly known, the uncertainty can be modeled as a Gaussian perturbation around mean estimates $\overline{\theta}^{(k)}$, $\overline{\lambda}_{\alpha}^{(k)}$, $\overline{\lambda}_{\beta}^{(k)}$, having covariance C, for simplification, C is restricted to be diagonal:

$$\begin{bmatrix} \theta^{(k)} \\ \lambda_{\alpha}^{(k)} \\ \lambda_{\beta}^{(k)} \end{bmatrix} = \begin{bmatrix} \overline{\theta}^{(k)} \\ \overline{\lambda}_{\alpha}^{(k)} \\ \overline{\lambda}_{\beta}^{(k)} \end{bmatrix} + \delta^{(k)} \tag{10}$$

$$\delta^{(k)} \sim N(0, C)$$

$$p\left(\theta^{(k)}, \lambda^{(k)}\right) = \left(\frac{\left|C^{-1}\right|}{(2\pi)^n}\right)^{\frac{1}{2}} \exp\left\{-\frac{1}{2} \delta^{(k)T} C^{-1} \delta^{(k)}\right\}. \tag{11}$$

A Huber prior is assumed to characterize the directional image gradients Dx in the super-resolution image x (in the horizontal, vertical, and two diagonal directions),

$$p(x) = \frac{1}{Z_x} \exp\left\{-\frac{\nu}{2} \rho(Dx, \alpha)\right\} \tag{12}$$

$$\rho(z, \alpha) = \begin{cases} z^2 & if \quad |z| < \alpha \\ 2\alpha|z| - \alpha^2 & otherwise \end{cases} \tag{13}$$

where α is a parameter of the Huber potential function, and ν is the prior strength parameter. It belongs to a family of functions often favored over Gaussians for super-resolution image priors [30], because the Huber distribution's heavy tails mean that image edges are less severely penalized.

Regardless of what exact forms of these probability distributions are, probabilistic super-resolution algorithms can be interpreted in one of the following ways:

Search for a MAP estimate, usually using an iterative scheme, by maximizing $p\left(x \mid \left\{y^{(k)}, \theta^{(k)}, \lambda^{(k)}\right\}\right)$ with respect to *x*, where

$$p\left(x \mid \left\{y^{(k)}, \theta^{(k)}, \lambda^{(k)}\right\}\right) = \frac{p(x)\prod_{k=1}^{K} p\left(y^{(k)} \mid x, \theta^{(k)}, \lambda^{(k)}\right)}{p\left(\left\{y^{(k)}\right\} \mid \left\{\theta^{(k)}, \lambda^{(k)}\right\}\right)} \tag{14}$$

and the denominator is unknown.

Tipping and Bishop's approach takes an ML estimate of the registration by marginalizing over *x*, then calculates the super-resolution estimate as in (9). Originally Tipping and Bishop ignored the photometric model, its inclusion leads to the following cost function (to be maximized with respect to θ and λ);

$$p\left(\left\{y^{(k)}\right\} \mid \left\{\theta^{(k)}, \lambda^{(k)}\right\}\right) = \int p(x) \prod_{k=1}^{K} p\left(y^{(k)} \mid x, \theta^{(k)}, \lambda^{(k)}\right) dx\,. \tag{15}$$

Note that in Tipping and Bishop's work the analogous to [51] data likelihood expression has been used, which forced them to select a Gaussian form for *p*(*x*), instead of a more suitable image prior, in order to keep the integral tractable.

In [51] *x* is found through marginalizing over θ and λ, so that a MAP estimate of *x* can be obtained by maximizing simply $p\left(x \mid \left\{y^{(k)}\right\}\right)$. This is achieved by finding

$$p\left(x \mid \left\{y^{(k)}\right\}\right) = \frac{p(x)}{p\left(\left\{y^{(k)}\right\}\right)} \int \prod_{k=1}^{K} p\left(\theta^{(k)}, \lambda^{(k)}\right) p\left(y^{(k)} \mid x, \theta^{(k)}, \lambda^{(k)}\right) d\{\theta, \lambda\}. \tag{16}$$

Note that the integral does not involve the prior, *p*(*x*). As can be seen, apart of MAP the maximum likelihood (ML) estimation has been applied to the SR reconstruction, too. Obviously, the ML estimation is a special case of MAP estimation without a priori knowledge, nevertheless, due to the ill-posed nature of SR-type inverse problems, MAP estimation is usually preferred.

Tom and Katsaggelos [9, 76] proposed ML solution to estimate subpixel shifts, the noise variances of each image, and the HR image simultaneously. The ML estimation problem is solved by the expectation-maximization (EM) algorithm. The SR reconstruction from an LR video sequence using the MAP technique was proposed by Schultz and Stevenson [64]. They proposed a discontinuity preserving MAP reconstruction method using the Huber-Markov Gibbs prior model, resulting in a constrained optimization problem with a unique minimum. Here, they used the modified hierarchical block matching algorithm to estimate subpixel displacement vectors. They also considered independent object motion and inaccurate motion estimates that are modeled by Gaussian noise. A MAP framework for the joint estimation of image registration parameters and the HR image was presented by Hardie et al. in [28]. The registration parameters, horizontal and vertical shifts in this case, are iteratively updated along with the HR image in a cyclic optimization procedure. Cheeseman et al. applied the Bayesian estimation with a Gaussian prior model to the problem of integrating multiple satellite images observed by the Viking orbiter [76, 13].

Robustness and flexibility in modeling noise characteristics and a priori knowledge about the solution are the major advantages of the stochastic SR approach. Assuming that the noise process is white Gaussian, a MAP estimation with convex energy functions in the priors ensures the uniqueness of the solution. Therefore, efficient gradient descent methods can be used to estimate the HR image. It is also possible to estimate the motion information and the HR image simultaneously.

3.3.2 Deterministic Regularized Super-Resolution

If estimates of registration parameters are known, the observation model from Chapter 1.2 can be formed. As the super-resolution image reconstruction is an ill conditioned numerical problem, it can be solved via a deterministic regularized approach, which solves the inverse problem by using the a priori information about the solution. For example, constrained least-squares (CLS) [75] technique can be formulated as minimization of the Lagrangian [6]:

$$\sum_{n=1}^{p} \left\| y_n - W_n x \right\|^2 + \alpha \left\| Cx \right\|^2 , \tag{17}$$

where the operator C denotes a high-pass filtering. A priori knowledge about the desired solution is a smoothness constraint, as most images are naturally smooth. As a consequence the constraint minimizes high-frequency power of the reconstructed image. Additionally, α is the Lagrange multiplier, which is a regularization parameter controlling the tradeoff between data fidelity (expressed by $\left\| y_n - W_n x \right\|^2$) and smoothness of the solution (expressed by $\left\| Cx \right\|^2$).

A larger value of α leads to a smoother solution. It can be convenient when only a small number of LR images is available (the problem is underdetermined) or the quality of data is very low (i.e. due to motion estimation error and noise). Alternatively, if many LR images are available and the noise is weak, small α will lead to a sharp image. The cost functional in (17) is convex and differentiable with the use of a quadratic regularization term $\left\| Cx \right\|^2$. Therefore, we can find a unique estimate image $\hat{x}$ which minimizes the cost functional (17). A rather basic deterministic iterative technique reduces (17) to

$$\left[\sum_{n=1}^{p} W_n^T W_n + \alpha C^T C \right] \hat{x} = \sum_{n=1}^{p} W_n^T y_n$$

which leads to the following iterative procedure for finding $\hat{x}$:

$$\hat{x}^{k+1} = \hat{x}^k + \beta \left[\sum_{n=1}^{p} W_n^T \left(y_n - W_n \hat{x}^k \right) - \alpha C^T C \hat{x}^k \right]$$

where β is the convergence parameter and W_n^T describes an upsampling operator defining blur and warping.

Katsaggelos et al. [75, 9] presented a multichannel regularized SR method where the regularization functional is applied to regularization parameter calculation without any a priori knowledge. Later, Kang proposed the generalized multichannel deconvolution method including the multichannel regularized SR approach [39]. The SR reconstruction method based on minimization of a regularized cost functional was proposed by Hardie et al. [28]. Authors defined an observation model that incorporated knowledge about optical system and detector array (sensor PSF). They applied an iterative gradient-based registration algorithm and considered both gradient descent and conjugate-gradient methods to minimize the cost functional. Bose et al. [8] focused on the important role of the regularization parameter and proposed constrained least-squares SR reconstruction which generates the optimum value of regularization parameter, using the L-curve method [75].

3.3.3 Projection onto Convex Sets Approach

The Projection onto Convex Sets (POCS) method is a one more approach to incorporation of a priori knowledge about the solution into the reconstruction process. Provided the motion parameters are known, the POCS simultaneously solves restoration and interpolation of the SR problem.

Stark and Oskoui [28], [75] proposed a POCS based SR technique that takes into account both the blur introduced by a sensors, as well as effects of undersampling. Later, the idea was extended by Tekalp et al. by taking into account observation noise [7].

The POCS solution is searched for in each closed convex set C_i that is defined by a set of vectors satisfying a particular property [75]. It is found in their intersection $C_s = \cap_{i=1}^{m} C_i$, provided the intersection is nonempty. The intersection is also a convex set. The solution is usually searched for by iteratively projecting temporary solutions onto consecutive convex sets.

It is assumed that an estimate of the high resolution image at time $k = t_r$ is desired [75]. A family of closed, convex constraint sets can be defined, one for each pixel of the low-resolution image sequence

$$C_{t_r}(m_1, m_2, k) = \left\{ y(n_1, n_2, t_r) : \left| r^{(y)}(m_1, m_2, k) \right| \le \delta_0 \right\}$$

where

$$r^{(y)}(m_1, m_2, k) = g(m_1, m_2, k) - \sum_{n_1, n_2} y(m_1, m_2, t_r)h_{t_r}(n_1, n_2, m_1, m_2, k)$$

represents the residual associated with an arbitrary member, y, of the constraint set h_{t_r}, which defines combination of PSF and relative motion of object and sensor. The quantity δ_0 denotes an a priori bound reflecting the statistical confidence with which the actual image, y, is a member of the set $C_{t_r}(m_1, m_2, k)$. This constraint set is referred to as data consistency constraints. As usual in POCS techniques, the HR image estimate is calculated iteratively.

Patti et al. [2] presented a POCS SR method which takes into consideration space varying blur, nonzero aperture time, nonzero physical dimension of each individual sensor element, sensor noise, and irregular sampling patterns. Tekalp et al. extended this approach by considering multiple moving objects in a scene. They introduced concepts of validity and segmentation maps [7].

The main advantages of POCS are its simplicity, easy inclusion of a priori knowledge, and very flexible spatial domain observation model. The POCS disadvantages are non-uniqueness of solution, slow convergence, and high computational cost.

3.3.4 MAP/ML-POCS Hybrid Super-Resolution

The ML-POCS hybrid super-resolution image reconstruction approach seeks SR estimates by minimizing an ML (or MAP) cost functional while constraining the solution within certain sets. This concept had been introduced by Schultz and Stevenson in [63]. Authors proposed constrained MAP optimization, projection-based constraints were implied. In that paper, the constraint set ensured that the down-sampled version of the HR image matched the reference image from the LR sequence.

Then, Elad and Feuer [15, 75] described a very flexible hybrid SR image reconstruction algorithm which combined stochastic approach advantages and the POCS idea. Simple ML (or MAP) constraints for POCS were applied simultaneously by defining a new convex optimization problem as follows:

$$\min \varepsilon^2 = \left\{ \left[y_n - W_n\, x \right]^2 R_k^{-1} \left[y_n - W_n\, x \right] + \alpha \left[Sx \right]^T V \left[Sx \right] \right\},$$

where $\{x \in C_n, 1 \le n \le M\}$, R_k is the autocorrelation matrix of noise, S denotes the Laplacian operator, V is weighting matrix controlling smoothing strength at each pixel, and C_n represents an additional constraint. A benefit from the hybrid approach is that all a priori knowledge is combined, and this ensures a single optimal solution, which is unusual for the POCS approach.

3.3.5 Optimal and Adaptive Filtering

Inverse filtering approaches to SR reconstruction have been proposed, however these techniques are limited in terms of inclusion of a priori constraints as compared with POCS or Bayesian methods and are mentioned only for completeness. Techniques based on adaptive filtering, have also seen application in SR reconstruction [75]. These methods are based on least mean square error (LMSE) estimators, which do not include non-linear a priori constraints.

3.3.6 Tikhonov Arsenin Regularization

Due to ill-posed nature of SR reconstruction, Tikhonov-Arsenin regularized SR reconstruction methods have been examined [4]. The regularizing functionals characteristic of this approach are typically special cases of Markov Random Fields (MRF) priors in the Bayesian framework.

4 Quasi Super-Resolution Algorithms – Single Frame Image Resolution Enhancement

The majority of SR methods that have been proposed concern the problem of combining multiple low resolution images of the same scene into a high-resolution one, called multiple-frame super resolution. Only few techniques estimate a high resolution image from a single low-resolution one, with a help of one or more training images of similar, or even not-so-similar scenes. This approach is referred to as single-frame super-resolution, or quasi super-resolution. An obvious limitation of the single-frame SR approach is that it can be effective only if the input database contains a high resolution image which is "similar" to that being enhanced. This is the reason why statistical learning is preferred here. Then, many quasi SR methods, mainly in spatial domain, use some kind of knowledge classification in order to optimize their results.

Learning-based image resolution enlargement techniques have been proposed by different researchers [38, 41, 37, 24]. The idea of these methods is to use a training set of high resolution images and their low resolution counterparts to build a co-occurrence model (stored either directly as image patches, or as coefficients of some alternative representations). Then, such quasi SR algorithm learns fine details that correspond to various image regions seen at a low-resolution, and then uses the learned relationships to predict fine details in other images [38]. The methods usually add real image details, but unfortunately may also add visual artifacts. Graphical models can be applied in SR, too [38]. In graphical models statistics of natural images are used to define compatibility functions between pixels, we can use them as a form of a priori knowledge in SR methods.

Jojic [38] presented a novel method named "epitome", epitome being a miniature of an image. Epitome has significantly smaller size than the original, but contains its most important components. The approach has been generalized in [49], it has been shown that patches having different scales and perspectives can be used in HR reconstruction, too. In general, an epitome is a condensed digital representation of an ordered dataset, such as matrices representing images, audio signals, videos, or genetic sequences. The epitome has been proposed to be an efficient representation of images and video sequences. Being a condensed version of the image, the epitome exhibits potentials in many applications, such as image mosaicing, super-resolution, image compression, etc.

Some advanced enhancement methods based on smoothing and interpolation techniques for denoising, called quasi super-resolution ones, have been implemented in various image processing applications. Smoothing has been usually done by applying spatial filters such as Gaussian, Wiener, and median ones. Widely used interpolation methods, such as spline interpolation [38], may approximate a set of image pixels values quite well. They result in a better performance than simple smoothing methods, but the resulting images often have blurred edges. Image sharpening masks have been proposed to adjust the results of these techniques [38]. Their application leads to sharper images, but may introduce haloing artifacts. Note that giving quite good effects deconvolution and image sharpening methods only enhance features that are present in the unprocessed low resolution image.

This section contains a survey of single frame Super-Resolution methods proposed in recent years. Their drawbacks and difficulties linked with their implementation, especially the neighborhood issue, are pointed out.

4.1 Training Set Generation and the Neighborhood Problem

To obtain a training set we begin with a collection of high-resolution images, then we degrade them in a manner conforming with a degradation model. Usually blurring and subsampling operators are applied. All what is needed as a training set is the collection of differences between cubic-spline interpolated and the original high-resolution images [39]. Images are divided into partially overlapping image patches, see Figure 10.

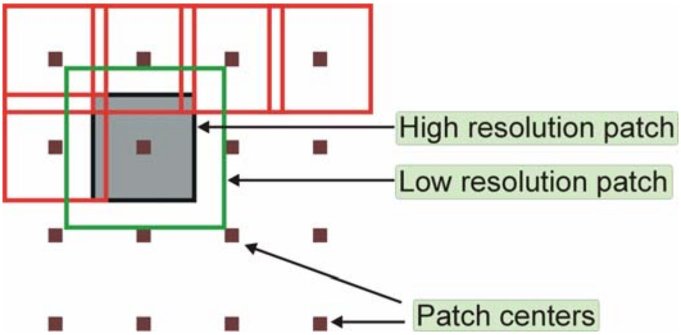


Fig. 10 Training set generation scheme. Images are divided into small patches: – 5x5 (HR), 7x7 (LR).

4.1.1 The Neighborhood Issue [39]

The single-frame super-resolution problem can be formulated as follows. Given a low-resolution image x_l as the input, the desired high-resolution image x_k is estimated with the help of a set of training low resolution images y_l^i and corresponding high-resolution images y_h^i, where *i = 1.. M, M* is the number of training image pairs.

Each low- or high-resolution image can be represented as a set of small image patches with or without overlap. A low-resolution image y_l^i and a corresponding high-resolution one y_h^i have the same number of patches. We denote low-resolution patches in x_l and appropriate high-resolution ones in x_k $\{x_l^p\}_{p=1}^{N_x}$ and $\{x_h^p\}_{p=1}^{N_x}$ respectively, where N_x is the number of patches in the input image. Low- and high-resolution patches in training image pairs are represented as $\{y_l^q\}_{q=1}^{N_y}$ and $\{y_h^q\}_{q=1}^{N_y}$

respectively, N_y is the total number of patches in training images. Obviously, N_x and N_y depend on a patch size and the degree of overlap between adjacent patches. In super-resolution for a low-resolution patch x_l^p from the input image x_l, we estimate x_h^p with the help of training image patches $\left\{y_l^q\right\}_{q=1}^{N_y}$ and $\left\{y_h^q\right\}_{q=1}^{N_y}$. The intuitive way is to find y_h^q which is the most similar to x_h^p from $\left\{y_h^q\right\}_{q=1}^{N_y}$, and let $x_h^p = y_h^q$, or to find several high-resolution patches in $\left\{y_h^q\right\}_{q=1}^{N_y}$ that are closest to x_h^p in high-resolution patch space, called nearest neighbors of x_h^p, and estimate x_h^p from them. Unfortunately, the nearest neighbors of x_h^p in $\left\{y_h^q\right\}_{q=1}^{N_y}$ cannot be easily determined, as we don't know x_h^p, hence, we cannot use any kind of distance metric to determine the nearest neighbors. We name this problem the neighborhood issue. Chang et al. [41] tried to address this issue by applying the new manifold learning method: locally linear embedding (LLE).

4.1.2 Locally Linear Embedding

Locally linear embedding (LLE) is a promising manifold learning method that has aroused a lot of interest in machine learning. It consists in computing low-dimensional neighborhood-preserving embeddings of high-dimensional inputs, while allowing reconstruction of global nonlinear structure from locally linear fits [39]. High-resolution image patches can represent points in a high dimensional data space, and the corresponding low resolution image patches are points in a low dimensional data space. In such case LLE is used to estimate low-resolution patches, given the corresponding high-resolution patches, hence, LLE can be thought as a reversed procedure of solving the super-resolution problem. In LLE, nearest neighbors of data point in high dimensional space will still be nearest neighbors in low dimensional space, the neighborhood will be preserved in both high and low dimensional spaces. Chang et al.'s work [41] is based on the assumption that the high dimensional data points whose corresponding low dimensional data points are neighbors will still be neighbors in the high dimensional space. However in super-resolution, this is not true. If some low-resolution image patches are neighbors, their corresponding high-resolution patches are not, and the opposite is also true.

4.2 Quasi SR Methods

If local image information alone were sufficient to predict the missing high-resolution details, we would be able to use training set patches themselves for super-resolution. For a given input image to be enlarged we would apply some preprocessing steps, break the image into patches, and simply look-up for the missing high resolution details [39]. Unfortunately, that approach doesn't work. Assume that for a given low-resolution input patch we search a training database of patches to find a small set of closest matches to it. Each of them looks fairly similar to the input patch, but the corresponding high-resolution items usually look fairly different one from the other. This means that local patch information alone is insufficient for super resolution, and that spatial neighborhood effects should be taken into account.

Two different approaches to exploit neighborhood relationships in super-resolution algorithms have been explored. In the first a Markov network to probabilistically model relationships between high- and low-resolution patches, and between neighboring high-resolution patches, has been used. An iterative usually quickly converging algorithm has been applied. The second approach has been a one-pass algorithm that used the same local relationship information as the Markov network, and has been able to approximate fast its output.

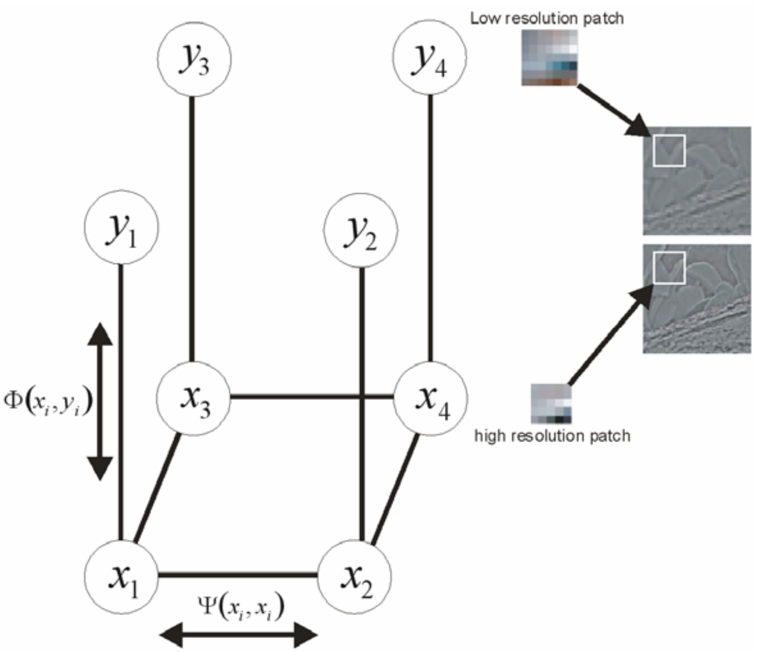


Fig. 11 Markov network model for the superresolution problem. Low-resolution patches at each node y_i are the observed input. The high resolution patch at each node x_i is the quantity to be estimated.

4.2.1 Markov Network Based Quasi SR Approach

In this section spatial relationships between image patches are modeled using a Markov network, having many well-known uses in image processing [39]. In the Figure 11 circles represent network nodes, and lines indicate statistical dependencies between nodes. Let low-resolution image patches be observation nodes *y*. Freeman et al. selected 16 or so best matching patch examples to each input patch as states of hidden nodes x to be estimated. The probability of the association of a given high-resolution patch with a network node is:

$$P(x/y)=\frac{1}{Z}\prod_{(ij)}\psi_{ij}(x_i,x_j)\prod_{i}\phi_i(x_i,y_i)$$

where compatibility matrices ψ relate possible states of each pair of neighboring hidden nodes, and vectors ϕ relate each observation with the underlying hidden state, Z is a normalization constant, and the first product is over all neighboring pairs of nodes, y_i and x_i are the observed low-resolution and estimated high-resolution patches at node i, respectively.

To specify the Markov network's $\psi_{ij}(x_i,x_j)$ functions, authors [39] applied a simple trick. They form input image nodes in such a way that high-resolution patches overlap by one or more pixels. In the overlap region the pixel values of compatible neighboring patches should agree. Distances $d_{ij}(x_i,x_j)$ are measured, being sums of squared differences between patch candidates x_i and x_j in their overlap regions at nodes i and j. The compatibility matrix between nodes i and j is then

$$\psi_{ij}(x_i,x_j)=\exp\left(-\frac{d_{ij}(x_i,x_j)}{2\sigma^2}\right)$$

where σ is a noise parameter. They use a similar quadratic penalty on differences between the observed low resolution image patch y_i, and the candidate low-resolution patch found from the training set x_i, to specify the Markov network compatibility function $\varphi_{i_i}(x_i,y_i)$.

The optimal high-resolution patches at each node are collections that maximize Markov network MAP probability. Finding the exact solution can be computationally intractable, but good results have been achieved by calculating an approximate solution obtained from a fast, iterative algorithm called belief propagation. The belief-propagation algorithm updates "messages" $m_{i_i j}$ from a node i to a node j, being vectors having dimensions equal to that of states estimated at node j. Let us use notation $m_{ij}(x_j)$ to indicate a component of vector $m_{i_i j}$ corresponding to the patch candidate x_j, then the rule for updating the message from node i to node j becomes

$$m_{ij}(x_j)=\sum_{x_j}\phi_{ij}(x_i,x_j)\prod_{k\neq j}m_{ki}(x_i)\phi_i(x_i,y_i)$$

Yedidia, Freeman, and Weiss [39] have shown a connection between this estimate and an approximation used in physics by Bethe. Freeman, Pasztor, and Carmichael provided details of the belief-propagation implementation.

4.2.2 One Pass Algorithm

The fact that belief propagation converged to a solution of the Markov network very quickly prompted authors of the method to consider a much simpler approach.

They developed a one-pass algorithm that gives results that are nearly as good as those of the iterative algorithm based on the Markov network [39]. In the one-pass algorithm only high-resolution patch compatibilities for neighboring high-resolution patches that are already selected are computed, typically the patches above and to the left in raster-scan order processing. The simplification circumvents various steps in setting up and solving the Markov random field (MRF) of the previous approach: finding the candidate set at each node, finding the compatibility matrices between all pairs of nodes, and using the iterative belief-propagation algorithm.

In the simplest terms, one-pass super-resolution generates the missing high-frequency content of a zoomed image as a sequence of predictions from local image information. The proposed approach divides the input image into low-frequency patches that are traversed in raster-scan order, see Figure 12. At each step, a high-frequency patch is selected by the nearest neighbor search from the training set based on the local low-frequency details and adjacent, previously determined high-frequency patches.

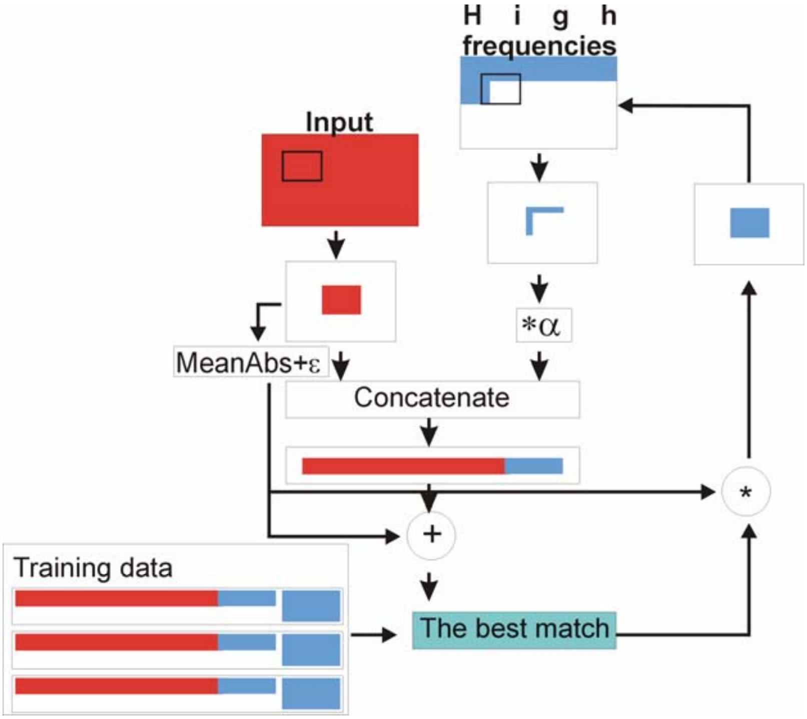


Fig. 12 Block diagram showing raster-order per-patch processing. At each step, local low- and high-frequency details (in red and blue, respectively) to search the training set for a new high frequency patch are used.

5 Application of Super-Resolution Algorithms to Real Data

5.1 Intuitive Super-Resolution

In the experiment cropped test images ‘Fun’, ‘Boats’ and ‘Cameraman’ have been used, Figure 13. Every image has been sub-sampled by a factor of 2, rotated and shifted. Four prepared in this way images have been combined into a super-resolution one, having two times higher sampling density. Four algorithms has been compared, see Figure 13.

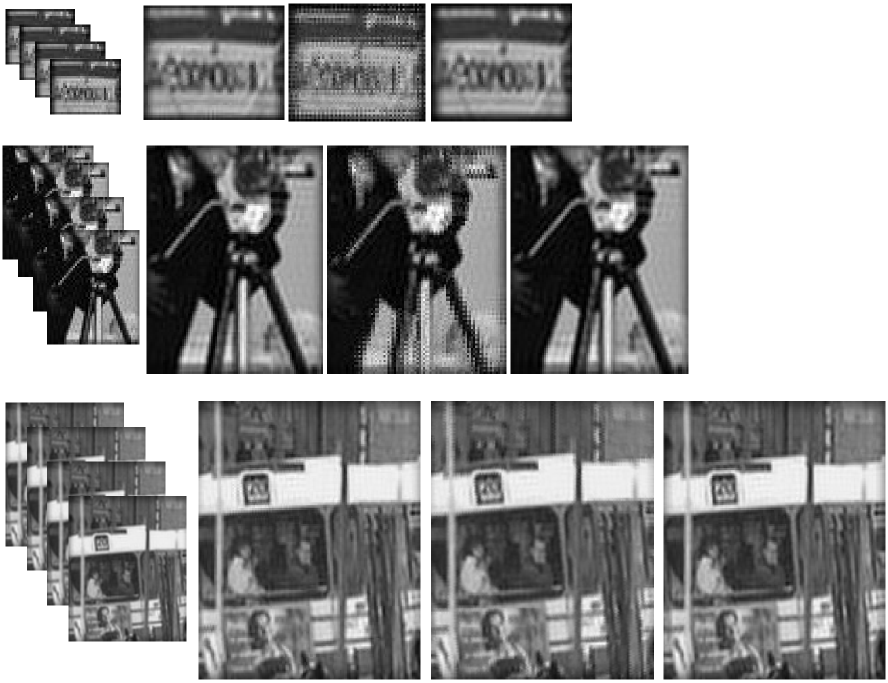


Fig. 13 Images before and after SR restoration: for each row, from the left to the right side: the sequence of low resolution images, MSE motion estimation, frequency domain algorithm [61], two-domain algorithm.

Note that a frequency-domain image parameter analysis is not accurate when a spectrum is degraded by aliasing. In such situations it is better to use Taylor series expansion. That is why the fourth technique resulted in improved registration and reconstruction stages and overcome majority of limitations of other methods, which is illustrated in Figure 13. The improvement is due to the two-domain frequency-space motion estimator and due to replacement of straight bicubic interpolation by iterative POCS based reconstruction algorithm. Indeed, the motion estimation problem is much more important than those linked with the reconstruction step.

5.2 Regularized Methods

As it has been shown in [40], the Bayesian SR approach allows estimation of the unknown point spread function, while being tractable due to the introduction of a Gaussian prior over images. Results indicate a significant improvement over techniques based on MAP (maximum a-posteriori) point optimization of the high resolution image and associated registration parameters.

The main advantage of the POCS-SR [29] is its simplicity, moreover, it implements a powerful spatial domain observation model. The approach allows for a convenient inclusion of a priori information. The algorithm disadvantages are non-uniqueness of solution, slow convergence, and high computational cost.

5.3 Back-Projection Algorithm and Optical Flow Super-Resolution

Irani and Peleg proposed an SR algorithm [34] that minimizes a cost function in an error back-projection scheme inspired by computer aided tomography. The residual images (obtained by subtraction of simulated images from the observed LR ones) are convolved with a back-projection function, averaged and fed back into the computational loop, see Figure 8. The robustness of the method depends on the choice of back-projection function (BPF). Since each update to the super-resolution estimate is simply a linear combination of BPF kernels, if the BPF is smooth, the resulting estimate tends to be smooth, too. The algorithm is unable to introduce high-frequency noise components that tend to dominate the unconstrained ML estimates.

Theoretically, super-resolution optical flow method can be used to enhance resolution of any video sequence, however, in practice its performance relies on robustness of optical flow determination [51]. Every frame of tested video sequence was downgraded. Images were spatially down-sampled 4 times. Only 20 frames have been processed. The number of frames has been chosen to prevent occluding artifacts. What should be noted, temporal resolution have not been limited and camera was running with “natural shaking” of hand-recording. The prepared in this way video has been processed by a generalized Iterative Back-Projection algorithm from [35].

It can be seen that the super-resolved single frame is very similar to a one from the input sequence, see Figure 14. The bilinear interpolated images suffer from blur and details are difficult to see.

An example of back-projection method application for improving performance of Optical Recognition System can be found in [50], modified Irani and Peleg's algorithm has been used. Namely, the technique has been optimized for enhancing low-resolution text images from handheld devices, see Figure 15. Modification of initial guess estimation has resulted in improvement of accuracy and convergence speed-up.

5.4 Quasi Super Resolution

The performance of improved epitome based algorithm on appropriately processed Lena test image is evaluated. The image has been downsampled and divided

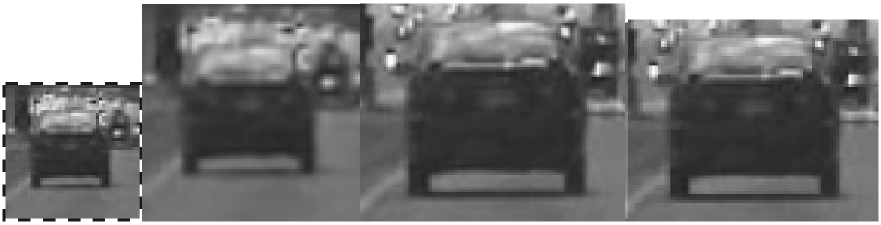
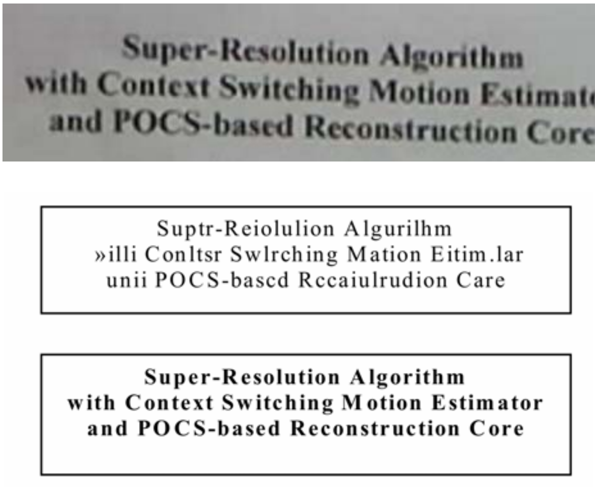


Fig. 14 From left to right: reference frame of the LR sequence, bilinear interpolated reference frame, generalized IBP scheme output image after 15 iterations, single frame taken from the test sequence, [48]



Super-Resolution Algorithm
with Context Switching Motion Estimator
and POCS-based Reconstruction Core

Suptr-Reiolulion Algurilhm
»illi Conltsr Swlrching Mation Eitim.lar
unii POCS-based Rccaiulrudion Care

Super-Resolution Algorithm
with Context Switching Motion Estimator
and POCS-based Reconstruction Core

Fig. 15 Top – example of LR image. Middle – poor result of directly applied OCR. Down - result of OCR reconstruction of auper-resolved image, [50]

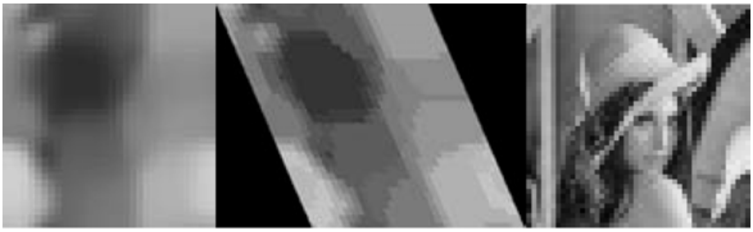


Fig. 16 Input data. Two image patches and miniature test image [49].

into 70 overlapping image regions, here called “patches”. Additionally, the training set has been also transformed using the eight parameter projection motion model. Besides, patches did not represent the same zoom level. In this way natural

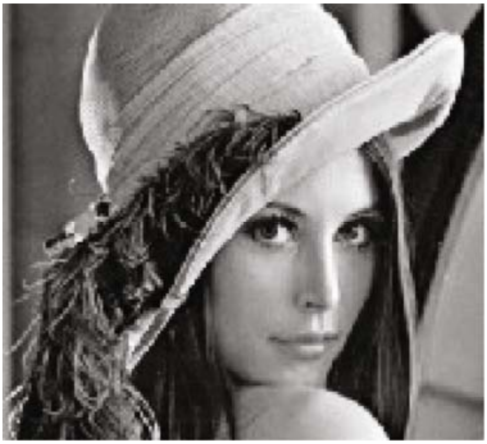


Fig. 17 Image registration and reconstruction results. Registration by Mann's algorithm is very accurate. Down-sampled areas have been eliminated (whole image plane has been overlapped) [16].

photographer motion has been simulated. Mini-scale Lena image and the set of patches have been used, Figure 16. Result can be seen in Figure 17.

An interesting example of face image enhancement has been shown in [21]. Namely, the training set for estimation consisted of different child and women photographs, while the reconstructed image was that of a man having a beard. Despite this, the resulting zooms were significantly sharper than those from cubic-spline interpolation, sharp edges and image details were preserved.

6 Conclusion

In the chapter an overview of super-resolution techniques is presented. The advantages and disadvantages of majority of super-resolution algorithms are pointed out. The most important SR methods are multi-frame ones, discussed in section 3, nevertheless, interesting results have been obtained for single frame based quasi SR techniques, too, section 4. In this survey multi-frame methods are divided into three major groups: intuitive, i.e. algorithms following the direct reverse of the observation model presented in section 2, back-projection, and regularized. Then, regularized techniques can be split into MAP/LM, least-squares and POCS based ones. The single-frame methods described in this survey are linked with application of image training sets to image enhancement. The techniques can be used both to SR reconstruction, as well as image mosaicing. In the last section experimental results for some SR algorithms are presented. The algorithms have been chosen from all main groups: intuitive, back-projection, optical flow, and single-frame SR. The examples illustrate typical visual effects obtained through SR reconstruction.

In Table 1 a quick overview of the most important multi-frame algorithms can be found. Due to wide variety of data models used in SR methods (i.e. different

Table 1 Properties of super-resolution algorithms

	Performance Measurements [73]													Source Type	
	Subj.	SNR	PSNR	MSE	Other	Motion models	Degradation model	Noise model	Computation req.	A-priori info	Regularization	Extensibility	Applicability	Still	Video
Nonuniform Sampling Based															
Komatsu [44]	√	√				global	limited, LSI	limited, SI	high	limited	good	good	limited	√	
Frequency Domain															
Kim and Bose [8]	√					global	limited, LSI	limited, SI	low	limited	limited	poor	limited	√	
Kim and Su [43]	√	√				global	limited, LSI	limited, SI	low	limited	limited	poor	limited	√	
Rhee and Kang [72]	√		√			global	limited, LSI	limited, SI	low	limited	limited	poor	limited	√	
Deterministic Regularization															
Hong, Kang, and Katsaggelos [31]	√		√				LSI or LSV		high	limited	very good	limited	medium		√
Alam et al. [2]	√						LSI or LSV		high	limited	very good	limited	medium		√
Bose and Koo [7]	√		√				LSI or LSV		high	limited	very good	limited	medium	√	
Stochastic Regularization															
Tom and Katsaggelos [9]	√						LSI or LSV		high	Prior PDF Easy to incorporate No hard constraints	excellent	limited	medium	√	
Schultz and Stevenson [63]	√	√					LSI or LSV		high	Prior PDF Easy to incorporate No hard constraints	excellent	limited	medium	√	
Hardie, Barnard, and Armstrong [28]					MAE		LSI or LSV		high	Prior PDF Easy to incorporate No hard constraints	excellent	limited	medium	√	
Cheeseman et al. [13]	√						LSI or LSV		high	Prior PDF Easy to incorporate No hard constraints	excellent	limited	medium	√	

Table 1 (*continued*)

	Performance Measurements													Source Type	
	Subj.	SNR	PSNR	MSE	Other	Motion models	Degradation model	Noise model	Computation req.	A-priori info	Regularization	Extensibility	Applicability	Still	Video
Projection onto Convex Sets															
Tekalp, Ozkan, and Sezan [7]	√						LSI or LSV		high	Convex Sets Easy to incorporate	very good	limited	medium	√	
Patti, Sezan, and Tekalp [2]	√						LSI or LSV		high	Convex Sets Easy to incorporate	very good	limited	medium		√
Eren, Sezan, Tekalp [16]	√						LSI or LSV		high	Convex Sets Easy to incorporate	very good	limited	medium		√
Patti, and Altunbasak [1]	√			√			LSI or LSV		high	Convex Sets Easy to incorporate	very good	limited	medium	√	
POCS & Maximum a posteriori hybrid															
Elad and Feuer [15]	√			√			LSI or LSV		high	Convex sets & prior PDF	very good	limited	medium	√	
Optical Flow															
Baker and Kanade [4]	√					optical flow	LSI or LSV		very high	limited	limited	good	excellent	√	
Zhao and Sawhney [81]	√					optical flow	LSI or LSV		very high	limited	limited	good	excellent	√	
Malczewski and Stasinski [5]	√					optical flow	LSI or LSV		very high	limited	limited	good	excellent	√	√
Other Methods															
Irani and Peleg [37]	√					global translation	LSI or LSV		high	limited	good	excellent	medium	√	
Baker and Kanade [19]					RMS		LSI or LSV		very high	limited	limited	limited	medium	√	

motion models, assumptions concerning image distortions, etc.) it is hard to formulate an unbiased suggestion what SR methods are appropriate for a given task. This is a reason why so many authors provide only subjective tests of their techniques consisting in comparison of HR images to upsampled and interpolated LR ones. The next group of features is linked with assumed observation model (motion, degradation, noise), note, however that in many cases such assumptions are not done. Then, important factors determining method usability are provided: computational requirements, possibility of including a priori information, applying regularization, and generalizing an approach, method applicability, finally, what type of source have been considered in the technique description – still image, or video. Hopefully, this table would be used as a guide when searching for an appropriate method for solving a super-resolution problem.

References

[1] Alakuijala, J., Laitinen, J., Sallinen, S., Helminen, H.: New Efficient Image Segmentation Algorithm: Competitive Region Growing of Initial Regions, 17 IEEE EMBS, Montreal, Canada (1995)

[2] Alam, M.S., Bognar, J.G., Hardie, R.C., Yasuda, B.J.: Infrared image registration and highresolution reconstruction using multiple translationally shifted aliased video frames. IEEE Trans. Instrum. Meas. 49, 915–923 (2000)

[3] Baker, S., Kanade, T.: Limits on super-resolution and how to break them. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(9), 1167–1183 (2002)

[4] Baker, S., Kanade, T.: Super resolution optical flow,Tech. Rep. CMU-RI-TR-99-36, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA (October 1999)

[5] Bergen, J.R., Anandan, P., Hanna, K.J., Hingorani, R.: Hierarchical model-based motion estimation. In: Sandini, G. (ed.) ECCV 1992. LNCS, vol. 588, pp. 237–252. Springer, Heidelberg (1992)

[6] Borman, S., Stevenson, R.: Super-resolution fromimage sequences - A review. In: Proceedings of the1998 Midwest Symposium on Circuits and Systems, Notre Dame, IN, USA, August 1998, pp. 374–378 (1998)

[7] Bose, N.K., Lertrattanapanich, S., Koo, J.: Advances in superresolution using l-curve. In: Proc. Int. Symp. Circuits and Systems, vol. 2, pp. 433–436 (2001)

[8] Bose, N.K., Kim, H.C., Valenzuela, H.M.: Recursive implementation of total least squares algorithm for image reconstruction from noisy, undersampled multiframes. In: Proc. IEEE Conf. Acoustics, Speech and Signal Processing, Minneapolis, MN, April 1993, vol. 5, pp. 269–272 (1993)

[9] Capel, D.: Image Mosaicing and Super-resolution (Distinguished Dissertations). Springer, Heidelberg (2004)

[10] Capel, D., Zisserman, A.: Computer vision applied to super-resolution. IEEE Signal Processing Magazine 20(3), 75–86 (2003)

[11] Chang, H., Yeung, D.Y., Xiong, Y.: Super- Resolution Through Neighbor Embedding. In: Proc. of IEEE Conf. CVPR (2004)

[12] Charbonnier, P., Blanc-F´eraud, L., Aubert, G., Barlaud, M.: Two deterministic half-quadratic regularization algorithms for computed imaging. In: Proceedings of IEEE ICIP, vol. 2, pp. 168–172 (1994)

[13] Cheeseman, P., Kanefsky, B., Kraft, R., Stutz, J., Hanson, R.: Super-resolved surface reconstruction from multiple images, NASA Ames Research Center, Moffett Field, CA, Tech. Rep. FIA-94-12 (December 1994)

[14] Dalley, G., Freeman, B., Marks, J.: Single-frame text super-resolution: a bayesian approach. In: ICIP 2004, pp. V3295–V3298 (2004)

[15] Elad, M., Feuer, A.: Restoration of a single super-resolution image from several blurred, noisy, and undersampled measured images. IEEE Transactions on Image Processing 6(12), 1646–1658 (1997)

[16] Eren, P.E., Sezan, M.I., Tekalp, A.M.: Robust, object-based high-resolution image reconstruction from low-resolution video. IEEE Trans. Image Processing 6, 1446–1451 (1997)

[17] Farneback, G.: Very High Accuracy Velocity Estimation using Orientation Tensors, Parametric Motion, and Simultaneous Segmentation of the Motion Field. In: Proc. Eighth ICCV, Vancouver, Canada, July 2001, vol. 1, pp. 171–177 (2001)

[18] Farsiu, S., Robinson, M.D., Elad, M., Milanfar, P.: Fast and robustmultiframe super-resolution. IEEE Transactions on Image Processing 13(10), 1327–1344 (2004)

[19] Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM 24(6), 381–395 (1981)

[20] Foroosh, H., Zerubia, J.B., Berthod, M.: Extension of phase correlation to subpixel registration. IEEE Transactions on Image Processing 11(3), 188–200 (2002)

[21] Freeman, W.T., Jones, T.R., Pasztor, E.C.: Example based super resolution. IEEE Computer Graphics and Applications (2002)

[22] Freeman, W.T.: Median filter for reconstructing missing color samples U.S. Patent No. 5,373,322 (1988)

[23] Freeman, W.T., Pasztor, E.C., Carmichael, O.T.: Learning low-level vision. International Journal of Computer Vision 40(1), 25–47 (2000)

[24] Freeman, W.T.: Example-based Super-resolution. IEEE Computer Graphics & Applications (2001)

[25] Galatsanos, N.P., Katsaggelos, A.K.: Methods for choosing the regularization parameter and estimating the noise variance in image restoration and their relation. IEEE Transactions on Image Processing 1, 322–336 (1992)

[26] Gluckman, J.: Gradient field distributions for the registration of images. In: Proceedings of IEEE International Conference on Image Processing (ICIP 2003), Barcelona, Spain, September 2003, vol. 3, pp. 691–694 (2003)

[27] Greenspan, H., Anderson, C., Akber, S.: Image enhancement by nonlinear extrapolation in frequency space. IEEE Trans. on Image Processing 9(6) (2000)

[28] Hardie, R.C., Barnard, K.J., Armstrong, E.A.: Joint map registration and high-resolution image estimation using a sequence of undersampled images. IEEE Transactions on Image Processing 6(12), 1621–1633 (1997)

[29] He, Q., Schultz, R.R.: Efficient Super-Resolution Image Reconstruction Applied to Surveillance Video Captured by Small Unmanned Aircraft Systems. In: Proceedings of the 2008 SPIE Defense and Security Symposium (Signal Processing, Sensor Fusion, and Target Recognition XVII), Orlando, Florida, March 16-20 (2008)

[30] Hendriks, C.L., van Vliet, L.: Improving resolution to reduce aliasing in an undersampled image sequence. In: Proceedings SPIE Electronic Imaging 2000 Conference San Jose, January 2000, vol. 3965, pp. 214–222 (2000)

[31] Hong, M.C., Kang, M.G., Katsaggelos, A.: An iterative weighted regularized algorithm for improving the resolution of video sequences. In: Proc. Int. Conf. Image Processing, vol. 2, pp. 474–477 (1997)

[32] Hou, H.H., Andrews, H.C.: Cubic splines for image interpolation and digital filtering. IEEE Trans. Acoust. Speech Signal Processing, ASSP 26(6), 508–517 (1978)

[33] Irani, M., Anandan, P.: About direct methods. In: Triggs, B., Zisserman, A., Szeliski, R. (eds.) ICCV-WS 1999. LNCS, vol. 1883, pp. 267–277. Springer, Heidelberg (2000)

[34] Irani, M., Peleg, S.: Super resolution from image sequences. In: ICPR, June 1990, vol. 2, pp. 115–120 (1990)

[35] Irani, M., Rousso, B., Peleg, S.: Computing occluding and transparent motions. International Journal of Computer Vision 12(1), 5–16 (1994)

[36] Jahanbin, S., Naething, R.: Super-resolution Image Reconstuction Performance (2005)

[37] Jia, K., Gong, S.: Hallucinating multiple occluded face images of different resolutions. Pattern Recognition Letters 27(15), 1768–1775 (2006)

[38] Jojic, N., Frey, J.J., Kannan, A.: Epitomic analysis of appearance and shapeIn. In: Proc. of ICCV (2003)

[39] Kang, M.G.: Generalized multichannel image deconvolution approach and its applications. Opt. Eng. 37(11), 2953–2964 (1998)

[40] Katsaggelos, A.K. (ed.): Digital Image Restoration, vol. 23. Springer, Heidelberg (1991)

[41] Keren, D., Peleg, S., Brada, R.: Image sequence enhancement using sub-pixel displacements. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 1988), Ann Arbor, Mich, USA, June 1988, pp. 742–746 (1988)

[42] Keys, R.G.: Cubic convolution interpolation for digital image processing. IEEE Transactions on Acoustics, Speech, and Signal Processing 29(6), 1153–1160 (1981)

[43] Kim, S.P., Su, W.Y.: Recursive high-resolution reconstruction of blurred multiframe images. IEEE Trans. Image Processing 2, 534–539 (1993)

[44] Komatsu, T., Igarashi, T., Aizawa, K., Saito, T.: Very high resolution imaging scheme with multiple different-aperture cameras. Sinal Processing: Image Commun. 5, 511–526 (1993)

[45] Landweber, L.: An iteration formula for Fredholm integral equations of the first kind. Amer. J. Math. 73, 615–624 (1951)

[46] Lee, S.H., Cho, N.I., Park, J.I.: Directional regularisation for constrained iterative image restoration. Electronics Letters 39(23), 1642–1643 (2003)

[47] Lin, F., Fookes, C., Chandran, V., Sridharan, S.: Investigation into Optical Flow Super-Resolution for Surveillance Applications. In: APRS Workshop on Digital Image Computing, Brisbane, Australia, February 21, pp. 73–78 (2005)

[48] Malczewski, K., Stasinski, R.: Generalized Iterative Back- Projection Algorithm for Super- Resolution Moving and Static Object Extraction. In: Proc. IWSSIP 2006 (CD) (2006)

[49] Malczewski, K., Stasinski, R.: High resolution image reconstruction using multiscale projected image patches integration. In: Proc. of EUSIPCO 2007 (2007)

[50] Malczewski, K., Stasinski, R.: Optical Character Recognition of Low Resolution Text Sequences From Hand-Held Device Supported by Super-Resolution. In: Proc. Multimedia Signal Processing and Communications, 48th International Symposium ELMAR-2006, Zadar, Croatia (2006)

[51] Mann, S., Picard, R.W.: Video Orbits of the Projective Group: A New Perspective on Image Mosaicing. MIT Media Lab Perceptual Computing Section Technical Report No. 338, Cambridge, MA (1995)

[52] Marshall, F., Tappen, F., Russell, B.C., Freeman, W.T.: Efficient Graphical Models for Processing Images. In: Proc. of IEEE Conf. CVPR (2004)

[53] Matthies, L.H., Szeliski, R., Kanade, T.: Kalman Filter-Based Algorithms for Estimating Depth from Image Sequences. Int. J. Computer Vision 3, 209–236 (1989)

[54] Morse, B., Schwartzwald, D.: Image magnification using level set reconstruction. In: Proc. International Conf. Computer Vision (ICCV), pp. 333–341 (2001)

[55] Nguyen, N., Milanfar, P., Golub, G.: Efficient generalized cross-validation with applications to parametric image restoration and resolution enhancement. IEEE Trans. Image Processing 10, 1299–1308 (2001)

[56] Ni, K., Kumar, S., Vasconcelos, N., Nguyen, T.Q.: Single Image Superresolution Based on Support Vector Regression. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-2006, Toulouse, France, May 2006, vol. 2, pp. 601–604.

[57] Papoulis, A.: Generalized sampling expansion. IEEE Transactions on Circuits Systems 24(11), 652–654 (1977)

[58] Park, S.M., Park, M.K., Kang, M.G.: Super-Resolution Image Construction: A Technical Overview. IEEE signal processing magazine (2003)

[59] Patti, A.J., Sezan, M.I., Tekalp, A.M.: Super-resolution video reconstruction with arbitrary sampling lattices and nonzero aperture time. IEEE Transactions on Image Processing 6(8), 1064–1076 (1997)

[60] PattiA, J., Altunbasak, Y.: Artifact reduction for set theoretic super resolution image reconstruction with edge adaptive constraints and higher-order interpolants. IEEE Trans. Image Processing 10, 179–186 (2001)

[61] Pickup, L.C., Capel, D.P., Roberts, S.J., Zisserman, A.: Bayesian image super-resolution, continued. In: Advances in Neural Information Processing Systems, Cambridge, Mass, USA, December 2006, vol. 19, pp. 1089–1096 (2006)

[62] Pickup, L.C., Roberts, S.J., Zisserman, A.: Optimizing and learning for super-resolution. In: Proceedings of the British Machine Vision Conference (2006)

[63] Rhee, S.H., Kang, M.G.: Discrete cosine transform based regularized high-resolution image reconstruction algorithm. Opt. Eng. 38, 1348–1356 (1999)

[64] Robinson, D., Milanfar, P.: Fundamental performance limits in image registration. IEEE Transactions on Image Processing 13(9), 1185–1199 (2004)

[65] Sandwell, D.T.: Biharmonic spline interpolation of GEOS-3 and SEASAT altimeter data. Geophys. Res. Lett. 14, 139–142 (1987)

[66] Schultz, R.R., Meng, L., Stevenson, R.L.: Subpixel motion estimation for super-resolution image sequence enhancement. Journal of Visual Communication and Image Representation 9(1), 38–50 (1998)

[67] Stark, H., Oskoui, P.: High resolution image recovery from image-plane arrays, using convex projections. J. Opt. Soc. Am. A 6, 1715–1726 (1989)

[68] Su, K., Tian, Q., Xue, Q., Sebe, N., Ma, J.: Neighborhood issue in single-frame image super-resolution. In: ICME 2005, pp. 1122–1125.

[69] Sun, J., Zheng, N.N., Tao, H., Shum, H.Y.: Image Hallucniation with Primal Sketch Priors. In: CVPR, vol. (2), pp. 729–736 (2003)

[70] Tang, F.: Example-based Super-resolution EE264 Project Report (2005)

[71] Tekalp, A.M., Ozkan, M.K., Sezan, M.I.: High-resolution image reconstruction from lower-resolution image sequences and space varying image restoration. In: Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), San Francisco, CA, March 1992, vol. 3, pp. 169–172 (1992)

[72] Tikhonov, A., Arsenin: Solution of Ill-posed Problems. Winston&Sons, Washington (1977)

[73] Tipping, M., Bishop, C.: Bayesian image super-resolution. In: Neural Information Processing Systems - NIPS 2002 Vancouver (2002)

[74] Tom, B.C., Katsaggelos, A.K.: Reconstruction of a high-resolution image by simultaneous registration, restoration, and interpolation of low-resolution images. In: Proc. 1995 IEEE Int. Conf. Image Processing, Washington, DC, October 1995, vol. 2, pp. 539–542 (1995)

[75] Tsai, R.Y., Huang, T.S.: Multiframe image restoration and registration. In: Advances in Computer Vision and Image Processing, ch. 7, vol. 1, pp. 317–339. JAI Press, Greenwich (1984)

[76] Ur, H., Gross, D.: Improved resolution from sub-pixel shifted pictures. In: CVGIP: Graphical Models and Image Processing, March 1992, vol. 54, pp. 181–186 (1992)

[77] Vandewalle, P., Süsstrunk, S., Vetterli, M.: A Frequency Domain Approach to Registration of Aliased Images with Application to Super-Resolution. EURASIP Journal on Applied Signal Processing 2006, p. 14, Article ID 71459 (2006)

[78] Zhao, W., Sawhney, H.S.: Is super-resolution with optical flow feasible? In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2350, pp. 599–613. Springer, Heidelberg (2002)

[79] Zomet, A., Rav-Acha, A., Peleg, S.: Robust superresolution. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001), Kauai, Hawaii, USA, December 2001, vol. 1, pp. 645–650 (2001)

Flexible Motion Models for Scalable and High Definition Video Coding

Marta Mrak and Toni Zgaljic

Abstract. In this chapter the technologies that enable efficient handling of the richest multimedia content - videos, are presented. Application of videos in multimedia system has become possible because of effective signal-processing methods that enable high data compression. Key technology for video compression is motion compensation which is in centre of this chapter. Here, a flexible motion model that is suitable for scalable and high definition video coding is presented. The motion model is based on the application of variable-size blocks which supports highly efficient motion compensation and high degree of adaptability to the input video. The presented model can be also represented in a layered way, thus, providing scalability of motion information. In order to efficiently compress such motion information, predictive coding and adaptive entropy coding are applied. In predictive coding motion information of the current block is predicted from adjacent blocks. The prediction error is subjected to entropy coding, which utilises information about adjacent blocks and/or the current block in creating contexts to provide improved probability estimates for arithmetic coding. Overall, the presented techniques are highly flexible and are suitable for next-generation multimedia applications.

1 Introduction

In recent years, there have been rapid developments in technologies for transmission of and access to video content. When video is being delivered to the user it usually needs to traverse network paths with very different traffic capacities: from very high bandwidth on dedicated glass fibre connections to very low bit-rate connectivity for wireless transmissions. Furthermore, the same content needs to be accessible from a variety of devices at the user side. Therefore, new video coding techniques have to provide both efficient video compression and a high adaptability of compressed content.

New trends in video coding applications require good quality and very high resolution videos with scalability functionalities [1]. The scalability refers to the ability of a compressed video bit-stream to be adapted in a low-complexity fashion, by simple bit-stream parsing. As the adaptation is performed by scaling of

Marta Mrak
Centre for Communication Systems Research, University of Surrey, GU2 7XH,
Guildford, UK
e-mail: m.mrak@surrey.ac.uk

Toni Zgaljic
School of Electronic Engineering and Computer Science, Queen Mary,
University of London, Mile End Road, E1 4NS, London, UK
e-mail: toni.zgaljic@elec.qmul.ac.uk

M. Grgic et al. (Eds.): Rec. Advan. in Mult. Sig. Process. and Commun., SCI 231, pp. 209–243.
springerlink.com

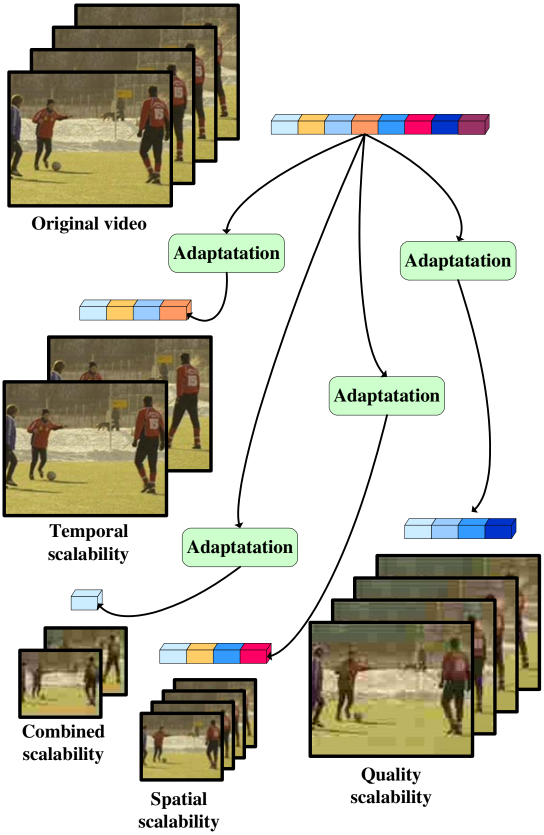


Fig. 1 Examples of basic scalabilities

compressed video parameters, the scalable bit-stream has to be encoded in such a way so that the bit-stream parts are hierarchically encoded according to these parameters. Basic types of scalability or adaptation parameters are spatial (resolution), quality (often referred to as signal-to-noise ratio scalability, or SNR scalability) and temporal (frame rate) scalability. In spatial scalability the encoded bit-stream enables extraction of the sequence containing frames of different resolution. For example, if the original sequences is of 4CIF resolution (704 × 576 pixels), then removing one resolution layer produces a sequence of CIF (352 × 288) resolution. Consequently, removing two resolution layers produces a sequence of QCIF (176 × 144) resolution. Similarly, for temporal scalability one-level extraction halves the frame rate, e.g. from 30 Hz (or 30 fps - frames per second) to 15 Hz. An example of the original sequence and scaled sequences is shown in Fig. 1. The content is encoded only once producing layered bit-stream. The resulting bit-stream is of the maximum required quality which can be quasi-lossless or even lossless. The example from Fig. 1 also demonstrates how different spatio-temporal resolutions and qualities are interleaved in the bit-stream. While light portions of the bit-stream indicate lower temporal resolution data, darker portions indicates additional data that is needed for the higher resolution. The bit-stream can be truncated at multiple points. When low quality, low spatio-temporal resolution is needed, the bit-stream parts that are associated with high resolutions are truncated, leading to combined scalability.

In recent years several scalable video coding solutions have been developed. The most recent video coding standard H.264/AVC [2] has also been extended to support scalable coding. In literature the term "scalable video coding" often refers only to the extension of H.264/AVC standard. In this chapter this terms has wider meaning and it represent the concept itself. Alternative solutions for scalable video coding are based on wavelet transform, as opposite to block-based spatial transforms used in MPEG/H.26x standards. Efficient 3D wavelet codecs that produce embedded bit-streams have been developed [3]-[6]. The common characteristic of these codecs is wavelet transform performed in the spatial and temporal domain, which facilitates spatial and temporal scalability. Important aspects of this technology have been adopted in the development of the aceSVC codec [4], [8]-[12]. aceSVC has been adopted in the content management and surveillance applications [13], [14]. Through this chapter the experiments related to the presented technologies have been executed using the aceSVC codec.

Scalability is of the utmost importance since different bandwidths and multiple video display devices are usually targeted. In order to adapt conventional video compression algorithms to these new requirements the design of each module of a video coder has to support high compression and wider range of compression settings, also enabling complexity scalability. In this chapter the advanced approaches to temporal decorrelation and entropy coding in video compression are addressed and the solutions described which support various scalability and high compression requirements.

The key of efficient video compression is temporal decorrelation of video frames that is performed using temporal prediction and motion compensation. Traditional motion compensation in conventional video coding is generally based on models with limited flexibility and small motion units (blocks). However, because of a need for higher flexibility of compression, new challenging applications require innovative motion compensation tools that have higher degree of adaptability.

In the approach presented in this chapter high degree of adaptability is introduced by application of the flexible motion model, which is based on variable-size blocks. This model also supports arbitrary motion precision and layered motion coding. The partitioning of motion units, during the employed optimisation, can be adaptively chosen according to the sequence content and other requirements. Additionally, variable-size motion models can be represented in a layered, i.e. scalable way.

Motion model used at the encoder needs to be transmitted to the decoder along with decorrelated texture information. Motion information assigned to adjacent blocks tends to be highly correlated. Therefore, in addition to predictive coding, its size in the compressed bit-stream can further be reduced by an application of the adaptive entropy coding. As probabilities of symbols that enter entropy coding module are not known prior to encoding, accurate on-the-fly probability estimation plays the key role in efficient compression of motion information. At this stage improved probability estimation can be achieved through the application of context modelling. Moreover, in addition to the efficient compression, the entropy coding module should also preserve scalability of the employed motion model.

This chapter starts with a brief overview of temporal prediction in video coding in Sect. 2. It discusses state-of-the-art motion models used in temporal prediction stage of temporal prediction. Sect. 3 focuses on a flexible motion model that supports wide range of video content and frame resolution as well as video scalability. The advantages of the flexible motion model and its performance evaluation in different scenarios are demonstrated using aceSVC, a state-of-the-art scalable wavelet-based video codec, whose fine performance is based on the presented motion model. In the subsequent section (Sect. 4) a detailed description of the basic principles applied in entropy coding is presented, followed by the description of its application in scalable motion coding. Conclusions are given in Sect. 5.

2 Temporal Prediction and Motion Models in Video Coding

A great part of efficiency in video compression depends on proper exploitation of temporal redundancies. Since neighbouring frames in a video sequence are similar, a frame can be represented using data from already encoded neighbouring frames. In this process, which is usually referred to as motion compensation, a significant part of the temporal redundancy can be removed. A frame that is being predicted is divided into motion units (blocks), each of which is associated with suitable descriptors of prediction - motion modes and motion vectors. Motion models are in this context used to define a prediction method for each motion unit, and also the configuration of the motion units themselves. The prediction descriptors define the mode of prediction - either the temporal prediction (inter modes) or spatial or no prediction (intra modes). Motion vectors, as descriptors of motion units' displacements between frames, are associated to units in the inter modes.

Block-based and triangular mesh models are popular representations of motion. In block models the motion units are of rectangular shape. Such models are used in all video-coding standards because the algorithms they are based on are of low complexity. Although the most efficient motion compensation is utilised with smaller units, providing for higher precision in motion description, the balance between final motion and texture rate has to meet the targeted coding demands. Moreover, certain applications require robust coding that is less sensitive to possible errors that can occur in, for example, network transmission. Therefore some motion estimation techniques do not aim to find a motion representation that achieves the best prediction at the encoder side, but rather aim for higher robustness of coded video at the expense of the decreased coding efficiency [15]. In video applications that require high degree of adaptability the main requirements for motion are scalability of related algorithms and high compression.

In early video coding standards (MPEG-1 [16], MPEG-2 [17], H.261 [18], H.263 [19]) motion blocks were of fixed size (8 × 8 pixels). Recently, more efficient frame partitioning into motion blocks has been standardised in the most recent video coding standard, H.264/AVC [1]. It takes advantage of variable block sizes and enables more efficient motion compensation and overall video compression.

However, in variable block size schemes the balance between the gain introduced by more precise motion estimation and the amount of data describing motion must be kept. Therefore, a goal of an encoder is to find the best matching motion constrained with the available bit budget.

Variable size block models have been also applied in scalable video frameworks [3], [4], [20]. In [20] a motion model based on one applied in H.264 is used, while in [3] a variable size block matching [21], in its refined version from [22] is utilised.

In H.264/AVC the block sizes are varied between 16 × 16 and 4 × 4 pixels, and for block dimensions only multiples of 4 pixels are used [23]. A model that offers higher flexibility of frame partitioning, and therefore better adaptation to the actual content, is often referred to as Binary Partition Tree (BPT) model [24]. This method that enables adaptive partitioning of video frames is originally motivated by rate-distortion optimisation requirements in compression. Recently it has been shown that this method is very suitable for application in 3D video coding [25], especially for depth-map images, since BPT model enables excellent adaptability to the actual frame content.

In the schemes based on variable size block models the block sizes are not predefined. Therefore, block sizes can be varied to optimise the trade-off between number of bits used to encode motion vectors and residual texture. To achieve this, the BPT motion model is fully adaptable to target rate-distortion requirements. It uses the partitioning of a frame into blocks described with a tree-structure. The partitioning is achieved using the two-step algorithm. First step is the growing the tree for partitioning of a frame (top-down approach). Second step is the pruning the tree which finds the optimal partitioning with the respect to given requirements (bottom-up approach).

The BPT model has demonstrated promising results in video coding. Its main advantage is its capability to partition the frame along actual motion boundaries. This is achieved by splitting a frame in a way that maximizes block-matching efficiency. Although BPT model is the most flexible block-based motion model, in order to achieve the highest compression performance it requires computationally highly costly estimation. This is because BPT model supports any block size, which is determined in a numerous motion estimation steps.

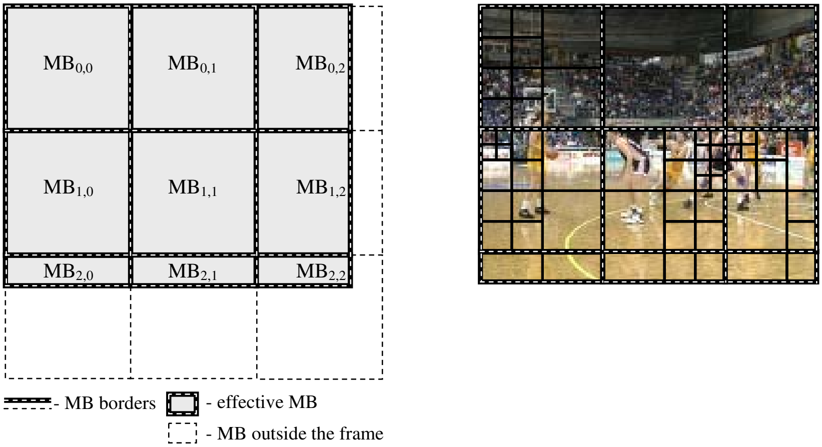
A flexible motion model that can be used for various scalability scenarios while preserving low complexity can still preserve low complexity by restricting the possibilities for block sizes. A detailed description of such model as well as its application in scalable video coding is given in the following sections.

3 A Flexible Block-Based Motion Model

In the flexible block-based motion model, as used in aceSVC, the basic motion units are macroblocks (MB), which cover a frame that has to be motion compensated, Fig. 2 a). The size of macroblocks can be chosen so that a macroblock covers a larger frame area (e.g. 64 × 64 or 128 × 128 pixels). As frame dimensions are generally not multiples of macroblock dimensions, parts of macroblocks may lie outside the frame (e.g. $\text{MB}_{2,i}$ and $\text{MB}_{i,2}$ in Fig. 2 a).

Partitioning of a frame into macroblocks is a first step towards the main objective - finding an optimised partitioning of the frame. First, in the case of complex motion, the smaller the final blocks are, the more precisely the motion between frames can be described, and consequently the better prediction can be obtained. If macroblock size is large, high amount of bits may be spent on macroblock partitioning information. On the other hand, using small blocks can increase the average bit-rate of motion vectors, since a higher number of motion vectors is associated to the compensated frame. Therefore, the macroblock size has to be carefully selected. It has to allow using large enough blocks to avoid possibly high average bit-rate of motion vectors and optimise the size of partitioning information. Second, using larger macroblocks is beneficial for sequences (or segments of a sequence) containing panning motion or a static background. As each motion block possesses motion properties, e.g. motion vector values, which have to be included in the compressed video bit-stream, using larger blocks can be beneficial for compression in this case. Therefore, in order to have a uniform motion model, a compromising decision has to be made on the macroblock size, which is usually done on per-sequence basis. In Fig. 2 b) an example of frame partitioning is shown where it is demonstrated that MBs in static or smooth panning areas are not divided. Optimal partitioning is decided in the motion estimation process.

Partitioning of MBs in the flexible block-based motion model is based on regular division of MBs into four square blocks, which can be further divided using the same principle. The division is described by the tree model, as shown in the following section. Partitioning of basic motion units is also important property of motion model applied since it allows for imposing a layered motion structure. Although the following discussion focuses on block-based models, it can be extended to different shapes of motion units.



a) frame partitioning into macroblocks

b) macroblock partitioning into smaller blocks depending on frame content

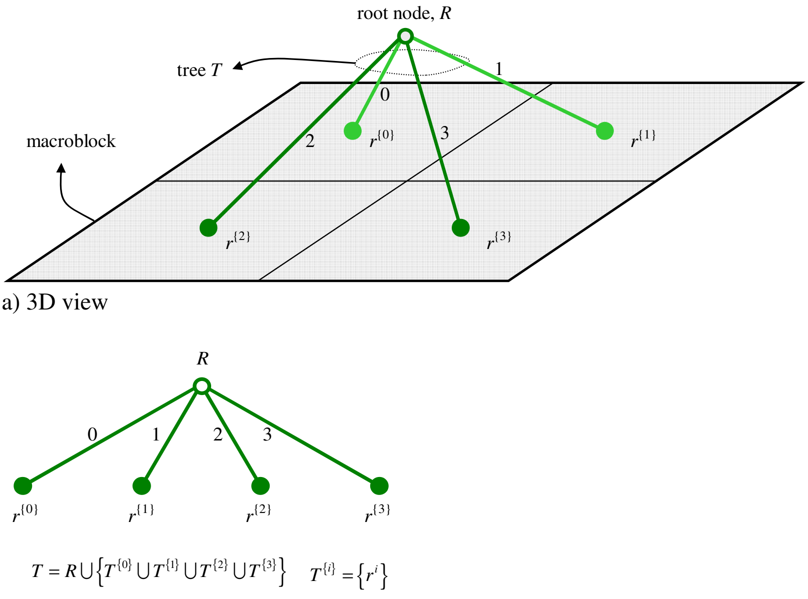
Fig. 2 An example of frame partitioning into macroblocks $MB_{i,j}$.and blocks

3.1 Motion Tree Structures

Partitioning of basic motion units, macroblocks, into smaller blocks is defined by a quadtree structure associated to each macroblock of a frame that has to be motion compensated. A quadtree T is a set of nodes, which can be partitioned into four subtrees - its subsets, each of which is also a tree. It can be defined recursively with the following relation:

$$T^{\{P(r)\}} = r \cup \left\{ T^{\{P(r),0\}} \cup T^{\{P(r),1\}} \cup T^{\{P(r),2\}} \cup T^{\{P(r),3\}} \right\}.$$

Node r is a root node of a current tree $T^{\{P(r)\}}$ and $P(\cdot)$ defines uniquely a path from an initial root node R to the current tree root. R is a root node of the initial tree that is not a subset of any other tree, and can be represented with $T^{\{\varnothing\}}$. The path consists of branches that can be understood as connections between nodes. Each branch is labelled with a number $i = 0,..., 3$, so that the path is defined as a sequence of these labels required to reach a particular node. The depth d of a node r in the tree is defined by the cardinal number of $P(r)$, $d = |P(r)|$, and represents the tree level distance from the initial tree root R.



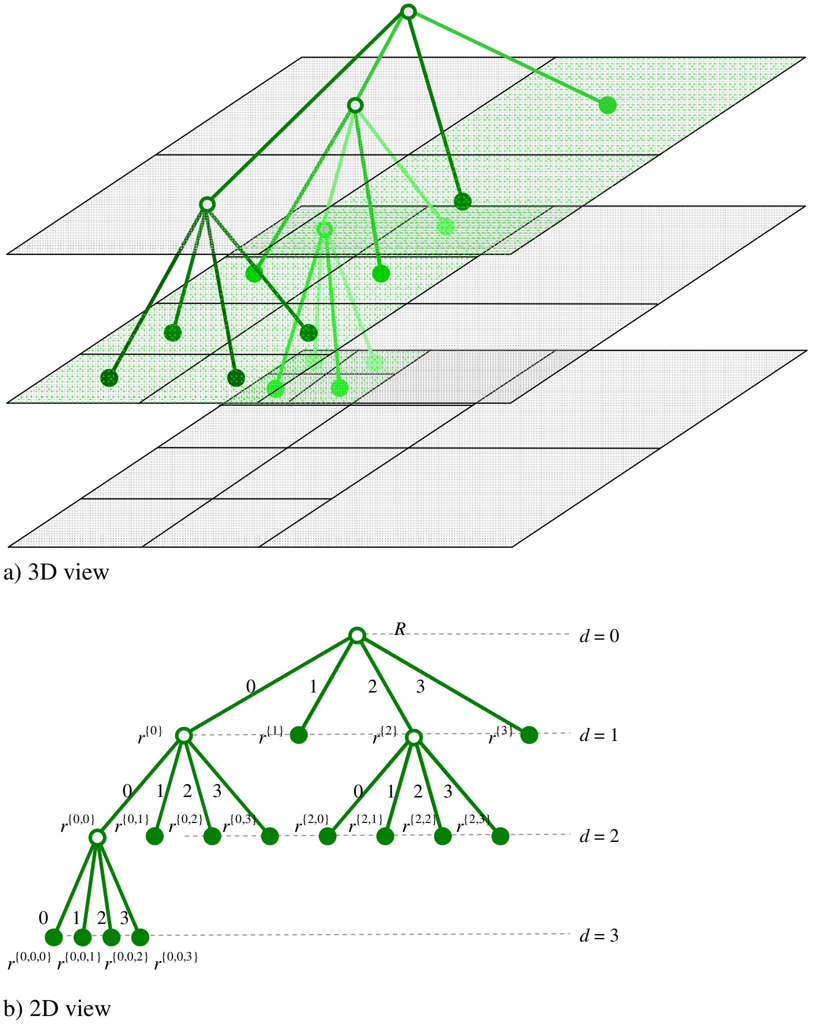
a) 3D view

b) 2D view

Fig. 3 Partitioning of a MB into 4 blocks described by the tree structure

A tree root R is assigned to a whole MB, as in Fig. 3, and the roots of $T^{\{i\}}$ subtrees are associated with four blocks this MB is partitioned into. In the example from Fig. 3 the subtrees consist of only one node (root node), which is indicated as

$T^{\{i\}} \equiv r^{\{i\}}$. In Fig. 4, further partitioning of blocks is shown for the case of $\mathrm{MB}_{1,0}$ from Fig. 2. In the following text some definitions concerning the motion quadtrees will be introduced.


a) 3D view

b) 2D view

Fig. 4 Description by tree structure of partitioning of $\mathrm{MB}_{1,0}$ from Fig. 2 b)

Each final motion block found in the process of motion estimation and that is used in motion compensation corresponds to a tree leaf. Such final motion blocks are also referred to as the external nodes. All other motion tree nodes are internal nodes. As some MBs lie partially outside the frame, some leaves are associated with blocks that are outside the frame area, Fig. 2. These nodes are referred to as empty nodes and an example of a tree that has empty nodes is shown in Fig. 5.

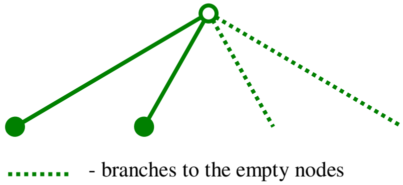


Fig. 5 Tree structure for the $MB_{2,2}$ from Fig. 2

Node r_c is a child node of node r_p if $\{P(r_p), i\} \equiv P(r_c)$, for any node r_c, where $i \in \{0, 1, 2, 3\}$. Node r_p is a parent node of nodes r_c for all such i. The size of a tree can be reduced by pruning, i.e. by removing the subtrees, Fig. 6. T_B is a pruned tree of the original tree T_A if

$$R_A \equiv R_B \text{ and } T_B \subseteq T_A.$$

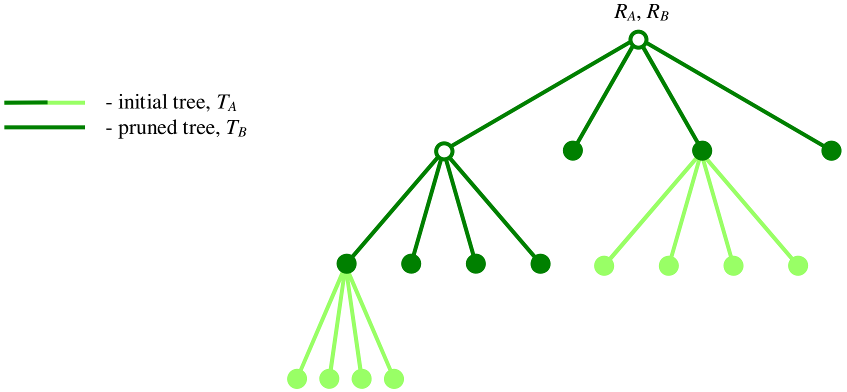
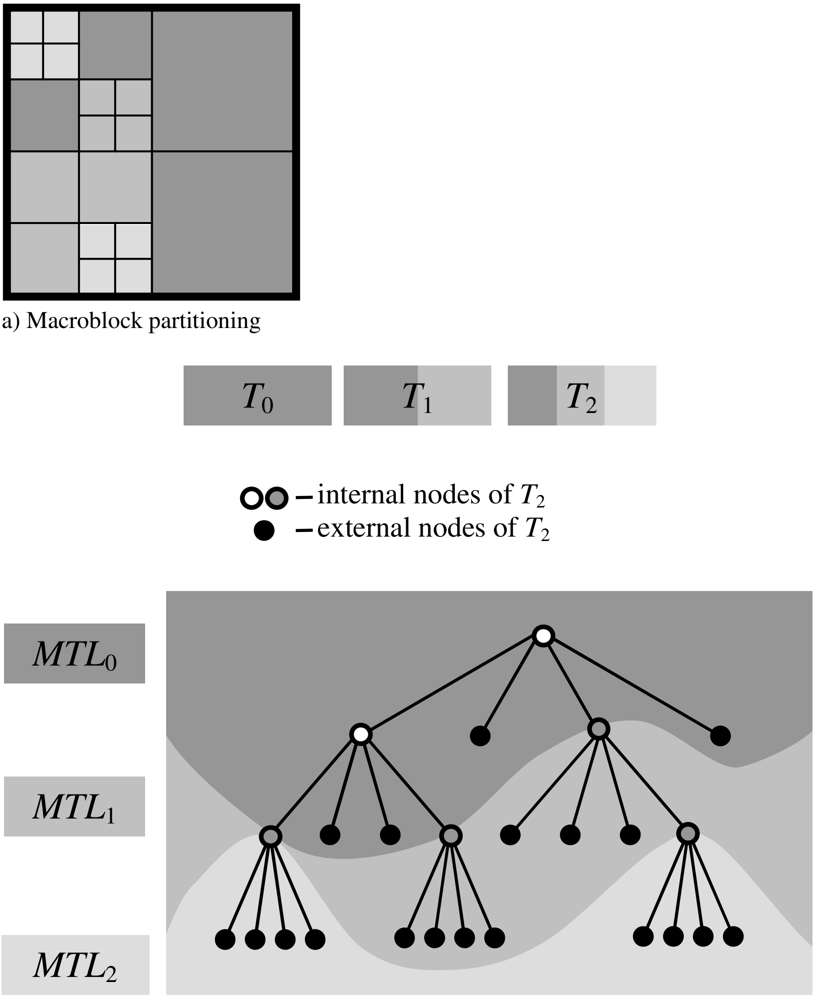


Fig. 6 Initial and pruned trees

Motion properties of each final block of a frame are associated with leaves of the trees, except for empty nodes. These motion properties basically refer to the motion compensation modes and motion vectors. They can also include different distortion measures used at the encoder, various tags needed for layered representation and other parameters required by the codec.

3.2 Motion Syntax

Motion parameters that define motion compensation and coding of motion information are motion vector precision, motion vector range, macroblock size, maximal allowed depth of motion trees and motion tree parameters which include motion



a) Macroblock partitioning

b) Corresponding motion tree

Fig. 7 Layered motion structure

block parameters. Scalable coding can be imposed on a motion tree structure and motion vector component values. Motion tree parameters are therefore defined in a way that scalable coding can be used.

In a non-scalable scenario the estimated motion tree is used in both motion compensation and inverse motion compensation, so that the tree structure and motion vectors associated to tree nodes are equal at both encoder and decoder sides. The motion tree parameters in this case include quadtree structure description, which can be completely represented with decisions on node splitting, and parameters of tree leaves / blocks - block mode (intra, inter and subclasses of inter mode) and motion vector values (one or two pairs for each inter block accordingly to the inter subclass). Node splitting parameter is a binary value that defines if the current node is an internal or external node. This parameter has to be transmitted for each tree node, except for the leaf nodes.

If scalability is imposed on the tree structure, motion tree layers have to be defined so that motion compensation uses T_{LS} - 1 tree and the inverse motion compensation uses one of T_0,..., T_{LS} - 1 trees, where *LS* is number of layers. For these trees holds:

$$T_0 \subseteq T_1 \subseteq \ldots \subseteq T_{LS\text{ - }1}.$$

Motion tree layers *MTL* are then defined as

$$MTL_q = \begin{cases} T_q, & \text{if } q = 0 \\ T_q \setminus T_{q-1}, & \text{if } q = \{1, \ldots, LS - 1\} \end{cases}.$$

MTL_0 is a base tree structure layer, and MTL_q, $q > 0$, are enhancement layers. The base layer is encoded as a non-scalable motion tree structure. Motion tree T_{LS} - 1 is an initial tree and is used in motion compensation.

Connection between enhancement layer MTL_q and the previous layer MTL_q - 1 is realised using a binary node growing parameter. It describes if an external node from layer MTL_q - 1 is a root node for a subtree in MTL_q layer. Other motion parameters that exist in the base layer are also specified in enhancement layers. An example of scalable motion structure is shown in Fig. 7, for 3-layer tree structure.

Other way to impose scalability on motion is by precision limiting of motion vector values. In that case two types of layers exist - base layer that contains precision limited (PL) values and refinement layers that consist of higher precision bit-planes. Full-precision motion vectors are used for motion compensation, while decoder can use motion vectors of any precision. Specialised coding techniques developed for this precision limited motion vector values will be further discussed in Sect. 4.4.

3.3 Motion Modelling and Rate-Distortion Optimisation

The main purpose of motion estimation is to find suitable motion displacement vectors that facilitate efficient compensation. Also, it must be driven by an optimisation algorithm that controls the balance between different motion partitioning and modes with respect to motion and texture rates. Such optimisation algorithm is of crucial importance when motion models with variable block sizes are used. However, since in fully scalable video coding besides targeted decoding rates all other rates must be available, the motion rate optimisation can negatively influence the performance in those remaining points. Since this problem in certain cases, for instance low bit-rates, can be tackled by other compression techniques, such as scalable motion coding, the presented rate-distortion optimisation algorithm is designed to support various values of corresponding parameters. This provides adaptation to different requirements and mainly depends on the content of the targeted sequence and targeted decoding points. In addition to the architectural part of a video codec, the algorithm for motion optimisation is the second most important module that defines the overall video codec performance.

Although the algorithm presented here is for the sake of clarity described as applied at the original resolution, in scalable video coding it can be used in motion estimation at any resolution scale.

Motion modelling and the associated rate optimisation take place in motion estimation module of the encoder. Since the motion model is based on hierarchical blocks, motion estimation determines macroblock partitioning. By selection of macroblock partitioning, block modes and motion vectors, motion estimation influences the resulting motion rate as well as the overall compression performance.

In fixed rate coding the motion rate is optimally allocated using advanced video rate control techniques [26]. The corresponding algorithms have also been proposed for application in SVC, [22], [27]. All these techniques focus on motion rate allocation in environments with variable block sizes and have therefore been considered in the design of the proposed algorithm. Since the flexibility of presented motion model enables wide range of settings, this algorithm for motion modelling follows the structure of the underlying motion model.

The main goal of the optimisation algorithm is to find the relationship between the rate of the motion information and corresponding distortion introduced to the frame to be compensated. In that way the algorithm is content adaptive but so is the resulting motion rate. This can pose a problem if the resulting motion rate is above targeted overall rate for transmission, which can be addressed in two ways. Firstly, before encoding the Lagrangian multiplier λ can be adjusted in order to meet the underlying requirements and thus to increase the resulting distortion while decreasing the rate of motion information. Alternatively, motion scalability can be employed, where the motion information of required precision is selected and preserved.

For a given motion estimation stage, the goal is to jointly minimise the distortion D and motion rate R for a given frame that needs to be compensated:

$$\min_{R}\left(D+\lambda\cdot R\right)\cdot \tag{1}$$

Here $J = D + \lambda \cdot R$ denotes a Lagrangian cost function for a given frame. In order to decrease the complexity of such global frame-wise optimisation, it is common to break the optimisation into processing of smaller units, and thus perform a local optimisation. It is therefore assumed that (1.1) can be broken into macroblocks $\mathrm{MB}_{\mathbf{i}}$:

$$\min_{R}\left(D+\lambda\cdot R\right)=\min_{R}\left(\sum_{\mathbf{i}}\left(D_{\mathbf{i}}+\lambda\cdot R_{\mathbf{i}}\right)\right)=\sum_{\mathbf{i}}\min_{R_{\mathbf{i}}}\left(D_{\mathbf{i}}+\lambda\cdot R_{\mathbf{i}}\right),$$

where **i** denotes coordinates of a macroblock in a frame. As the motion vector values are coded in a predictive fashion, the final frame rate R depends on the coding order. For this reason the Lagrangian cost function for each macroblock, $J_{\mathbf{i}} = D_{\mathbf{i}} + \lambda \cdot R_{\mathbf{i}}$, has to be accurately predicted. Since motion rate $R_{\mathbf{i}}$ of each macroblock depends on the previously encoded macroblocks, the optimisation order has to follow the coding order. Calculation of the distortion $D_{\mathbf{i}}$ of each macroblock depends on the selected block modes within the macroblock. The block modes within the macroblock are adaptively chosen in such way to minimise $J_{\mathbf{i}}$.

The optimisation at macroblock level follows the recursive algorithm which adaptively chooses modes for the corresponding blocks. The available block modes depend on the chosen temporal prediction complexity. Supported modes are:

- bidirectional mode, BIDIR, prediction by averaging signals from both previous and subsequent frames; requires two motion vectors per block;
- unidirectional modes; require one motion vector per block,
- forward mode, FORW; prediction from previous (forward) frame,
- backward mode, BACK, prediction from subsequent (backward) frame,
- intra mode, INTRA, without prediction.

Modes available for optimisation depending on selected temporal prediction mode are summarised in Table 1. The modes from Table 1 are available for all frames to be predicted, except for the last frame at the particular motion compensation level and for intra frames. Intra frames are those frames that are not motion compensated. The last frame in the selected coding unit can be predicted only from previous frames.

Table 1 Motion modes and associated codes for different prediction modes in motion compensation

-	-	BIDIR	FORW	BACK	INTRA
Unidirectional prediction	Availability	No	Yes	No	Yes
	Code	-	1	-	0
Bidirectional Prediction	Availability	Yes	Yes	Yes	Yes
	Code	1	00	010	011

Depending on prediction mode chosen for temporal decomposition, different codes are associated to the motion modes. The codes are also displayed in Table 1. For prediction modes for which backward (including bidirectional) prediction is not supported, only two modes are possible - intra and forward modes. Therefore the motion mode can be signalled with one bit. For other cases, where all modes are supported, variable length codes are designed. The design of these codes depends on the expected probability of a certain mode. For example, intra mode is in general least probable and is therefore coded with 3 bits ("011" in Table 1). On the other hand, bidirectional prediction is a usual result of the optimisation process and therefore its code word consists of one bit only ("1" in Table 1). Estimation of the encoded motion value rate follows the principles of the motion coder, Sect. 3.4. Additional elements contributing to the motion rate are the node splitting flags which describe macroblock partitioning. The node splitting flag is associated with all levels of motion trees except the lowest level of the tree, where nodes cannot be split.

The distortion d for each compensated block $\mathbf{B}_{2 \cdot k + 1}(\mathbf{0})$ from $(2 \cdot k + 1)$-frame is calculated taking into account corresponding block in i-th reference frame which is shifted by motion vector **mv**:

$$d(\mathbf{mv}) = \left|\mathbf{B}_{2 \cdot k+1}(\mathbf{0}) - \mathbf{B}_i(\mathbf{mv})\right|^p ,$$

where p is usually chosen from $\{1, 2\}$. Since both choices give similar results, in video coding applications $p = 1$ is selected as the default option because of its lower complexity. As the distortion equivalent for intra modes, i.e. uncompensated blocks, the following measure is used:

$$d_{\text{intra}} = \left|\mathbf{B}_{2 \cdot k+1} - \bar{b} \cdot \mathbf{I}\right|^p ,$$

where $\bar{b}$ is the average pixel value for the given block and **I** is the unit matrix.

Block matching and mode decision (BM) are performed according to the following steps:

– Step BM.1 - Block matching
  Motion estimation is performed for the current block. Estimation consists of two independent steps - estimation of forward and backward motion vectors, if bidirectional motion is used. Otherwise only forward motion vector is estimated. For each direction a motion vector is selected taking into account the distortion, i.e. the prediction residual, as well as the corresponding motion rate:

$$\mathbf{mv}_m = \arg\min_{\mathbf{mv}_m} \left(d(\mathbf{mv}_m) + \lambda \cdot r(\mathbf{mv}_m) \right) ,$$

  where $\mathbf{mv}_m$ is motion vector for mode $m \in$ {FORW, BACK }.
– Step BM.2 - Mode decision
  For each possible coding mode a corresponding Lagrangian cost function is calculated and the mode corresponding to the minimal one is selected for the current block:

$$m = \arg\min_{m} \left(j_m \right) ,$$

  where j_m is the Lagrangian cost function for the current block with m selected between available block modes.

The optimisation algorithm and its following coding stages in follow the raster scan across the macroblocks, Fig. 8. The actual optimisation at the macroblock level is performed in a top-down fashion, i.e. starting from the motion tree root

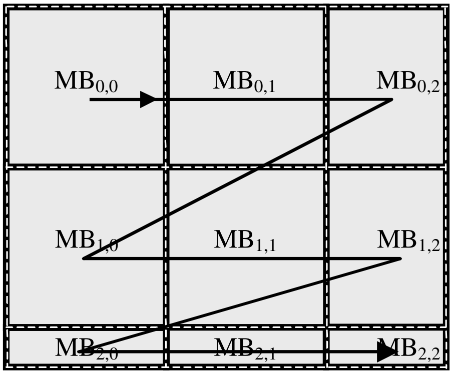


Fig. 8 Processing and coding scan from macroblocks in a frame

towards the leaves of the motion tree. This tree growing (TG) process follows the steps TG.0 – TG.2:

- Step TG.0 Initialisation
  An initial motion tree is created consisting of the root node only: $T = R$. According to the predefined input setting, the tree will be allowed to grow until certain depth d.

  BM is performed for the current macroblock. If $d > 0$, the algorithm proceeds with Step TG.1 in which the macroblock is treated as any other block. Otherwise the algorithm terminates at this point, with motion parameters obtained in BM associated to this root node.
- Step TG.1 Block matching for child nodes
  BM is performed for blocks that correspond to all immediate child nodes of the current block. For a faster convergence towards the optimal motion vectors, the block search at this stage starts in the directions of motion vectors from the parent node. When motion parameters are found for each of the child nodes, the evaluation of tree growing is performed as defined in Step TG.2.
- Step TG.2 Decision on node splitting
  It is supposed that a subtree $T = r$ can grow such that it becomes $T'^{\{P(r)\}} = \{r \cup \{r^{\{0\}}, r^{\{1\}}, r^{\{2\}}, r^{\{3\}}\}\}$. The Lagrangian cost functions of child nodes are compared to the cost function of the current node. If the sum of the cost function of the current node is larger or equal then the sum of the cost functions of the child nodes, i.e.

$$j(r) \le \sum_{c=0}^{3} j\left(r^{\{c\}}\right),$$

  the algorithm terminates for this branch, while in the opposite case this node is split into up to four child nodes. The tree growth consists of setting the tree growing flag for the current node, creation of branches of the new nodes and associating the motion parameter to the newly created nodes. If newly created nodes are at the tree level different then the maximum allowed level, the algorithm continues with Step TG.1 for each of the child nodes. Otherwise the growth of the current branch terminated at this point.

The described steps define a "greedy" tree growing algorithm. An alternative, globally optimal, approach would be to perform full growth with subsequent pruning. However, such approach would considerably increase the complexity.

As described through the algorithmic steps of the tree growing and block matching algorithm, the same Lagrangian multiplier is used in all steps. This parameter is predefined and in aceSVC can be set for each temporal decomposition level.

Presented technique is evaluated in terms of objective video quality. In image and video coding research communities the most widespread measure that is used for the objective evaluation of end results is the Peak Signal to Noise Ratio (PSNR). Because of its popularity, PSNR has become the most common tool for the comparison of different algorithms across different publications.

PSNR measures distortion of given signal, usually with respect to the original, uncompressed content. Therefore it evaluates closeness between two signals by exploiting the differences of sample values. The main component of PSNR calculation is measurement of the Mean Square Error (MSE) for each frame f of the video sequence:

$$MSE_f = \frac{1}{N} \cdot \sum_{n=0}^{N-1} \left(x_f(n) - \hat{x}_f(n) \right)^2 ,$$

where x_f and $\hat{x}_f$ are pixels of the original frame and the decoded frame, respectively, and each frame consists of N pixels. PSNR of a frame f is then defined as:

$$PSNR_f = 10 \cdot \log \frac{\left(2^k - 1\right)^2}{MSE_f} = 10 \cdot \log \frac{255^2}{MSE_f} ,$$

where k is the number of bits per pixels for given frame ($k = 8$ for commonly used test material). PSNR for the whole sequence can be evaluated in several ways. In this work the most common way of PSNR computation for the whole video is used, where PSNR is computed as the mean value of PSNRs of individual frames:

$$\overline{PSNR} = \frac{1}{F} \cdot \sum_{f=0}^{F-1} PSNR_f .$$

Average PSNR ($\overline{PSNR}$) is in video coding evaluations denoted as PSNR and it corresponds to a PSNR of the complete video sequence.

The benefits of having an adaptive choice for motion vector modes in motion compensation are demonstrated in Fig. 9 and Fig. 10 where the test results, in terms of average PSNR, are summarised for popular test video sequences. For

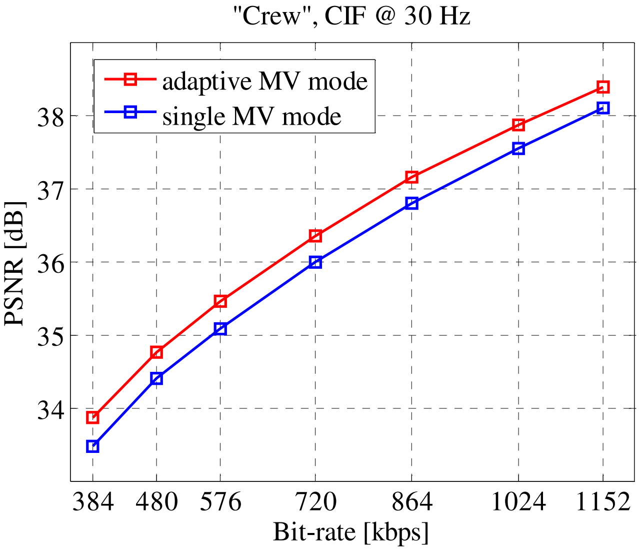


Fig. 9 Comparison of the adaptive and single MV mode compression performance, "Crew" sequence

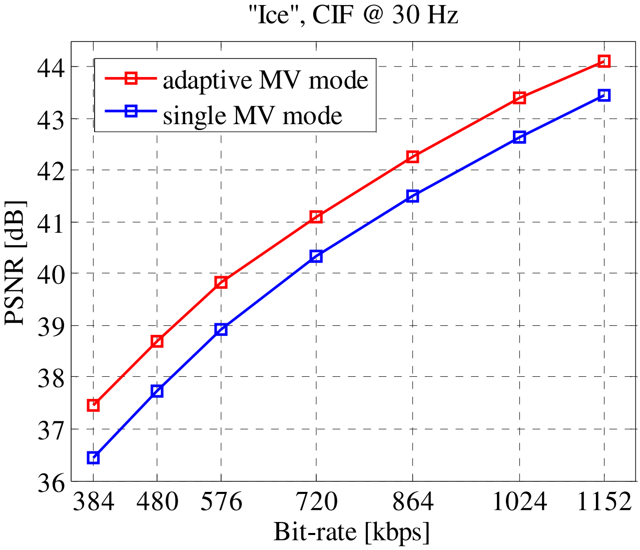


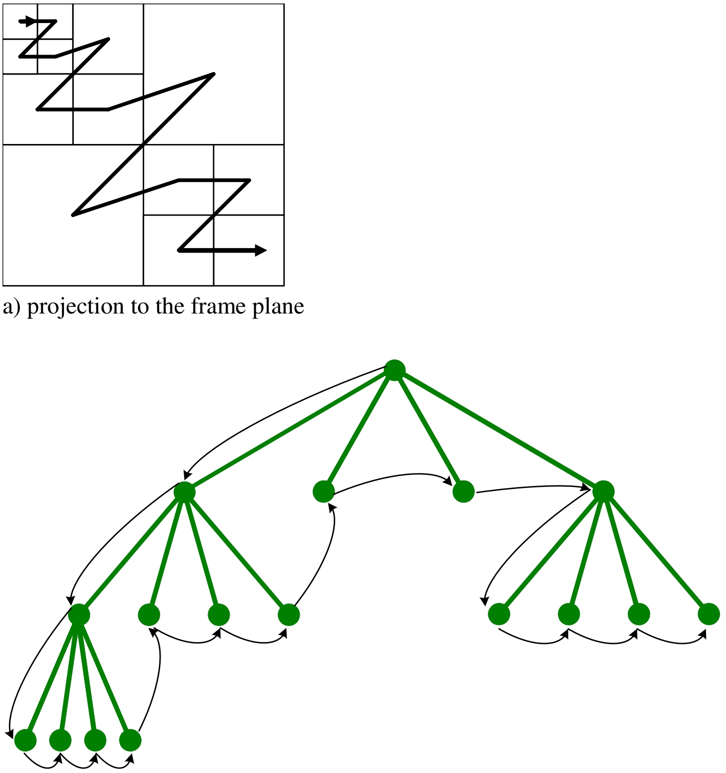
Fig. 10 Comparison of the adaptive and single MV mode compression performance, "Ice" sequence

both sequences a CIF resolution is used. A single motion mode refers to fixing the motion mode to the bidirectional for all motion blocks in a frame, while the adaptive refers to selecting the optimum in terms of the prediction error between all available motion modes. The higher performance gap in favour of the adaptive MV mode selection for the "Ice" sequence, compared to the "Crew" sequence, is due the higher amount of object occlusions in the sequence, which are more efficiently predicted using unidirectional motion modes.

While this section focused on the optimisation part of motion estimation, the following section provides more details on practical implementation related to the actual motion coding.

3.4 Hierarchical Motion Coding

The motion information of each compensated frame is encoded in a raster scan order across the macroblocks, as depicted in Fig. 8. For each macroblock the corresponding parameters are encoded. These include partitioning flags of the macroblock, which describe the motion tree structure, as well as the motion mode information for each motion tree leaf node followed by motion vector values for inter blocks only. A recursive algorithm encodes all the parameters within one macroblock in the Z-scan, Fig. 11.



a) projection to the frame plane

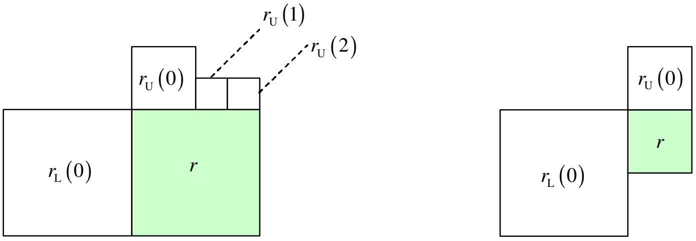
b) motion tree describing macroblock partitioning

Fig. 11 Encoding scan for blocks within a macroblock

Motion mode information for each final block used in motion compensation is encoded according to Table 1. In the Z-scan order, the prediction of motion vectors associated to the current block, i.e. current tree leaf *r* of corresponding macroblock, can be performed using the blocks from its top and left sides. In this scheme only neighbouring blocks that have pixels in the same rows, or columns as the current block are taken into account. If the number of blocks ($r_L(i)$) on the left side of the current block is denoted as *LB* and the number of blocks ($r_U(i)$) above the current block is denoted as *UB* then the predicted value of forward motion vector, $\mathbf{pmv}_{\text{FORW}}(r)$, is computed as:

$$\mathbf{pmv}_{\text{FORW}}(r) = 2^{-p(L-1)} \cdot \left[\frac{\sum_{i=0}^{UB-1} bl\left(r, r_{\text{U}}(i)\right) \cdot \mathbf{mv}_{\text{FORW}}\left(r_{\text{U}}(i)\right) + \sum_{i=0}^{LB-1} bl\left(r, r_{\text{L}}(i)\right) \cdot \mathbf{mv}_{\text{FORW}}\left(r_{\text{L}}(i)\right)}{blx(r) + bly(r)} \cdot 2^{p(L-1)} \right], \tag{1.2}$$

where $bl(r, r(i))$ is the number of neighbouring pixels between blocks *r* and $r(i)$ blocks in inter modes, $blx(r)$ and $bly(r)$ are number of pixels at the border between *r* and blocks in inter modes used in prediction, and [·] represents operation of rounding to the nearest integer. For neighbouring intra blocks $\mathbf{mv}_{\text{FORW}}(r(i))$ is set to 0. If a neighbouring block does not have forward motion vector, prediction is performed from its inverted backward motion vector ($-\,\mathbf{mv}_{\text{BACK}}(r(i))$). To ensure that predicted values fall within motion vector value range and that the predicted value is of the same precision, rounding to the nearest integer and normalisation to the highest precision $p(L - 1)$ is used. Fig. 12 shows two examples for prediction of motion vectors.



$$\mathbf{pmv}(r) = \frac{\mathbf{mv}\left(r_{\text{L}}(0)\right)}{2} + \frac{\mathbf{mv}\left(r_{\text{U}}(0)\right)}{4} + \frac{\mathbf{mv}\left(r_{\text{U}}(1)\right)}{8} + \frac{\mathbf{mv}\left(r_{\text{U}}(2)\right)}{8} \qquad \mathbf{pmv}(r) = \frac{\mathbf{mv}\left(r_{\text{L}}(0)\right)}{2} + \frac{\mathbf{mv}\left(r_{\text{U}}(0)\right)}{2}$$

Fig. 12 Prediction of motion vectors; motion vector directions are here omitted

As the prediction defined in this way can be performed at the decoder, it is sufficient to encode prediction error only **emv**, whose value is:

$$\mathbf{emv}_{\text{FORW}}(r) = \mathbf{mv}_{\text{FORW}}(r) - \mathbf{pmv}_{\text{FORW}}(r) \cdot \tag{1.3}$$

The prediction of backward motion vector is done in the same way, i.e. by substitution of $\mathbf{pmv}_{\mathrm{FORW}}(r)$ and $\mathbf{mv}_{\mathrm{FORW}}(r(i))$ in (1.2) with $\mathbf{pmv}_{\mathrm{BACK}}(r(i))$ and $\mathbf{mv}_{\mathrm{BACK}}(r(i))$, and inversion of forward motion vectors ($-\,\mathbf{mv}_{\mathrm{FORW}}(r(i))$) if backward motion vectors are not available. Indices FORW in (1.3) are substituted with BACK for backward motion vector prediction error. The prediction error is encoded using context-dependent binary arithmetic coder, described in Sect. 4.

For encoding the motion tree structure, information from neighbouring blocks can also be utilised to improve compression. This is because it can be expected that a certain amount of correlation exists between the splitting information of neighbouring blocks. For instance, if a frame area contains complex motion then smaller block sizes are likely to be used in that area. Thus it may be expected that blocks of large size will not be present in this area.

Splitting information is the binary information and therefore it can be efficiently encoded using context-dependent binary arithmetic encoder. The order of encoding is from the motion tree root to the motion tree leaves, following the "in-depth first" approach, as depicted in Fig. 11. If a motion tree leaf is located at the maximum allowed tree depth, there is no need to encode its splitting information. More information about entropy coding of motion structure is given in Sect. 4.

3.5 Experimental Evaluation of the Flexible Motion Model

To demonstrate effectiveness of the presented flexible block-based motion model, experimental results are presented for two sequences commonly used for evaluation in video processing: "City" and "Crew". Both sequences were encoded at high definition (HD) resolution (1280 × 720 pixels) and frame rate of 60 Hz using the aceSVC codec [13]. For each sequence the coding was performed four times, each time with different settings used for motion estimation and temporal prediction. In the first coding flexible block-based motion model was used to represent motion. In the second and third coding blocks of fixed size of 64 × 64 and 8 × 8 pixels, respectively, were used to represent motion. In the fourth coding the sequences were compressed without performing temporal prediction, i.e. intra-only coding was used. Compressed scalable bit-stream resulting from each encoding was adapted to nine different bit-rates and decoded. PSNR results for decoded luminance component are shown in Fig. 13 and Fig. 14.

It can be observed that for both sequences temporal prediction using flexible motion model gives the best PSNR performance. On the other hand, as expected, skipping temporal prediction during the compression provides the worst PSNR results. The performance gap between these two modes of coding is around 9 dB for the "City" sequence and 2.5 dB for the "Crew" sequence. These results clearly demonstrate the importance of performing the temporal prediction in video coding. Among the three modes of coding that perform temporal prediction, using blocks of 8 × 8 pixels to represent motion provides the lowest performance. This is because relatively large amount of bits is spent on encoded motion parameters, since the overall number of blocks is large. Using blocks of 64 × 64 pixels provides satisfactory performance in temporal prediction, but at the same time the number of bits needed to encode motion parameters is much lower than in the case of 8 × 8 blocks. Flexible motion model provides the best results since it adaptively balances between the effectives of temporal prediction and number of bits spent on encoded motion parameters. The results also show that the large block sizes (64 × 64 pixels) are suitable for HD video coding. Such large blocks are not suitable for compression of sequences of lower resolution, which has also been reflected to the popular video coding standards, e.g. H.264/AVC [2] which uses block sizes in a range of 4 × 4 and 16 × 16.

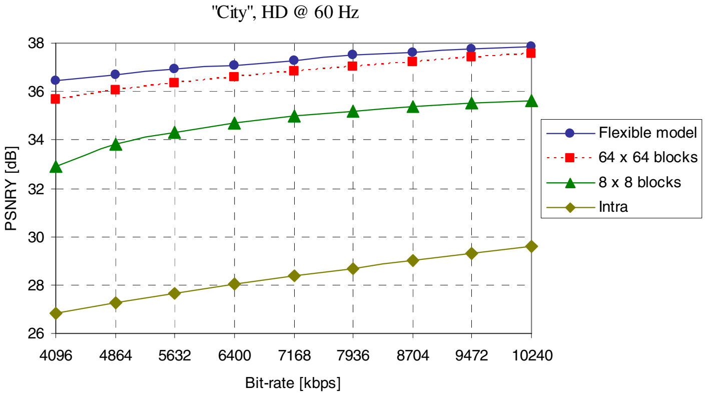


Fig. 13 Evaluation of motion models, “City” sequence

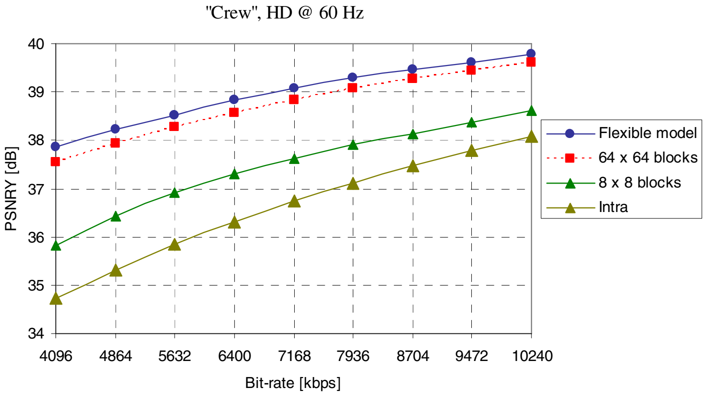


Fig. 14 Evaluation of motion models, “Crew” sequence

4 Entropy Coding of Motion Information

As already mentioned in Sect. 3, the flexible block-based motion model consists of three main data types: splitting information, motion modes and motion vectors’ prediction differences. Between the symbols of these data types a certain amount of correlation is present. To reduce that correlation and therefore to improve the compression efficiency, entropy coding has to be used. This section deals with the application of entropy coding to the motion information coding. The section starts with entropy coding fundamentals relevant to this application. Then the practical realisation of motion information entropy coding is given.

4.1 Entropy Coding Fundamentals

Generally, symbols at the input of motion information entropy encoder can be modelled as a binary source, which generates a stochastic process

$$\mathbf{X} = [X_1, X_2, \ldots, X_N], \tag{1.4}$$

where X_i is a random variable occurring at time instance $i = \{1, 2, \ldots, N\}$ and takes a value of either zero or one. Let message $\mathbf{x} = [x_1, x_2, \ldots, x_N]$ be a realisation of such stochastic process, where x_i is the realisation of the corresponding binary random variable. The minimum theoretical number of bits required to encode this message is determined by self-information, $I(x_1, x_2, ..., x_N)$, where

$$I(x_1, x_2, ..., x_N) = -\log_2 p(\mathbf{x}) = -\log_2 p\left(x_1, x_2, ..., x_N\right). \tag{1.5}$$

Here, $p\left(x_1, x_2, ..., x_N\right)$ is defined by the joint probability mass function, $\Pr\left[\left(X_1, X_2, ..., X_N\right) = \left(x_1, x_2, ..., x_N\right)\right] = p\left(x_1, x_2, ..., x_N\right)$, $\left(x_1, x_2, ..., x_N\right) \in \chi^N$, where χ^N is the alphabet (binary in this case) of the stochastic process and N is a positive integer. The minimum average number of bits required to encode the messages generated by stochastic process $\mathbf{X}$ is defined by the entropy, $H(X_1, X_2, ..., X_N)$ [28]:

$$H(X_1, X_2, ..., X_N) = -\sum_{\forall(x_1, x_2, ..., x_N)} p\left(x_1, x_2, ..., x_N\right)\log_2 p\left(x_1, x_2, ..., x_N\right). \tag{1.6}$$

A common interpretation of the entropy described by (1.6), is that the entropy is an average measure of uncertainty in the stochastic process **X**. If this uncertainty is high, lower compression can be achieved on average and vice versa. The equation (1.6) describes the N-th order entropy, which takes into account the probability of the occurrence of messages consisting of N symbols. If the random variables generated by the stochastic process are statistically independent, then the entropy of the source is equal to the sum of each random variable's first-order entropy. In practice, this is rare, as some correlation between symbols usually exists.

Entropy coding of motion information aims at compressing data resulting from motion estimation so that the minimum theoretical average code length, as defined by the entropy, is reached. Indeed, for a specific realisation of a stochastic process, the ideal entropy coder would place the number of bits defined by (1.5) into the output bit-stream; for an infinite number of stochastic process realisations this would result in the average code length defined by the entropy.

In practice it is not easy to achieve the optimal average code length. There are mainly two reasons for this:

- Any realistic entropy coder requires an integer number of bits to encode a message generated by a stochastic process. Since (1.5) generally results in a real number, a loss in compression performance can be expected due to the integer-length limitation of the code. It can be shown that the coding loss is negligible when all input symbols are encoded jointly, as in that case the average optimal per-symbol code length, ℓ^*_N, is determined by the following bounds [28]:

$$\frac{H(X_1, X_2, ..., X_N)}{N} \le \ell^*_N < \frac{H(X_1, X_2, ..., X_N)}{N} + \frac{1}{N}. \tag{1.7}$$

- The statistical distribution of data at the input of entropy coder is usually not known prior to encoding and can vary in time, i.e. the input source is generally a non-stationary stochastic process. To exploit source statistics during the encoding, probabilities of symbols can be estimated either on-the-fly or a predefined probability model can be used. In majority of cases, both methods result in probabilities that do not exactly match the actual probability model of the data that are encoded. This is mainly due to latency in matching the source characteristics, i.e. some time is required for the applied model to adapt to the time-changing source statistics. Thus, a loss in compression performance can be expected due to incorrect probabilities applied during the entropy coding.

As suggested by (1.6) and (1.7), to obtain the optimal average per-symbol code length, the entire message generated by the stochastic process has to be encoded at once. In other words, the input symbols of which the message consists have to be blocked and encoded jointly. This is impractical from both transmission and coding point of view. First, in many applications streaming is required and therefore video content needs to be transmitted incrementally. Blocking input symbols can introduce severe coding delays. Second, in practice it is very difficult to estimate the probability of the occurrence of the entire message.

The two problems described in the previous paragraph can be tackled by applying the chain rule for entropy. The chain rule suggest that joint entropy of random variables $X_1, X_2, ..., X_N$ can be expressed as a sum of conditional entropies in time instances $i = 1, 2, \ldots, N$:

$$H(X_1, X_2, ..., X_N) = \sum_{i=1}^{N} H(X_i \mid X_{i-1}, X_{i-2}, ..., X_1),$$

where

$$\begin{aligned} &H(X_i \mid X_{i-1}, X_{i-2}, ..., X_1) = \\ &\quad - \sum_{\forall(x_1, x_2, ..., x_i)} p\left(x_1, x_2, ..., x_i\right) \log_2 p\left(x_i \mid x_{i-1}, x_{i-2}, ..., x_1\right). \end{aligned} \tag{1.8}$$

From the above analysis it is clear that entropy coding can be broken into two phases: probability modelling and actual coding. Since the source statistics are usually not known prior to encoding, in the first phase, conditional probabilities have to be estimated on the fly, i.e. during the encoding. Efficient estimation of these probabilities is critical, as it directly influences the compression performance. Moreover, using the whole set of random variables $\{X_{i-1}, X_{i-2}, ..., X_1\}$ for estimating conditional probabilities, as specified in (1.8), can result in some negative effects. Specifically, in the case of finite sources such as motion information in video coding, the resulting data may be expanded, instead of compressed. Therefore identifying a proper subset of random variables that will be used for the estimation of conditional probabilities is also a critical issue in this approach. The second phase consists of actual mapping of the input symbols to the output code, where the idea is to represent the symbols with high probability of the occurrence with smaller number of output bits than the symbols with the low probability of occurrence. This can be done by arithmetic coding, a popular approach to practical entropy coding, which can encode input symbols incrementally without the restriction that each input symbol must be represented by an integer number of bits at the output. As it is used in encoding of motion information, binary arithmetic coding is briefly described in the following section.

4.2 Binary Arithmetic Coding

Arithmetic coder is a practical realisation of entropy coding, which can asymptotically reach the theoretical minimum length of the code assigned to the input sequence of symbols, given that the probabilities of the input symbols are known for each symbol entering the arithmetic encoder. The main idea behind arithmetic coding is to represent the message to be encoded by an interval whose width is proportional to the probability of the occurrence of the message. The longer the message is, smaller is its probability of the occurrence and therefore of the interval used for its representation. In this section, the basic operations performed only during arithmetic encoding are explained, as the decoder generally follows operations inverse to the ones performed at the encoder.

Without loss of generality, let the process defined by (1.4) be an independent and identically-distributed (i.i.d.) stochastic process, where the random variable X_i can obtain values from the binary alphabet, i.e. $x_i \in \chi = \{0, 1\}$. At this point it is assumed that the probability of the occurrence of each symbol at the input of arithmetic coder is known. More information about probability estimation will be given in Sect. 4.3. The cumulative distribution function $F(x_i)$ in time instance i is defined as

$$F(x_i) = \sum_{\forall y \le x_i} p(y),$$

where $p(y)$ is defined by probability mass function $\Pr[X_i = y] = p(y)$, $y_i \in \chi$. Assuming that $p(x_i) > 0$ for every $x_i \in \chi$, then $F(x_i) > F(x_i - 1)$ holds. Note that $F(x_i)$ is constant over time i, in case of i.i.d. process.

Encoding starts with a unitary coding interval, [0, 1), which is divided into subintervals, each of which represents one symbol from the source alphabet. The width of a subinterval is equal to the probability of its corresponding symbol. In each subsequent step i of the encoding process, the coding interval can be described by

$$I_{\mathrm{C}}^{i} = \left[B_{\mathrm{L}}\left(I_{\mathrm{C}}^{i}\right), B_{\mathrm{H}}\left(I_{\mathrm{C}}^{i}\right)\right),$$

where $B_{\mathrm{L}}\left(I_{\mathrm{C}}^{i}\right)$ and $B_{\mathrm{H}}\left(I_{\mathrm{C}}^{i}\right)$ denote the lower and upper bound of the coding interval I_{C}^{i} respectively. The bounds can be obtained using the following relations:

$$B_{\mathrm{L}}\left(I_{\mathrm{C}}^{i}\right) = B_{\mathrm{L}}\left(I_{\mathrm{C}}^{i-1}\right) + F\left(x_i - 1\right) \cdot W_{\mathrm{C}}\left(I_{\mathrm{C}}^{i-1}\right),$$
$$B_{\mathrm{H}}\left(I_{\mathrm{C}}^{i}\right) = B_{\mathrm{L}}\left(I_{\mathrm{C}}^{i-1}\right) + F\left(x_i\right) \cdot W_{\mathrm{C}}\left(I_{\mathrm{C}}^{i-1}\right),$$

where $W_{\mathrm{C}}\left(I_{\mathrm{C}}^{i-1}\right) = B_{\mathrm{H}}\left(I_{\mathrm{C}}^{i-1}\right) - B_{\mathrm{L}}\left(I_{\mathrm{C}}^{i-1}\right)$ is the width of the interval I_{C}^{i-1}, $B_{\mathrm{L}}\left(I_{\mathrm{C}}^{0}\right) = 0$ and $B_{\mathrm{H}}\left(I_{\mathrm{C}}^{0}\right) = 1$. Encoding continues iteratively in this fashion until the final symbol has been encoded. In order to decode the message, it is enough to send to the decoder any number within the final interval I_{C}^{N}. Hereafter, this number is called interval tag. Note that the analysis above can easily be extended to non i.i.d. sources by considering conditional cumulative probabilities.

The basic operations of the arithmetic coder described in the above text are usually enhanced in practice because of the following reasons [29]:

- It requires an infinite arithmetic precision. With each encoded symbol, the coding interval becomes narrower. If the message is long enough the width of the interval will eventually become narrower than the smallest number that can be represented by the processor on which the encoding is executed (defined by machine epsilon), and the encoded message will be impossible to decode.
- Even when short messages are encoded and an infinite precision problem can be avoided, the arithmetic encoder still uses floating point operations, which are generally more complex than integer operations. This is especially critical if the employed processor does not support floating point arithmetic.
- Data are not transmitted incrementally; the whole message needs to be encoded before the tag value from the resulting interval is sent to the decoder.

An efficient arithmetic coder that circumvents the inefficiencies listed above uses integer arithmetic with so-called interval rescaling [29]. Instead of the unitary interval, the initial coding interval is set as $[0, 2^W - 1)$, where W represent the desired number of bits for representing the interval. In each step of the encoding process, low and high bounds of the coding interval are represented by integer numbers. Encoding is performed as explained in the above example. However, each time the coding interval is entirely positioned either in the upper or lower half of the initial interval, a bit indicating the corresponding half is sent to the decoder and the coding interval is rescaled (its width is roughly multiplied by factor two). The rescaling is possible because positioning the coding interval in the upper or lower half of the initial interval implies positioning the interval tag in the corresponding half. Therefore, the information about which half contains the interval tag can be sent to the decoder immediately, and the interval can be expanded in order to avoid its constant width reduction. This ensures incremental transmission of the compressed data and avoids the need for infinite precision.

The smallest permitted width of the coding interval in any coding step is determined by the width of probability interval, which is used to represent probabilities of the input symbols. In [29] the probability interval is set as $[0, 2^{W-2} - 1)$. In this case the smallest permitted width of the coding interval in any coding step is 2^{W-2}. This can easily be justified by examining the following example. Consider the width lower than 2^{W-2}, say $2^{W-2} - 1$. If a symbol to be encoded has the smallest possible probability, $1 / 2^{W-2}$, then the width of the new coding interval is determined by multiplication $(2^{W-2} - 1) \times (1 / 2^{W-2})$. It is easy to see that this multiplication will produce a zero value when an integer multiplication is used. Therefore, the encoding has to terminate, as a unique interval tag cannot be assigned to future symbols. Hence, every time when the width falls below one quarter of the input interval (2^{W-2}), the interval is rescaled until its width is higher than the minimum allowed width. Each rescaling due to the critical interval size is recorded and sent to the decoder when a "regular" rescaling occurs (rescaling due to the positioning of the coding interval in lower or in upper half of the initial coding interval).

Although such an arithmetic coder based on integer implementation achieves nearly optimum performance in terms of length of resulting codes, it is still fairly complex. This complexity stems mainly from the need for integer multiplications and divisions while calculating the coding interval for each encoded symbol. The complexity can be reduced by quantising probability values assigned to the symbols in each coding step, i.e. representing the probabilities by state machines where the states are stored in the form of look-up tables. In each subsequent coding step the next probability state is assigned to a symbol based on its state in the current coding step. By using this approach, the width of the coding interval can be approximated for each state; therefore, there is no need to undertake complex multiplication and division. Although this results in a reduced compression performance, the complexity of such arithmetic coders is much lower and, consequently, they are more appropriate for hardware implementations. This approach is used in many image and video coding standards, for example in M-coder implemented in the H.264/AVC standard [30], and its scalable extension SVC, and MQ-coder, implemented in the JPEG 2000 still image coding standard [31].

Until this point it has been assumed that the probabilities of symbols entering arithmetic coding are known. The following section explains how these probabilities can be efficiently estimated.

4.3 Context Modelling

Consider the stochastic process described by (1.4) and let x_i be a realisation of the binary random variable at time instance i, i.e. the symbol at the input of an entropy coder at time instance i. If the sequence of all previously encoded symbols is denoted as $x^i = x_{i-1}x_{i-2}...x_1$, an ideal entropy coder would place $-\log_2 p\left(x_i \mid x^i\right)$ bits into the output bit-stream for encoding x_i. This ideal code length can be asymptotically achieved by the application of binary arithmetic coding. Here, the value of x^i forms the modelling context in which the symbol x_i occurs. The critical issue with this approach is that the conditional probability of the symbol x_i, with respect to its context, is not known prior to encoding and therefore it has to be estimated on-the-fly. However, if all past symbols are directly considered for modelling of conditional probabilities within a context, the probability estimation will certainly fail, as the number of past symbols rises with each encoded symbol. Therefore, in each time instance the modelling context is different from any other context that has occurred in the past. Thus, it is impossible to perform efficient probability estimation within a context.

A general solution that can be used to solve the above problem is to approximate the random process to a Markov process [32] and to consider only a subset of m past symbols during the probability estimation, i.e. a subset $\underline{x}^i = z_k z_{k-1} \ldots z_{k-m+1}$ (m-th order modelling context). Here, the new notation, z_k, is used to denote the past symbol, as symbols of which $\underline{x}^i$ consists do not necessarily form a prefix of x^i [32]. Symbols z_k, which are used in the formation of $\underline{x}^i$, are commonly called context elements and the set of temporal locations (or spatio-temporal locations in case of image and video coding) where they occur is called context template.

Although application of contexts designed in this way can significantly improve compression due to the improved probability estimation, the number of contexts can still be too large, depending on the size of context template. For m binary context elements there will be 2^m contexts. This is a critical issue in case of compression of finite sources, such as video signals, as it may result in a context dilution. Context dilution is a phenomenon which occurs when a number of symbols within given contexts is to low to obtain a good probability estimate. Due to the poor probability estimation in this case, the resulting code can represent an expanded version of the input signal, instead of a compressed one. On the other hand, if too few contexts are used, redundancies between symbols are not exploited properly, which results in less efficient compression. Therefore, it is desirable to optimise the contexts carefully.

Selection of final contexts that are used during encoding can be performed through a context quantization. In other words, probabilities that drive entropy encoding process are estimated as $p(x_i \mid ctx))$, where $ctx = Q(\underline{x}^i)$ represents a context used in encoding/decoding, which is obtained by applying a quantiser Q to the value of sequence of previously encoded symbols, $\underline{x}^i$. Context quantisation aims at merging those contexts in which the source shows similar statistical properties, and thus reducing the number of final contexts.

Contexts quantiser can be optimised either adaptively during the encoding process [33], or off-line and then "hard-coded" into encoder / decoder and used for encoding an arbitrary signal [34]. The first approach intuitively yields in shorter code length associated to a particular source, as contexts are tailored to the underlying source statistics. Although this approach can provide best results in terms of compression, its drawback is an increased encoding complexity. On the other hand using predefined contexts provides a good trade-off between the computational complexity and compression efficiency. Here, contexts can be stored in the form of look-up tables.

A popular off-line approach for selecting optimised context quantiser uses so-called "heuristic context quantization". This means that contexts used for encoding are not obtained automatically through an application of any particular context optimisation algorithm, but rather heuristically, by exploiting an expected efficiency of using different values of context elements in predicting the symbol that is currently encoded. Using heuristic optimisation of contexts is motivated by observation that some neighbouring symbols are expected to show similar correlation to the symbol that is encoded and therefore can be treated jointly. Heuristically optimised contexts are hard-coded into encoder / decoder. Some examples of image / video coders that use this approach are: Context Adaptive Binary Arithmetic Coding (CABAC) [30] used in the H.264/AVC, the EBCOT algorithm [35] used in the JPEG 2000 standard [31], entropy coding in EZBC [36] and 3D-ESCOT [37] algorithms.

Probability estimates in each context can be obtained by using counts of symbols that previously appeared in that context. Following this approach, a common way of obtaining symbol probabilities at each coding instance *i* for binary coding is expressed as:

$$p\left(x_i \mid ctx\right)=\frac{c_{i-1}\left(x_i \mid ctx\right)}{c_{i-1}\left(0 \mid ctx\right)+c_{i-1}\left(1 \mid ctx\right)+2}, \qquad (1)$$

where $c_{i-1}(x_i \mid ctx)$) represent number of occurrences of the symbol x_i in the contexts *ctx* at the time instance i - 1, $c_{i-1}(0 \mid ctx)$) represent the number of occurrences of zeros in the contexts *ctx* at the time instance i - 1 and $c_{i-1}(1 \mid ctx)$) represent the number of occurrences of ones in the contexts *ctx* at the time instance i - 1. This approach to probability estimation is used in the entropy coding of flexible block-based motion models presented in the following section.

4.4 Arithmetic Coding of Flexible Block-Based Motion Models Using Context Modelling

Entropy coding of motion parameters supporting scalable structure coding and precision limited coding (PLC) [12] is depicted in Fig. 15. For each layer of motion structure joint encoding with PLC is performed. The base layer of *q*-th structure layer consist of motion structure related parameters and corresponding precision limited (PL) components of motion vectors. The PL component of a motion vector is obtained by quantisation of full-precision motion vectors. The applied quantisation step depends on desired number of refinement sub-layers (*Ri*), which consist of lower bit-planes of motion vector values. Each refinement sub-layer contains bits from one bit-plane of full-precision motion vector values. For instance refinement layer R_1 contains prediction errors for least significant bits (LSBs) of all motion vector values within the observed motion structure layer *q*. PL values are differentially predicted from already encoded PL values, as explained in Sect. 3.4. The same prediction technique is used also for data in each refinement sub-layer.

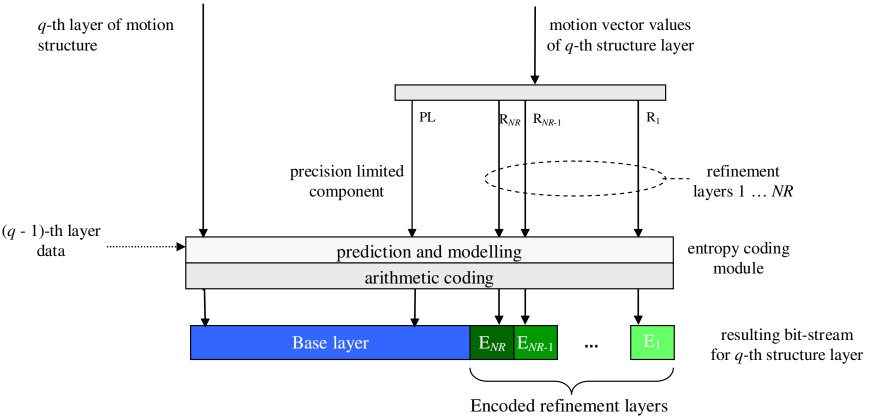

q-th layer of motion structure
motion vector values of q-th structure layer
PL
R_NR
R_NR-1
R_1
precision limited component
refinement layers 1 … NR
(q - 1)-th layer data
prediction and modelling
arithmetic coding
entropy coding module
Base layer
E_NR
E_NR-1
...
E_1
resulting bit-stream for q-th structure layer
Encoded refinement layers

Fig. 15 Entropy coding of motion vectors and motion structure supporting scalability

Arithmetic encoding engine takes the probability estimate of the symbol to be encoded according to the modelling context. The probability model within the context is updated with each occurring symbol, according to (1). Symbols at the encoder input can belong to one of the data types shown in Table 2. For each data type the number of employed contexts and size of the alphabet is shown. All contexts used in entropy coding were determined heuristically and are hard-coded into the encoder / decoder.

Table 2 Data types and coding settings

Data type	Number of contexts	Alphabet size
mv_diff_row	5	m-ary
mv_diff_col	5	m-ary
mv_map	3	binary
mv_ref_diff_row	2	binary
mv_ref_diff_col	2	binary

The *mv_map* data type represents motion tree structure, i.e. partitioning of blocks into sub-blocks. Therefore it contains information if the current block of the current structure layer is divided into the smaller blocks or not. The context modeller uses three contexts determined by the following rules:

$$ctx_{\text{map}} = \begin{cases} 0, \text{ if } \sigma(r_{\text{L}}) + \sigma(r_{\text{U}}) = 0 \\ 1, \text{ if } \sigma(r_{\text{L}}) + \sigma(r_{\text{U}}) = 2 \\ 2, \text{ otherwise} \end{cases},$$

where $\sigma(r_{\text{L}}) = 1$ if the block r_{L} positioned left of the current block r at the same motion tree depth is further divided in the current motion structure layer; 0 if not. Similarly,

$\sigma(r_U) = 1$ if the block r_U positioned above of the current block r at the same motion tree depth is further divided in the current motion structure layer; 0 if not.

The *mv_diff_row* and *mv_diff_col* data types represent the difference between the predicted and the actual precision limited values of motion vectors' vertical and horizontal component respectively. The alphabet size for these data types depends on number of refinement bits. First the values of *mv_diff_row* and *mv_diff_*col are processed in order to be suitable for binary arithmetic coding. Processing and encoding is performed for both data types in the same way. Following steps show processing / encoding order for *mv_diff_row* value of the current block *r, mv_diff_row*(*r*):

1. Significance coding. Information if *mv_diff_row*(*r*) = 0 is encoded. i.e. whether PL component has been correctly predicted from neighbouring PL components.
2. Sign coding. If *mv_diff_row*(*r*) is not zero then a bit indicating sign is encoded.
3. Coding of binarised values. If *mv_diff_row*(*r*) is not zero then the absolute value of *mv_diff_row*(*r*) is binarised and each bin is subjected to entropy coding. A popular binarisation scheme that can be utilised here is the unary binarisation scheme [30]. In this scheme, each unsigned integer *k* is represented by *k* leading bits of value one and a trailing zero bit.

Contexts for *mv_diff_row* data type can be selected in the following way:

$$ctx_{\mathrm{PL},row} = \begin{cases} 0, & \text{significance coding} \\ 1, & \text{sign coding} \\ 2, & \text{coding of binarised non-zero value, first bin} \\ 3, & \text{coding of binarised non-zero value, second bin} \\ 4, & \text{coding of binarised non-zero value, all other bins} \end{cases},$$

Contexts for *mv_diff_col* are selected in the same way.

The binary data types *mv_ref_diff_row* and *mv_ref_diff_col* are used to signal if a vertical and horizontal component of a motion vector refinement bit has been correctly predicted or not. Prediction is performed from the same blocks as for *mv_diff_row* and *mv_diff_col* data types. For the *mv_ref_diff_row* data type the modeller uses two contexts determined by following rules:

$$ctx_{\mathrm{ref},row} = \begin{cases} 0, \text{if } mv\_diff\_row(r) = 0 \\ 1, \text{otherwise} \end{cases}.$$

For *mv_ref_diff_col* data type, context is determined in the same way but considering *mv_diff_col* value instead of *mv_diff_row*.

4.5 Experimental Evaluation of Motion Information Entropy Coding

The multi-component scalable motion and entropy coding scheme presented in the previous section have been integrated in aceSVC. The tests were performed on two popular test sequences: "Basket" (4CIF, 30 Hz) and "Mobile" (CIF, 30 Hz).

Each encoded sequence contained 3 refinement layers and 4 motion structure layers. Layers of the motion structure were selected in such way that layer 0, layer 1 and layer 2 contained 10 %, 20 % and 30 % nodes of the full motion structure nodes respectively. Motion block size of 8 × 8 to 64 × 64 and 1/8-pixel precision were used.

Two tests have been performed. First test measures the compression efficiency for each refinement layer *n* and motion map data with and without employed contexts shown in the previous section. Results are shown in Table 3. Compression ratio (*CR*) is defined as

$$CR = \left(1 - \frac{out\_bits}{in\_bits}\right) \cdot 100\% ,$$

where *out_bits* and *in_bits* are the numbers of compressed bits and uncompressed bits respectively for each data shown in Table 3. It can be seen that the higher compression is obtained with context modelling. Furthermore, compression ratio of lower enhancement layers is lower than compression ratio of higher enhancement layers. Since the lower enhancement layers are used to represent very precise motion vector values, the correlation within these layers is relatively small and therefore high compression is not obtained.

Table 3 Compression efficiency of the binary data types

Data being compressed	Contexts employed [yes/no]	*CR* [%]	
-	-	“Mobile”	"Basket"
Refinement layer *n* = 1	No	1.65	1.97
	Yes	1.81	3.04
Refinement layer *n* = 2	No	15.07	8.72
	Yes	16.76	13.21
Refinement layer *n* = 3	No	35.03	17.40
	Yes	39.92	27.47
Motion map	No	24.11	17.53
	Yes	24.63	25.72

In the second test the influence of the removal of the refinement layers on the decoding quality is examined. Results are shown in Table 4. In this test, the sequences were decoded with full texture in order to avoid the influence of rate-distortion method applied in the real scenarios and to show the influence of lossy received motion information on the decoding quality. However, the MSE caused by loss in motion information can be almost linearly added to the MSE caused by loss of texture data. Results for removal of refinement layers of motion vector values ("bit-planes") are displayed for full motion structure. On the other hand, the results for removal of motion structure layers ("map"), all bit-planes of motion vector values have been decoded. Finally, in this experiment the refinement layers from motion vectors were removed when only the coarsest layer of motion structure remained to obtain combined scalability. It can be seen that that removing the motion structure layers gives finer scalability than removing of refinement bits. However, it is also shown that even finer granularity of motion information can be achieved by applying both strategies.

Table 4 PSNR after removing different layers

Sequence	Type of scalability	PSNR$_Y$ after one removed layer [dB]	PSNR$_Y$ after two removed layers [dB]	PSNR$_Y$ after three removed layers [dB]
"Mobile"	bit-plane	38.76	30.04	23.25
	map	43.45	41.60	39.18
	combined	35.64	29.21	22.96
"Basket"	bit-plane	40.30	32.02	25.32
	map	42.68	38.17	31.83
	combined	31.24	28.92	24.47

5 Conclusions

In this chapter an important part of the efficient video codec has been presented - a flexible block-based motion model. High flexibility is achieved through the application of variable-size blocks for representation of motion in consecutive frames. The employed model provides optimised balance between the efficiency of temporal prediction and amount of data spent on encoding of motion information. Consequently, the model automatically adapts to motion activity in different temporal segments of the input video. It has been shown that the presented model can be efficiently applied to the scalable coding of high-definition sequences.

To support high compression and scalability of motion information the efficient prediction and entropy coding schemes have been designed. In the presented approach the motion vectors are predicted from the neighbouring blocks and the prediction residual is encoded using arithmetic coding. Probability estimates of the symbols at the input of arithmetic encoder are improved by the application of a specially designed context modelling algorithm. Moreover, the presented coding scheme supports scalable motion information coding. Scalability is achieved through careful organisation of motion structure layers and motion vectors' refinement layers into the scalable bit-stream. The influence of different motion data layers' removal on the decoded video quality has been also studied. It has been shown that removing a large number of nodes in the motion structure performs better than removing large number of refinement bits. However, it is also shown that finer granularity of motion information can be achieved by applying both strategies.

References

[1] Mrak, M., Izquierdo, E.: Scalable video coding. In: Furht, B. (ed.) Encyclopedia of Multimedia. Springer, Heidelberg (2006)

[2] Coding of audiovisual objects - Part 10: Advanced video coding, ISO/IEC 14 496-10 (ITU-T Recommendation H.264)

[3] Hsiang, S.-T., Woods, J.W.: Embedded video coding using invertible motion compensated 3-D subband/wavelet filter bank. Signal Processing: Image Communication 16(8), 705–724 (2001)

[4] Sprljan, N., Mrak, M., Abhayaratne, G.C.K., Izquierdo, E.: A scalable coding framework for efficient video adaptation. In: Proc. 6th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2005) (April 2005)

[5] Secker, A., Taubman, D.: Lifting-based invertible motion adaptive transform (LIMAT) framework for highly scalable video compression. IEEE Transactions on Image Processing 12(12), 1530–1542 (2003)

[6] Andreopoulos, Y., Munteanu, A., Barbarien, J., van der Schaar, M., Cornelis, J., Schelkens, P.: In-band motion compensated temporal filtering. Signal Processing: Image Communication 19(7), 653–673 (2004)

[7] Valentin, V., Cagnazzo, M., Antonini, M., Barlaud, M.: Scalable context-based motion vector coding for video compression. In: Proc. Picture Coding Symposium 2003 (PCS 2003), April 2003, pp. 63–70 (2003)

[8] Mrak, M., Sprljan, N., Abhayaratne, G.C.K., Izquierdo, E.: Scalable generation and coding of motion vectors for highly scalable video coding. In: Proc. 24th Picture Coding Symposium 2004 (PCS 2004) (December 2004)

[9] Mrak, M., Sprljan, N., Izquierdo, E.: Evaluation of techniques for modeling of layered motion structure. In: Proc. IEEE International Conference on Image Processing (ICIP 2006) (October 2006)

[10] Sprljan, N., Mrak, M., Izquierdo, E.: Motion driven adaptive transform based on wavelet transform for enhanced video coding. In: Proc. 2nd International Mobile Multimedia Communications Conference (Mobimedia 2006) (September 2006)

[11] Zgaljic, T., Sprljan, N., Izquierdo, E.: Bit-stream allocation methods for scalable video coding supporting wireless communications. Signal Processing: Image Communication 22(3), 298–316 (2007)

[12] Zgaljic, T., Mrak, M., Sprljan, N., Izquierdo, E.: An entropy coding scheme for multi-component scalable motion information. In: Proc. 31st IEEE International Conference on Acoustics, Speech, and Signal Processing 2006 (ICASSP 2006), May 2006, vol. 2, pp. 561–564 (2006)

[13] Sprljan, N., Mrak, M., Zgaljic, T., Izquierdo, E.: Software proposal for Wavelet Video Coding Exploration group, ISO/IEC JTC1/SC29/WG11/MPEG2005, M12941, 75-th MPEG Meeting (April 2006)

[14] Zgaljic, T., Ramzan, N., Akram, M., Izquierdo, E., Caballero, R., Finn, A., Wang, H., Xiong, Z.: Surveillance centric coding. In: Proc. International Conference on Visual Information Engineering (VIE 2008) (July 2008)

[15] Wan, S., Izquierdo, E., Yang, F., Chang, Y.: End-to-end rate-distortion optimized motion estimation. In: Proc. IEEE International Conference on Image Processing (ICIP 2006), October 2006, pp. 809–812 (2006)

[16] Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s, Part 2: Video, ISO/IEC, ISO/IEC 11 172-2

[17] Generic coding of moving pictures and associated audio information - Part 2: Video, ITU-T and ISO/IEC JTC1, ITU-T Recommendation H.262 - ISO/IEC 13818-2 (MPEG-2)

[18] Video codec for audiovisual services at p 64 kbit/s, ITU-T Recommendation H.261

[19] Video coding for low bit rate communication, ITU-T, Recommendation H.263

[20] Luo, L., Wu, F., Li, S., Xiong, Z., Zhuang, Z.: Advanced motion threading for 3D wavelet video coding. Signal Processing: Image Communication 19(7), 601–616 (2004)

[21] Chan, M.H., Yu, Y.B., Constantinides, A.G.: Variable size block matching motion compensation with application to video coding. IEEE Proceedings, Pt. I 137, 205–212 (1990)

[22] Choi, S.-J., Woods, J.W.: Motion-compensated 3-D subband coding of video. IEEE Transactions on Image Processing 8(2), 155–167 (1999)

[23] Wien, M.: Variable block-size transforms for H.264/AVC. IEEE Transactions on Circuits and Systems for Video Technology 13(7), 604–613 (2003)

[24] Servais, M., Vlachos, T., Davies, T.: Motion Compensation using Variable -Size Block-Matching with Binary Partition Trees. In: Proc. of IEEE International Conference on Image Processing, Genova (September 2005)

[25] Kamolrat, B., Fernando, A., Mrak, M., Kondoz, A.: Flexible motion model with variable size blocks for depth frames coding in colour-depth based 3D video coding. In: Proc. of IEEE International Conference on Multimedia & Expo. (ICME 2008) (June 2008)

[26] Wiegand, T., Schwarz, H., Joch, A., Kossentini, F., Sullivan, G.J.: Rate-constrained coder control and comparison of video coding standards. IEEE Transactions on Circuits and Systems for Video Technology 13(7), 688–703 (2003)

[27] Chen, P., Woods, J.W.: Bidirectional MC-EZBC with lifting implementation. IEEE Transactions on Circuits and Systems for Video Technology 14(10), 1183–1194 (2004)

[28] Cover, T.M., Thomas, J.A.: Elements of Information Theory. John Wiley and Sons, Chichester (1991)

[29] Witten, I.H., Neal, R.M., Cleary, J.G.: Arithmetic coding for data compression. Communications of the ACM 30(6), 520–540 (1987)

[30] Marpe, D., Schwarz, H., Wiegand, T.: Context-based adaptive binary arithmetic coding in the H.264/AVC video compression standard. IEEE Transactions on Circuits and Systems for Video Technology 13(7), 620–636 (2003)

[31] Taubman, D., Marcellin, M.W.: JPEG2000 image compression: fundamentals, standards and practice. Kluwer Academic Publishers, Dordrecht (2002)

[32] Wu, X.: Context quantization with Fisher discriminant for adaptive embedded wavelet image coding. In: Proc. Data Compression Conference (DCC), March 1999, pp. 102–111 (1999)

[33] Mrak, M., Marpe, D., Wiegand, T.: A context modeling algorithm and its application in video compression. In: Proc. IEEE International Conference on Image Processing (ICIP) (September 2003)

[34] Zgaljic, T., Mrak, M., Izquierdo, E.: Optimised compression strategy in wavelet-based video coding using improved context models. In: Proc. IEEE International Conference on Image Processing 2007 (ICIP 2007) (September 2007)

[35] Taubman, D.: High performance scalable image compression with EBCOT. IEEE Transactions on Image Processing 9(7), 1158–1170 (2000)

[36] Hsiang, S.T.: Highly scalable subband / wavelet image and video coding, PhD thesis, Rensselaer Polytechnic Institute, Troy, New York (January 2002)

[37] Li, S., Xu, J., Xiong, Z., Zhang, Y.-Q.: Three dimensional embedded subband coding with optimal truncation (3D-ESCOT). Applied and Computational Harmonic Analysis 10, 290–315 (2001)

Video Transcoding Techniques

Sandro Moiron, Mohammed Ghanbari, Pedro Assunção, and Sérgio Faria

Abstract. This chapter addresses the most recent advances in video transcoding and processing architectures for multimedia content adaptation. The first section provides an overview of the state of the art, where video transcoding is described as a key solution to enable seamless interoperability across diverse multimedia communication systems and services. Then, a detailed analysis of relevant transcoding functions is presented, addressing their implementation, coding performance and computational complexity. The different processing architectures that can be used for video transcoding are also described, according to their respective functionalities and application scenarios. Other solutions capable of providing adaptation of coded video to heterogeneous environments, such as scalable video coding and multiple description video coding are also discussed and their main advantages and disadvantages are highlighted. A case study is also presented to illustrate the need and effectiveness of video transcoding in modern applications.

1 Introduction

Video transcoding is a process of converting a compressed video stream into another compressed video stream. The set of differences between the original and the transcoded streams define the type of transcoding, which is intrinsically related to the target application where such function is necessary. The need for video transcoding is a consequence of an increasing diversity in communications technology along with significant changes in the way multimedia is consumed by users. Therefore, this is mainly an adaptation process of coded video in order to cope with heterogeneous communication environments, either networks, user terminals, applications

Sandro Moiron and Mohammed Ghanbari,
School of Computer Science and Electronic Engineering, University of Essex,
Wivenhoe Park, Colchester. CO4 3SQ. United Kingdom
e-mail: {smoiro,ghan}@essex.ac.uk

Pedro Assunção, Sérgio Faria
Instituto de Telecomunicações, Polytechnic Institute of Leiria, Campus 2, Morro do Lena - Alto do Vieiro - 2411-901 Leiria - Portugal
e-mail: {assuncao,sfaria}@estg.ipleiria.pt

M. Grgic et al. (Eds.): Rec. Advan. in Mult. Sig. Process. and Commun., SCI 231, pp. 245–270.
springerlink.com

and services or even user demand when interactivity is allowed. A generic transcoder is obtained by a simple cascade of a decoder, an optional processing system and an encoder. The decoder produces a sequence of uncompressed pictures which may undergo some kind of processing through the optional system (e.g., noise filtering, resolution reduction, etc.), and finally the resulting video signal is reencoded using a new set of coding parameters, and possibly another type of encoding format. However, such a scheme completely ignores the encoding information embedded in the original bitstream, which is usually the result of smart rate-distortion decisions, aiming at encoding each block with the highest possible efficiency. Also, the use of independent cascaded equipment is in general more expensive than standalone transcoders. Therefore the generic transcoding system has the potential to evolve into more efficient architectures and processing algorithms by merging decoding/encoding functions and optimising each combined system according to its target application.

2 Context of Video Transcoding

In the past, video transcoding was mostly investigated to cope with network bandwidth constraints [3, 36], which is also referred to as transrating, since its main purpose is to modify a compressed video stream to a lower rate. Such type of processing on a coded video stream may be done either with the purpose of matching static network constraints, i.e., Constant Bit Rate (CBR) [5, 21], or dynamic bitrate conversion [4, 15] to cope with Variable Bit Rate (VBR) transmission. This is the case, for example, of packet networks and joint transcoding of several video streams for statistical multiplexing into a constrained transport stream [38, 42]. The various processing architectures devised for this type of transcoding have a common objective of producing efficient video streams at minimum complexity [40]. Another network-level application of video transcoding is to adapt error protection mechanisms embedded in a coded bitstream, in order to increase their robustness to a communication channel with rather different error characteristics than those initially foreseen at the first encoding. This is the case, for example, of video delivery over wired high quality networks (e.g. optical) where the last mile is comprised of error-prone wireless channels requiring stronger protection. Such video transcoders combine Unequal Error Protection (UEP) in the original video stream by using error resilience coding tools and bitrate control imposed by network constraints [14, 30].

The recent evolution of multimedia services and applications led to an increasing number of video display formats, namely the spatio-temporal resolutions. From full high definition format, targeted for large screens, to sub-QCIF (Quarter Common Intermediate Format), for small handheld devices, there are nowadays a plethora of different terminals willing to consume the same content. In order to avoid having multiple coded versions of the same source material, video transcoders may be used to match high quality and high resolution video streams to specific terminal constraints. This type of transcoding includes service adaptation such as conversion from High Definition Television (HDTV) to Standard Definition Television (SDTV)

[43] or to mobile video [6, 34]. Efficient methods for down-conversion of compressed video, including fast computation of motion information and coding decision modes have been devised in [44, 47]. In [39] an object based approach to video transcoding is proposed to deal with object based visual content. This approach is significantly different from others because it may use metadata information with information about the semantic relevance of each object. A greater flexibility can be achieved with such type of content since in addition to all parameters that are usually processed in natural video, there is another degree of freedom for adaptation at object level to comply with different network conditions or user requirements.

The increasing number of video coding standards has also been a driving force for developing transcoding technology. In fact, whenever a new standard appears on the market, the problem of interoperability with the other existing standards comes with it because it is not viable to develop any sustained multimedia business based on a single standard. This rises the need for transcoding between different standards, which is referred to as heterogeneous transcoding [31]. Since different standards still use some common techniques, eg., motion estimation/compensation, it is always possible to reuse some information from the original video stream to develop optimised processing architectures [11]. Several transcoding techniques have been proposed in the recent past for most of the available video coding standards. For example, transcoding from MPEG-2 to H.263 [18], H.263 to H.264/AVC *(Advanced Video Coding)* [45], MPEG-2 to H.264/AVC [19, 29] and from H.264/AVC to MPEG-2 [10, 24] have been thoroughly investigated for achieving high efficiency at reduced computational complexity. Another dimension exploited in transcoding techniques is the transform domain operation whenever this adds some advantage to the system, either in lowering the transcoding complexity, transcoding delay or increasing the compression efficiency [32, 35].

Video transcoding is also useful in content adaptation, and customisation at the application level. This is the case where a service provider wants to insert a logo into a compressed video stream [27] or introduce a Picture-in-Picture (PIP) [25]. Also, in order to comply with diverse user requirements, transcoding has been found as an efficient solution to either meet home network conditions [13, 23] or user specific needs such as interoperability with different storage formats [9, 20]. More recently the range of transcoding research topics have broaden by including more video content related variables, such as the user context [12, 46], merging different types of web content [17] or making use of proxy caching [26, 28]. In [1] the authors highlight research directions, some of them are still hot topics since there are a number of issues that need further investigations. The remaining of this chapter offers a deeper analysis of the most relevant video transcoding techniques currently used for various applications described above.

3 Transcoding Architectures

As already pointed out, the simplest transcoding architecture consists of a decoder-encoder cascade, where transcoding is performed by fully decoding the input video,

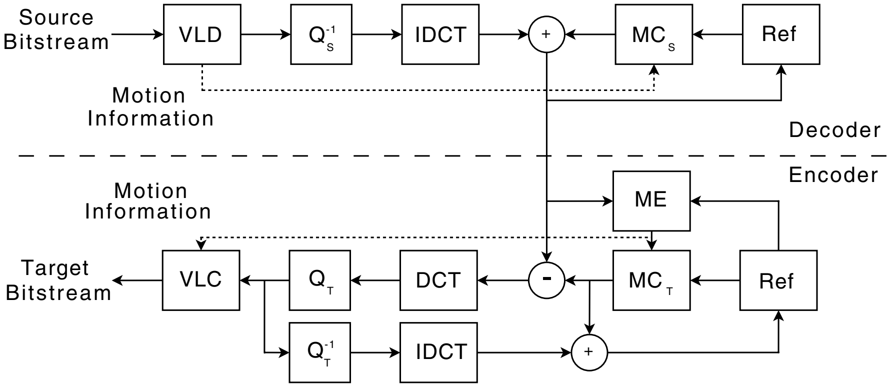


Fig. 1 Cascade video transcoder

up to the pixel domain and reencoding, as illustrated in figure 1. The flexibility of this simple architecture allows modifying the video format without introducing significant distortion in the image quality, therefore, it can be used as a benchmark transcoder for comparison of the performance of other architectures. To simplify the computational complexity several techniques have been developed such that, while they can maintain the quality as close as possible to the benchmark cascade transcoder, the computational complexity is greatly reduced. It should be emphasised that the modifications introduced by a transcoder should result in a minimal video quality loss, as it was coded by a video encoder.

In this architecture, the source bitstream is firstly entropy decoded by the variable length decoder (VLD). The decoded Discrete Cosine Transform (DCT) coefficients are then inversely quantised (Q_s^{-1}) and transformed (IDCT), producing a copy of the original coded pixels. For the intra coded images, these pixels represent the final reconstructed image and for inter coded images, the decoded pixels represent the residual information, which are added to the respective reference frame after proper motion compensation routines (MC). Then this is stored in the decoded frame buffer (Ref) to keep a local copy of the frame. The resulting uncompressed video is fed to be re-encoded with new coding parameters defined for the target bitstream.

3.1 Open-Loop Architecture

The open-loop architecture, as illustrated in figure 2, is the fastest and simplest form of video transcoding compared to other architectures. In this transcoder, the bitstream is passed through the entropy decoder to extract the quantised DCT coefficients, the macroblock type information and the residual prediction parameters. After the entropy decoding, the DCT coefficients are inverse quantised. The subsequent requantisation of the DCT coefficients with a large quantisation step size, forces more coefficients to zero, reducing the number of bits needed to reencode it. The requantiser step size Q_T is dictated by the target bitrate. Other method

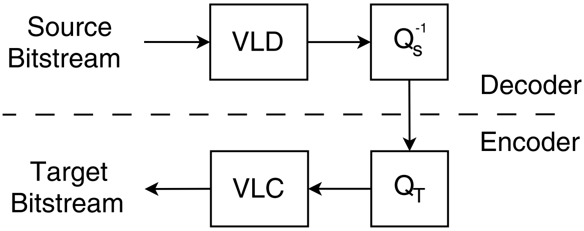


Fig. 2 Open-loop transcoder

commonly used is the high order DCT coefficient discard [36], applied to each macroblock to fine tune the bitrate. These two methods should be used with great care to achieve a good trade-off between distortion and bitrate reduction. This architecture is suitable for intracoded pictures like Motion JPEG.

Open-loop systems are commonly associated with picture drifts, mainly caused by the removal of high-frequency DCT coefficients in the residual information, producing error accumulation in the decoder. To avoid drift error, the image reconstruction at the decoder must be similar to the encoding loop of the encoder. Since this open loop architecture removes some coefficients and does not compensate for their loss, the mismatch between the encoder and the decoder predictions introduces drift error. The quality loss due to drift can generate errors in predicted frames and accumulates in the loops until a non-predicted picture is decoded, setting the drift error to zero.

Despite these disadvantages, for interframe coded pictures with more frequent I-pictures, the open loop architecture might be a good choice when low complexity is the main issue, and some loss in the image quality is tolerated.

3.2 Closed-Loop Architecture

This type of architecture represented in figure 3 is characterised by the ability to remove the mismatch between the residual and the predicted image, overcoming the problems of the open-loop architecture. The main difference between the closed-loop architecture of figure 3 and the cascaded decoder-encoder is the reconstruction loop operating in the pixel domain. While the former has only one DCT/IDCT pair, the latter needs two pairs of DCT/IDCT to operate.

Although this closed-loop architecture assumes the motion compensation as a linear function, such approximation does not allow significant loss in the final quality [2]. Such closed-loop architecture minimises the drift distortion inherent to the open-loop architecture. The resulting image quality of the closed loop architecture is significantly higher than that of the open loop, as well as its computational complexity.

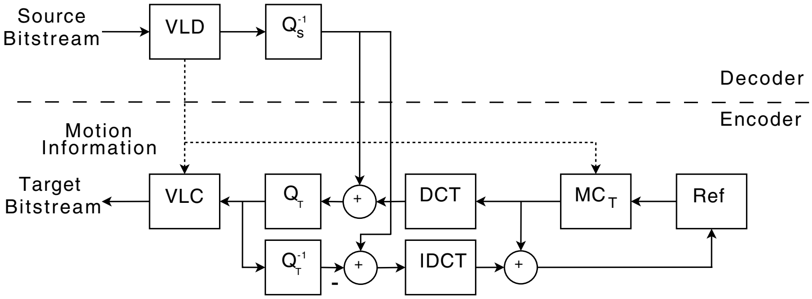


Fig. 3 Closed-loop transcoder

4 Homogeneous Transcoding

In homogeneous architectures, the multimedia content is recompressed within the same compression standard. This transcoding type provides several functions, including adjustment of bit rate and picture format conversion, where the video characteristics such as spatio-temporal resolutions or bitrates are changed. Such type of conversion can be performed by several techniques, as described in the following sections.

Moreover, multimedia communications over heterogeneous networks may use transcoding in order to match the particular constraints of each channel/user. Each channel may have different bandwidth limitations and different target decoders, for instance Digital Video Broadcasting - Terrestrial (DVB-T) and Handheld (DVB-H) terminals.

4.1 Bitrate Reduction

The use of a transcoder for transrating is one of the first transcoding applications, namely for matching network bandwidth resources of compressed video rates [4]. The bitrate reduction, while maintaining the spatio-temporal resolution, can be performed by using four different techniques [36]:

- **Coefficient truncation** - This technique relies on the unequal energy distribution of the DCT block coefficients along the frequency spectrum [15]. Since most of the coefficients' energy is concentrated at the low frequency band, the high frequency coefficients have minimal impact on video quality. Such scheme avoids inverse quantisation and requantisation, but needs to be carefully implemented, otherwise blocking artifacts will be visible in the reconstructed image. Figure 4 illustrates an architecture where the input bitstream is parsed and some coefficients are removed to match the target bitrate.
- **Requantisation** - Quantisation is the main tool to perform bitrate control during encoding, by varying the quantisation step to match the target bitrate. The most

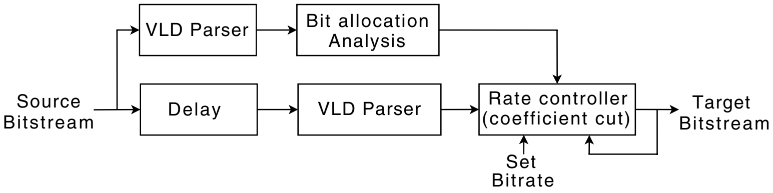


Fig. 4 Coefficient discard transcoder architecture

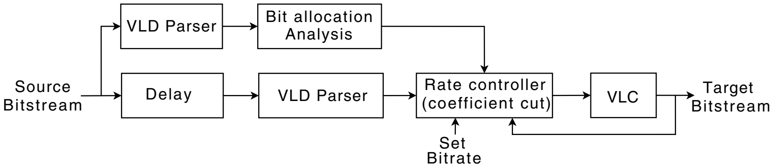


Fig. 5 Coefficient requantisation transcoder architecture

common method for bitrate reduction is to increase the quantisation step size, resulting in a higher compression ratio, caused by the decrease of the number of representation levels of the transformed coefficients. This also increases the number of zero coefficients. Thus, requantisation decreases the number of symbols to be encoded, increasing the compression. Due to its characteristics, this mechanism is also suitable to control the bitrate [48]. Figure 5 illustrates the transrating process, where the coefficients are requantised and entropy coded.

- **Reencoding with reuse of motion vectors and mode decisions** - This technique reencodes the video by reusing the original motion vectors and mode decisions embedded in the bitstream. Compared with the previous techniques it eliminates the error drift, as the reference frames are reconstructed and the residual information recompressed. Due to elimination of the mismatch, the image quality increases in the reference frames but additional computational complexity is added. However, most of the calculation is still avoided, since there is no need for new motion estimation and no mode decision is required. Figure 6 illustrates this transrating process where the *Reencoder* block receives the previously decoded motion vectors and coding modes that are reused, thus avoiding the recalculation.

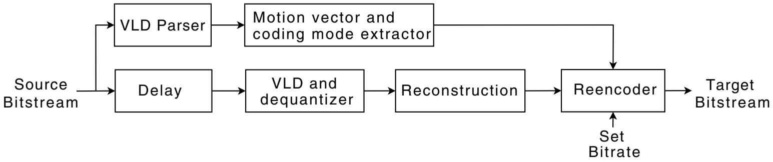


Fig. 6 Reencoding reusing motion vectors and coding modes

- **Reencoding reusing input motion vectors** - This technique is an extension of the previous one, where the coding modes may be changed. For higher bitrate reductions, the motion information overhead becomes too high, which constrains the bitrate, then the residual information is poorly encoded. This technique reuses the motion information, but modifies the coding modes to achieve new optimal coding decisions based on the output bitrate. The additional computational complexity further approximates this technique to the classical cascade decoder-encoder transcoder, but the heavy motion estimation is still avoided. In order to improve the process, the input motion vectors can be used as a starting point to a fine tune refinement with a small increase in computational complexity.

4.2 Spatial Resolution Reduction

Spatial resolution reduction of compressed bitstreams can also reduce the bit rate and can be achieved by using various techniques. In the past few years the diversity of the mobile devices has been significantly increased, as well as their functional characteristics. Such diversity raises the need to develop multimedia content adapted to each of these devices. Since in general content adaptation for mobile devices implies resolution changes, transcoding can also be used for this purpose. Picture resolution reduction is normally done at a 4:1 size reduction, where each horizontal/vertical directions of the pictures is divided by 2. For example, HD to SD and CIF to QCIF conversions where every 4 blocks are converted to a single block. This problem can be solved either in the frequency or spatial domain [1].

- **Frequency domain** - The main issue of down-conversion in the frequency domain is to find efficient ways of merging four IDCT and one DCT. This can be achieved using only the low frequency coefficients from the four original blocks to produce a new resized block. During this operation the input coefficients are filtered by a set of frequency domain filters, to achieve the equivalent operations in the spatial domain. Figure 7 illustrates a frequency domain transcoder architecture (FDTA).

  In order to complete this conversion, the associated motion vectors are also downscaled to meet the new block properties. For computation of the new motion vectors, several algorithms can be used. The most simple one is averaging the four motion vectors and scaling, according to the new frame size. This technique only produces good results when all motion vectors have identical directions, otherwise it will result in a poor approximation. Another method is to select the motion vector based on the prediction error energy. The best motion vector replacement should be selected according to the worst prediction error, that represents the dominant overall prediction error. However, if the residual information is low, then the motion vector should be selected according to the best prediction error. Another well known downscaling scheme is the weighted average, where all four motion vectors are computed to produce a unique equivalent motion vector. Figure 7 illustrates a frequency domain transcoder architecture (FDTA), where the *Converter* contains the motion vector converter/refinement

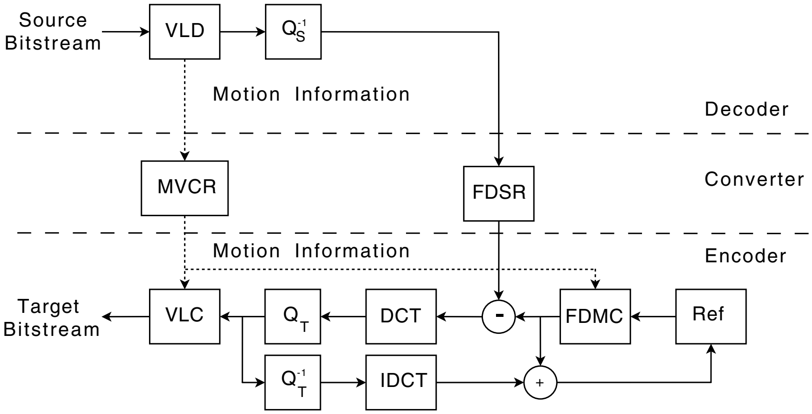


Fig. 7 Frequency domain transcoder architecture

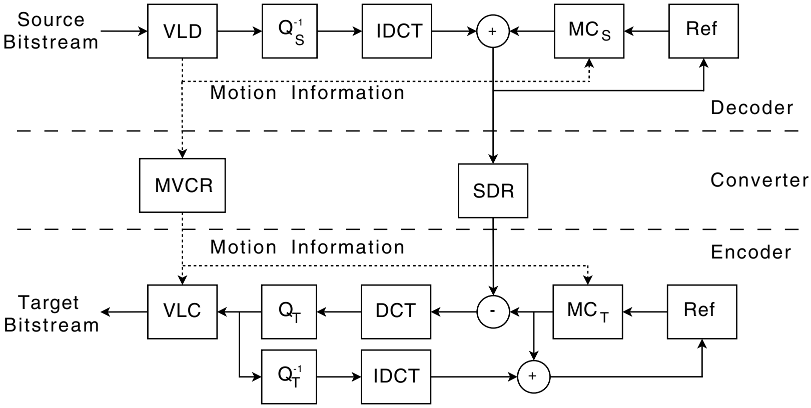


Fig. 8 Spatial domain transcoder architecture

(MVCR) block and the frequency domain spatial reduction (FDSR) block. Since all operations are performed in the frequency domain, the embedded encoder of this transcoder also contains a frequency domain motion compensation (FDMC) block.

- **Spatial domain** - The spatial resolution reduction transcoding in the pixel domain provides high quality drift free transcoding. Spatial domain transcoding is based on a cascade decoder-encoder, where the decoded output is downsampled before entering to the encoder. New motion vectors are computed through a mapping function whereby the input motion vectors are reused in the downconverted signal. When compared with the previous architecture, the spatial domain transcoding architecture (SDTA), in figure 8, presents a more complex operation in the decoder side. The *Converter* contains two functional blocks: the MVCR to convert the motion vectors and another to perform the spatial domain resolution reduction (SDR).

4.3 Temporal Resolution Reduction

Temporal resolution reduction is a very useful transcoder for low power devices, due to its low computational complexity. Reducing the frame rate of a sequence can be simply performed by skipping some frames in the bitstream. However, when removing frames from the bitstream, it is necessary to readjust the decoder buffer controller to avoid underflow and overflow, since the corresponding constraints were initially computed for a different stream. Frame dropping introduces several problems in the motion vector information due to the new broken references. If a dropped frame was used as a reference in a prediction, then the associated motion vector must be recomputed by using only the existing frames after the temporal subsampling. The simplest method is to perform new motion estimation. Nevertheless, such a method is suboptimal because it completely ignores any previously computed information. Therefore, methods that use a combination of motion vector conversion with some form of post-refinement provide better efficiency results. In this process, despite the information provided by the converted motion vector, the additional refinement aims to optimise the prediction accuracy.

5 Heterogeneous Transcoding

Heterogeneous transcoding is used to perform format conversion between different compression standards, such as MPEG-2 to H.263, H.264/AVC to MPEG-2, etc. It can also be combined with some forms of the homogeneous transcoding. This type of transcoding implies syntax conversion between the input and output standards, and may additionally change any format parameters to match the terminal capabilities. Compared with homogeneous transcoders, the heterogeneous transcoders exhibit a significant increase in complexity, due to the asymmetry between the decoder and the encoder. The heterogeneous transcoder illustrated in figure 9 performs a full decompression, up to the pixel domain, to perform the syntax conversion (SC). It may change the picture type, resolution, direction of MVs and frame rate. The decoder stores all the embedded information as motion information, picture and macroblock type to reuse them at the encoder (MVCR), which in turn are used by the encoder module of the transcoder. Whenever the bitstream parameters need to be modified, the converter module operates the necessary adaptation of the corresponding parameters, either spatial or temporal resolution (STR). Due to spatio-temporal subsampling, and different encoding formats of the output video, the decoder and decoder motion compensation tasks in a heterogeneous transcoder are more complex.

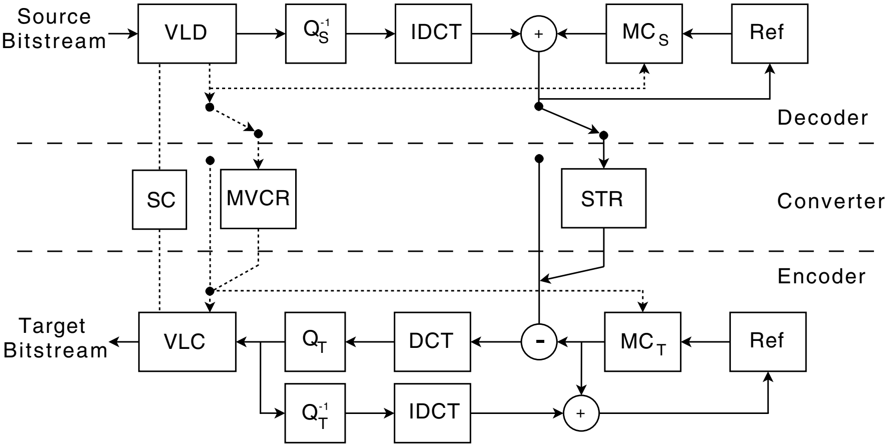


Fig. 9 Hybrid heterogeneous transcoder

Error Resilience

With the ever growing demand of multimedia services and applications for video delivery, the demand for transmitting video over a variety of channels has significantly increased. When using a mixed transport environment of wired and wireless channels, either dedicated or shared by several users, the need for matching the compressed streams to the channel characteristics is particularly relevant. Wireless channels are usually more error prone than the wired ones, affecting therefore the signal quality. In order to deal with such problems, error resilience techniques are often used to adapt the signal to the channel characteristics along the transport links, while maintaining the image quality within an acceptable range.

At the video coding level, increasing the error resiliency can be obtained by using one of the following strategies [41]:

- Localisation - removing the spatial and temporal dependency to constraint the error propagation such that the subsequent blocks are not affected.
- Data partitioning - splitting the coded data into several data blocks and applying unequal protection schemes, according to the image quality impact under a packet loss condition.
- Redundant coding - introducing duplicate versions of those data blocks that can potentially induct higher quality penalty if lost.
- Concealment-driven - enabling error concealment techniques in video encoding, that allow the decoder to use post processing tools to recover from the loss of some data blocks.

6 Other Content Adaptation Methods

In the past few years, there has been an overall explosion of new services and applications in the multimedia field. A wide diversity of network types and terminals with different video display capabilities is growing everyday. Providing quality content users and matching the constraints of both networks and terminals has revealed to be a challenge. Transcoding and scalable encoding are among the most common options to deal with such type of problems. As pointed out before, in the case of transcoding, the video is firstly decoded and then reencoded with the desired parameters in order to match the requirements of the network and/or receiver. This process is repeated every time a new constraint is identified in the network or when the terminal equipment needs it. In the scalable coding, as opposed to transcoding, the video is encoded once using a flexible syntax which allows the receiver to partially decode the bitstream according to several different levels previously defined at the encoding stage.

6.1 Scalable Video Coding

The main goal of scalable video coding (SVC) is to provide on-the-fly adaptation with minimum complexity by allowing truncation of the bitstream according to the necessity of the network or terminal requirements. The built-in scalable concept of SVC is very network friendly due to the seamless conversion that can be provided [37]. This allows content adaptation between different terminals and networks without the use of transcoders, reducing the computational resource requirements for the operators and broadcasters while adapting the content to their networks.

Scalable video coding is implemented through the encoding of several layers which can be combined to achieve a video sequence with higher quality. In this strategy, the video is composed by a base layer and one or more enhancement layers. Furthermore, the layered coding strategy has a hierarchical dependency which means that decoding of any enhancement layer requires previous successful decoding of all lower level layers, as illustrated in figure 10. As a result, the base layer is essential in order to decode the following layers. Therefore, the base layer is usually a very constrained version of the video with very low quality, frame rate and bitrate. This should be enough to guarantee that the base layer will meet the minimum user quality requirements and matches any network and terminal while the content is being delivered. In addition to the base layer, several enhancement layers can be introduced to gradually improve the video quality in several aspects described as follows.

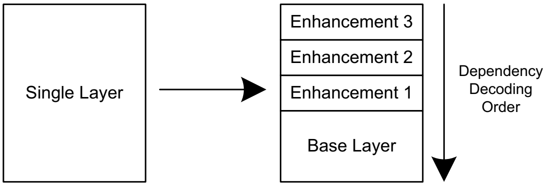


Fig. 10 SVC layers

6.1.1 Types of Scalability

The most common scalability modes are the spatial, temporal and fidelity improvements. Each individual mode is usually included in a different layer to provide a gradual and progressive increase of quality as more layers are decoded. However, several types of scalability can also be combined into a single layer in order to create a single enhancement that includes a significant quality increase.

Temporal Scalability

In temporal scalability the temporal resolution is higher in upper layers (e.g. from 15 fps to 30 fps) by inserting frames in between those of lower layers. The enhancement frames may be coded with reference to either themselves or their prediction may come from the lower layer frames. This, in addition to simplify the implementation of the enhancement frames, allows dropping pictures without affecting the quality of the base layer. Furthermore, the inclusion of B frames in the enhancement layer enables a simple and efficient scalability solution due to the higher compression efficiency usually obtained in B frames.

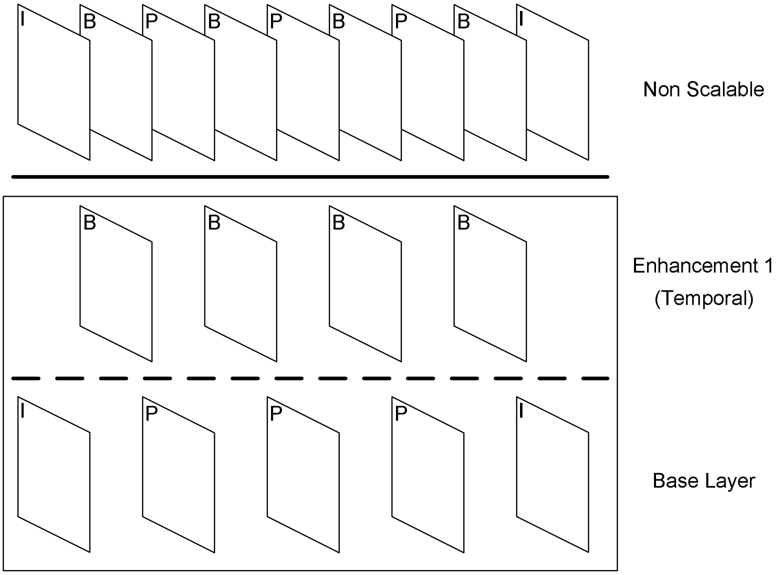


Fig. 11 Temporal Scalability

Spatial Scalability

Spatial scalability provides an efficient coding method to increase the spatial resolution (e.g. from QCIF to CIF) as more layers are decoded. In this case, the base layer is composed by a spatially reduced version of the whole video. This downsize operation can be performed either in the pixel or in the compressed domain in order to save some computational resources at the encoding stage. When encoding the enhancement layers, different procedures are used according to the frame type. The most simple one is the encoding of I frames belonging to the enhanced layer. In this case, the enhanced I frame is coded using an inter-layer predictor which consists of an upsampled version of the temporally coincident frame from the base layer. The subsequent predicted frames can be coded using two different methods. The first method uses the same inter-layer prediction technique while the second method uses the previously coded images belonging to the same layer combined with the common motion estimation algorithms as illustrated in the figure 12.

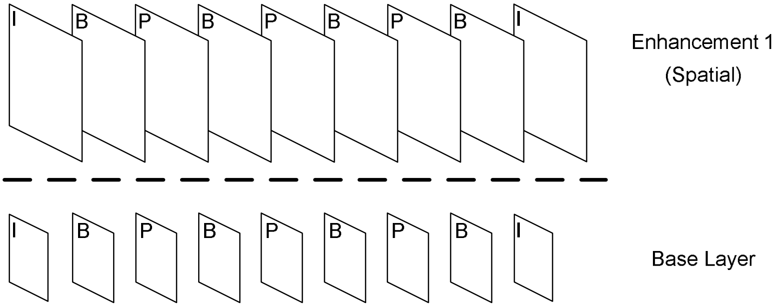


Fig. 12 Spatial Scalability

SNR scalability

The quality scalability, also known as SNR scalability, aims to provide an additional layer with an increased quality. In this scalability type, pictures of the same spatial resolutions are produced at different quality. In this situation, the enhanced layer uses a smaller quantisation parameter which results in a new residual texture corresponding to the distortion previously introduced in the lower layer as illustrated in figure 13.

6.1.2 Applications

SVC is particularly suited to application scenarios which include the transmission over heterogeneous networks and terminals with different hardware capabilities as illustrated in figure 14. The embedded scalability function allows the bitrate reduction by truncating the bitstream in order to accommodate the variable channel conditions of the network. A similar procedure can be adopted to match the terminal requirements such as display size, computational capabilities or network bandwidth.

6.1.3 Multi Layer Transcoding

Besides the various scalability types described above, it may be a better option to convert a single layer coded bitstream to a multi stream. One of the advantages of

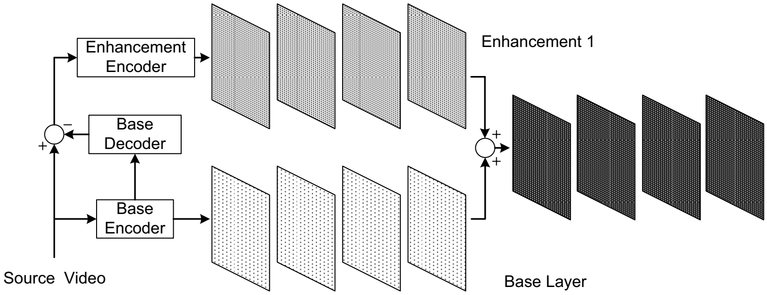


Fig. 13 Quality Scalability

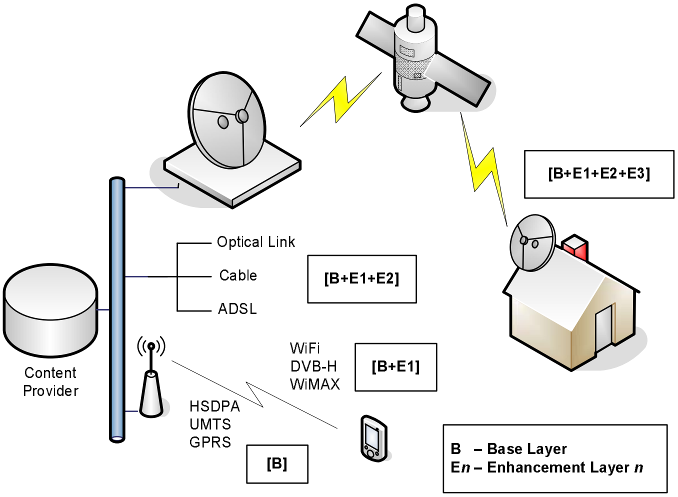


Fig. 14 Application Scenario

multilayer transcoder over the multilayer coding is that, the bitrate of each layer of a multilayer transcoder can be adapted to the underlying network constraints, while those of the scalable coders are predefined and fixed. Furthermore, in a network scenario with several bandwidth constraints, the exclusive use of video transcoder to perform bandwidth adaptation may lead to an excessive number of transcoding operations. The consecutive transcoding operations performed in this scenario does not only introduce significant computational complexity but also introduces a continuous drop in the resulting video quality [48]. Therefore, multilayer transcoding has a better network efficiency than scalable coder. Figure 15 shows an architecture of a three layer SNR scalable transcoder proposed in [33]. However, other variations can also be implemented to include spatial and temporal scalability.

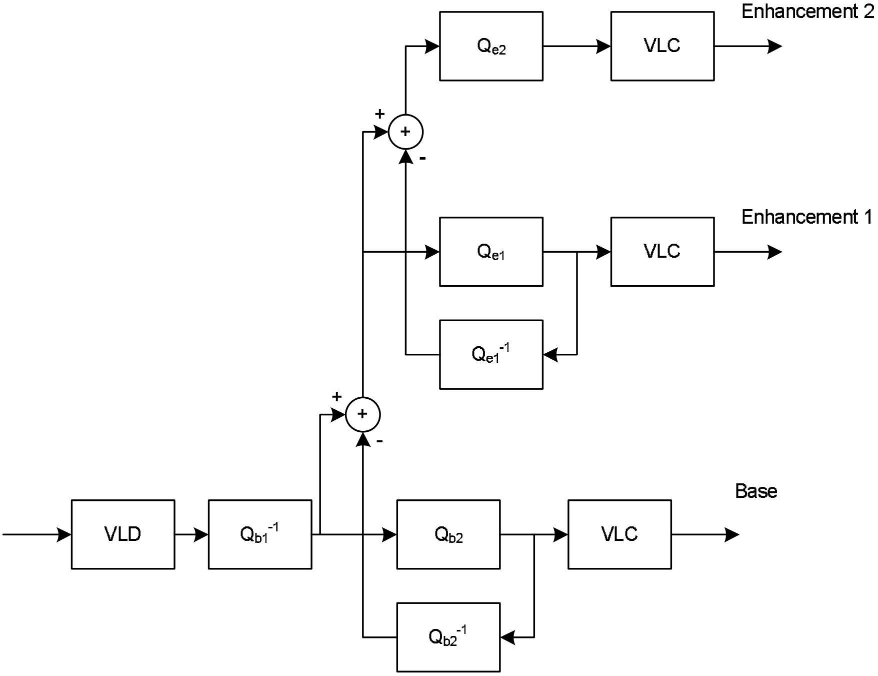


Fig. 15 Multilayer transcoder

6.2 Multiple Description

Multiple description is other technique capable of providing scalability video adaptation in network streaming environments. Despite the network bandwidth scalability provided by SVC, the performance of the overall system can be compromised if the network is subject to high packet loss or disruption of the link. The hierarchical layer dependency of SVC introduces a fragile point where all the upper layers become useless if the base layer is not successfully decoded. In order to solve this problem, multi source systems can be implemented through the distribution of the content over a variety of servers to maximise the number of different routes to the potential receivers. Replicating original versions of the same content over different servers only improves the redundancy in transmission of the content. This is useful in multipath lossy channels where the receiver has a higher probability of always receiving at least one of the streams, which ensures a minimum quality without disruption.

6.2.1 Implementation

The multiple description coding techniques go a step forward by combining resiliency and quality scalability into a single system. In this technique, each server stores a different version of the bitstream. However, there is a major difference when compared with SVC. In multiple description coding each stream can be independently decoded which guarantees a significant improvement against link disruption. Furthermore, combining more than one stream leads to an increase of the video quality.

6.2.2 Applications

Figure 16 shows an application scenario where the use of multiple description introduces significant improvements in video quality and resiliency against errors and link disruption.

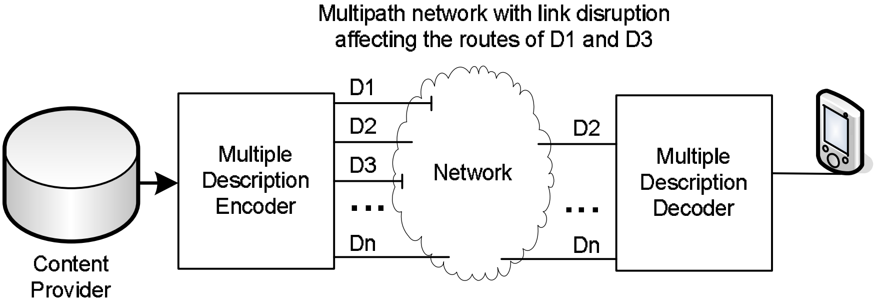


Fig. 16 Application scenario for a multiple description system

6.2.3 Multiple Description Transcoding

Similar to multilayer transcoding, a single bitstream can also be transcoded for multiple descriptions [22][16]. This can be useful in a network with path diversity where the multiple description coding (MDC) is mainly used as a resilience method against link disruption. At the origin, the single layer is transcoded to several multiple descriptions and sent to the network over several network routes to increase its resilience to packet losses and link disruption. At the destination, the streams that arrived successfully are transcoded back to single layer to deliver it to the final receiver over a more reliable network.

7 Case Study of an H.264/AVC to MPEG-2 Video Transcoder

The H.264/AVC standard is currently the most powerful video coding algorithm, offering 2 to 3 time's higher compression efficiency than its MPEG-2 counterpart for the same image quality. However, MPEG-2 is widely used and supported by most end-user equipment, namely by digital television receivers. On the other half, in Europe, it is decided that future high definition (HD) video to be encoded with

H.264/AVC standard. Giving the MPEG-2 Standard Definition (SD) users the opportunity to watch HD programmes, H.264/AVC to MPEG-2 transcoders look a viable tool. Also, simple decoders that may not have the sufficient processing power to decode the heavy processing demanding H.264/AVC may find it easier to decode MPEG-2 format video. Moreover, it provides a peaceful technology migration by industry, providing a smooth transition between different standards as the one illustrated in figure 17.

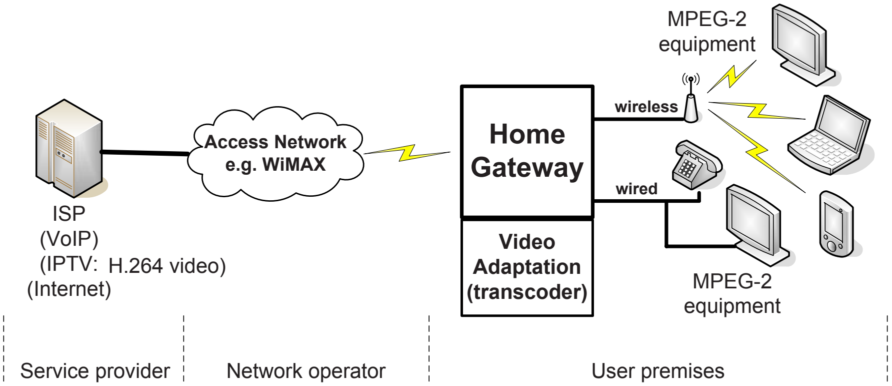


Fig. 17 Application scenario

This case study presents a video transcoding system from H.264/AVC to MPEG-2 Video for the ”Main Profile”, focusing particularly on the development of an efficient transcoding architecture. It exploits the information embedded in the H.264/AVC bitstream, extracting its parameters and converting them into MPEG-2 compatible ones, in order to create an MPEG-2 bitstream using a low complexity approach. The main focus of this study is the conversion of B slices. Other research topic comprises the conversion of prediction types for B slices, even for H.264/AVC macroblocks encoded with tools that allow the motion estimation beyond the picture boundaries.

The algorithm performs the conversion of multi-partitioned H.264/AVC macroblocks into a unique MPEG-2 macroblock, supporting the use of H.264/AVC multiple-reference frame prediction. Additionally, motion estimation beyond picture boundaries in B slices, used by H.264/AVC, is converted to MPEG-2 by exploiting their bidirectionally predicted motion vectors.

Such transcoder is able to achieve a conversion ratio of approximately 95% of the total H.264/AVC inter coded macroblocks. For the tested sequences, the overall computational complexity reductions can go by up to 60%, without significant objective quality penalty.

7.1 Transcoding Architecture

In previous sections it was pointed out that straightforward transcoding architecture is comprised of a cascade of a decoder-encoder. In the context of H.264/AVC to MPEG-2 transcoding this means an H.264/AVC decoder followed by an MPEG-2 encoder, as illustrated in figure 18. As already mentioned, using a generic approach completely discards the H.264/AVC encoding information embedded in the bitstream and fully reencodes the whole video signal without the benefit from the first encoding step. For example, the motion estimation is once more required to be performed using a full search window, which is computationally very intensive and consumes most of the processor resources by the encoder. Therefore, developing an alternative transcoding method is preferable, in order to reduce the computational complexity.

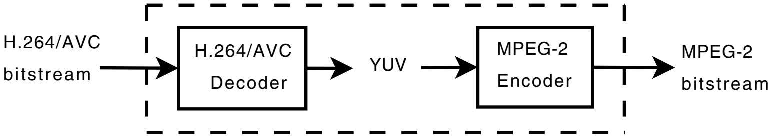


Fig. 18 Cascade transcoding architecture (reference transcoder)

By exploiting the information embedded in the H.264/AVC bitstream it is possible to design a transcoding architecture capable of reducing the MPEG-2 encoding complexity. The coded data contains several parameters which were previously computed in the H.264/AVC encoder, therefore they can be exploited to minimise the computational complexity of the MPEG-2 encoder module, as symbolically shown in figure 19.

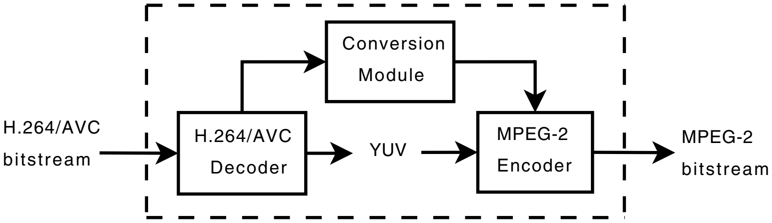


Fig. 19 Modified transcoding architecture

Figure 20 presents a detailed diagram of the cascaded architecture, where in addition to the decoder-encoder modules, the transcoder includes a conversion module. It interacts directly with the MPEG-2 encoder in order to modify its original operation scheme, based on the H.264/AVC parameters, allowing to short-cut the motion

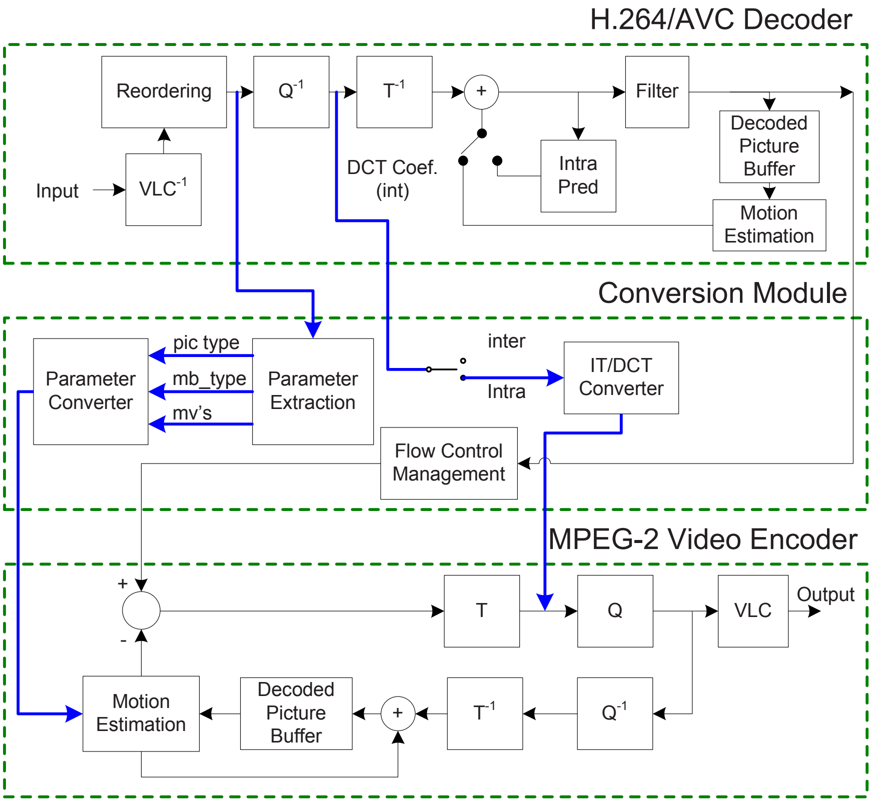


Fig. 20 Detailed transcoding architecture

estimation routines. It is important to refer that only inter frame prediction is addressed.

The motion estimation is one of the most time consuming tasks at the encoder, which makes it a good starting point for improvement [1]. Since both standards share the same coding principals in temporal domain, i.e., a block based approach, the motion estimation process in the transcoder can use most of the H.264/AVC motion vectors as good candidates for MPEG-2 video encoding, saving a great deal of computational resources.

The H.264/AVC encoder, analyses the video sequence in detail and each macroblock is efficiently encoded regarding the best R-D *(Rate-Distortion)* point, thus, the optimal H.264/AVC R-D coding modes may be reused by the MPEG-2 video encoder module. This is done in the *Conversion Module* by reusing the H.264/AVC parameters, such as, macroblock information, picture type, motion estimation, among others. These parameters are then processed and converted to MPEG-2 video format. In the MPEG-2 video encoder this information is used to modify the *Motion Estimation* procedure, thus avoiding a large number of operations. Nevertheless, the parameter extraction procedure can also be applied to intraframe coding by converting the Discrete Cosine Transform coefficients to Integer Transform coefficients [7].

7.2 H.264/AVC Decoder

Among the three main modules of figure 20, the *H.264/AVC Decoder* plays an important role, as it decodes the pre-processed information that will be used by the *MPEG-2 video encoder*. The *H.264/AVC Decoder* is fully implemented, as this is necessary to fully decode the bitstream, in order to generate the reference frames to be used during the encoding process. From the encoder point of view, full decoding is mandatory to reconstruct the source video when there is no one-to-one relevance between the coding modes of H.264/AVC and MPEG-2. The extracted information is given to the *Conversion Module* and consists of parameters like macroblock, partition type, motion vectors, reference images and prediction modes.

7.3 Conversion Module

This module is responsible for all data extraction, analysis and conversion processing. The *Conversion Module* is the interface between the decoder and the encoder, which converts a set of H.264/AVC parameters into MPEG-2 video format. These parameters are analysed and converted wherever they are useful for reducing the computational complexity. Additionally, this module is divided into three submodules: the Parameter Extraction and Conversion, the IT/DCT *(Integer Transform)* coefficient conversion and the Flow Control Management [8].

The *Parameter Extraction* block intercepts the decoding process in several points of the H.264/AVC decoder to extract the embedded parameters needed for the macroblock mode conversion. The extracted parameters are then processed in the *Parameter Converter* block and converted to an MPEG-2 compatible format. They are also used to determine which action will be performed at the encoder. The *Parameter Extraction* module analyses the source macroblock features, according to the restrictions imposed by the MPEG-2 format, and defines the best macroblock mode. The macroblock mode mapping has three operation modes: direct conversion, complex conversion and full recode. The direct conversion mode is used when the source macroblock is fully compatible with MPEG-2 or requires only few adaptations, as it happens with the SKIP and 16×16 modes, respectively. When the input macroblocks are segmented into multiple partitions and/or use long distance reference frames, the operation mode requires a complex conversion approach to transform the embedded information into a reusable format for the encoder. When the input macroblock is incompatible with the MPEG-2 format constraints, and there is no alternative solution, the encoder is free to convert it by using the classical full recode method.

The *Flow Control Management* block is placed between the H.264/AVC decoder output and the MPEG-2 encoder input. This block handles the decompressed bitstream produced by the decoder before entering into the encoder. Since the

H.264/AVC decoder can decode frames at a higher frame rate than the encoder, due to its unbalanced load, the use of this management block is mandatory to maintain a synchronised process and to avoid idle times at the encoder. This intermediate block controls the data flow between the decoder and the encoder through a circular memory buffer, while the decoder continues to decode the remaining frames and the encoder processes the previous ones. By using such scheme the encoder avoids to stop running while the decoder processes the following frame. The decoder fills in the intermediate buffer with future frames to feed the encoder with frames to process.

7.4 MPEG-2 Video Encoder

The modified MPEG-2 video encoder has also the capability to interface with the *Conversion Module* to receive a set of parameters and several control signals to improve some MPEG-2 encoder functions, namely the motion estimation. The received parameters, after being processed by the *Parameter Converter*, are delivered to the *Motion Estimation* block in order to modify the full search algorithm and use the final motion vectors provided by the H.264/AVC decoder. This procedure avoids carrying out the motion estimation function, thus resulting in significant computational complexity gains.

The *Motion Estimation* module, checks whether the current macroblock can be represented by the H.264/AVC converted information. If the macroblock can be efficiently encoded based on the extracted parameters, the encoder uses the converted motion vector as the final motion vector, without computing any motion estimation, which is called Fast Mode Conversion. However, there are several differences between both standards that make some macroblocks non-compatible for fast mode

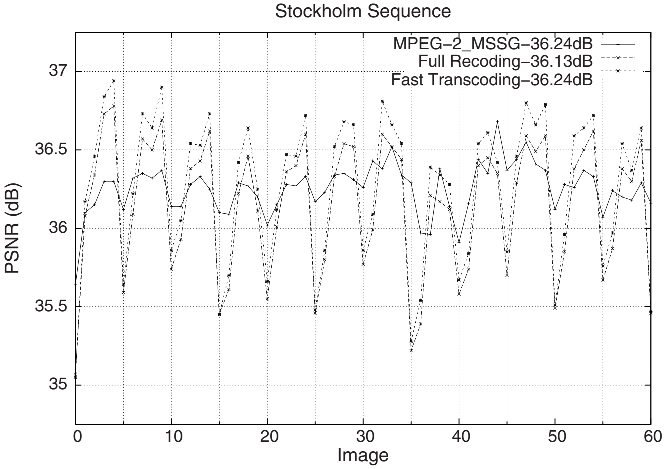


Fig. 21 Objective quality comparison between direct MPEG-2 encoding, reference and fast transcoder

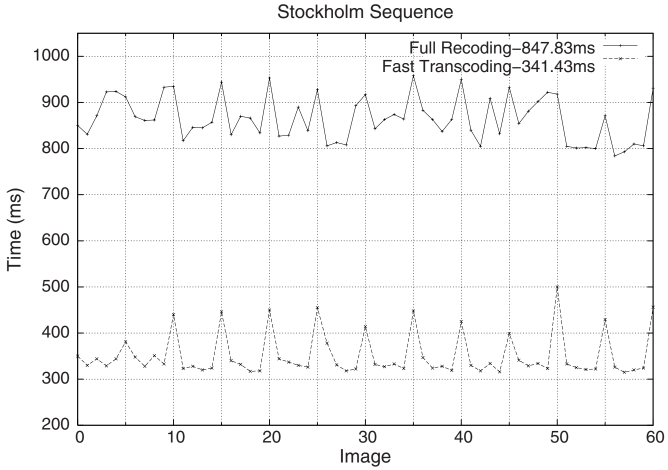


Fig. 22 Time savings comparison between the reference and fast transcoder

conversion. For these non-compatible macroblocks, the transcoder resorts to classical motion estimation at the expense of higher computational complexity.

Figure 21 and 22 show the performance comparison for the Stockholm sequence between the full recoding and the fast transcoder architecture, for objective quality and time savings respectively. These results show that reusing the input information embedded in the bitstream can significantly reduce the computational complexity without introducing any quality penalty.

8 Conclusions

In this chapter we have discussed the importance and flexibility of video transcoding describing functions and capabilities that can be provided. Open loop and closed loop architectures were presented and compared, highlighting the main advantages and disadvantages of each implementation. Several video transcoding techniques were also presented for each conversion function, describing the general procedure and identifying the assumptions in each case, and their impact in terms of video quality and conversion performance. Concurrent approaches were also highlighted such as SVC and MDC.

A case study of a heterogeneous transcoding system was also presented, aiming the conversion of video content from H.264/AVC into MPEG-2. Then, a brief description of each video compression standards was introduced, highlighting the main differences and conversion issues between them.

Acronyms

CBR	Constant Bit Rate
VBR	Variable Bit Rate
UEP	Unequal Error Protection

QCIF	Quarter Common Intermediate Format
HDTV	High Definition Television
SDTV	Standard Definition Television
PIP	Picture-in-Picture
DCT	Discrete Cosine Transform
DVB-T	Digital Video Broadcasting - Terrestrial
DVB-H	Digital Video Broadcasting - Handheld
SVC	Scalable Video Coding
AVC	Advanced Video Coding
MDC	Multiple Description Coding

References

[1] Ahmad, I., Wei, X., Sun, Y., Zhang, Y.-Q.: Video Transcoding: An Overview of Various Techniques and Research Issues. IEEE Transactions on Multimedia 7(5), 793–804 (2005)

[2] Assuncao, P., Ghanbari, M.: Post-processing of mpeg2 coded video for transmission at lower bit rates. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, May 1996, vol. 4, pp. 1998–2001 (1996)

[3] Assuncao, P., Ghanbari, M.: Transcoding of single-layer MPEG video into lower rates. IEE Proceedings on Vision, Image and Signal Processing 144(6), 377–383 (1997)

[4] Assuncao, P., Ghanbari, M.: A frequency domain transcoder for dynamic bit rate reduction MPEG-2 bit streams. IEEE Transactions on Circuits and Systems for Video Technology 8(8), 953–967 (1998)

[5] Assuncao, P., Ghanbari, M.: Buffer Analysis and Control in CBR Video Transcoding. IEEE Transactions on Circuits and Systems for Video Technology 10(1), 83–92 (2000)

[6] Correia, P.F., Silva, V.M., Assuncao, P.: A method for improving the quality of mobile video under hard transcoding conditions. In: IEEE International Conference on Communications, ICC 2003, vol. 2, pp. 928–932, May 11-15 (2003)

[7] Marques, R., Faria, S., Assuncao, P., Silva, V., Navarro, A.: Fast Conversion of H.264/AVC Integer Transform Coefficients into DCT Coefficients. In: International Conference on Signal Processing and Multimedia Applications, August 2006, pp. 5–8 (2006)

[8] Moiron, S., Faria, S., Assuncao, P., Silva, V., Navarro, A.: H.264/AVC to MPEG-2 Video Transcoding Architecture. In: Proceeding of Conference on Telecommunications - ConfTele, May 2007, pp. 449–452 (2007)

[9] Moiron, S., Faria, S., Assuncao, P., Silva, V., Navarro, A.: Low-complexity video content adaptation for legacy user equipment. In: III International Mobile Multimedia Communications Conference (August 2007)

[10] Moiron, S., Faria, S., Assuncao, P., Silva, V., Navarro, A.: Fast Interframe Transcoding from H.264 to MPEG-2. In: IEEE International Conference on Image Processing (September 2007)

[11] Bjork, N., Christopolous, C.: Transcoding Architectures for Video Coding. IEEE Transactions on Consumer Electronics 44, 88–98 (1998)

[12] Bolla, R., Repetto, M., De Zutter, S., Van de Walle, R., Chessa, S., Furfari, F., Reiterer, B., Hellwagner, H., Asbach, M., Wien, M.: A context-aware architecture for QoS and transcoding management of multimedia streams in smart homes. In: IEEE International Conference on Emerging Technologies and Factory Automation, September 2008, pp. 1354–1361 (2008)

[13] Caron, F., Coulombe, S., Wu, T.: A Transcoding Server for the Home Domain. In: IEEE International Conference on Portable Information Devices, May 2007, pp. 1–5 (2007)

[14] Dogan, S., Cellatoglu, A., Uyguroglu, M., Sadka, A.H., Kondoz, A.M.: Error-Resilience transcoding for robust inter-network communications using GPRS. IEEE Transactions on Circuits and Systems for Video Technology 12, 453–464 (2002)

[15] Eleftheriadis, A., Anastassiou, D.: Constrained and general dynamic rate shaping of compressed digital video. In: IEEE International Conference on Image Processing, October 1995, vol. 3, pp. 396–399 (1995)

[16] Essaili, A.E., Khan, S., Kellerer, W., Steinbach, E.: Multiple Description Video Transcoding. In: IEEE International Conference on Image Processing, June 2002, vol. 6, pp. 77–80 (2002)

[17] Hsiao, J.-L., Hung, H.-P., Chen, M.-S.: Versatile Transcoding Proxy for Internet Content Adaptation. IEEE Transactions on Multimedia 10(4), 646–658 (2008)

[18] Feamester, N., Wee, S.: An MPEG-2 to H.263 transcoder. In: SPIE International Symposium on Voice, Video and Data Communications (September 1999)

[19] Fernandez-Escribano, G., Kalva, H., Cuenca, P., Orozco-Barbosa, L.: RD-Optimization for MPEG-2 to H.264 Transcoding. In: IEEE International Conference on Multimedia and Expo., July 2006, pp. 309–312 (2006)

[20] Jinghui, C., Lu, W., Liu, Y., Song, Y.: Xiaowei Song; Sile Yu; H.264/MPEG-2 transcoding based on personal video recorder platform. In: Proceedings of the Ninth International Symposium on Consumer Electronics, June 2005, pp. 438–440 (2005)

[21] Keesman, G., Hellinghuizen, R., Hoeksema, F., Heidman, G.: Transcoding of MPEG streams. In: Signal Processing: Image Communication, vol. 8 (September 1996)

[22] Kim, I.K., Cho, N.I.: Video Transcoding for Packet Loss Resilience Based on the Multiple Descriptions. In: IEEE International Conference on Multimedia and Expo., July 2006, pp. 109–112 (2006)

[23] Kim, J.-W., Kwon, G.-R., Kim, N.-H., Morales, A., Ko, S.-J.: Efficient video transcoding technique for QoS-based home gateway service. IEEE Transactions on Consumer Electronics 52(1), 129–137 (2006)

[24] Kunzelmann, P., Kalva, H.: Reduced Complexity H.264 to MPEG-2 Transcoder. In: International Conference on Consumer Electronics, ICCE 2007. Digest of Technical Papers, January 2007, pp. 1–2 (2007)

[25] Li, C.-H., Lin, H., Wang, C.-N., Chiang, T.: A fast H.264-based picture-in-picture (PIP) transcoder. In: IEEE International Conference on Multimedia and Expo., June 2004, vol. 3, pp. 1691–1694 (2004)

[26] Liu, D., Chen, S., Shen, B.: Modeling and Optimization of Meta-Caching Assisted Transcoding. IEEE Transactions on Multimedia 10(8), 1444–1454 (2008)

[27] Panusopone, K., Chen, X., Ling, F.: Logo insertion in MPEG transcoder. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, May 2001, vol. 2, pp. 981–984 (2001)

[28] Park, Y., Lee, Y., Kim, H., Kim, K.: Hybrid Segment-Based Transcoding Proxy Caching of Multimedia Streams. In: IEEE International Conference on Computer and Information Technology Workshops, July 2008, pp. 319–324 (2008)

[29] Raman, A., Singh, K., Mohan, M.A., Shigihalli, N., Sethuraman, S., Supreeth, B.S.: MPEG-2 to H.264 transcoder on TI TMS320DM642. In: IEEE International Conference on Multimedia and Expo. (2004)

[30] Reyes, G., Reibman, A., Chang, S.-F., Chuang, J.C.I.: Error Resilience transcoding for video over wireless channels. IEEE Journal in Selected Areas in Communications 18, 1063–1074 (2000)

[31] Shanableh, T., Ghanbari, M.: Heterogeneous video transcoding to lower spatio-temporal resolutions and different encoding formats. IEEE Transactions on Multimedia 2(2), 101–110 (2000)

[32] Shanableh, T., Ghanbari, M.: Transcoding architectures for DCT-domain heterogeneous video transcoding. In: International Conference on Image Processing, October 2001, vol. 1, pp. 433–436 (2001)

[33] Shanableh, T., Ghanbari, M.: Multilayer transcoding with format portability for multicasting of single-layered video. IEEE Transactions on Multimedia 7(1), 1–15 (2005)

[34] Shen, B., Tan, W.-T., Huve, F.: Dynamic Video Transcoding in Mobile Environments. IEEE Multimedia 15(1), 42–51 (2008)

[35] Siu, W.-C., Fung, K.-T., Chan, Y.-L.: A compressed-domain heterogeneous video transcoder. In: IEEE International Conference on Image Processing, October 2004, vol. 4, pp. 2761–2764 (2004)

[36] Sun, H., Kwok, W., Zdepski, J.: Architectures for MPEG compressed bitstream scaling. IEEE Transactions on Circuits and Systems for Video Technology 6, 191–199 (1996)

[37] Schwarz, H., Wien, M.: The Scalable Video Coding Extension of the H.264/AVC Standard [Standards in a Nutshell]. IEEE Signal Processing Magazine 25(2), 135–141 (2008)

[38] Takahashi, T., Kasai, H., Hanamura, T., Sugiura, M., Tominaga, H.: MPEG-2 multi-program transport stream transcoder. In: IEEE International Conference on Multimedia and Expo., August 2001, pp. 423–426 (2001)

[39] Vetro, A., Sun, H., Wang, Y.: Object-based transcoding for adaptable video content delivery. IEEE Transactions on Circuits and Systems for Video Technology 11(3), 387–401 (2001)

[40] Vetro, A., Christopolous, C., Sun, H.: Video Transcoding Architectures and Techniques: An Overview. IEEE Signal Processing Magazine, 18–28 (March 2003)

[41] Vetro, A., Xin, J., Sun, H.: Error resilience video transcoding for wireless communications. IEEE Wireless Communications 12(4), 14–21 (2005)

[42] Xin, J., Sun, M.T., Kan, K.: Bit allocation for joint transcoding of Multiple MPEG coded video streams. In: IEEE International Conference on Multimedia and Expo., August 2001, pp. 8–11 (2001)

[43] Xin, J., Sun, M.-T., Choi, B.S., Chun, K.W.: An HDTV to SDTV spatial transcoder. IEEE Transactions on Circuits and Systems for Video Technology 12(11), 998–1008 (2002)

[44] Xin, J., Sun, M.-T., Chun, K., Choi, B.S.: Motion re-estimation for HDTV to SDTV transcoding. In: IEEE Symposium Circuits and Systems, Scottsdale, May 2002, vol. 4, pp. 715–718 (2002)

[45] Ye, T., Tan, Y.-P., Xue, P.: A low complexity H.263 to H.264 transcoder. In: IEEE International Symposium on Circuits and Systems, p. 5290 (2006)

[46] Yeh, C.-H., Chen, S.-M., Chern, S.-J.: Content-Aware Video Transcoding via Visual Attention Model Analysis. In: International Conference on Intelligent Information Hiding and Multimedia Signal Processing, August 2008, pp. 429–432 (2008)

[47] Yin, P., Wu, M., Lui, B.: Video transcoding by reducing spatial resolution. In: IEEE International Conference on Image Processing, October 2000, pp. 972–975 (2000)

[48] Werner, O.: Requantization for transcoding of MPEG-2 intraframes. IEEE Transactions on Image Processing 8(2), 179–191 (1999)

Cross-Layer Approach for Reliable Transmission of Wavelet Coded Images over Portable Multimedia Devices

Ekram Khan, Athar A. Moinuddin, and Mohammed Ghanbari

Abstract. Transmission of compressed images over portable multimedia devices is a challenging task due to high error rate wireless channels, fluctuating and limited bandwidth availability and low energy requirements. Wavelet-based embedded (or progressive) image coders are most suited to cope with time varying channel bandwidth of wireless networks. These coders have excellent rate-distortion performance but they are extremely sensitive to channel errors. Several error resilient image coding techniques have been proposed in order to minimize the effects of transmission errors. Among these, unequal error protection (UEP) of coded image bitstream is one of the most successful techniques, where important bits have a higher protection than the rest of the bitstream. Conventionally, the forward error correction (FEC) based UEP is applied at the application layer. Alternatively, UEP can also be provided using hierarchical modulation approach at the physical layer. In this chapter, we discuss the cross-layer design methodology for UEP that rely on interaction between the application layer and the physical layer to achieve reliable and high quality end-to-end performance in wireless environments. The discussion is mainly focused on set partitioning in hierarchical trees (SPIHT) image coder, but it is applicable to other progressively coded bitstreams as well.

1 Introduction

Recent advances in data compression and networking technologies has facilitated the wide spread use of imaging services and applications. Use of image databases for scientific and medical applications, remote sensing and satellite images, and image archiving has increased. Moreover, with the growing popularity of wireless network, there is a growing trend to access these services over wireless networks. However, high error rates, time varying and limited bandwidth, and low energy budget make the delivery of high quality multimedia a difficult and challenging

Ekram Khan and Athar A. Moinuddin
Department of Electronics Engineering,
Aligarh Muslim University, Aligarh, India

Mohammed Ghanbari
School of Computer Science and Electronics Engineering
University of Essex, Colchester, UK

M. Grgic et al. (Eds.): Rec. Advan. in Mult. Sig. Process. and Commun., SCI 231, pp. 271–294.
springerlink.com

task. Therefore, an efficient image compression and error control scheme is required for effective wireless communication.

For image data compression, wavelet transform has proven to be the most effective tool for redundancy reduction. Over the years, a number of highly efficient wavelet-based image compression algorithms have been proposed [1]-[14]. Some of the well known algorithms include embedded zerotree wavelet (EZW) [2], set partitioning in hierarchical trees (SPIHT) [3], set partitioning embedded block (SPECK) [8] and embedded block coding with optimized truncation (EBCOT) [10] which is adapted in the JPEG2000 [15] standard. These algorithms have excellent rate-distortion performance, while generating embedded bitstream. Combined with the progressive picture build up of the embedded coded images, good quality images at the earlier stages of the transmission is becoming an important element of these type of codecs. This is particularly important, if the web pictures are browsed over the wireless lines.

Even though the standard JPEG2000 codec generates feature rich bitstream, however, use of layered block coding, fractional bitplanes, block-based rate-distortion (R-D) optimizations, and context-based arithmetic coding makes it highly complex. On the other hand, SPIHT is a state-of-the-art image coding algorithm that gives competitive performance at considerably lower complexity [16]. Low complexity algorithms are essential for wireless image communication using handheld portable devices having limited processing power.

The bitstream of embedded wavelet coders such as SPIHT is extremely sensitive to bit errors, and therefore, not suitable for error prone wireless channels. To deal with this problem proper error control is required. In principle, error control for embedded bitstream may employ forward error correction (FEC) [17]-[26], robust image coding [27]-[34], or joint source-channel coding (JSCC) [35]-[42]. In FEC-based error control, source coder bits are appended to the channel coding bits to facilitate the error detection and correction at the receiver. Sherwood and Zeger were the first to use such a scheme [17]. In their work, embedded source bitstream is divided into fixed size packets and protected each packet equally by using a concatenation of a cyclic redundancy check (CRC) outer code for error detection and a rate-compatible punctured convolutional (RCPC) [43] inner code for error correction. Later, it was extended for fading channel by using a product channel code structure with Reed-Solomon (RS) code between the packets [18]. However, one of the nicest features of the embedded coding is that bits are hierarchically organized from most important to least. Therefore, it seams natural to use unequal error protection (UEP), whereby earlier bits are given more protection against error than the later. In literature, UEP-based protection for image data transmission is widely reported [19]-[26]. In most of these works, the source coded bitstream is partitioned in to a number of groups according to their error sensitivity for UEP. First group bits have the highest error sensitive and these are the bits in which error causes global effects; therefore, they are heavily protected. The bits in which error has local effect are grouped into second group and they are lightly protected as compared to the first group bits. And no protection is provided for the remaining least significant bits. An excellent theoretical analysis on UEP for progressive image transmission is presented in [26].

On the other hand, in robust image coding the wavelet coefficients are encoded into packets that can be decoded independently of each other [27]-[34] . The advantage is that the effect of error is localized within the packets where errors occur, which can be estimated at the decoder by using suitable error concealment method [29], [33]. Some researchers have also used a hybrid of robust source coding and FEC for improved error resilient transmission [27], [31], [34].

Alternative to the above two methods of error control, joint design of source and channel, that requires co-ordination between source and channel encoders, is another approach for error control which has attracted a lot of attention [35]-[42]. In JSCC, source and channel coding is formulated as an optimization problem. The main design issue is then optimal allocation of the available bandwidth resource between the source and the channel coding so as to achieve the best possible end-to-end performance in the noisy channel. To reduce the optimization complexity, some fast algorithms have also been introduced [37]-[38], [41] .

In order to benefit from JSCC in real systems, control information needs to be transferred through the network and system layers. Unfortunately, the impact of the network and networking protocols are quite often discarded while presenting the JSCC systems. Network protocol currently follows the layered architecture in which each layer provides a separate solution to the challenges offered by the modern wireless networks. However, this layered strategy does not always result in an optimal overall performance for multimedia transmission. In order to optimize the overall system performance, interaction among the various protocol layers is needed. This is known as *cross-layer* approach. The cross-layer design of multimedia transmission is aimed to improve overall systems performance by jointly considering the multiple protocol layers. The cross-layer design approach for image/video transmission system has gained a lot of attention in recent past [44]-[53].

For robust and reliable communication of image/video, cross-layer design approach considers the error control components from multiple layers. The idea is to jointly consider error protection strategies at various network layers, in order to improve the transmission efficiency in terms of protection, bandwidth and resource consumption. The error control strategies can be implemented at the various layers:

- *Physical layer:* adaptive modulation and channel coding
- *Media Access Control (MAC) layer:* Automatic repeat request (ARQ)
- *Application layer:* FEC and ARQ

Some work has been carried out in order to provide cross-layer protection strategies for video streaming over wireless network, such as combining the adaptive selection of application layer FEC and MAC layer ARQ as presented in [44][45]. But most of the research efforts in the area of robust wireless transmission have mainly focused on enabling adaptive error-control strategies at the application layer. Though, applying protection strategies at the application layer leads to a higher system complexity, but it has the advantage that it can be more specifically targeted toward the content characteristics, necessary levels of protection, etc.

In this chapter, we concentrate on UEP mechanism provided by application and physical layers. At the application layer, UEP can be achieved by adding redundancies for FEC to source coded bits. Use of FEC at the application layer for protection increases bandwidth requirement. Alternatively, UEP can also be achieved by bandwidth efficient modulation scheme at the physical layer. The hierarchical QAM (HQAM) [54] modulation is one possible scheme to provide UEP by protecting important information at the cost of other information. Though higher order HQAM is bandwidth efficient, but protection of lesser important information often needs the increase in transmission power. However, in portable devices, due to limited battery power, increase in transmission power may not be possible in many cases. Cross-layer design methodologies that rely on interaction between different protocol layers hold great promises for addressing these challenges and for providing reliable and high-quality end-to-end performance in wireless multimedia communications. This may be possible by enabling the information exchange between application layer (robust image encoding, FEC) and physical/link layers (data partitioning and hierarchical modulation). This is the main motivation of this chapter.

The remaining part of the chapter is organized as follows. Section 2 reviews the wavelet-based embedded image coders in general and SPIHT algorithm in particular. The error resilient analysis of SPIHT coded bitstream will be presented in Section 3. The FEC-based UEP is discussed in Section 4. Use of HQAM for UEP of embedded bitstream is presented in Section 5. The cross-layer approach for UEP is introduced in Section 6. Finally the chapter is concluded in Section 7.

2 Wavelet-Based Embedded Image Coders

Image coding techniques achieve compression by reducing the spatial redundancies from the source image. In practice, this redundancy reduction is facilitated by using signal transform which compacts most of the energy of the source signal into a relatively small number of transform coefficients. Efficient quantization of these transform coefficients gives high degree of compression performance. Though, large numbers of signal transforms have been proposed, however, modern image compression algorithms including JPEG2000 are based on the discrete wavelet transform (DWT) [55].

The two-dimensional (2-D) DWT for images consists of a recursive decomposition of the lowest resolution subband [56]. In this dyadic decomposition, the original image is decomposed into four subbands each being one fourth the size of the original image, and the lowest-resolution subband is recursively decomposed. This gives a pyramidal structure in which coefficients are correlated in frequency within subbands and spatially across the subbands. This energy clustering property has been used in designing a number of highly successful image compression algorithms [1]-[16].

Tree-based algorithms exploit spatial correlations by grouping the wavelet coefficients corresponding to the same spatial location and orientation to form a spatial orientation tree (SOT). This allows the prediction of insignificance of the coefficients across scales. EZW [2] was the first algorithm to combine this data structure with bitplane-based encoding to generate embedded bitstream. Later, SPIHT algorithm [3] improved upon this concept by adding a number of sorted lists that contain sets of coefficients and individual coefficients. Combined with efficient set partitioning, SPIHT substantially improves over EZW in both speed and compression performances and is widely recognized as the benchmark wavelet coder in the image coding community. Ever since, the tree-based coding framework has been intensively studied and utilized in various image/video coding algorithms [1], [4]-[7]. Alternatively, block-based algorithms such as SPECK and EBCOT exploit within subband correlations, whereby a transformed image is recursively partitioned into subblocks, achieving compression by single symbol encoding of insignificant blocks. Some attempt has also been made to combine the good features of both tree- and block-based algorithms for more efficient image coding [11]-[14].

The important feature of these coding algorithms is that they generated embedded bitstream, whereby bits are hierarchically organized from high to low distortion reduction capability. Usually, such an embedded bitstream is generated by quantizing wavelet coefficient using a bitplane-based coding. In bitplane-based coding, first the most significant bit of all coefficients is coded, followed by the next most significant bit, and so forth. This can be further elaborated with the help of the SPIHT algorithm, which works as follow. To exploit the self similarity among the wavelet coefficient magnitudes in different scales, the coefficients are grouped into a SOT. The organization of coefficients into SOT is based on relating each coefficient at a given scale to a set of four coefficients with the same orientation at the next finer scale. These coefficient trees are grouped into sets and magnitude ordered from the highest bitplane. The ordering information is encoded with a set partitioning algorithm which is facilitated by the use of three lists: a list of insignificant pixels (LIP), a list of insignificant sets (LIS), and a list of significant pixels (LSP). At the initialization step, the pixels in the lowest band (the highest pyramid level) are added to LIP, and those with descendents also are added to LIS. The LSP starts as an empty list.

The coding process starts with the most significant bitplane and proceeds towards the lowest bitplane. For each bitplane there are two passes; sorting and refinement. In the sorting pass, the encoder goes through the lists LIP followed by LIS for locating and coding the significant coefficients. For each pixel in LIP, one bit is used to describe its significance. If the pixel is not significant, it remains in LIP and no more bits will be generated; otherwise, the sign bit is produced and the pixel is moved to LSP. Similarly, each set in LIS requires one bit for significance information. Insignificant sets remain in LIS while significant sets will be partitioned into subsets to locate and code the significant coefficients. The significant coefficients so found will be moved to LSP. In the refinement pass, each pixel in LSP, except those just added at the current bitplane, is refined with one bit to increase its precision. Such a progressive refinement generates fully embedded bitstream. The algorithm then repeats the above procedure for the next bitplane.

3 Error Sensitivity Analysis of Embedded Bitstream

The embedded image coding algorithms such as SPIHT encode an image in bitplanes, with sorting and refinement passes for each bitplane as described in the previous section. The bits so generated have different degree of vulnerability to the errors. The effect of errors in some bits is more severe, damaging the image globally by disturbing the synchronization between the encoder and decoder. Also some bits turn out to be more important than the others in terms of their contribution to the reconstructed image quality. Therefore, an analysis of these bits for error sensitivity will help in designing a proper error control strategy. Here, we present an error sensitivity analysis of embedded bitstream. Though the present analysis is specific to SPIHT, however, with slight modification it can also be used for a number of other similar algorithms such as [2], [4]-[5], [8], [11] and [14].

As discussed previously, the SPIHT algorithm generates three different types of bits; significant, sign and refinement bits. Their degree of error sensitivity for SPIHT coded '*Lena*' image at a rate of 1.0 bits/pixel (bpp) as shown in Fig. 1 (a) is described as follows:

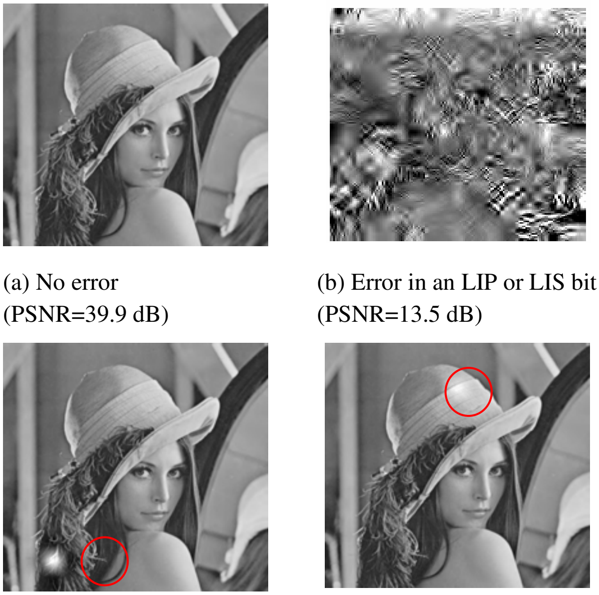
Significance bits: During the sorting pass, coefficients and sets of coefficients are checked for their significance. Error in the significant bits will result in the propagation of error down to the bitstream during the decoding process. A significant coefficient or set may be deemed insignificant and vice-versa due to errors. This will cause the decoder to update the lists with the wrong nodes and can cause a fast degradation in the decoded image quality. The effect of single error in the bits generated during the LIP or LIS test are shown in Fig. 1 (b). Therefore a single bit error in the significant bits has global effect and may damage the entire image.

Sign bits: The sign bit of a wavelet coefficient is generated immediately after it is found significant in the sorting pass. Any error in this bit will change only the phase of the coefficient and does not disturb the synchronization between the encoder and the decoder. Therefore, the effect of an error in the sign bit corrupts the image locally only during the inverse wavelet transform. The effect of error in one of the sign bits is shown in Fig. 1 (c).

Refinement bits: The refinement bits are generated during the refinement pass. An error in these bits simply changes the magnitudes of the decoded coefficients. During the inverse wavelet transform, the changed coefficients affect the neighboring pixels only. Therefore, errors in the refinement bits distort the image only locally. The effect of an error in a refinement bit for the decoded image is shown in Fig. 1 (d).

Based on the degree of their sensitivity, the bits generated by the SPIHT algorithm can be classified into the following two classes: critical and non-critical bits. The critical bits are those, which causes the loss of synchronization between the encoder and decoder. A single bit error in the critical bit causes the failure of the reconstruction process after that point. It consists of significant bits generated during LIP and LIS tests. The non-critical bits on the other hand cause less severe errors. The effect of error in a non-critical bit is limited to a single coefficient or in its neighborhood and does not disturb the progression of the decoding process after it occurs. The non-critical bits consist of the sign and refinement bits.

In order to study the effect of errors in critical and non-critical bits on the performance of the SPIHT decoder, typical R-D curves obtained by decoding the bitstream received from an SPIHT encoder under noise free and noisy channel



(a) No error (PSNR=39.9 dB)

(b) Error in an LIP or LIS bit (PSNR=13.5 dB)

(c) Error in a sign bit (PSNR=34.1 dB)

(d) Error in a refinement bit (PSNR=36. 8 dB)

Fig. 1 Effect of errors in different bits on the reconstructed image ‘*Lena* ’at 1.0 bpp (PSNR=39.9 dB without error)

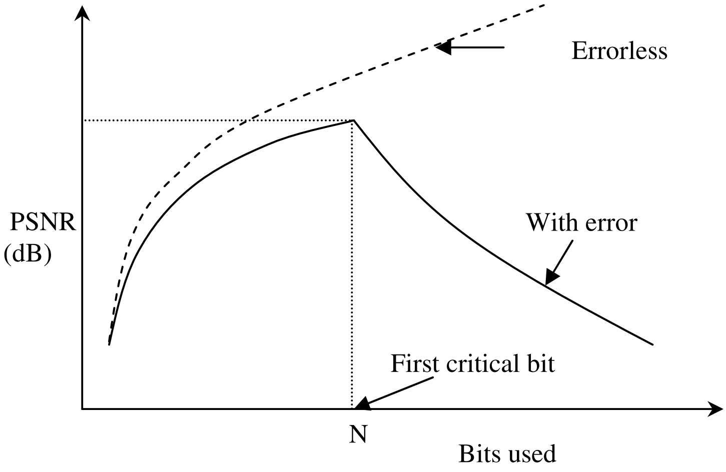


Fig. 2 A typical R-D curve of PSNR versus bits used for SPIHT coded bitstream in noisy channel

conditions are considered. Fig. 2 shows a typical progression of peak signal-to-noise ratio (PSNR) of the reconstructed image as more and more bits from the embedded bitstream are used [20]. The dashed curve shows the errorless case. In the presence of noise, the R-D curve degrades to the solid one. As long as there are no critical bit errors, and the error in the non-critical bits are below a reasonable bound, the PSNR curve continues to rise. The rise is of course not as much as in the noiseless case due to the non-critical bit error affecting the reconstruction. However, as soon as a critical bit error is encountered, say at the N^{th} bit, synchronization is lost. The PSNR progression becomes highly uncertain beyond this point and typically degrades. The bit N, at which the first error occurs, therefore to a large extent, determines the quality of the reconstructed image. A major consideration in the design of an inherently noise resilient SPIHT coder is then to increase the expected value of N, as much as possible.

A simplistic computation of the expected number of bits that are decoded in a received bitstream, before the first critical failure occurs, is now presented. This statistical estimation is similar to the one presented in [20]. The critical bits occur in the sorting passes only and refinement passes generate entirely the non-critical bits. Assume α is the probability that a given bit is a critical bit. Since critical and non-critical (sign) bits are mixed in the sorting pass, α is the fraction of the total bits in the sorting pass that are critical. If the bitstream is transmitted over a binary symmetric channel with a bit error rate (BER) of β, then the probability that a critical bit is received in error is $\alpha\beta$. If the sorting pass bit number at which the first critical bit error occurs is denoted by P, then the expected value of P is given as

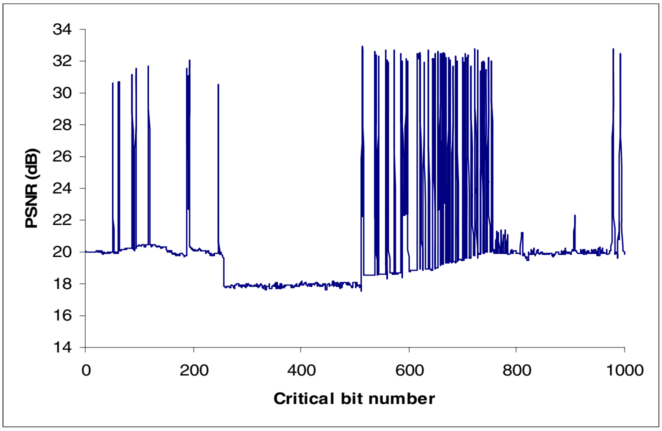
$$E(P) = \sum_{m=0}^{\infty} {}^{m}C_1(\alpha\beta)(1-\alpha\beta)^{m-1} = \frac{1}{\alpha\beta} \quad (1)$$

Furthermore, the incoming bit will either be from the sorting or the refinement passes. If σ is the probability that the incoming bit is from the sorting pass, then the expected value of N, can be approximated as

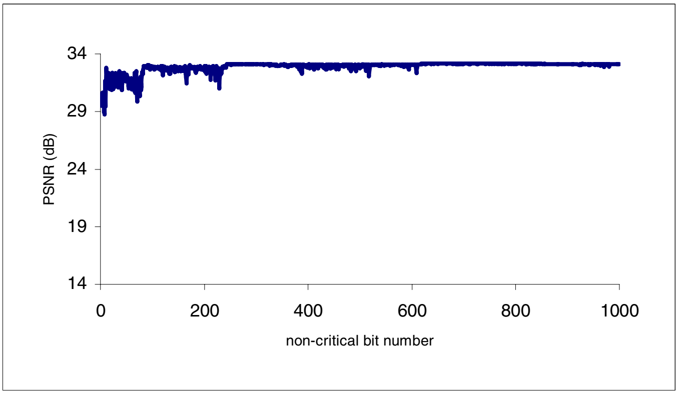
$$E(N) = \frac{E(P)}{\sigma} = \frac{1}{\alpha\beta\sigma} \quad (2)$$

Thus for a given channel BER β, the product $\gamma=\alpha\sigma$, determines the expected waiting time for a critical bit error to occur. This quantity can be defined as the average fraction of the total bits used that are critical, as seen by an arriving sorting pass. Note that this quantity is close to the overall average fraction of bits that are critical, although not exactly the same. The lower the value of γ, the larger the value of *E(N)*, and vice-versa. The larger the value of *E(N)*, the less severe the effect of errors in the N^{th} bit will be.

We have considered image '*Lena*' to measure the bit error sensitivity of the SPIHT bitstream. The bit error sensitivities of first 1000 critical and non-critical bits are shown in Fig. 3 (a) and (b) respectively. It can be seen that in general the critical bits are highly sensitive to the channel errors as compared to the non-critical bits. However, there are some critical bits (as explained above), having the error sensitivity of the same level as that of the non-critical bits. These correspond to impulsive rise in PSNR in Fig. 3 (a). Table 1 shows the distribution of critical and non-critical bits and γ values for the SPIHT coder at various bit rates for the '*Lena*' image.



(a)



(b)

Fig. 3 Bit error sensitivity of SPIHT coded image '*Lena*' for the 1000 (a) critical bits and (b) non-critical bits

Table 1 Distribution of critical and non-critical bits and γ values in SPIHT for the '*Lena*' image at different bit rate

bits/pixel (bpp)	0.1	0.2	0.4	0.6
No. of non-critical bits	6393	13547	28693	39281
No. of critical bits	19821	38882	76165	118005
γ	0.756	0.742	0.726	0.750

4 Unequal Error Protection (UEP) Using Forward Error Correction (FEC)

The analysis presented in the previous section suggests that due to different error sensitivity of different bits, it is possible to devise an UEP scheme in which bits are protected according to their vulnerability to channel errors. Many UEP schemes with different channel codes for SPIHT coded bitstream are suggested in the literature [19]-[26], which are reviewed in section 1. Here we present an UEP scheme that is based on re-organization of bits of each pass into critical and non-critical sub-streams, similar to one reported in [21].

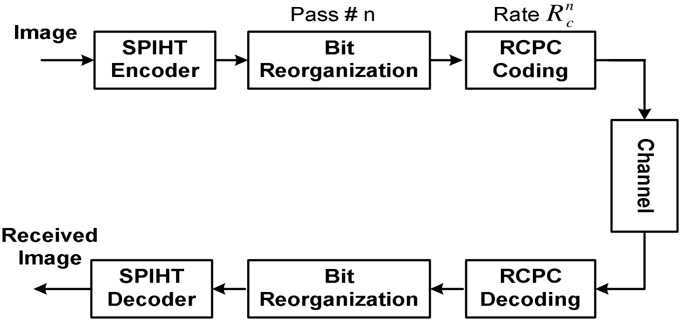


Fig. 4 Image transmission system for n^{th} bitplane of SPIHT bitstream with RCPC rate of R_c^n

Fig. 4 shows an image transmission system with UEP using RCPC codes for one bitplane. Since the SPIHT encoder is progressive and encodes an image bitplane wise, the bits of each bitplane are reorganized to separate critical and non-critical bits of that bitplane, as shown in Fig. 5. The extra headers are embedded in the beginning so that decoder can separate critical and non-critical sub-stream of each bitplane. The separated critical bits are then channel coded using RCPC code to give error robustness, while non-critical bits are transmitted without any protection. The amount of protection (or coding rate) for critical bits reduces as the encoder progresses towards the lower significant bitplanes. This is due to the fact that bits generated in lower bitplanes have lesser effect in reconstruction of image compared to that of earlier passes (or higher bitplanes). Thus SPIHT bitstream are unequally error protected.

In general, for most of the images, non-critical bits of each bitplane comprises at most for 25% of the total bits in that bitplane and the majority (about 75%) of the bits is critical bits. Therefore, within each pass, non-critical bits sub-stream is transmitted before the critical bits sub-stream so that the receiver can maintain progressiveness after buffering non-critical bits portion of the bitstream. Some decoding delay is inevitable, which will depend on the size of non-critical bits sub-stream.

The partitioning and re-organization of SPIHT bitstream presented here differs from the other similar methods [17], [19], [21]. These methods divide the entire SPIHT bitstream into important and unimportant bits and then use the channel coding (e.g. block codes or convolutional codes) to give them protection against the channel noise. These methods have two problems. Firstly, it can not be used to encode the image in real time i.e. we have to wait for all bits to be generated, before applying the channel coding. Secondly, progressiveness feature of SPIHT algorithm is compromised as it is necessary to receive all the encoded bits, important and unimportant, before the image can be reconstructed. However, presented partitioning scheme preserves the scalability feature due to its bitplane-wise partitioning nature.

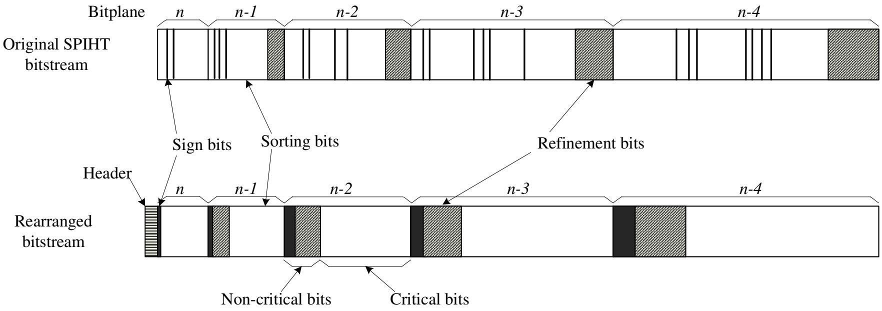


Fig. 5 The partitioning and re-organization of SPIHT bitstream in UEP scheme

The RCPC codes are variants of the convolutional codes in which different coding rates are accomplished by puncturing the same low rate 1/*m* convolutional code by different amounts. The puncturing process may be described as periodically deleting the selected bits from the output of the encoder, thus, creating a periodically time varying trellis code. It begins with a rate 1/*m* parent code and defines a puncturing period *P*, corresponding to *P* input information bits to the encoder. Hence, in one period, the encoder outputs *mP* coded bits. Associated with the *mP* encoded bits is a puncturing matrix *P* of the form

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1p} \\ p_{21} & p_{22} & \cdots & p_{2p} \\ \vdots & \vdots & & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mp} \end{bmatrix} \tag{3}$$

Where each column of *P* corresponds to *m* possible bits from the encoder for each input bit and each element of *P* is either 0 or 1. When $p_{ij} = 1$, the corresponding output from the convolution coder is transmitted. When $p_{ij} = 0$, the corresponding output bit from the encoder is deleted. Thus, the code rate is determined by the period *P* and the number of bits deleted. If *N* bits out of *mP* are deleted, the code rate is *P/(mP-N)*, where *N* may take any integer value in the range of 0 to *(m-1)P-1*. Hence, the achievable code rates are

$$R_c = P/(P+M), \quad M = 1, 2, \dots.(m-1)P \tag{4}$$

4.1 Performance of FEC-Based UEP Method

The performance of the UEP scheme using RCPC code is evaluated. The base convolution coder used has the following specification:

Code rate = 1/3, i.e. $m = 3$
Constraint length, K = 3
Generator matrices are [100], [101] and [111]

In order to generate RCPC codes, the puncturing matrices with period $P = 8$ are considered. This gives the range for code rates as:

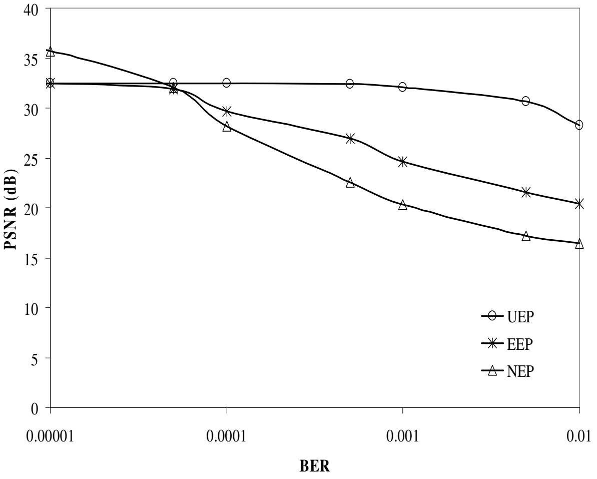
$$R_c = \frac{8}{(8+M)} \tag{5}$$

where M = 16, 14, 12, 10, 8, 6, 4, 2, 1.

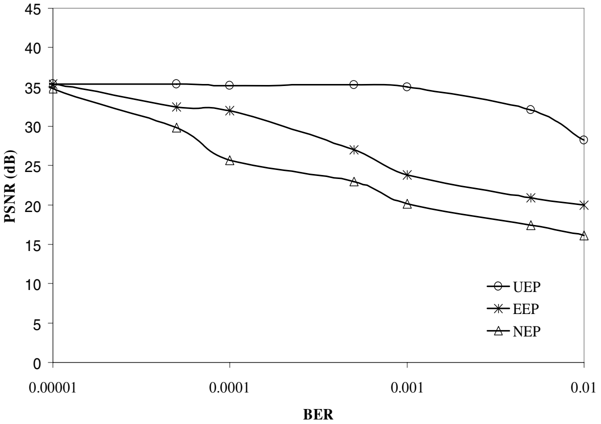
The critical bits generated in the first pass (most significant bitplane) are given the highest protection while critical bits of the last pass (least significant bitplane) are given the least protection. For critical bits of intermediate bitplanes, the RCPC code rate is decreased uniformly. For the protection purpose, the two consecutive bitplanes are clubbed together and they are protected uniformly. That is, the critical bits of the first two passes are protected using M=16, critical bits of the next two passes are protected with M=14, the next two with 12 and so on up to the last pass. Only the critical bits in each pass are protected, while the non-critical bits are transmitted without any protection. This is due to the fact that errors in non-critical bits have only local effect and protecting these bits may increase the quality marginally at the expense of much increased bit budget. Further, since the number of non-critical bits is only 25% of the total generated bits, therefore the probability of errors in these bits is only 25% as compared to that of 75% in the critical bits.

The performance of FEC-based UEP scheme is measured and compared with equal error protection (EEP) as well as no error protection (NEP), at two different bit rates over a binary symmetric channel (BSC). Fig. 6 shows the simulation results for '*Lena*' image. For EEP, the code rate R_c is taken as 4/7 to protect all the critical bits, irrespective of bitplane they belong to. All results of Figs. 6 are averaged over 20 independent channel conditions.

It can be seen that as the BER increases, the PSNR value reduces drastically in case of NEP. This is due to propagation of errors in critical bits in the bitstream, causing the loss of synchronization between the encoder and decoder. It is obvious that SPIHT algorithm is not suitable for noisy channels. Although EEP scheme performs better than the NEP, as all critical bits are protected, but still there is problem of error propagation. The selected coding rate is not sufficient to correct all the critical bits in the early passes, which may result in the degradation of the quality at the given target bit rate.



(a)



(b)

Fig. 6 Performance comparisons of the proposed UEP scheme in terms of PSNR versus BER with EEP and NEP for '*Lena*' image at (a) 0.2 bpp and (b) 0.5 bpp

On the other hand, the UEP scheme has much better robustness against the channel errors as compared to EEP and NEP. It is due to the fact that bits generated in the later passes have less effect on reconstructing the image, and are lesser protected, thereby saving the overall bits, while in EEP, all the critical bits are protected with the same level of protection. Therefore it is better to protect bits according to their importance in the reconstruction of images. From that point of view, bits generated in early sorting passes are more important and hence are protected strongly in the UEP scheme. It has been observed that UEP scheme results in the performance gain up to 5-10 dB and 10-15 dB over EEP and NEP respectively. This gain increases as the BER increases.

Although UEP improves the reconstructed image quality (in presence of errors) but the results of UEP and EEP are inferior to that of NEP in some cases when channels are in good conditions (BER=10^{-5} to 10^{-4}) and at lower bit rates as evident from Fig. 6(a). The reason for this is as follows. At lower bit budgets and when channel is in a good condition (with lower BER), on an average only few bits are likely to be corrupted. Further, when averaged over 20 channel conditions, it is possible that most of the times either non-critical bits or critical bits of later passes are in errors, therefore the averaged overall PSNR will be as good as in error free channel. The channel coding in such situations is simply the wastage of the bit-budget and hence the bandwidth.

5 Unequal Error Protection (UEP) Using Hierarchical Modulation

Modulation is one of the key components in transporting information over wireless networks. Due to high bit rate requirements of image/video based applications, it is desirable to choose a bandwidth efficient modulation technique such as *M-ary* QAM (*M*=16, 64, or even higher). These higher order modulation schemes increase the transmission capacity by assigning more bits to each transmitted symbol. However, modulation schemes that allow a larger number of bits per symbol, have symbols closer to each other in the constellation diagram, and small errors can result in erroneous decoding. To overcome this problem, the uses of non-uniform signal space constellations are suggested to give different degrees of error protection [57]. The idea of dividing the broadcasted messages into two or more classes and to give every class a different degree of error protection was introduced firstly in [58] and later used in DVB-T where it is known as hierarchical modulation [59]. The hierarchical modulation is also known as multi-resolution or asymmetrical or non-uniform constellation. The basic philosophy of using hierarchical modulation as an alternative to FEC-based UEP is that important information is protected without any additional bandwidth requirement. This is in contrast to FEC-based approach where added parity bits either increase the channel bandwidth (bit rate) or for constant bit rate the quality is sacrificed.

In HQAM, the signal constellation is partitioned so that specific blocks of the partition contain message points with the maximum possible Euclidean distance between them at the expense of message points in other blocks which are separated by a much smaller minimum distance. The coding of the message points is chosen so that the specific bits assigned to the widely separated points have lower error probabilities than the others. The data stream is split into two separate substreams so that the most sensitive bits are assigned higher priority and are known as a “high priority (HP)” data. The remaining bits are assigned lower priority and are called “low priority (LP)” data. As discussed in section 3, in SPIHT coded bitstream of images, some critical bits (generated in early passes) are highly sensitive to channel errors and therefore may be assigned as higher priority. The remaining critical bits and all non-critical bits, which are less sensitive to errors, may be assigned lower priority. Block diagram of an image transmission system that prioritizes data streams to protect them unequally (according to their error sensitivity) using HQAM is shown in Fig. 7.

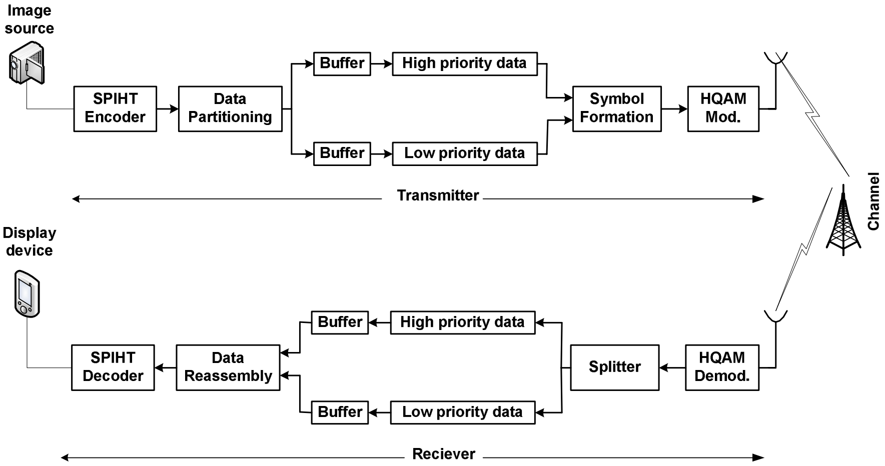


Fig. 7 Image communication system to achieve UEP using HQAM

5.1 Hierarchical Quadrature Amplitude Modulation (HQAM)

Hierarchical modulations were initially proposed to provide different classes of data to users in different wireless reception conditions [60]. Users within the coverage range can reliably receive basic information, while users under more favorable conditions can receive additional refinement information. The conventional HQAM with signal constellation size M (M-HQAM), defined in the digital video broadcast standard [59], and offers two levels of priority. HP data occupies the first two most significant bits (MSBs) of each point (Fig. 8 for 16- and 64-HQAM). Gray code labeling permits all points belonging to the same quadrant to have the same HP bits. This means that if the received point is de-mapped erroneously to a neighboring point but remains within its constellation quadrant, the HP bits will remain uncorrupted. LP data occupies the rest of the bits in the point label. For M-point constellation, the number of LP bits in each symbol is given by $\log_2 M - 2$.

Fig. 8 (a) and (b) depicts the constellation diagram of 16- and 64-HQAM respectively. The parameters 'a' and 'b' represent the minimum distance between quadrants and the minimum distance between points inside each quadrant respectively. To achieve UEP for HP and LP bits, 'a' and 'b' ($a > b$) are adjusted such that their ratio a/b controls the degree of protection. Let α (also known as modulation parameter)

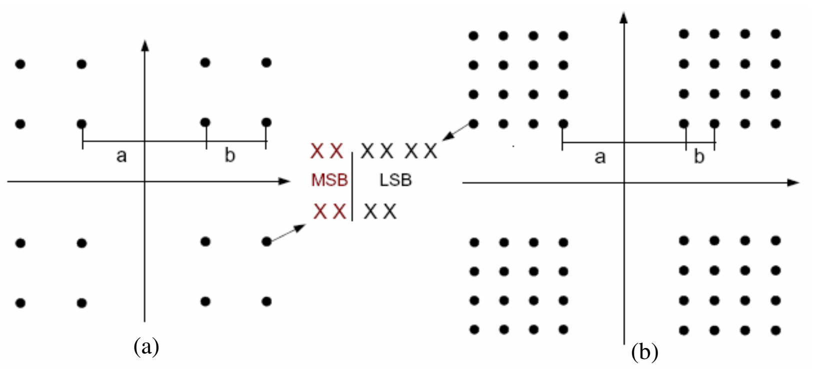


Fig. 8 Two level hierarchical constellation diagrams for (a) 16-HQAM and (b) 64-HQAM

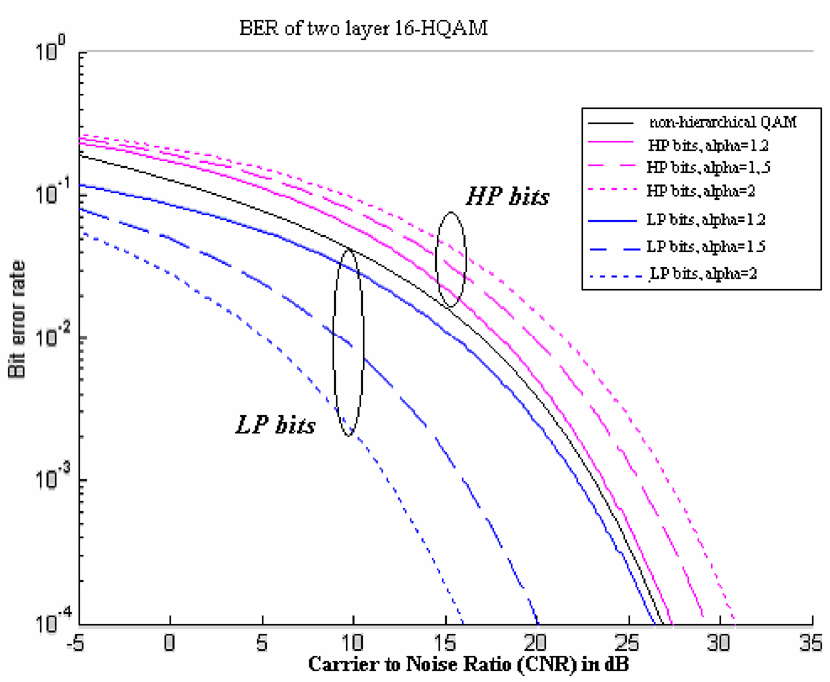


Fig. 9 BER versus CNR for 16-HQAM when α=1.2, 1.5 and 2 in an AWGN channel (central black line shows the performance of non-hierarchical 16-QAM)

be the ratio between minimum distance between quarters and minimum distance between points inside each quarter and is given by

$$\alpha = \frac{a}{b} \tag{6}$$

when $\alpha = 1$, the signal constellation becomes a conventional rectangular M-QAM with each layer having the same reliability. On the other hand, if $\alpha > 1$, the signal constellation becomes a hierarchical M-QAM. By controlling the value of α, it is possible to control the BER of the HP and LP data streams. Increasing α will increase the robustness of HP data stream against channel noise at the expense of robustness of the LP data stream for a constant average constellation power. Fig. 9 illustrates a BER performance of the 16-HQAM in an additive white Gaussian noise (AWGN) channel as a function of channel signal-to-noise ration (CNR). It can be seen from this figure that hierarchical modulation improves the BER performance of the HP data transmission by sacrificing the BER performance of LP data. Although HQAM does not cause any significant reduction in the source data rate when offering UEP, it seems to only work satisfactorily at high channel SNR. In fact, HQAM can be considered as a means of UEP at the physical layer.

5.2 Bitstream Partitioning and Priority Assignment

As discussed in the previous section, by controlling the modulation parameter α of HQAM, the BER of HP bitstream can be improved at the cost of that of LP bitstream. In order to utilize this property of HQAM, encoded bitstream should be partitioned such that error in LP bits should have minimal effects in the overall quality of the reconstructed image. As discussed in section 3, SPIHT bitstream can be partitioned into critical and non-critical sub-streams based on their bit error sensitivity. The critical bits may be assigned higher priority than those of non-critical bits. Also, it is observed that critical bits generated in the later passes have relatively smaller error sensitivity as compared to those of early passes of SPIHT. Since, the number of critical bits is generally 2-3 times larger than those of non-critical bits. And the use of HQAM requires fewer number of HP bits as compared to those of LP bits. For example, from Fig. 8(a), it can be observed that for 16-HQAM, every symbol is formed by multiplexing two HP bits (as MSBs) and two LP bits (as LSBs). Therefore, the number of bits in the HP and LP bitstream should be the same for 16-HQAM, and hence, the critical bits and non-critical bits need to be re-organized accordingly.

In order to have nearly equal number of bits in each sub-stream, some of the critical bits (those generated at later passes) have to be treated as part of LP bitstream. Therefore for 16-HQAM, $\left\lfloor N_n - \left(\frac{N_c + N_n}{2}\right) \right\rfloor$ of the critical bits of later passes are appended in the LP bitstream along with all the non-critical bits, where N_n and N_c are the number of non-critical and critical bits generated by SPIHT encoder at the specified bit-budget respectively. The bitstream having critical bits of early passes are assigned higher priority and form the HP bitstream whereas, the remaining bits are assigned lower priority, forming the LP bitstream. This bitstream partitioning allows matching the bit-error sensitivity of the transmitted symbols in terms of mapping in the constellation diagram of H-QAM. For 16-HQAM, each of the 4-bits symbols are formed by combining two bits from each bitstreams with HP bits as the MSBs and LP bits as LSBs of the symbols. These symbols are mapped to the non-uniform signal constellation of 16-HQAM. By adjusting the modulation parameter $\alpha = a / b$, the relative priority of two bitstreams can be changed, depending on the noise conditions in the channel.

5.3 Performance of UEP Scheme Using HQAM

A 512×512 gray-scale *'Lena'* image is considered to measure the performance of UEP based on HQAM (error protection at physical layer). The performance is evaluated in terms of PSNR of the reconstructed image transmitted over AWGN channel as shown in Fig. 10. The channel conditions are specified in terms of peak CNR. The carrier power is assumed to be constant. It can be observed that the increase in α increases the quality of the received image under the similar noise conditions. This is because with increasing 'α' greater protection is provided to the HP bits and hence the important bits in the image are protected.

Fig. 11 compares the subjective quality of the reconstructed *'Lena'* image coded with SPIHT encoder at a bit rate of 0.1 bpp. The encoded bitstream is rearranged, prioritized, modulated using 16-HQAM and transmitted over an AWGN channel having CNR equal to 20 dB. Fig. 11 includes (in order) the original image, reconstructed images for α =1 (non-hierarchical QAM), 1.5 and 1.8. The advantage of HQAM over conventional QAM over erroneous channel is very obvious from this figure.

From these results, it can be concluded that under the power constrained environments, hierarchical modulation offers a greater protection of the HP bitstream. As the modulation parameter α is increased a tradeoff between the BER for HP and LP bits is observed. For higher values of α, HP bits are protected more

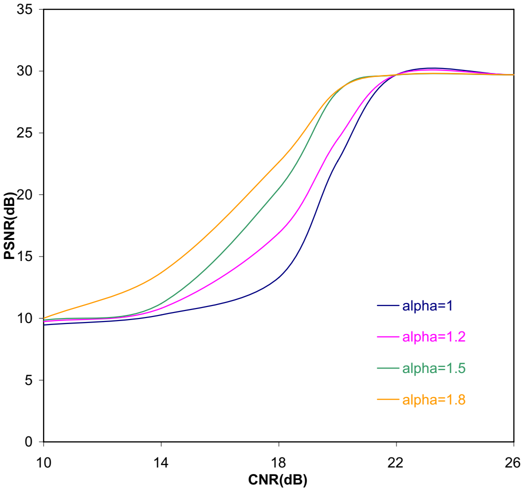
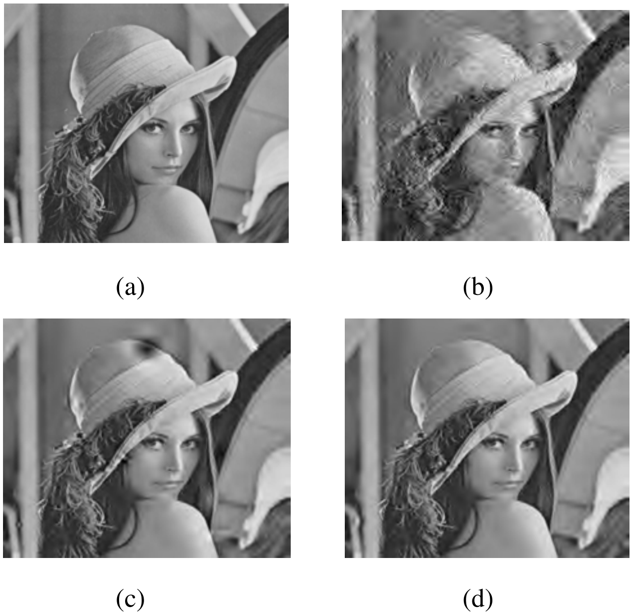


Fig. 10 Performance of HQAM-based UEP for '*Lena*' image coded at a rate of 0.1 bpp



(a) (b)

(c) (d)

Fig. 11 Subjective quality of decoded '*Lena*' image at 0.1 bpp and CNR=20 dB (a) original image (b) with QAM (c) with H-QAM for α=1.5 and (d) with H-QAM for α=1.8

strongly whereas LP bits are assigned lesser protection. Hierarchical modulation is thus an effective method for transmitting the layered source coded image. It should be noted that HQAM provides UEP to source coded data without any additional bandwidth requirements.

6 Unequal Error Protection (UEP) Using Cross-Layer Approach

The cross-layer design of multimedia transmission is aimed to improve overall systems performance by jointly considering the multiple protocol layers. We will consider the error control components from multiple layers. The idea is to jointly consider error protection strategies at various network layers, in order to improve the transmission efficiency in terms of protection, bandwidth and resource consumption.

Progressively coded image bitstreams are usually of different importance (as discussed in previous sections), therefore the network recourses should be utilized according to their importance in order to provide optimal unequal error protection. Two approaches of unequal error protections discussed in the previous two sections have their own limitations. The FEC-based UEP performed at the application layer is bandwidth inefficient. It offers a trade-off between quality and bandwidth.

On the other hand, HQAM-based UEP performed at the physical layer achieves protection of high priority bits without additional bandwidth, but at the cost of increased errors in the lower priority bits. In order to improve the overall quality of the reconstructed image, lower priority bits also need to be protected, which can be achieved by increased the transmitter power only. That is, the physical layer protection offers a tradeoff between quality and power. Transmission of images over mobile devices is generally constrained by both powers (due to limited battery life) as well as bandwidth of wireless channels.

The cross-layer approach can be used to transmit progressively coded images over portable devices under power and bandwidth constrained environments to achieve the best overall quality. That is to achieve better quality, the HQAM need to be combined with channel coding.

In cross-layer approach of UEP, SPIHT bits are first partitioned into two priority streams, namely HP (critical bits of early passes) and LP streams (critical bits of later passes and non-critical bits). HP streams are protected using RCPC code with pass-by-pass varying coding rates. But this protection is relatively lower than that used in Section 3. This allows the part of LP bits also protected, keeping the overall bit budget the same. The parts of LP bitstream (critical bits of later passes) are lightly protected with RCPC channel codes. During the channel coding phase, care should be taken to keep the total number of bits in two streams to be the same. The resulting HP and LP streams (after channel coding) are then multiplexed and mapped to the 16-HQAM. The value of α in HQAM and RCPC codes for two streams can be adjusted appropriately to achieve best overall quality.

Further, the performance of HQAM can be improved by increasing the number of priority layers in the higher order HQAM (64-QAM). In general, three layers 64-HQAM will be better than two layers 16-HQAM, both in terms of bandwidth efficiency as well as more margin of protection with slight increase in power.

7 Conclusions

In this chapter UEP of progressively coded image was investigated with RCPC coding where we showed that in order to have a high protection for the HP layer, its source rate need to be largely reduced if bandwidth is restricted. This significantly decreases the quality of the image. To address this problem, cross-layer approach is introduced, in which high priority bits are also protected by HQAM, hence lower RCPC coding redundancy is required for the same level of protection. Consequently, more bits can be protected with higher channel rates therefore increasing the quality of service.

Acknowledgments

The authors would like to thank Mr. Mohd. Ayyub Khan, Ms. Priya Agrawal and Mrs. Shahwar Nishat for their help.

References

[1] Fowler, J.E., Pesquet-Popescu, B.: An overview on wavelets in source coding, communications, and networks. EURASIP Journal on Image and Video Processing 2007, 1–27 (2007)

[2] Shapiro, J.M.: Embedded image coding using zero trees of wavelet coefficients. IEEE Trans. on Acoustic, Speech and Signal Processing 41, 3445–3462 (1993)

[3] Said, A., Pearlman, W.A.: A new fast and efficient codec based on set partitioning in hierarchical trees. IEEE Trans. on Circuits and Systems for Video Technology 6, 243–250 (1996)

[4] Khan, E., Ghanbari, M.: Very low bit rate video coding using virtual SPIHT. IEE Electronics Letters 37, 40–42 (2001)

[5] Mukherjee, D., Mitra, S.K.: Vector SPIHT for embedded wavelet video and image coding. IEEE Trans. on Circuits and Systems for Video Technology 13, 231–246 (2003)

[6] Danyali, H., Mertins, A.: Highly scalable image compression based on SPIHT for network applications. IEEE Int. Conf. on Image Processing (ICIP'02) 1, 217–220 (2002)

[7] Su, C.Y., Wu, B.F.: A low memory zerotree coding for arbitrarily shaped objects. IEEE Trans. on Image Processing 12, 271–282 (2003)

[8] Pearlman, W.A., Islam, A., Nagaraj, N., Said, A.: Efficient low-complexity image coding with set-partitioning embedded block coder. IEEE Trans. on Circuits and Systems for Video Technology 14, 1219–1235 (2004)

[9] Hsiang, S.T., Woods, J.W.: Embedded image coding using zeroblocks of subband/wavelet coefficients and context modeling. In: IEEE Int. Symposium on Circuits and Systems (ISCS 2000), pp. III:662–665 (2000)

[10] Taubman, D.: High performance scalable image compression with EBCOT. IEEE Trans. on Image Processing 9, 1158–1170 (2000)

[11] Wheeler, F.W., Pearlman, W.A.: Combined spatial and subband block coding of images. In: IEEE Int. Conf. on Image Processing (ICIP 2000), pp. III:861-864 (2000)

[12] Arora, H., Singh, P., Khan, E., Ghani, F.: Memory efficient image coding with embedded zero block-tree coder. In: IEEE Int. Conf. on Multimedia and Expo. (ICME 2004), pp. I:679-682 (2004)

[13] Yin, X.W., Fluery, M., Downton, A.C.: Prediction and adaptive scanning in block-set wavelet image coder. IEE Proceedings-Vision Image Signal Processing 153, 230–236 (2006)

[14] Moinuddin, A.A., Khan, E., Ghanbari, M.: Efficient algorithm for very low bit rate embedded image coding. IET Image Processing 2, 59–71 (2008)

[15] Information Technology-JPEG2000 image coding system-part 1: Core coding system. ISO/IEC 15444-1 (2000)

[16] Oliver, J., Malumbres, M.P.: Low-complexity multiresolution image compression using wavelet lower trees. IEEE Trans. on Circuits and Systems for Video Technology 16, 1437–1444 (2006)

[17] Sherwood, P.G., Zeger, K.: Progressive image coding on noisy channels. IEEE Signal Processing Letters 4, 189–191 (1997)

[18] Sherwood, P.G., Zeger, K.: Error protection for progressive image transmission over memoryless and fading channels. IEEE Trans. on Communications 46, 1555–1559 (1998)

[19] Man, H., Kossentini, F., Smith, M.J.T.: A family of efficient and channel error resilient wavelet/subband image coders. IEEE Trans. on Circuits and Systems for Video Technology 9, 95–108 (1999)

[20] Mukherjee, D., Mitra, S.K.: A vector set partitioning noisy channel image coder with unequal error protection. IEEE Journal on Selected Areas in Communications 18, 829–840 (2000)

[21] Alatan, A.A., Zhao, M., Akansu, A.N.: Unequal error protection of SPIHT encoded image bit streams. IEEE Journal on Selected Areas in Communications 18, 814–818 (2000)

[22] Mohr, A., Riskin, E., Ladner, R.: Unequal loss protection: Graceful degradation of image quality over packet erasure channel through forward error correction. IEEE Journal on Selected Areas in Communications 18, 819–828 (2000)

[23] Yap, C.W., Ngan, K.N.: Error resilient transmission of SPIHT coded images over fading channels. IEE Proceedings-Vision Image Signal Processing 148, 59–64 (2001)

[24] Khan, E., Ghanbari, M.: Error resilient virtual SPIHT for image transmission over noisy channels. In: European Association for Signal, Speech, and Image Processing Conference (EUSIPCO 2002), pp. II:369–372 (2002)

[25] Kim, J., Mersereau, R.M., Altunbasak, Y.: Error-resilient image and video transmission over the internet using unequal error protection. IEEE Trans. on Image Processing 12, 121–131 (2003)

[26] Cao, L.: On the unequal error protection for progressive image transmission. IEEE Trans. on Image Processing 16, 2384–2388 (2007)

[27] Man, H., Kossentini, F., Smith, M.J.T.: Robust EZW image coding for noisy channels. IEEE Signal Processing Letters 4, 227–229 (1997)

[28] Creusere, C.D.: A new method of robust image compression based on the embedded zerotree wavelet algorithm. IEEE Trans. on Image Processing 6, 1436–1442 (1997)

[29] Rogers, J.K., Cosman, P.C.: Wavelet zerotree image compression with packetization. IEEE Signal Processing Letters 5, 105–107 (1998)

[30] Yang, S.H., Cheng, T.C.: Error resilient SPIHT image coding. Electronic Letters 36, 208–210 (2000)

[31] Cosman, C., Rogers, J.K., Sherwood, P.G., Zeger, K.: Combined forward error control packetized zerotree wavelet encoding for transmission of images over varying channels. IEEE Trans. on Image Processing 9, 982–993 (2000)

[32] Kim, T., Choi, S., Van Dyck, R.E., Bose, N.K.: Classified zerotree wavelet image coding and adaptive packetization for low bit rate transport. IEEE Trans. on Circuits and Systems for Video Technology 11, 1022–1034 (2001)

[33] Khan, E., Ghanbari, M.: Error detection and correction of transmission errors in SPIHT coded images. In: IEEE Int. Conference on Image Processing (ICIP 2002), pp. II:689–692 (2002)

[34] Boulgouris, N.V., Thomos, N., Strintzis, M.G.: Transmission of images over noisy channels using error-resilient wavelet coding and forward error correction. IEEE Trans. on Circuits and Systems for Video Technology 13, 1170–1181 (2003)

[35] Chande, V., Farvardin, N.: Progressive transmission of images over memoryless channels. IEEE Journal on Selected Areas in Communications 18, 850–860 (2000)

[36] Stankovic, V., Hamzaoui, R., Xiong, Z.: Real-time error protection of embedded codes for packet erasure and fading channels. IEEE Trans. on Circuits and Systems for Video Technology 14, 1064–1072 (2004)

[37] Hamzaoui, R., Stankovic, V., Xiong, Z.: Fast algorithm for distortion-based error protection of embedded image codes. IEEE Trans. on Image Processing 14, 1417–1421 (2005)

[38] Hamzaoui, R., Stankovic, V., Xiong, Z.: Optimized error protection of scalable image bit streams. IEEE Signal Processing Magazine 22, 91–107 (2005)

[39] Etemadi, F., Jafarkhani, H.: An Efficient progressive bitstream transmission system for hybrid channels with memory. IEEE Trans. on Multimedia 8, 1291–1298 (2006)

[40] Pan, X., Banihashemi, A.H., Cuhadar, A.: Progressive transmission of images over fading channels using rate-compatible LDPC codes. IEEE Trans. on Image Processing 15, 3627–3635 (2006)

[41] Fresia, M., Lavagetto, F.: Determination of optimal distortion-based protection in progressive image transmission: A heuristic approach. IEEE Trans. on Image Processing 17, 1654–1662 (2008)

[42] Yao, L., Cao, C.: Turbo codes-based image transmission for channels with multiple types of distortion. IEEE Trans. on Image Processing 17, 2112–2121 (2008)

[43] Hagenauer, J.: Rate-compatible punctured convolutional (RCPC) codes and their applications. IEEE Trans. on Communication 36, 389–400 (1988)

[44] Schaar, M.V., Krishnamachari, S., Choi, S., Xu, X.: Adaptive cross-layer protection strategies for robust scalable video transmission over 802. 11 WLANs. IEEE Journal on Select. Areas in Communication 21, 1752–1763 (2003)

[45] Schaar, M.V., Shankar, S.N.: Cross-layer wireless multimedia transmission: challenges principles, and new paradigms. IEEE Trans. Wireless Communication 12, 50–58 (2005)

[46] Wu, D., Hou, T., Zhang, Y.Q.: Transporting real-time video over the Internet: Challenges and approaches. Proc. IEEE 88, 1855–1875 (2000)

[47] Costa, C., Granelli, F., Katsaggelos, A.K.: A Cross-layer approach for energy efficient transmission of progressively coded images over wireless channels. IEEE Int. Conf. on Image Processing (ICIP 2005) 1, 213–216 (2005)

[48] Kumwilaisak, W., Hou, Y.T., Zhang, Q., Zhu, W., Kuo, C.C.J., Zhang, Y.Q.: A cross-layer quality-of-service mapping architecture for video delivery in wireless network. IEEE J. Sel. Areas Communication 21, 1685–1698 (2003)

[49] Ksentini, A., Naimi, M., Gueroui, A.: Toward an improvement of H.264 video transmission over IEEE 802.11e through a cross-layer architecture. IEEE Commun. Mag. 44, 107–114 (2006)

[50] Zhang, Y.Q., Yang, F., Zhu, W.: Cross-layer QoS support for multimedia delivery over wireless Internet. EURASIP Journal on Applied Signal Processing 2, 207–219 (2005)

[51] Foh, C.H., Zhang, Y., Ni, Z., Cai, J., Ngan, K.N.: Optimized cross-layer design for scalable transmission over IEEE 802.11e network. IEEE Trans. Circuits and Systems for Video Technology 17, 1665–1678 (2007)

[52] Huusko, J., Vehkapera, J., Amon, P., et al.: Cross-layer architecture for scalable video transmission in wireless networks. Signal Processing: Image Communication 22, 317–330 (2007)

[53] Granelli, F., Costa, C.E., Katsaggelos, A.K.: A Study on the usage of cross-layer power control and forward error correction for embedded video transmission over wireless links. In: Advances in Multimedia, special issue on Cross-layer Optimized Wireless Multimedia Communications: Hindawi, Article ID 95807 (2007)

[54] Mirabbasi, S., Martin, K.: Hierarchical QAM: a spectrally efficient DC-free modulation scheme. IEEE Communication Magazine 38, 140–146 (2000)

[55] Usevitch, B.E.: A tutorial on modern lossy wavelet image compression: Foundations of JPEG 2000. IEEE Signal Processing Magazine 18, 22–35 (2001)

[56] Mallat, S.G.: A theory for multiresolution signal decomposition: The wavelet representation. IEEE Trans. on Pattern Analysis and Machine Intelligence 11, 674–693 (1989)

[57] Hanzo, L.: Bandwidth-Efficient Wireless Multimedia Communication. Proc. IEEE 86, 1342–1382 (1998)

[58] Cover, T.: Broadcasts channels. IEEE Trans. on Information Theory IT 18, 2–14 (1972)

[59] DVB-T standard: ETS 300 744 (2001) Digital Broadcasting Systems for Television, Sound and Data Services: Framing Structure, Channel Coding and Modulation for Digital Terrestrial Television. ETSI Draft, 1.2.1:EN300 744

[60] Wei, L.F.: Coded modulation with unequal error protection. IEEE Trans. on Communications 41 (1993)

HD Video Communication at Very Low Bitrates

Ulrik Söderström and Haibo Li

1 Introduction

This chapter describes compression of facial video, i.e., video where the face or head-and-shoulders of a person is the important information. This kind of video is used for communication between people and HD quality of this video makes it much more useful. When it is wearable so that the user is free to move it becomes an even more appealing application.

Approximately 65% of the communication between people is determined by non-verbal cues such as facial expressions and body language [2]. Therefore, face-to-face meetings are indeed essential and important for the understanding and agreement between humans. It is found that face-to-face meetings are more personal and easier to understand than phone or email. It is easy to see that face-to-face meetings are clearer than email since you can get direct feedback; email is not real-time communication. Face-to-face meetings are also seen as more productive and the content is easier to remember. But, face-to-face does not need to be in person. Distance communication through video conference equipment is a human-friendly technology that provides the face-to-face communications that people need in order to work together productively, without having to travel. Using technology rather than travelling to meetings has benefits beyond the environmental savings. A business trip for a two-hour meeting typically entails spending six hours out of the office - time that can be used to boost productivity in the workplace as well as a better work-life balance. The removal of these trips will mean a huge saving in hours, fuel bills and running costs. The technology also allows people who work at home or teleworkers to collaborate as if they actually were in the office. Video conferencing allows a level of engagement which can remove the loneliness, or isolation, factor which often is

Ulrik Söderström
Digital Media Lab, Dept. Applied Physics and Electronics, Umeå University SE-90187, Umeå, Sweden
e-mail: ulrik.soderstrom@tfe.umu.se

Haibo Li
Digital Media Lab, Dept. Applied Physics and Electronics, Umeå University SE-90187, Umeå, Sweden
e-mail: haibo.li@tfe.umu.se

M. Grgic et al. (Eds.): Rec. Advan. in Mult. Sig. Process. and Commun., SCI 231, pp. 295–314.
springerlink.com

associated with distance working. Workers may be "out of site" but they are certainly not "out of sight." Even if there are several benefits with video conferencing it is not very popular. In most cases, video phones have not been a commercial success, but there is a market on the corporate side. Several large companies are using video conference systems, in hopes of gaining reduced travel bills and other advantages with high quality visual contact. Video conferencing with HD resolution can give the impression of face-to-face communication even over networks. HD video conference can essentially eliminate the distance and make the world connected. On a communication link with HD resolution you can look people in the eye and see whether they follow your argument or not.

One reason that video conference applications aren't used very often is because of the low quality in most conference systems. Communication isn't close to face-to-face communication because the image is too small and with too low quality. The bitrate needed for HD video communication is also a big problem. Communication requires bitrates that cannot be provided over several mobile networks. According to the company Polycom a bitrate above 1 Mbps is needed for HD video conferencing; 2 Mbps is recommended [33].

Two key expressions for video communication are anywhere and anytime. Anywhere means that communication can occur at any location, regardless of the available network, and anytime means that the communication can occur regardless of the surrounding network traffic or battery power. To achieve this there are several technical challenges:

1. The usual video format for video conference is CIF (352x288 pixels) with a framerate of 15 fps. 1080i video (1920x1080 pixels) has a framerate of 25 fps. Every second there is $\approx$ 26 times more data for a HD resolution video than a CIF video.
2. The bitrate for HD video grows so large that it is impossible to achieve communication over several networks. Even with a high-speed wired connection the bitrate may be too low since communication data is very sensitive to delays.
3. Most of the users want to have high mobility; having the freedom to move while communicating.

A solution for HD video conferencing is to use the H.264 [19, 32] video compression standard. This standard can turn HD video into high quality compressed video. There are however two major problems with H.264:

1. The complexity for H.264 coding is quite high. High complexity means high battery consumption; something that is becoming a problem with mobile battery-driven devices. The power consumption is directly related to the complexity so high complexity will increase the power usage and reduce the battery time.
2. The bitrate for H.264 encoding is very high. The vision of providing video communication anywhere cannot be fulfilled with the bitrates required for H.264. The transmission power is related to the bitrate so low bitrate will also save battery power.

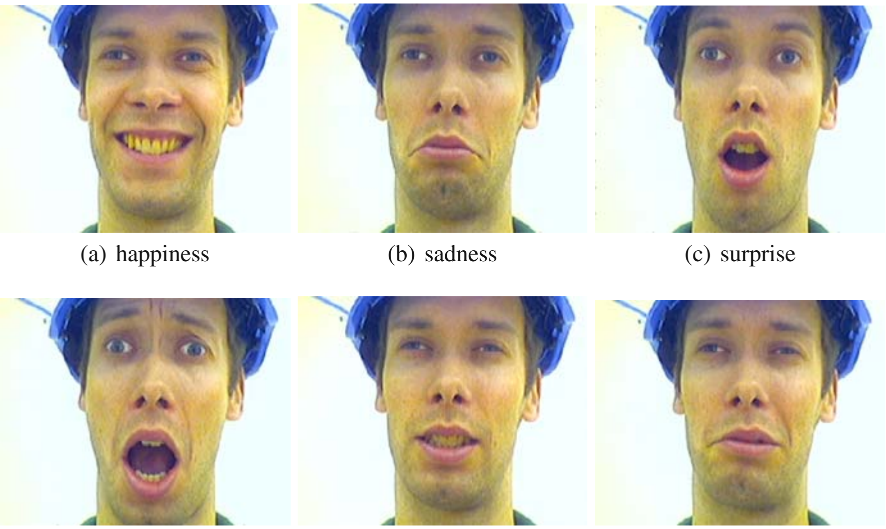
H.264 encoding cannot provide video neither anywhere or anytime. But can principal component analysis (PCA) [8] video coding [22, 25] provide high quality video anywhere and anytime? Instead of using regular PCA it is possible to use of an extension to PCA, called asymmetrical PCA (aPCA) [24, 26]. aPCA decreases the computational complexity and the bitrate plus that it allows encoding to focus on semantically important facial areas and still decode the entire face.

The bitrate for video coding through PCA can be as low as below 5 kbps with a high video quality and spatial resolution. But the complexity for PCA encoding is linearly dependent on the number of pixels in the frames; when high resolution, e.g., HD resolution, is used the complexity will increase and consume power. When PCA is extended into aPCA the complexity is reduced for encoding and transmission of basis functions. aPCA encodes the video by using only a part of the pixels in the original image but decodes the entire image. By combining a subset of pixels and entire frames it is possible to reduce the decoder complexity as well. For PCA and aPCA it is essential that the facial features are positioned on approximately the same pixel positions in all frames. Wearable video equipment will not only provide the user with freedom but also remove much of the need for a computationally heavy normalization of the facial position.

Since aPCA can encode entire frames based on only some features the selected features are given higher quality with aPCA compared to regular PCA for an entire frame. There is a loss of quality in the areas which aren't used for encoding. Since these parts of the frame aren't considered important this loss can be considered negligible. We will show how aPCA outperforms encoding with discrete cosine transform (DCT) of the video when it comes to quality for the selected region. The rest of the frame will have poorer reconstruction quality with aPCA compared to DCT encoding. For H.264 video coding it is also possible to protect a specific area by selecting a region of interest (ROI); similarly to aPCA. For encoding of this video the bitrate used for the background is very low and the quality of this area is reduced. So the bitrate for H.264 can be lowered without sacrificing quality for the important area but not to the same low bitrate as aPCA. Video coding based on PCA has the benefit of a much lower complexity for encoding and decoding compared to H.264 and this is a very important factor. The reduced complexity can be achieved at the same time as the bitrate for transmission is reduced. This lowers the power consumption for encoding, transmission and decoding.

2 Principal Component Analysis

Principal component analysis (PCA) [8] can be used to create compact representations of human faces. This enables PCA to be used for highly efficient video coding and other image processing tasks. All human faces can be represented, recognized and synthesized by using a model of faces called the face space. This space can be extracted with PCA. The faces in the face space all have the same facial expression; often a neutral expression. It is also possible to use PCA to create a space which only contain one persons face but with different facial expressions. This is referred to as the personal face space, facial mimic space or personal mimic space [12, 13]. Instead of different human faces this space contain the face of the same person



(a) happiness (b) sadness (c) surprise

(d) fear (e) anger (f) disgust

Fig. 1 The six basic emotions

but with several different facial expressions. According to the American psychologist Paul Ekman it is enough to model six basic emotions to actually model all facial expressions [4, 5]. The six basic emotions; happiness, sadness, surprise, fear, anger and disgust (Fig. 1), are blended in different ways to create all other possible expressions.

The blending of basic emotions is not directly applicable to a linear combination of digital images, so the space needs more than six dimensions. In [25] we examine how many dimensions that are needed to reach a certain representation quality.

Efficient use of PCA for modeling of any data requires that the global motion is removed from the data set. For facial video this motion corresponds to motion of the entire head, e.g., positional shift and facial rotation. The motion that is modelled with PCA is the local motion, i.e., the changes in the face, the facial mimic. The global motion can be removed with hardware techniques, e.g., hands-free video equipment [22] or software implementations such as facial detection and feature tracking.

PCA provides a natural way for scaling video regarding quality. For the same encoding the decoder can select how many dimensions of the space that are used for decoding and thus scale the quality of the reconstructed video. The built-in scalability of PCA is easily utilized in video compression.

All operations with PCA involves all pixels in a frame K. When PCA is used for video compression the complexity for encoding and decoding is linearly dependent on K. It is desirable to have a low complexity for encoding but it is also desirable to have a high spatial resolution on the decoded video.

PCA extracts the most important information in the data based on the variance of the data, i.e., the variance of the individual pixels throughout the sequence of video

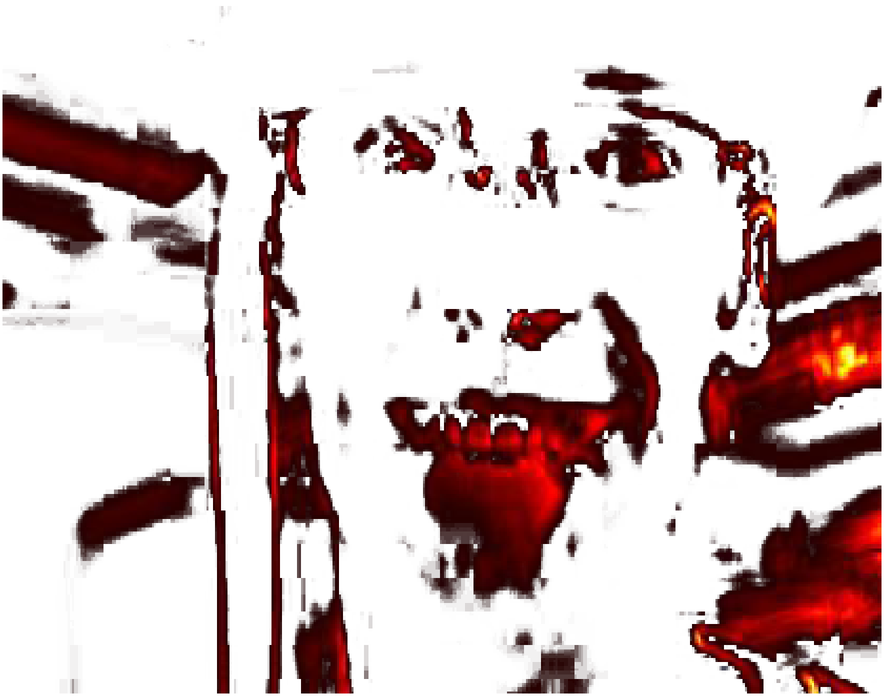


Fig. 2 Variance image of the individual pixels. *White = low variance Red = high variance*

frames. In an ideal case this variance is only dependent on the facial mimic for facial video. But as can be seen in Fig. 2 the background and semantically unimportant parts of the face have high variance.

Pixels with high variance but no semantic importance for the facial mimic will degrade the model efficiency for the facial mimic.

To prevent that these high variance semantically unimportant pixels have effect on the model a region of interest (ROI) can be cropped or extracted from the video frames. But when the video should be decoded it is not interesting to see only parts of the face so the entire face should be decoded. aPCA provides a simple solution to this.

Wearable video equipment allows the user to roam freely and together with HD resolution this enables users to communicate through a high quality medium with both hands free.

In the next section there is an overview of research which is related to this work, including very low bitrate facial representation ans scalable video coding. In section 4 we introduce video coding based on PCA. Section 5 describes the capacity of facial mimic modelling through PCA and in section 6 there is an introduction to aPCA video coding. Section 7 describes wearable video equipment and HD video with aPCA and H.264 is described in section 8. The work is concluded in section 9.

3 Related Work

A representation of the facial mimic has previously been used to encode sequences of faces [28, 29] and head-and-shoulders [22, 23] of persons. All attempts aim at presenting facial images and sequences with high quality at low bitrate cost. Video coding in general does not use PCA. Face images can easily be represented by a combination of Discrete Cosine Transform (DCT)-bases. Most video codecs [19, 32] are based on DCT and this is regarded as state-of-the-art for arbitrary video coding. This representation does however require bitrates higher than most low bandwidth networks can provide. Video compression based on DCT does not provide sufficiently high compression by itself; DCT is combined with temporal compression through block matching (motion estimation). DCT and block-matching requires several DCT-coefficients to encode the frames and several possible movements of the blocks between the frames. Consequently the best codec available today does not provide high quality video at very low bitrates even if the video is suitable for high compression.

Another way to represent the images is as a collection of features from an alphabet. The idea can easily be visualized by the letters in the Arabic alphabet; 26 letters, or features, is sufficient to model all the words in the English language. By building an alphabet for video features it should be possible to model all video frames as a combination of these features. A technique that uses such an alphabet is Matching Pursuit (MP) [11]. The encoder divides the original video image into features from an alphabet and very low bitrate is achieved by only transmitting information about which features that are used between the encoder and decoder. The decoder uses the features to reconstruct the video frame.

Images of a face can also be represented in other ways. Several techniques make use of a wireframe to model faces. The wireframe has the same shape as a face and to make it look more natural it can be texture-mapped with a real image of a face. To make the face move and change appearance between video frames it is enough to transmit information about the changes in the wireframe. Techniques that make use of a wireframe to model facial images are for example MPEG4 facial animation [14] and model based coding [1, 6]. These techniques reach very low bitrates while retaining high spatial resolution and framerate. A method that relies on the statistical model of the shapes and gray-levels of a face is Active Appearance Model (AAM) [3]. All video compression techniques have drawbacks that are critical for efficient usage in visual communication. Pighin *et al.* provides a good explanation why high visual quality is important and why video is superior to animations [16]. The face simply exhibits so many tiny creases and wrinkles that it is impossible to model with animations or low spatial resolution. Therefore any technique based on animation or texture-mapping to a model is not sufficient. Some approaches have focused on retaining the spatial quality of the video frames at the expense of frame rate. Wang and Cohen presented a solution where high quality images are used for teleconferencing over low bandwidth networks with a framerate of one frame each 2-3 seconds [31]. The idea of using low framerate is however not acceptable since both high framerate and high spatial resolution is important for many visual tasks [10]. Any technique that want to provide video at very low bitrates must be able to provide video with high spatial resolution, high framerate and have natural-looking appearance.

Methods that are presented in Video coding (Second generation approach) [27] make use of certain features for encoding instead of the entire video frame. This idea is in line with aPCA since only part of the information is used for encoding in this technique.

Scalable video coding (SVC) has high usage for video content that is received by heterogenous devices. The ability to display a certain spatial resolution and/or visual quality might be completely different if the video is received by a cellular phone or a desktop computer. The available bandwidth can also limit the video quality for certain users. The encoder must encode the video into layers for the decoder to be able to decode the video in layered fashion. Layered encoding has therefore been given much attention in the research community. An review of the scalable extension for H.264 is provided by Schwarz *et.al.* [18].

4 Principal Component Analysis Video Coding

First, we introduce video compression with regular principal component analysis (PCA) [8]. Any object can be decomposed into principal components and represented as a linear mixture of these components. The space containing the facial images is called Eigenspace $\boldsymbol{\Phi}$ and there as many dimensions of this space as there are frames in the original data set. When this space is extracted from a video sequence showing the basic emotions it is actually a personal mimic space. The Eigenspace $\boldsymbol{\Phi}$=$\{\phi_1\ \phi_2\ ...\ \phi_N\}$ is constructed as

$$\phi_j = \sum_i b_{ij}(\mathbf{I}_i - \mathbf{I}_{\underline{0}}) \tag{1}$$

where b_{ij} are values from the Eigenvectors of the the covariance matrix $\{(\mathbf{I}_i - \mathbf{I}_{\underline{0}})^T(\mathbf{I}_j - \mathbf{I}_{\underline{0}})\}$. $\mathbf{I}_{\underline{0}}$ is the mean of all video frames and is constructed as:

$$\mathbf{I}_{\underline{0}} = \frac{1}{N}\sum_{j=1}^{N}\mathbf{I}_j \tag{2}$$

Projection coefficients $\{\alpha_j\}$=$\{\alpha_1\ \alpha_2\ ...\ \alpha_N\}$ can be extracted for each video frame through projection:

$$\alpha_j = \phi_j(\mathbf{I} - \mathbf{I}_{\underline{0}})^T \tag{3}$$

Each of the video frames can then be represented as a sum of the mean of all pixels and the weighted principal components. This representation is error-free if all N principal components are used.

$$\mathbf{I} = \mathbf{I}_{\underline{0}} + \sum_{j=1}^{N}\alpha_j\phi_j \tag{4}$$

Since the model is very compact many principal components can be discarded with a negligible quality loss and a representation of the image with fewer principal components M can represent the image.

$$\hat{\mathbf{I}} = \mathbf{I}_{\underline{0}} + \sum_{j=1}^{M} \alpha_j \phi_j \tag{5}$$

where M is a selected number of principal components used for reconstruction ($M < N$).

The extent of the error incurred by using fewer components (M) than (N) is examined in the next section. With the model it is possible to encode entire video frames to only a few coefficients $\{\alpha_j\}$ and reconstruct the frames with high quality. A detailed description and examples can be found in [21, 22].

PCA video coding provides natural scalable video since the quality is directly dependent on the number of coefficients M that are used for decoding. The decoder can scale the quality of the video frame by frame by selecting the amount of coefficients used for decoding. This gives the decoder large freedom to scale the video without the encoder having to encode the video into scalable layers. The scalability is built-in in the reconstruction process.

5 Theoretical Capacity of PCA

The performance of PCA video coding is dependent on the facial mimic model extracted through PCA. The efficiency of this model will directly decide how many dimensions that are needed to reach a certain representation quality. More than the six basic emotions is needed to represent the facial mimic with digital images. This representation is usually measured objectively in peak signal-to-noise ratio (PSNR) so the result will be a certain quality for a given number of model dimensions; not that it models all possible emotions. We have examined how compact representation that is needed to reach a reconstruction quality on images approximated by the model.

Kirby and Sirovich have previously stated that it is enough with 100 male Caucasian faces to model all possible male Caucasian faces [9, 20]. This means that it is enough with a space of 100 dimensions to model millions of faces. All these face do however have the same facial expression; usually a neutral expression where teeth aren't showing. This doesn't model any facial expression and the modelling of facial mimic isn't considered. The bound is affected by several factors so we have limited the boundary to the following circumstances:

- A spatial resolution of 240x176 pixels.
- A color depth of 8 bits per pixel (0-255).
- RGB color space in the original video (24 bits).
- Theoretical bounds are calculated individually for each person.
- The objectively quality is measured for the R, G and B-channel together.

The spatial resolution for these calculations is chosen to match the resolution of mobile phones. The standard with the highest quality that was used when these boundaries where calculated was called QVGA and has a total pixel number of approximately 77000 [17]. A resolution of 240x176 pixels is equal to 42420 pixels. The resolution can be different and the rate is still unchanged; the quality might be affected.

The representation quality of facial mimic is affected by the number of dimensions that are used for modelling of the mimic. The mean value for each pixel position $\underline{\mathbf{I}_0}$ also affects the modeling in a large extent. An error in representation occurs when less features, or components, ($M < N$) are used for reconstruction. The error is calculated as:

$$mse(opt) = \sum_{j=M+1}^{N} \lambda_{ij} \tag{6}$$

where λ_{ij} are the Eigenvalues for the principal components. The modeling efficiency is calculated as the sum of the error which is incurred by not using a number of Eigenvectors.

How the mean square error can be calculated for a high-dimensional source when the mean has been subtracted is explained by Fukunaga [7]. This will only explain how much information that is removed from mean-centered data. To calculate the quality of the total representation you also need to include the mean.

A mean square error bound can be calculated for the number of Eigenvectors ϕ_j that are used for image representation (equation 6) by varying the number M of dimensions of the model which are used for representation of the facial mimic. This is the distortion bound for representing the signal with the selected number of Eigenvectors. From the mean square error (mse) it is easy to calculate the PSNR. Even though the distortion bound is calculated individually for each person the average result of facial mimic from 10 video sequences (6 different persons) are shown in Table 1. This show the average of maximum quality that can be reached for facial mimic representation.

Table 1 Average PSNR values for 10 facial mimic video sequences

Nr of ϕ_j	5	10	15	20	25
PSNR [dB]	34.56	36.82	38.13	39.07	39.75

The bound will start at 0 Eigenvectors and continue above 25 as well. We have chosen to calculate the boundary from 5 to 25 Eigenvectors. This gives a clear view of how many dimensions that are needed for facial mimic representation.

6 Asymmetrical Principal Component Analysis Video Coding

There are two major issues with the use of full frame encoding:

1. The information in the principal components are based on all pixels in the frame. Pixels that are part of the background or are unimportant for the facial mimic

may have large importance on the model. The model is affected by semantically unimportant pixels.

2. The complexity of encoding, decoding and model extraction is directly dependent on the spatial resolution of the frames, i.e., the number of pixels in the frames. Video frames with high spatial resolution will require more computations than frames with low resolution.

When the frame is decoded it is a benefit of having large spatial resolution (frame size) since this provides better visual quality. A small frame should be used for encoding and a large frame for decoding to optimize the complexity and quality of encoding and decoding. This is possible to achieve through the use of pseudo principal components; information where not all the data is a principal component. Parts of the video frames are considered to be important; they are regarded as foreground $\mathbf{I}^f$.

$$\mathbf{I}^f = crop(\mathbf{I}) \tag{7}$$

The Eigenspace for the foreground Φ^f=$\{\phi_1^f\ \phi_2^f\ ...\ \phi_N^f\}$ is constructed according to the following formula:

$$\phi_j^f = \sum_i b_{ij}^f (\mathbf{I}_i^f - \mathbf{I}_{\underline{0}}^f) \tag{8}$$

where b_{ij}^f are the Eigenvectors from the the covariance matrix $\{(\mathbf{I}_i^f - \mathbf{I}_{\underline{0}}^f)^T(\mathbf{I}_j^f - \mathbf{I}_{\underline{0}}^f)\}$ and $\mathbf{I}_{\underline{0}}^f$ is the mean of the foreground. Encoding and decoding is performed as:

$$\alpha_j^f = (\phi_j^f)(\mathbf{I}^f - \mathbf{I}_{\underline{0}}^f)^T \tag{9}$$

$$\hat{\mathbf{I}}^f = \mathbf{I}_{\underline{0}}^f + \sum_{j=1}^{M} \alpha_j^f \phi_j^f \tag{10}$$

where $\{\alpha_j^f\}$ are coefficients extracted using information from the foreground $\mathbf{I}^f$. The reconstructed frame $\hat{\mathbf{I}}^f$ has smaller size and contains less information than a full size frame. A space which is spanned by components where only the foreground is orthogonal can be created. The components spanning this space are called pseudo principal components and this space has the same size as a full frame:

$$\phi_j^p = \sum_i b_{ij}^f (\mathbf{I}_i - \mathbf{I}_{\underline{0}}) \tag{11}$$

From the coefficients $\{\alpha_j^f\}$ it is possible to reconstruct the entire frame:

$$\hat{\mathbf{I}} = \mathbf{I}_{\underline{0}} + \sum_{j=1}^{M} \alpha_j^f \phi_j^p \tag{12}$$

where M is the selected number of pseudo components used for reconstruction. A full frame video can be reconstructed (Eq. 12) using the projection coefficients from only the foreground of the video (Eq. 9) so the foreground is used for encoding and the entire frame is decoded. It is easy to prove that

$$\hat{\mathbf{I}}^f = crop(\hat{\mathbf{I}}) \tag{13}$$

since $\phi_j^f = crop(\phi_j^p)$ and $\mathbf{I}_{\underline{0}}^f = crop(\mathbf{I}_{\underline{0}})$.

aPCA provides the decoder with a freedom to decide the spatial size of the encoded area without the encoder having to do anything more. Reduction in spatial resolution is not a size reduction of the entire frame; parts of the frame can be decoded with full spatial resolution. No quality is lost in the decoded parts; it is up to the decoder to choose how much and which parts of the frame it wants to decode. The same bitstream is exactly the same regardless of what video size the decoder wants to decode. With aPCA the decoder can scale the reconstructed video regarding spatial resolution and area.

6.1 Reduction of Complexity for the Encoder

The complexity for encoding is directly dependent on the spatial resolution of the frame that should be encoded. The important factor for complexity is $K * M$, where K is the number of pixels and M is the chosen number of Eigenvectors. When aPCA is used the number of pixels k in the selected area gives a factor of $n = \frac{K}{k}$ in resolution reduction.

6.2 Reduction of Complexity for the Decoder

The complexity for decoding can be reduced when a part of the frame is used for both encoding and decoding. In the formulas above we only use the principal components for the full frame ϕ_j^p for decoding but if both Φ^p and Φ^f are used for decoding the complexity can be reduced. Only a few principal components of Φ^p are used to reconstruct the entire frame. More principal components from Φ^f are used to add details to the foreground.

$$\hat{\mathbf{I}} = \mathbf{I}_{\underline{0}} + \sum_{j=1}^{L} \alpha_j^f \phi_j^p + \sum_{j=L+1}^{M} \alpha_j^f \phi_j^f \tag{14}$$

The result is reconstructed frames with slightly lower quality for the background but with the same quality for the foreground $\mathbf{I}^f$ as if only Φ_j^p was used for reconstruction. The quality of the background is decided by parameter L: a high L-value will increase the information used for background reconstruction and increase the decoder complexity. A low L-value has the opposite effect. A significant result is that spatial scalability is achieved naturally. The reduction in complexity (compression ratio *CR*) is calculated as:

$$CR = \frac{K(M+1)}{(1+L)K + (M-L)k} \tag{15}$$

When $k << K$ the compression ratio can be approximated to $CR \approx \frac{M+1}{L+1}$.

7 Wearable Video Equipment

Recording yourself with video usually requires that another person carries the camera or that you use a tripod to place the camera on. When the camera is placed on a tripod the movements that you can make are restricted since the camera cannot move; except for the movements that can be controlled remotely. A wearable video equipment allows the user to move freely and have both hands free for use while the camera follows the movements of the user. The equipment is attached to the back of the person wearing it so the camera films the user from the front. The equipment that we have used is built by the company Easyrig AB and resembles a backpack; it is worn on the back (Fig. 3). It consists of a backpack, an aluminium arm and a mounting for a camera at the tip of the arm.

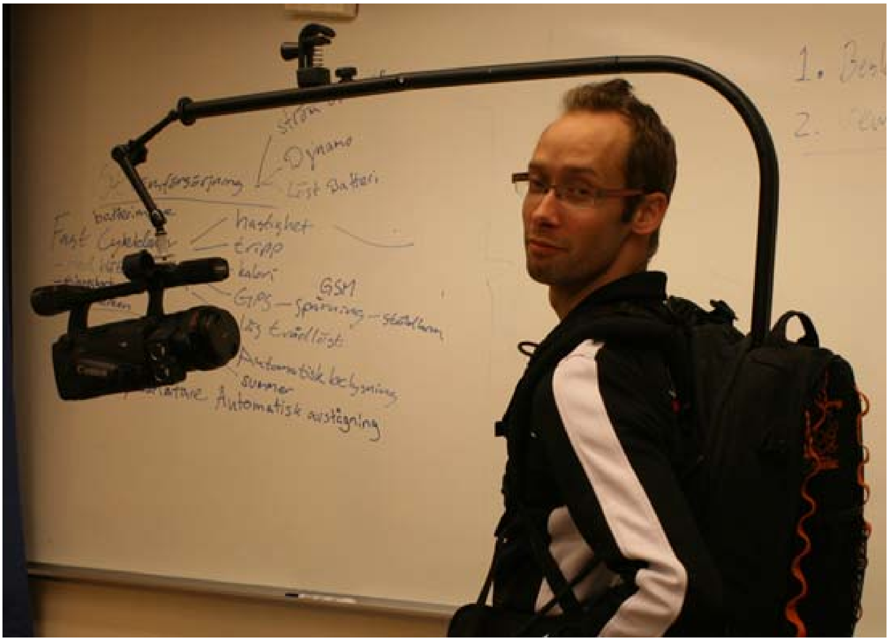


Fig. 3 Wearable video equipment

8 High Definition (HD) Video

High-definition (HD) video refers to a video system with a resolution higher than regular standard-definition video used in TV broadcasts and DVD-movies. The display resolutions for HD video are called 720p (1280x720), 1080i and 1080p (both 1929x1080). i stands for interlaced and p for progressive. Each interlaced frame is divided into two parts where each part only contains half the lines of the frame. The two parts contain either odd or even lines and when they are displayed the human eye perceives that the entire frame is updated. TV-transmissions that have HD resolution use either 720p or 1080i; in Sweden it is mostly 1080i. The video that we use as HD video has a resolution of 1440x1080 (HD anamorphic). It is originally recorded as interlaced video with 50 interlace fields per second but it is transformed into progressive video with 25 frames per second.

8.1 HD Video with H.264

As a comparison of HD video encoded with aPCA we encode the video sequence with H.264 as well. We use the same software for encoding of the entire video as for encoding of the Eigenimages; but we also enable motion estimation. The entire video is encoded with H.264 with a target bitrate of 300 kbps. To get this bitrate we encode the video with a quantization step of 29. We compare the quality of the foreground and background separately since they have different qualities when aPCA is used. With standard H.264 encoding the quality for the background and foreground are approximately equal.

The complexity for H.264 encoding is linearly dependent on the frame size. Most of the complexity for H.264 encoding comes from motion estimation through block matching. The blocks has to be matched for several positions and the blocks can move both in horizontal and vertical direction. The complexity for H.264 encoding is dependent on K and the displacement D in square (D^2). When the resolution is increased the number of displacements are increased. Imagine a line in a video with CIF resolution. This line will consist of a number of pixels, e.g., 5. If the same line is described in HD resolution the number of pixels in the line will increase to almost 19. If the same movement between frames is used in CIF and HD the displacement in pixels is much higher for HD video. When motion estimation is used for H.264 video the complexity grows high because of D^2. So even if the complexity is only linearly dependent on the number of pixels K the complexity grows more faster than linearly for high resolution video.

8.2 HD Video at Low Bitrates

aPCA can be utilized by the decoder to decode parts of the same frame with different spatial resolution. Since the same part of the frame $\mathbf{I}^f$ is used for encoding in both cases, the decoder can choose to decode either $\mathbf{I}^f$ or the entire frame $\mathbf{I}$. The decoded

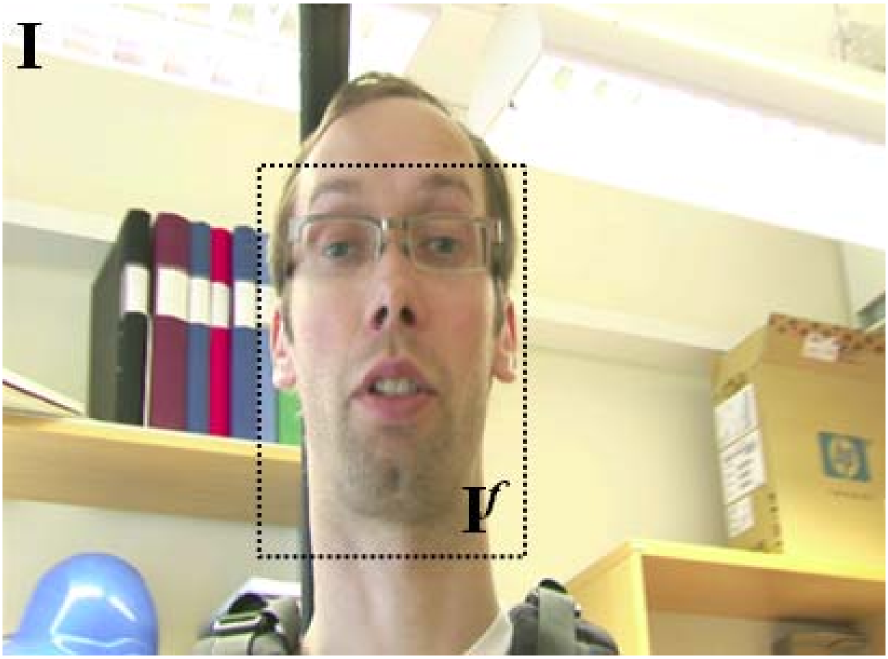


Fig. 4 Frame with the foreground shown

video can also be a combination of $\mathbf{I}^f$ and **I**. This is described in detail in [26]. How Φ^f and Φ^p are combined can be selected by a number of parameters, such as quality, complexity or bitrate. In this work we will focus on bitrate and complexity.

The bitrate that we select as a target for video transmission is 300 kbps. The video needs to be transmitted below this bitrate at all times. The frame size for the video is 1440x1080 (**I**). The foreground in this video is 432x704 ($\mathbf{I}^f$) (Fig. 4). After YUV 4:1:1 compression the number of pixels in the foreground is 456192. The entire frame **I** consists of 2332800 pixels and the frame area which is not foreground is 1876608 pixels. The video has a framerate of 25 fps but this has only slight impact on the bitrate for aPCA since each frame is encoded to a few coefficients. The bitrate for these coefficients is easily kept below 5 kbps. Audio is an important part of communication but we will not discuss this in our work. There are several codecs that can provide audio with good quality at a bitrate which can be used. We use 300 kbps for transmission of the Eigenimages (Φ^p and Φ^f) and the coefficients $\{\alpha_j^f\}$ between sender and receiver.

Transmission of the Eigenimages ϕ_j means transmission of images. The Eigenimages have too large size $\approx$ 7,5 MB (1440x1080 resolution minus the foreground) to be transmitted without compression. Since they are images we could use image compression but the images share large similarities in appearance; the facial mimic is independent between the images but it is the same face with similar background. Globally the images are not only uncorrelated but also independent and doesn't share any similarities. Image or video compression based on DCT divides the frames into blocks and encodes each block individually. Even though the frames are independent globally it is possible to find local similarities so to consider the images as a sequence will provide higher compression. We want to remove the complexity associated with motion estimation and only encode the images through DCT.

We use the H.264 video compression without any motion estimation; this encoding uses both intracoding and intercoding. The first image is intracoded and the subsequent images are intercoded but without motion estimation. The mean image is only one image so we will use the JPEG [15, 30] standard for compression of it. The mean image is in fact compressed in the same manner as in [25].

To make the compression more efficient we first use quantization of the images. In our previous article we discussed the usage of pdf-optimized or uniform quantization extensively and came to the conclusion that it is sufficient to use uniform quantization [25]. So, in this work we will use uniform quantization. In our previous work we also examined the effect of high compression and loss of orthogonality between the Eigenimages. To retain high visual quality on the reconstructed frames we will not use so high compression that the loss of orthogonality becomes an important factor. The compression is achieved through the following steps:

- Quantization of the Eigenimages. Φ^Q=Q(Φ)
- The Eigenimages are compressed into a video sequence. Φ^{Comp}=C(Φ^Q)
- Reconstruction of the Eigenimages from compressed video. $\hat{\Phi}^Q$=C′(Φ^{Comp})
- Inverse quantization mapping of the quantization values with the reconstruction values. $\hat{\Phi}$=Q′($\hat{\Phi}^Q$)

The mean image $\mathbf{I}_{\underline{0}}$ is compressed in a similar way but we use JPEG compression instead of H.264. We have 295 kbps for Eigenimage transmission and this is equal to $\approx$ 60 kB. The foreground $\mathbf{I}^f$ have a size of $\approx$ 1,8 MB when it is uncompressed. It is possible to choose from a wide range of compression grades when it comes to encoding with DCT. We select a compression ratio based on reconstruction quality that the Eigenimages provides and the bitrate which is needed for transmission of the video; the compression is chosen by the following criteria.

- A compression ratio that allow the use of a bitrate below our needs.
- A compression ratio that provide sufficiently high reconstruction of video when compressed Eigenimages are used for encoding and decoding of video.

The first criteria decides how fast the Eigenimages can be transmitted; e.g., how fast high quality video can be decoded. The second criteria decides the quality of reconstructed video.

8.3 aPCA Decoding of HD Video

The face is the most important information in the video so Eigenimages ϕ_j^f for the foreground $\mathbf{I}^f$ is transmitted first. The bitrate for the compressed Eigenimages

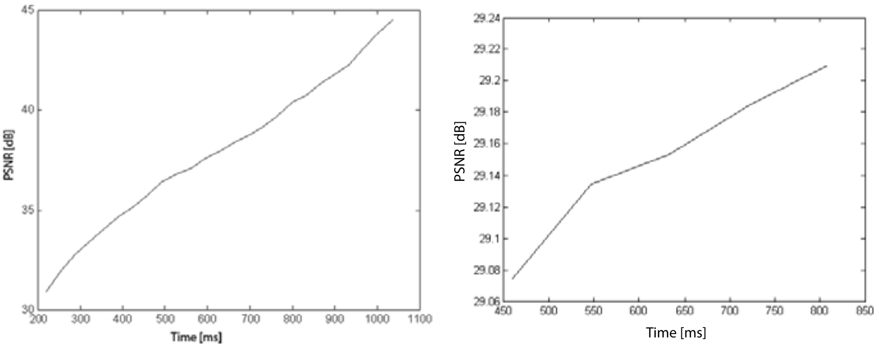
$\phi^{f^{Comp}}$ is 13 kbps but the bitrate for the first Eigenimage is higher since it is intra-coded. The background is larger in spatial size so the bitrate for this is 42 kbps. Transmission of 10 Eigenimages for the foreground ϕ^{Comp^f} , 1 pseudo Eigenimage for the background ϕ^{Comp^p} plus the mean for both areas can be done within 1 second. After $\approx$ 220 ms the first Eigenimage and the mean for the foreground is available and decoding of the video can start. All the other Eigenimages are inter-coded and a new image arrives every 34th ms. After $\approx$ 520 ms the decoder has 10 Eigenimages for the foreground. The mean and the first Eigenimage for the background needs $\approx$ 460 ms for transmission and a new Eigenimage for the background can be transmitted every 87th ms. The quality of the reconstructed video is increased as more Eigenimages arrive. There doesn't have to be a stop to the quality improvement; more and more Eigenimages can be transmitted. But when all Eigenimages that the decoder wants to use for decoding has arrived only the coefficients needs to be transmitted so the bitrate is then below 5 kbps. The Eigenimages can also be updated; something we examined in [25]. The Eigenspace may need to be updated because of loss of alignment between the model and the new video frames.

Table 2 Reconstruction quality for the foreground

	Rec. qual. PSNR [dB]		
Compression method	Y	U	V
H.264	36.4	36.5	36.5
aPCA	44.2	44.3	44.3

Table 3 Reconstruction quality for the background

	Rec. qual. PSNR [dB]		
Compression method	Y	U	V
H.264	36.3	36.5	36.6
aPCA	29.6	29.7	29.7



(a) Foreground quality

(b) Background quality

Fig. 5 Quality of the Y-channel over time

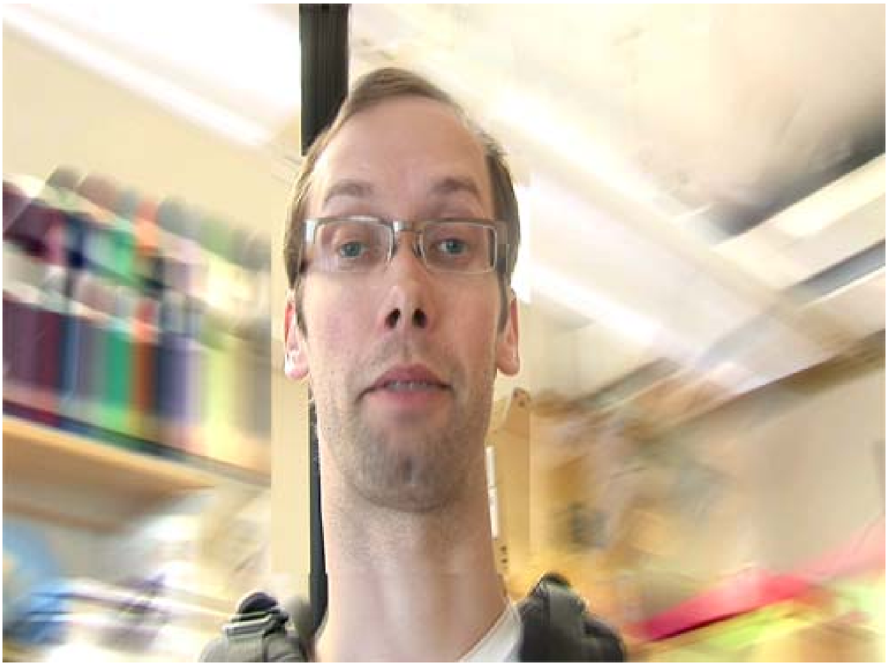


Fig. 6 Frame reconstructed with aPCA. (25 ϕ_j^f and 5 ϕ_j^p are used.)

The average results measured in psnr for the video sequences are shown in Table 2 and Table 3. Table 2 show the results for the foreground and Table 3 show the results for the background. The results in the tables are for full decoding quality (25 ϕ_j^f and 5 ϕ_j^p). Fig. 5 show how both the foreground and background quality of the Y-channel is increased over time for aPCA. An example of a frame reconstructed with aPCA is shown in Fig. 6. A reconstructed frame from H.264 encoding is shown in Fig. 7.

As it can be seen from the tables and the figures the background quality is always lower for aPCA compared with H.264. This will not change even if all Eigenimages are used for reconstruction; the background is always blurred. The exception is when the background is homogenous but the quality of this background with H.264 encoding is also very good.

The foreground quality for aPCA is better than H.264 already when 10 Eigenimages (after $\approx$ 1 second) are used for reconstruction and just improves after that.

That the quality doesn't increase linearly depends on the fact that the Eigenimages that are added to reconstruction have different mimics. The most important mimic is the first so it should improve the quality the most and the subsequent ones should improve the quality less and less. But the 5th expression may improve some frames with really bad reconstruction quality and thus increase the quality more than the 1st Eigenimage. It may also improve the mimic for several frames; the most

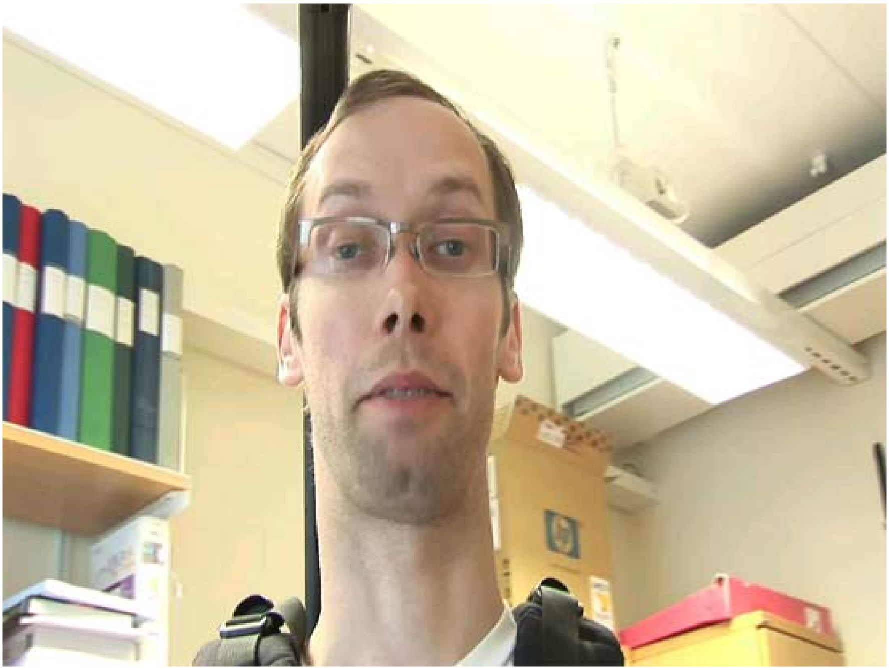


Fig. 7 Frame encoded with H.264

important mimic can be visible in fewer frames than another mimic which is not as important based on the variance.

9 Conclusions

Asymmetrical principal component analysis (aPCA) has large potential when it comes to compression of facial video sequences, especially HD video. The complexity for encoding can be reduced more than 10 times and the complexity for decoding is also reduced at the same time as the objective quality is lowered slightly, i.e., 1 dB (PSNR). aPCA is also very adaptive for heterogenous decoding since a decoder can select which size of video frames it wants to decode with the encoder using the same video for encoding. PCA provides natural scalability of the quality and aPCA also provides scalability in spatial resolution with the same encoding. The freedom of assembling the reconstructed frames differently also provide the decoder with the freedom to select different quality for different parts of the frame.

The use of aPCA for compression for video with HD resolution can reduce the bitrate for transmission vastly after an initial transmission of Eigenimages. The available bitrate can also be used to improve the reconstruction quality further. A drawback with any implementation based on PCA is that it is not possible to reconstruct a changing background with high quality; it will always be blurred due to motion.

Initially there are no Eigenimages available at the decoder side and no video can be displayed. This initial delay in video communication cannot be dealt with by buffering if the video is used in online communication such as a video telephone conversation. This shouldn't have to be a problem for video conference applications since you usually don't start communicating immediately. And a second is enough time to wait for good quality video. When existing Eigenspaces are used it is possible to have transmission of HD video below 5 kbps from the start of a communication session. Then transmission of HD video will require less bits than audio transmission. There are possibilities of combining PCA or aPCA with DCT encoding such as h.264 and this will be a hybrid codec. For an initial period the frames can be encoded with h.264 and transmitted between the encoder and decoder. The fames are available at both the encoder and decoder so they can both perform PCA for the images and produce the same Eigenimages. All other frames can then be encoded with the Eigenimages to very low bitrates with low encoding and decoding complexity.

References

[1] Aizawa, K., Huang, T.S.: Model-based image coding: Advanced video coding techniques for very low bit-rate applications. Proc. of the IEEE 83(2), 259–271 (1995)

[2] Argyle, M.: Bodily Communication. Methuen and Co., NewYork (1988)

[3] Cootes, T., Edwards, G., Taylor, C.: Active appearance models. In: Proc. European Conference on Computer Vision (ECCV), vol. 2, pp. 484–498 (1998)

[4] Ekman, P., Friesen, W.V.: Unmasking the face. A guide to recognizing emotions from facial clues. Prentice-Hall, Englewood Cliffs (1975)

[5] Ekman, P.: Emotion in the Human Face. Cambridge University Press, New York (1982)

[6] Forchheimer, R., Fahlander, O., Kronander, T.: Low bit-rate coding through animation. In: Proc. International Picture Coding Symposium PCS 1983, pp. 113–114 (1983)

[7] Fukunaga, K.: Introduction to statistical pattern recognition, 2nd edn. Academic Press Professional, Inc., San Diego (1990)

[8] Jolliffe, I.: Principal Component Analysis. Springer, New York (1986)

[9] Kirby, M., Sirovich, L.: Application of the karhunen-loeve procedure for the characterization of human faces. IEEE Transactions on pattern Analysis and Machine Intelligence 12(1), 103–108 (1990)

[10] Lee, J., Eleftheriadis, A.: Spatio-temporal model-assisted compatible coding for low and very low bitrate video telephony. In: Proceedings, 3rd IEEE International Conference on Image Processing (ICIP 1996), Lausanne, Switzerland, pp. II.429–II.432 (1996)

[11] Neff, R., Zakhor, A.: Very low bit-rate video coding based on matching pursuits. IEEE Transactions on Circuits and Systems for Video Technology 7(1), 158–171 (1997)

[12] Ohba, K., Clary, G., Tsukada, T., Kotoku, T., Tanie, K.: Facial expression communication with fes. In: International conference on Pattern Recognition, p. 1378 (1998)

[13] Ohba, K., Tsukada, T., Kotoku, T., Tanie, K.: Facial expression space for smooth telecommunications. In: FG 1998: Proceedings of the 3rd. International Conference on Face & Gesture Recognition, p. 378 (1998)

[14] Ostermann, J.: Animation of synthetic faces in mpeg-4. In: Proc. of Computer Animation, pp. 49–55. IEEE Computer Society, Los Alamitos (1998)

[15] Pennebaker, W.B., Mitchell, J.L.: JPEG Still Image Data Compression Standard. Van Nostrand Reinhold, New York (1993)

[16] Pighin, F., Hecker, J., Lishchinski, D., Szeliski, R., Salesin, D.H.: Synthesizing realistic facial expression from photographs. In: SIGGRAPH Proceedings, pp. 75–84 (1998)

[17] Rasmusson, J., Dahlgren, F., Gustafsson, H., Nilsson, T.: Multimedia in mobile phones - the ongoing revolution. Ericsson Review no. 01 (2004)

[18] Schwarz, H., Marpe, D., Wiegand, T.: Overview of the scalable video coding extension of the H.264/avc standard. IEEE Transactions on Circuits and Systems for Video Technology 17(9), 1103–1120 (2007)

[19] Schäfer, R., et al.: The emerging H.264 avc standard. EBU Technical Review 293 (2003)

[20] Sirovich, L., Kirby, M.: Low-dimensional procedure for the characterization of human faces. Journal of the Optical Society of America 4, 519–524 (1987)

[21] Söderström, U., Li, H.: Very low bitrate full-frame facial video coding based on principal component analysis. In: Signal and Image Processing Conference (SIP 2005), August 15-17 (2005)

[22] Söderström, U., Li, H.: Full-frame video coding for facial video sequences based on principal component analysis. In: Proceedings of Irish Machine Vision and Image Processing Conference 2005 (IMVIP 2005), August 30-31, pp. 25–32 (2005)

[23] Söderström, U., Li, H.: Eigenspace compression for very low bitrate transmission of facial video. In: IASTED International conference on Signal Processing, Pattern Recognition and Application, SPPRA (2007)

[24] Söderström, U., Li, H.: Asymmetrical principal component analysis for video coding. Electronics letters 44(4), 276–277 (2008)

[25] Söderström, U., Li, H.: Representation bound for human facial mimic with the aid of principal component analysis (2008)

[26] Söderström, U., Li, H.: Asymmetrical principal component analysis for efficient coding of facial video sequences (2008)

[27] Torres, L., Kunt, M.: Video Coding (The Second Generation Approach). Kluwer Academic Publishers, Dordrecht (1996)

[28] Torres, L., Delp, E.: New trends in image and video compression. In: Proceedings of the European Signal Processing Conference (EUSIPCO), Tampere, Finland, September 5-8 (2000)

[29] Torres, L., Prado, D.: A proposal for high compression of faces in video sequences using adaptive eigenspaces. In: Proceedings of 2002 International Conference on Image Processing, vol. 1, pp. I-189– I-192 (2002)

[30] Wallace, G.K.: The jpeg still picture compression standard. Communications of the ACM 34(4), 30–44 (1991)

[31] Wang, J., Cohen, M.F.: Very low frame-rate video streaming for face-to-face teleconference. In: DCC 2005: Proceedings of the Data Compression Conference, pp. 309–318 (2005)

[32] Wiegand, T., Sullivan, G.J., Bjontegaard, G., Luthra, A.: Overview of the H.264/avc video coding standard. IEEE Trans. Circuits Syst. Video Technol. 13(7), 560–576 (2003)

[33] `http://www.polycom.com` (2008-08-12)

Evolution of IPTV Architecture and Services towards NGN

Eugen Mikoczy and Pavol Podhradsky

Abstract. New technologies and advances enable the evolution of digital television and IPTV. Next step in IPTV evolution, which is actually quite frequently being discussed in the telecommunication community, industry and also standardization bodies like ETSI TISPAN or ITU-T, is the evolution of the Next Generation Network (NGN) architecture in order to provide the IPTV services over same NGN infrastructure. This chapter will explain main trends driving the IPTV evolution and highlight the most relevant standardization efforts in this area. Several concepts of IPTV architecture are presented as well as main differences and possible migration scenarios amongst various alternative solutions like non-NGN IPTV, NGN non-IMS, IMS based IPTV and the converged NGN IPTV architecture.

1 Introduction

The concept of NGN has been evolving for several years and first real standardization activities have been started by ITU-T (International Telecommunication Union) an NGN focus group and ETSI (European Telecommunications Standards Institute) TISPAN (The Telecoms & Internet Converged Services & Protocols for Advanced Networks). But the first version of NGN was more voice oriented and besides that the definition of NGN required the capability to provide multimedia services. Therefore only such NGN which is also capable to provide the Internet Protocol Television (IPTV) services fully meets the idea of the multiservice NGN architecture.

In fact, ITU-T defines the IPTV by the following definition [1]:

IPTV are multimedia services such as television/video/ audio/text/graphics/data delivered over IP-based networks managed to support the required level of QoS/QoE (Quality of Service/Experience), security, interactivity and reliability.

Eugen Mikoczy and Pavol Podhradsky
NGNlab, Department of Telecommunications, FEI, Slovak University of Technology,
Ilkovičova 3, 812 19, Bratislava, Slovakia
e-mail: {mikoczy,podhrad}@ktl.elf.stuba.sk

M. Grgic et al. (Eds.): Rec. Advan. in Mult. Sig. Process. and Commun., SCI 231, pp. 315–339.
springerlink.com

ETSI TISPAN employs IP Multimedia Subsystem (IMS) concept from TISPAN NGN Releases 1 (NGN R1) and later releases. Now IMS has been widely accepted as one of the key platforms for future NGN service composition, orchestration and delivery. IMS based NGN has been considered as a long term vision of telecom industry of how to develop unified service architecture using multi-service concept. There where services can be provided and controlled over a single common service control platform (e.g. based on IMS) independently from any access technology.

Therefore operators can move from vertical silo architecture where each type of service has dedicated access, transport, control, and application infrastructure per service to horizontally oriented architecture more independent form provided services (Fig.1).The main idea of NGN based IPTV is include functionalities and infrastructure required for any of multimedia NGN services specially here the IPTV type of services to NGN architecture.

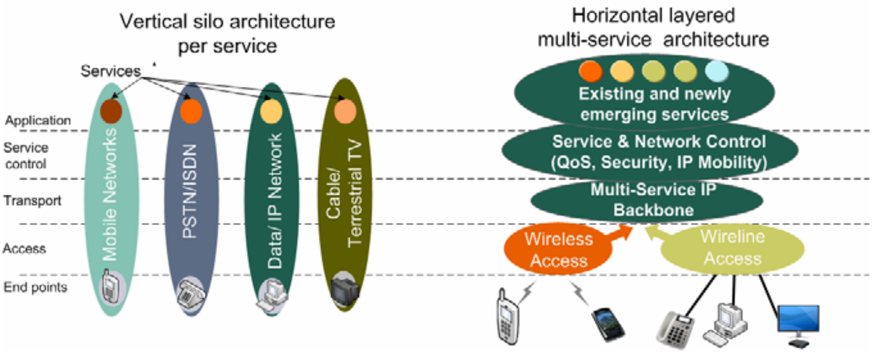

Vertical silo architecture per service
Services
Application
Service control
Transport
Access
End points
Mobile Networks
PSTN/ISDN
Data/ IP Network
Cable/ Terrestrial TV
Horizontal layered multi-service architecture
Existing and newly emerging services
Service & Network Control (QoS, Security, IP Mobility)
Multi-Service IP Backbone
Wireless Access
Wireline Access

Fig. 1 From vertical silos to horizontal NGN architecture

The digital television started as terrestrial or satellite digital television several years ago. The second generation of the digital television could be considered a new service known as the Internet Protocol based Television (IPTV). Service providers began to provide the IPTV only few years ago, mainly because of new types of broadband access networks and new advance video coding technologies. The IPTV solution [1] is defined as a composition of functions and interfaces needed to provide the IPTV service.

But the most actual question is whether the IPTV concept is capable to provide much more than a traditional broadcasted digital TV, for example interactivity, mobility and service personalization as providers promise. What is the actual state-of-art in the IPTV technologies and could the IPTV be integrated as part of Next Generation Networks (NGN) in the future? We will try to explain this in following sections of this chapter and give you also overview about NGN based architecture and services which could be potentially considered as “third generation” of digital television service.

2 IPTV State-of-art Overview

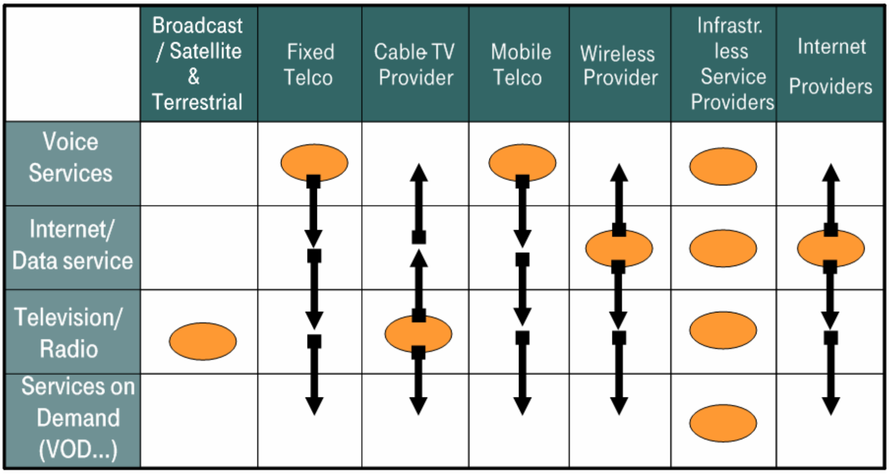
Several aspects have led the industry to the state-of-art in the IPTV technologies. First of all advances in technology and market trends enable the introduction of first generation of IPTV technology following by the standardization effort of most important standardization bodies in telecommunications.

2.1 IPTV Technology Aspects and Market Trends

During the past 5 years the IPTV technology has evolved from hi-end service to usual provider service offer as part of Triple play kind of product (Voice, Data, Video). In some cases there also quad play offers combining additionally mobile services to the product package. The preconditions to such a wide number of deployments in short time are responsible following aspects and advances in several fields of technology:

1. High speed networks – advances in technology of access and transport networks (xDSL - digital subscriber line, FTTX – Fiber to the X, GPON- Gigabit Passive Optical Network)
2. Advances in signal processing and video encoding (MPEG-4 AVC encoding bitrate improving compared to MPEG2 - Moving Picture Experts Group).
3. Digitalization of television in satellite or terrestrial transmission (some countries already replace analog TV with digital).
4. Lower cost of equipment and improvements in end devices (e.g. Single on Chip STBs).
5. Technology for mobile TV.
6. Hype in use of Internet and streaming or user generated content provided from Internet over the top of the internet access (e.g. web TV).
7. Development of IPTV multimedia services which enable additional features and services compared to Linear TV delivery (content on demand, personal video recording, trick play, interactive TV, etc.).

The market for providing digital content and multimedia services has also dramatically changed. All existing providers or new ones have to fight with newcomers and extension of traditional business areas of all market players because the unified converged market of telecommunication services required from the provider to cover wider portfolio of services (including voice services, data services, access to Internet, video/television services, mobile services). For example (Fig.2), traditional Service providers which originally provided only voice services have added to their portfolio the internet access and these days plenty of them provide also TV over IP networks. In opposite direction, but with the same aim, also Cable operators have extended their portfolio from TV over cable to provide also high speed internet access and voice services.



Note: ● historically, the main service provided over the providers platform

Fig. 2 Market trends and extension of service providers

2.2 *IPTV Standardization and Relevant Forums*

During the last years, many efforts have been made on the IPTV standardizations. We would like to mention the most significant ones to give the readers an overview as well as references to the relevant documents.

2.2.1 DVB

The Digital Video Broadcasting Project [DVB] is an industry-led consortium of broadcasters, manufacturers, network operators, software developers, regulatory bodies and others in over 35 countries committed to design open interoperable standards for the global delivery of digital media services (most of the specifications are published later as ETSI standards).

The DVB is dealing with several aspects of digital video broadcasting such as:

1. Transmission technology for satellite (DVB-S or second generation in DVB-S2), terrestrial (DVB-T, DVB-T2), cable (DVB-C) and mobile handheld (DVB-H).
2. Security aspects for digital television in area of Content Protection Copy Management and Conditional Access.
3. Multiplexing and metadata of digital content and related service information (e.g. subtitles, program guides, etc.)
4. Interactivity and middleware (including MHP - Multimedia Home Platform)
5. Delivery of DVB-Services over IP-based Networks (DVB-IPTV).

The DVB work, related to the IPTV is done on DVB-IPTV standards [DVB-IPTV] (formerly DVB-IPI as DVB - Internet Protocol Infrastructure), could be find in the DVB-IPTV BlueBook [2] which provides a set of technical specifications to cover the delivery of DVB MPEG 2 based services over bi directional IP networks, including specifications of the transport encapsulation of MPEG 2 services over IP and the protocols to access such services. Another important issue is the specification of the Service Discovery and Selection (SD&S) mechanism for DVB MPEG 2 based Audio/Video (A/V) services over bi directional IP networks to define the service discovery information, and its data format and the protocols.

2.2.2 ITU-T

International Telecommunication Union [ITU-T] is the leading United Nations agency for information and communication technologies. As the global focal point for governments and the private sector, ITU's role in helping the world communicate spans 3 core sectors: radiocommunication (ITU-R), standardization (ITU-T) and development (ITU-D). The main products of ITU-T are Recommendations (ITU-T Recs) – standards defining of how the telecommunication networks operate and interwork. ITU-T is in force on topics from service definition to network architecture and security, from broadband DSL to Gbit/s optical transmission systems to next-generation networks (NGN) and IP-related issues, all together fundamental components of today's information and communication technologies (ICTs).

ITU-T Focus Group IPTV evaluates the IPTV on various issues such as services, architecture, IPTV middleware, security, etc. Several WGs produce documents which have been transferred from ITU-T IPTV FG to SG13 (responsible primarily for NGN) and another relevant study group as part of ITU-T IPTV SGI (Global Standards Initiative) [IPTV SGI]. It is expected that IPTV documents will be evaluated and finalized for inclusion to the next ITU-T NGN recommendations. First standard published from the IPTV SGI is related to NGN based IPTV architecture: Recommendation ITU-T Y.1910 IPTV functional architecture [3].

2.2.3 ETSI

The European Telecommunications Standards Institute [ETSI] produces globally-applicable standards for Information and Communications Technologies (ICT), including fixed, mobile, radio, converged, broadcast and internet technologies. ETSI is officially recognized by the European Commission as a European Standards Organization.

Since its creation in 2003, ETSI TISPAN (Telecommunications and Internet converged Services and Protocols for Advanced Networking) [TISPAN] has been the key standardization body in creating the Next Generation Networks (NGN) specifications. NGN Release 1 was finalized in December 2005, provided the robust and open standards for the first generation of NGN systems. The NGN Release 1 specifications adopt the 3GPP IMS (IP Multimedia Subsystem) standard for SIP-based applications, but also add further functional blocks and subsystems to handle the non-SIP applications. Initially TISPAN worked on harmonizing the IMS core for both wireless and wireline networks (TISPAN NGN R1 with 3GPP R7 and TIPSAN NGN R2 with 3GPP R8). However, in early 2008, the common IMS specifications were transferred back to 3GPP so that one single standard organization is responsible for providing a Common IMS fitting of any network (fixed, 3GPP, CDMA2000, etc.). The NGN Release 2 was finalized early 2008, and added a key element to the

NGN such as the IMS and non IMS based IPTV, Home Networks and devices, as well as the NGN interconnected with Corporate Networks.

In TISPAN NGN contain several specifications which are addressing to the IPTV regarding service requirements [4] in stage 1 and architectures in stage 2 with non-IMS IPTV subsystem [5] as well as the IMS based IPTV [6]. In the stage 3 TISPAN IPTV specification deals with protocol and implementation details. We will focus on ETSI TISPAN NGN based IPTV architecture and present more details in section 4 of the chapter.

2.2.4 3GPP

The 3rd Generation Partnership Project [3GPP] unites telecommunications standards bodies world wide (including ETSI, ATIS, ARIB, CCSA, TTA, TTC).

The original scope of 3GPP was to produce Technical Specifications and Technical Reports for a 3G Mobile System based on evolved GSM core networks and the radio access technologies that they support (i.e., Universal Terrestrial Radio Access (UTRA) both Frequency Division Duplex (FDD) and Time Division Duplex (TDD) modes).

The scope was subsequently amended to include the maintenance and development of the Global System for Mobile communication (GSM) Technical Specifications and Technical Reports including evolved radio access technologies (e.g. General Packet Radio Service (GPRS) and Enhanced Data rates for GSM Evolution (EDGE)) as well as the third generation mobile network architecture called UMTS (Universal Mobile Telecommunications System). Just as GSM has become synonymous with the whole mobile system for 2G, UMTS is 3G, which includes the whole of the W-CDMA (Wideband Code Division Multiple Access) and HSPA (High-Speed Packet Access) specifications catalogue. 3GPP has published several releases (R99, R4, R5, R6, R7, R8 and most recently R9) which describe some aspects of UMTS services, architectures and protocols. For the core network domain and providing multimedia services is crucial a IP Multimedia Subsystem (IMS) [7] already accepted by the industry as the unified service control architecture for NGN (some other bodies like ITU-T or ETSI TISPAN also bring the concept of IMS based IPTV where IMS is used also for provided IPTV services). 3GPP is now working on Long Term Evolution (LTE), which will be built up on UMTS, as the Industry looks beyond 3G as well as related concept of Service Architecture Evolution (SAE). 3GPP will contribute to the ITU-R towards the development of the IMT-Advanced via its proposal for the LTE-Advanced.

3GPP specifies also MBMS (Multimedia Multicast/Broadcast Services) specifications [8] mainly to define an efficient way to deliver and control multicast and broadcast services over 3G networks. MBMS are limited at this moment to Mobile TV channel bandwidth which is up to 256kbit/s per MBMS Bearer Service. The Advantage of MBMS in comparison with the DVB-H is that the same IP-based common infrastructure is used for Mobile TV as for 3G data services.

2.2.5 OMA

The Open Mobile Alliance [OMA] introduces the concept service enablers to provide standardized components in order to create an environment in which services may be developed and deployed. The OMA enablers, the decomposition into these components and the interactions among them comprise the OSE (OMA Service Environment) framework architecture [9].

From the Mobile TV perspective most relevant work in the OMA is dealing with Mobile Broadcast Services Enabler [10] to address functional issues which are generic enough to be common to many Broadcast Services (OMA BCAST) which can be defined and implemented in a bearer-independent way. These functional issues are: Service Guide, File Distribution, Media Stream Distribution, Service Protection, Content Protection, Service Interaction, Service Provisioning, Terminal Provisioning and Notification etc. Generally it is expected that Mobile Broadcast Services should enable the distribution of rich, interactive, and bandwidth consuming media content to a large number of mobile audiences.

2.2.6 IETF

The Internet Engineering Task Force [IETF] is a protocol engineering and development arm of the Internet. The most of the IPTV standards are using existing protocols defined by the IETF from the Internet Protocol (IP) version 4 or 6 [11,12] itself to application protocols like the Session Initiation Protocol (SIP) for session control [13], the Real-Time Stream Protocol (RTSP) [14] for media control of CoD (content on demand) services as well as the Internet Group Management Protocol (IGMP) [15] for the IPv4 or Multicast Listener Discovery (MLD) [16] for the IPv6 multicast based services. The Real-Time Transport Protocol (RTP) is used for media delivery [17]. The Hypertext Transfer Protocol (HTTP) [18] is used for transfer of metadata or presentation of client graphical interface using web browser technologies.

There are also several internet drafts regarding the IPTV channel description to enable a unified IPTV service identification within the IETF.

2.2.7 ATIS

The Alliance for Telecommunications Industry Solutions [ATIS] is a standardization organization that develops technical and operational standards for the communications industry and is accredited by the American National Standards Institute (ANSI).

The ATIS (Alliance for Telecommunications Industry Solutions) initiates the IPTV Interoperability Forum (IIF) that develops the standards to enable the interoperability, interconnection and implementation of the IPTV systems and services, including video on demand and interactive TV services [19]. The ATIS analyzes within the IIF initiative several important aspects the of IPTV and define IPTV logical domains, IPTV reference architectures (IMS and non-IMS based IPTV, and their coexistence) [20], content delivery concepts with quality of experience, digital rights management (DRM) requirements, interoperability standards as well as testing requirements for components, reliability and robustness of service components.

2.2.8 CableLabs

Cable Television Laboratories, Inc. [PacketCable] is a non-profit research and development consortium founded by cable operating companies dedicated to pursue new cable telecommunications technologies. The PacketCable is one of CableLabs-led initiative to develop interoperable interface specifications for delivering advanced, real-time multimedia services over two-way cable plant. Built on top of the industry's highly successful DOCSIS (Data Over Cable Service Interface Specification) cable modem infrastructure (version 1.1 or greater), PacketCable networks use Internet protocol (IP) technology to enable a wide range of multimedia services, such as IP telephony, multimedia conferencing, interactive gaming, and general multimedia applications. In the architecture of the PacketCable 2.0 [21] the IMS concept is utilized, too.

2.2.9 Open IPTV Forum

The Open IPTV Forum [OIPF] is a pan-industry initiative with the purpose of producing end to end specifications for the development of end-to-end solution to allow any consumer's end-device, compliant to the Open IPTV Forum specifications, to access enriched and personalized IPTV services [22] either in managed or a non-managed networks (non-managed means without guarantee of QoS and over Open Internet). To that end, the Open IPTV Forum focuses on standardizing the user-to-network interface (UNI) both for managed and non-managed network with their NGN based architecture [23].

2.2.10 HGI

The Home Gateway Initiative [HGI] is an open forum launched by Telecommunication providers with the aim to release specifications of the home gateway. In addition to telecommunication providers, several manufacturers have joined the alliance and actually provide specification for NGN capable home gateways [24].

3 IPTV Domains and Services

The end user finally percepts the quality and IPTV service portfolio as well as the usability in order to satisfy his requirements. Several actors are responsible for the delivery of the content from their originators such as TV stations and studios but probably also from other users. In this section are described the IPTV domains and services together with the explanation of how the standardization develops from requirements to architecture.

3.1 IPTV Domains

End to end chain for delivery of the IPTV content to the end user usually contains these 4 main domains that are involved in the provision of an IPTV service (Fig.3):

- Content provider,
- Service provider,
- Network provider,
- End-user.

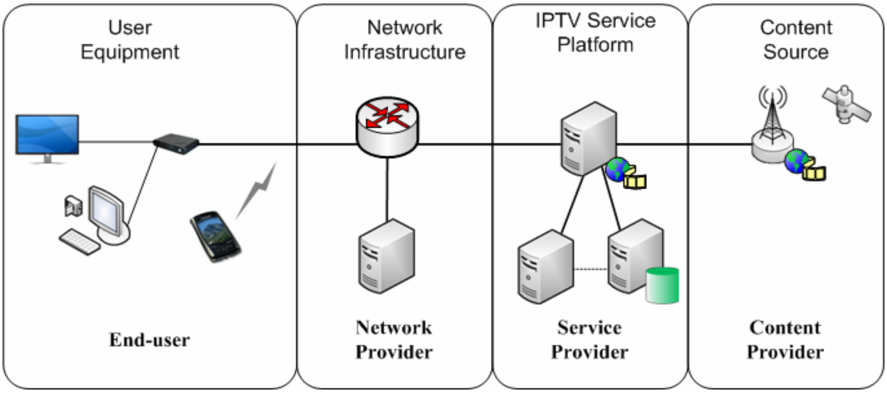


Fig. 3 IPTV Domains [3]

The four IPTV domains definitions could be provided by the ITU-T [3] or ETSI TISPAN specification [4].

3.2 From Requirements to IPTV Services and Architecture

Most of the standardization bodies follow the same schema to produce end to end solution specifications that apply also to the IPTV. First of all it is necessary to specify all requirements for the service but also from the UE and network capabilities point of view (stage 1). Secondly it is the specification of the functional architecture, functional entities and their task, relevant reference point among the functional entities as well as high level procedures for services (this is done usually in stage 2). In the final stage 3 is required to conclude all details needed from the implementation perspective as for example the protocol models and detailed protocol procedures.

3.3 IPTV Services

There are two main aspects of the IPTV. First one is technological one resulting to the IPTV architecture and second one is the user's perspective aspect which can be seen from the provided IPTV services and user experience.

From the user's perspective is not really important what architecture the IPTV service provider selects, but it is surely more important which services are provided. Most of the existing non-NGN solutions provide only basic set of services like linear TV (live TV channels), video on demand (VoD), and some of them also PVR (Personal Video Recording). New NGN based IPTV solution should therefore provide much more services, features but most important also new user experience in watching TV with more interactivity, personalization, mobility and last but not least comfort in consumption of the right content in the right time and right way.

ETSI TISPAN defines several groups of IPTV services in Release 2 [4]:

1) Entertainment services:

- Broadcast TV (with or with trick modes) – delivery of linearly broadcasted TV channels.
- Trick Modes – enable control playback and pause, forward, rewind content.
- Pay Per View (PPV) – user pay for example only for particular show or time period not whole TV channel or TV package.
- Content on Demand (CoD) – user request content consumption on demand.
- Near CoD (NCoD) – content is transmitted over several channels, each with a different start time, so user can start watching in defined period (e.g. same VOD asset start each 30 minutes).
- Interactive TV (iTV)– service providing interactivity between provider/broadcaster and end user or between several users.
- Push CoD – content is downloaded in advance to STB from where user can watch it locally.
- Personal Video Recording (PVR) – user can record content in network (network or n-PVR) or locally in STB (client or c-PVR).
- Audio – different type of audio program like radio, music shows, etc.

2) Advertising (Ads) – additional to traditional advertising IPTV enabling new advertising models and more targeted advertising (TAI).

3) Regulatory services like Emergency Information, Applications for the disabled, Content Advisories, Educational facilities.

4) Hybrid Services.

5) Third Party Content – content delivered from 3rd party content providers.

New TISPAN Release 3 identified much more entertaining services [4] like:

- User Generated Content (UGC) – content produced by the end user with the intention to share it with other users.
- Content Recommendation (CR) – service advisory for favorite shows based on user's preferences and behaviors.
- Time Shift TV (TsTV) – user can browse and play from past broadcasted content pre-recorded by provider.
- Personalized channel (PCh) – user specific list of programs that are scheduled as playlist for personalized preview.
- Targeted advertising (TAI) – advertising mechanism which is targeted to specified group of user based on his user profiles.
- Profiling and personalization – feature which enable personalized IPTV services based on user preferences and user profile. Provider can also use the information about user's behavior and content consumptions.
- IPTV and NGN Service Interaction (e.g. presence based game, incoming call notification, sharing the remote control).

Set of the services could vary from solution to solution and also need to be considered operator and subscriber preferences (see different standardization views Table 1).

Table 1 IPTV services specified in IPTV related specification

IPTV **Service & Feature**	**TISPAN**	**ITU-T**	**ATIS**	**DVB**	**OIPF**
	[4]	[1]	[20]	[2]	[22]
Linear/ Broadcast TV	X	X	X	X	X
Linear/ Broadcast TV with Trick Play	X	X	X		
Time Shifted TV	X	X			X
Content/Video on Demand (CoD/VoD)	X	X	X	X	X
Push VoD	X	X	X		X
Near VOD	X	X			
Network PVR	X	X			X
Local PVR	X	X			X
Audio	X		X		
Pay-Per-View	X	X	X		X
User Generated Content	X		X		
Advertising	X		X		X
Interactive TV	X		X		
Service Information (EPG)	X	X	X	X	X
Parental Control	X	X	X	X	X
Content recommendation	X				
Games	X		X		
Picture	X		X		
Bookmarks	X				
Personalized channel	X				
Personalized Stream Composition	X				
User Profile & profiling	X				X
Service Portability	X	X			X
Emergency Information.	X		X		
Applications for the disabled.	X		X		
Content Advisories	X		X		
Educational facilities.	X		X		
Hybrid services	X		X		
Communication/Messaging	X		X		X
Notification	X		X		X
IPTV Presence	X				
Interaction between users	X				
Interaction with NGN	X		X		
3rd Party content	X		X		

4 Evolution of IPTV Architectures towards NGN

This section will explain potential evolution of IPTV architecture with relation to NGN and main characteristics of systems. Each evolutional step usually effects capabilities of the system with additional functionalities and system features. Goal of each step is to provide new values for the IPTV services for example, to increase the Quality of Experience (QoE) for the end users and to converge the TV with other telecommunications and interactive multimedia services [25]. A quick and easy introduction of new service features and reduction of the operating costs may be another important motivation to evolve the IPTV systems (as shown in Fig. 4).

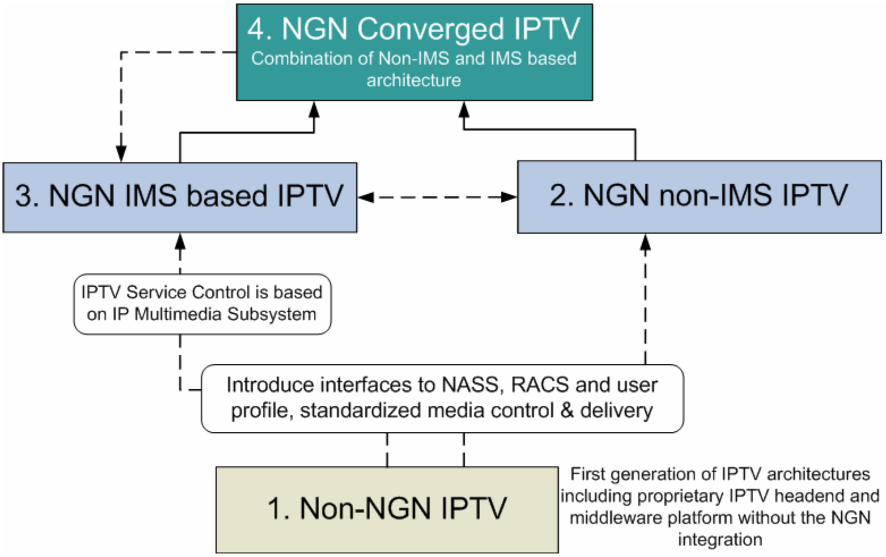

4. NGN Converged IPTV
Combination of Non-IMS and IMS based architecture
3. NGN IMS based IPTV
2. NGN non-IMS IPTV
IPTV Service Control is based on IP Multimedia Subsystem
Introduce interfaces to NASS, RACS and user profile, standardized media control & delivery
1. Non-NGN IPTV
First generation of IPTV architectures including proprietary IPTV headend and middleware platform without the NGN integration

Fig. 4 The potential IPTV migration paths and evolution steps for NGN based IPTV architectures [25]

In comparison with the proprietary IPTV solutions (Type 1) NGN based IPTV (Type 2) first comes with standardized IPTV control and media delivery functions. The NGN based IPTV Subsystem also enables additionally to non-NGN solution the integration with NGN important components like the User Profile Server Function (UPSF), Network Attachment Subsystem (NASS), Resource and Admission Control Subsystem (RACS). This allows NGN based IPTV generally (all type of NGN based IPTV – with or without IMS) to realize personalized value-added IPTV features, and to use network resources more efficiently [26].

The evolution to NGN IMS based IPTV (Type 3) or NGN converged IPTV (combined IMS and Non-IMS pros) is based on the observation that IMS as a unified service control platform is increasingly important for the future NGN services. IMS based IPTV can be therefore inherently integrated in the NGN IMS based service platforms. On the other side, however, we cannot expect all NGN services in the future will be only IMS based. So convergence and combination of IMS and Non-IMS IPTV to be a NGN converged IPTV can be foreseen in the future (Type 4).

4.1 General Architecture Non-NGN Based IPTV

The general Triple Play architecture usually consists of the following parts [27]:

- Service platform domain including IPTV middleware (non-NGN)
- Transport network
- Access network
- Home network and CPEs

The Triple Play service platform usually contains several less independent parts of complex service architecture:

- Content acquisition subsystem which allows to receive, process, and encode content from external sources to defined media coding and encapsulation (receiver and decoders infrastructure, IPTV headend, VoD import and preprocessing).
- Content distribution subsystem responsible for retrieving, protecting, distributing, storing and delivering of the content by preferred way to the end user's system (user equipment).
- IPTV middleware contains the application severs which control and manage the whole IPTV infrastructure (servers, databases, frontend, backend systems, interfaces to external systems e.g. OSS/BSS), users and services. Part of the application platform could be also additional IPTV applications or gateways allowing limited interaction with other systems (e.g. VoIP, NGN).
- Service selection and discovery subsystem which allow the user to browse and find via user TV portal an appropriate content or service information (metadata) which he would like to watch (could be part of IPTV middleware).
- VoD, nPVR or other subsystems – specialized subsystem infrastructure required for dedicated services (Video on Demand or network based personal video recording service).

For the Triple Play contains tree type of services – video, voice, data – the connection to internet services and voice service platform is required (e.g. over VoIP gateway).

There is no single approach to the IPTV service provisioning. Due to huge costs involved in the network equipment, operators usually follow incremental approaches to network upgrading, always relying on existing premises and procedures. Therefore the way a new NGN service is provisioned, it clearly depends on the history of the operator. Therefore there are a lot of differences from solution to solution and also to operator specific transport, access and home network design. NGN based IPTV standardization at least try to specified architecture of IPTV

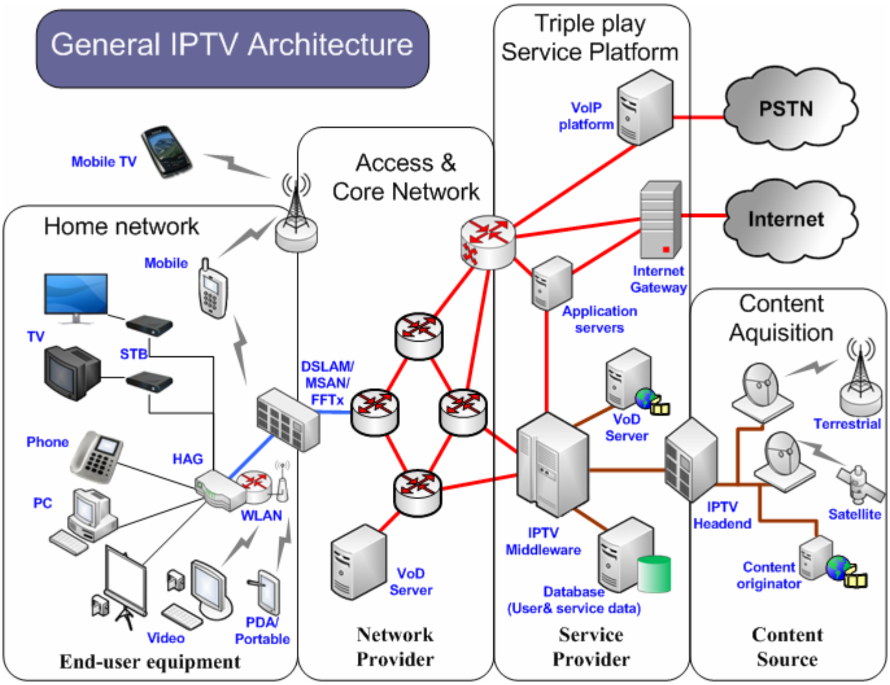


Fig. 5 General IPTV Architecture [27]

service provider to enable interoperability between elements and assure general functionality.

4.2 General Architectural Concepts of NGN Based IPTV

In the ETSI TISPAN NGN, several specification of stage 1 (requirements) and stage 2 (architecture) address the IPTV integration within TISPAN NGN standards:

- IPTV service requirements [4]
- NGN integrated IPTV subsystem architecture [5]
- IMS support for IPTV architecture [6]

The specifications about the implementation of the IPTV functions, interfaces, procedures, protocol recommendations of stage 3 have been finalized and more information can be found for NGN dedicated IPTV in ETSI TS 183 064 [28] or for IMS based IPTV in ETSI TS 183 063 [29].

4.3 NGN Integrated IPTV Architecture

Concept of NGN integrated IPTV subsystem architecture (formerly in Release 2 called NGN dedicated IPTV subsystem) is describing how to integrate of IPTV functions into the NGN architecture [5]. TISPAN NGN Integrated IPTV subsystem is what ITU-T or ATIS call non-IMS NGN based IPTV. The proposed architecture focuses on closer integration between IPTV services and features with NGN network and its subsystems (NASS, RACS, UPSF) but also migration scenarios from existing solutions (i.e. DVB-IPI, ATIS-IIF) into TISPAN NGN and common components. Several part of system are complies and base on existing standard like those from DVB-IPTV and in fact DVB recognize this solution as potential architecture for easier implementation and interworking as with IMS based IPTV [30].

Core IPTV function [28]:

- Service Discovery & Selection - SD&S
- IPTV Control - IPTV-C
- Client Facing IPTV Application - CFIA
- IPTV User Data Function - IUDF
- Media Control Function - MCF
- Media Delivery Function - MDF
- User Equipment - UE

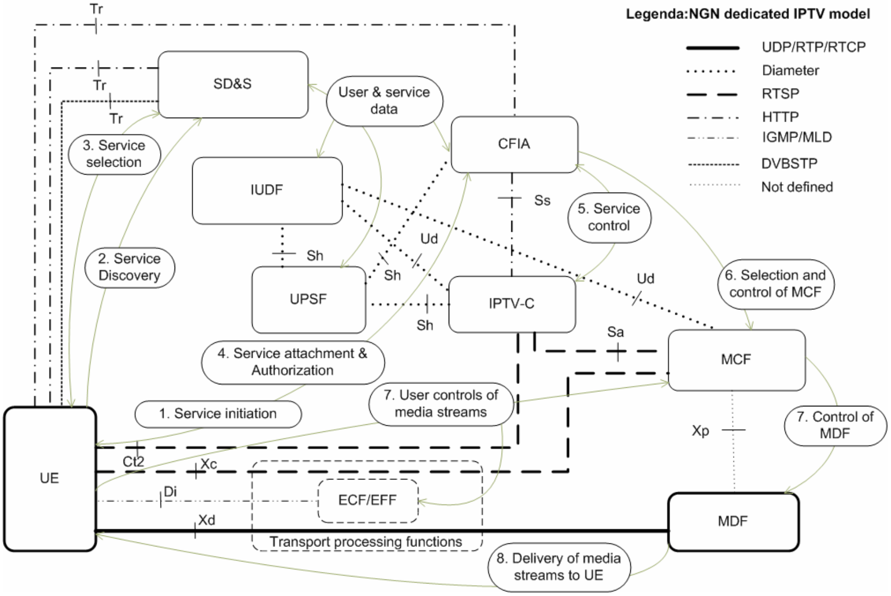


Fig. 6 NGN integrated IPTV – protocol model with FE relation [28]

Initially, the UE is required to start or boot (i.e. a set-top-box, PC, mobile or any device with an IPTV client) and perform the network attachment to obtain network parameters (i.e. an IP address, etc.).

1. After the network attachment the UE is required to start the service initiation steps.
2. The UE shall perform service provider discovery in order to enable SD&S procedure followed by the IPTV service selection and attachment as defined in [6]. SD&S can use HTTP over Tr or DVBSTP over Tr.
3. Then the UE shall perform the service selection procedures with CFIA via Tr (using HTTP over Tr) to receive service selection information.
4. At this stage the IPTV UE needs to acquire and use collected service selection information to establish an appropriate selected service via CFIA (in step 5).
5. The CFIA can control via IPTV-C service behavior. The IPTV-C is also able to initiate a resource reservation and allocation process for network resources needed by the IPTV service according to the capabilities of the UE and access network (using standardized transport control functions of NGN RACS available).
6. Following a successful session initiation, the IPTV-C informs the MCF via Sa reference point (or UE in some cases, i.e. BC) about identification of selected content from the Media Delivery Function (or ECF/EFF for BC services) to initiate delivery of the selected multimedia content (CoD, nPVR).
7. The UE may interactively control CoD media stream over the Xc reference point (between the UE and the MCF) via RTSP protocol. The UE may control BC media stream over the Di reference point (between the UE and the ECF/EFF) with IGMP/MLD protocol.
8. The MDF performs media delivery over the Xd interface using UDP/RTP stream delivery and several transport variants.

4.4 NGN IMS Based IPTV Architecture

Second concept for providing IPTV services over NGN architecture is described in TISPAN IMS based IPTV [6]. Main difference (other are mentioned in next subsection 4.2.3) as name said that IP Multimedia Subsystem is used to service and session control of IPTV services. Main advantage is reusing existing capabilities (IMS registration, authentification, session management, routing, service trigger, identity management, personalization, mobility, charging) of IMS and possibility to integrate service control layer to unified service control platform by utilization for IMS. Disadvantage of IMS based IPTV is higher complexity and less backward compatibility with existing standards (beside that several part of concept reused existing standard and protocols) [29], [32].

Functional entities [29]:

- Service Discovery Function - SDF
- Service Selection Function - SSF
- Service Control Function - SCF
- Core IMS elements (P-CSCF, S-CSCF, I-CSCF)
- Media Control Function - MCF
- Media Delivery Function - MDF
- User Equipment - UE

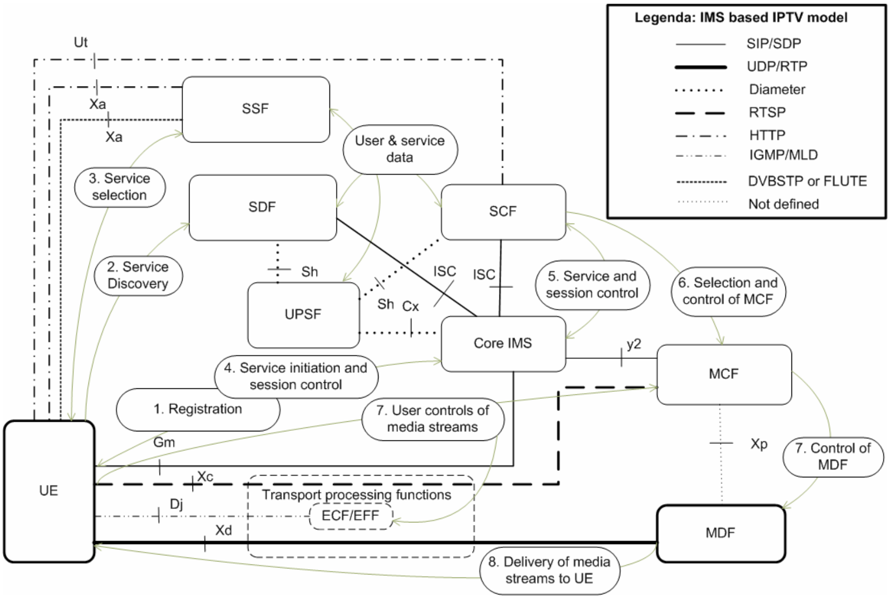

Legenda: IMS based IPTV model
SIP/SDP
UDP/RTP
Diameter
RTSP
HTTP
IGMP/MLD
DVBSTP or FLUTE
Not defined
Ut
Xa
Xa
SSF
User & service data
3. Service selection
SDF
SCF
2. Service Discovery
Sh
ISC
ISC
5. Service and session control
6. Selection and control of MCF
UPSF
Sh
Cx
Core IMS
y2
4. Service initiation and session control
MCF
1. Registration
7. User controls of media streams
Gm
Xp
7. Control of MDF
UE
Xc
Transport processing functions
Dj
ECF/EFF
Xd
MDF
8. Delivery of media streams to UE

Fig. 7 IMS based IPTV – protocol model with FE relation [29]

1. First of all it is needed to start or boot an UE (like a set-top-box, PC, mobile or any device with an IPTV client) and achieve a network attachment to obtain network parameters (like an IP address, P-CSCF address, etc.).
2. After the network attachment the UE initiate the IMS registration process with core IMS.
3. The UE will perform the IPTV service attachment functions including SIP based service discovery to perform SDF tasks.
4. Then the UE is able to initiate the service selection procedures with SSF via Xa (using HTTP over Xa or using DVBSTP or FLUTE) to receive the service selection information.
5. The IMS based IPTV UE needs to have and use the received service selection information in order to establish an appropriate multimedia session by generating SIP INVITE messages during service initiation procedure (over Gm towards home C-CSCF) send via IMS core to SCF.
6. After a successful session initiation, the SCF informs the MCF via core IMS and y2 interface (or UE in some case like BC) about identification of selected content from the Media Delivery Function (or ECF/EFF in case of BC services) to initiate start streaming the selected multimedia content (CoD, nPVR).
7. The UE may control CoD media stream over the Xc (interface (between the UE and the MCF) to control media delivery with the RTSP protocol. The UE may control the BC media stream over the Dj interface (between the UE and the ECF/EFF) to control media delivery with IGMP/MLD protocol.
8. The MDF performs media delivery over the Xd interface is based on UDP/RTP stream delivery and several transport variants.

The IMS based IPTV has number of advantages because IMS can act as unified service control subsystem for all NGN services instead of establishing an additional specialized subsystem (case of NGN dedicated/integrated IPTV subsystem). Additionally the IMS can more naturally support mobility, interaction with NGN service enablers (like messaging or presence), service personalization or quadruple play services (voice, data, video and mobile).

4.5 Protocols Comparison between IMS and Non-IMS Based IPTV

When the IMS based IPTV [29] is compared with NGN dedicated/integrated IPTV [28] main differences in protocols or solution characteristics between both are following (Table 2) [31]:

- Separation of service selection and discovery in the IMS based IPTV (SIP based discovery).
- Similar service selection specs, IMS based IPTV additionally support OMA BCAST ESG.
- SIP based service initiation and service control.
- Support direct RTP encapsulation.
- Related differences in interfaces and protocols (from Tr to Ut, Xa and change from http based Ct, Ss, Sa to SIP based Gm, ISC, y2 interfaces).

Generally, we propose possible evolution path based on analyses of several actual non-NGN solutions (which some of them follow existing DVB-IPI specifications) and the TISPAN release 2 specs. The NGN dedicated IPTV is based on main concepts of DVB specifications and can additionally provide more completed architecture, and also allows possible integration or migration towards the NGN architecture. New concept of the IMS based IPTV can be evolved by replacing the IPTV control with SIP based IMS service/session control, and implement additional inter-faces Gm instead of Ct2 or Ut, Xa interfaces instead of Tr.

Table 2 Comparison of characteristics for non-IMS and IMS based NGN IPTV concepts [31]

General characteristics	NGN dedicated IPTV architecture (NGN Non-IMS)	IMS based IPTV architecture (NGN IMS based)
ETSI TISPAN specification	ETSI TS 183 064	ETSI TS 183 063
	WI 3127 [28]	WI3137 [29]
1. SD&S	ETSI TS 102034 based SD&S model - separate SDF, SSF SIP based (Mandatory), HTTP (Optional) , DVBSTP (Optional) via Xa to SSF - HTTP based	ETSI TS 102034 based SD&S model - single SD&S HTTP based (Mandatory) DVBSTP (Optional) via Tr to SD&S - HTTP based

Table 2 (*continued*)

2. Service selection information (e.g. program guides)	DVB SD&S (ETSI TS 102034) DVB BCG (ETSI TS 102 539) OMA BCAST ESG TISPAN XML	DVB SD&S (ETSI TS 102034) DVB BCG (ETSI TS 102 539)
3. Multicast control - IGMP	SIP based initiation IGMP join to ECF/EFF IGMPv3, MDLv2	Pure IGMP based IGMP join to ECF/EFF IGMPv3, MDLv2
4. Unicast control - RTSP methods	SIP based initiation Mixture RTSP control (RFC 2326), partially ETSI TS 102034 based Method 1 – new coupled SIP/RTSP Method 2 – SIP and RTSP separated	RTSP based on ETSI TS 102034 Coupled, decoupled mode
5. Media Delivery	MPEG2TS over RTP MPEG2TS over UDP direct RTP encapsulation	MPEG2TS over RTP MPEG2TS over UDP
6. Service control (initialization, modification, teardown)	SIP based service control using IMS [10] Session based control	HTTP resp. RTSP based
7. Service configuration	Ut – XCAP	Tr – XCAP
8. Resource allocation & reservation	Via core IMS Gq' to RACS	IPTV-C Gq' to RACS
9. User profile, user data	Distributed UPSF (SSP located) SCF (SIP AS) SSP used (SSF, SDF via Sh)	Distributed UPSF (SSP located), IUDF CFIA (http AS), IPTVC SSP used (SD&S, IPTVC)

4.6 Possible Migration and Switch over Scenarios

Several scenarios are possible and really depend only on operator choice which solution and migration scenarios will select (if any).

One possible scenario was presented in TISPAN 18 bis meeting [26] showing also possible steps as example of the step by step evolution of the IPTV from the non-NGN IPTV to NGN integrated IPTV and later to the IMS based IPTV [5] R3 (Table 3). We can summarize proposed steps to following sequence:

1. If the non-NGN IPTV solution is at least partially based on some existing standards like DVB-IPI the easiest way of start with migration to NGN, with implementation of interfaces to standardized transport control subsystems to

NASS (e2) and to RACS (Gq’). This allows standardized resource reservation and standard dynamic bandwidth allocation as assurance of the QoS.

2. Next step is the introduction of standardized media delivery and control architecture (MCF, MDFs, and Sa, Xc, Xd) and its integration with middleware.
3. Change the internal architecture of the IPTV middleware to standardized NGN dedicated/integrated subsystem architecture (SD&S, CFIA, IUDF and specially interfaces to UE as Tr, Ct2) with integration to NGN user profile (UPSF).
4. Change the HTTP/RTSP based control to SIP based session oriented control and introduction of the IMS (with core elements and Gm interface). The CFIA is evolved to SIP application server SCF and moves user data to UPSF.
5. Last step should be performed by replacing the HHTP based service discovery and selection function (SD&S) with two separate SDF (SIP based application function for service discovery) and SSF (service selection function using HTTP interface Xa). Which in fact complete the migration to the IMS based IPTV.

Table 3 Possible migration scenarios [5b]

Evolution step	UE	Transport	MC&DF	Service control	Application
Non-NGN	STB	Content delivery network	IPTV middleware		
TISPAN NGN integrated IPTV	UE	Transport processing, NASS & RACS	MCF & MDF	IPTV-C/AS Interaction with IMS and other NGN subsystems	CFIA, SD&S
TISPAN IMS based IPTV	UE	Transport processing, NASS & RACS	MCF & MDF	IMS	SCF, SSF, SDF
Converged NGN based IPTV architecture	UE	Transport processing, NASS & RACS	MCF & MDF	This part is in discussion how merge IMS and integrated IPTV or just extend IMS based IPTV. See next section about possible solution.	

5 Concept of Converged NGN Based IPTV Architecture

Several scenarios are possible and really depend only on the operator’s choice which solution and migration scenarios are selected (if any). The previously mentioned ITU-T IPTV FG specify converged application framework where the non-IMS IPTV merges with the IMS based IPTV architecture, but just as purely a unity of all elements supporting all interfaces from both architectures that make no sense from complexity perspective, but also for both architectures if used in parallel, they are able to provide similar services. The ETSI TISPAN has proposed in release 3 [5] in the informative annex some possible migration scenarios (similar to those in previous sub-section) where were mentioned converged NGN based IPTV. The following section describes a potential concept of such architecture with additional goal describing also a way of adaptation with multiple types of content sources but also with multiple access and distribution networks converged to a single functional architecture (Fig.8).

The proposed Converged NGN based IPTV architecture (CN-IPTV) is an evolution of the IMS based IPTV where combined IPTV service control function (ISCF) are used for service control, which can use the IMS for specific cases but not all signaling traffics need to pass the core IMS (which improve performance and shorter delays).

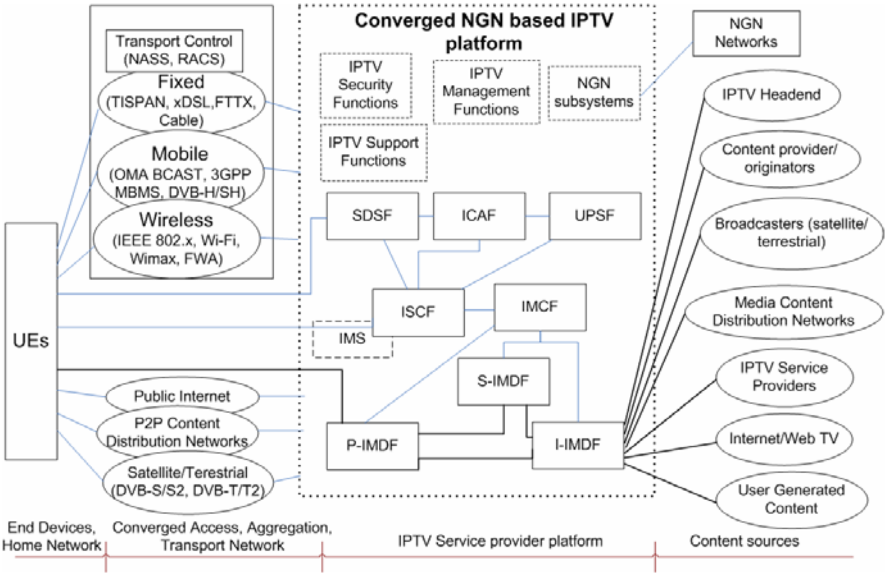

Transport Control (NASS, RACS)
Fixed (TISPAN, xDSL,FTTX, Cable)
Mobile (OMA BCAST, 3GPP MBMS, DVB-H/SH)
Wireless (IEEE 802.x, Wi-Fi, Wimax, FWA)
Converged NGN based IPTV platform
IPTV Security Functions
IPTV Management Functions
NGN subsystems
IPTV Support Functions
SDSF
ICAF
UPSF
ISCF
IMCF
IMS
S-IMDF
P-IMDF
I-IMDF
UEs
Public Internet
P2P Content Distribution Networks
Satellite/Terestrial (DVB-S/S2, DVB-T/T2)
NGN Networks
IPTV Headend
Content provider/ originators
Broadcasters (satellite/ terrestrial)
Media Content Distribution Networks
IPTV Service Providers
Internet/Web TV
User Generated Content
End Devices, Home Network
Converged Access, Aggregation, Transport Network
IPTV Service provider platform
Content sources

Fig. 8 Proposed conceptual architecture for Converged NGN based IPTV

The concept of media control and delivery (MC&D) was extended with a proposal of specialized types of IMDF (organized in hierarchical MC&D architecture) [25] with three architecture IMDFs elements described as follows:

1. Interconnection - IPTV Media Delivery Function (I-IMDF): this element handles the media import and ingress of content from multiple content sources (ingress of media, metadata, content provider information and interconnection to external domains):

   - IPTV Headend or from content providers/originators or broadcasters.
   - From other IPTV service providers in case of interconnection or roaming or as offer of the content from service provider playing a role of content aggregator.
   - From the Internet sources like the Web based TV or from the end users like the user generated content.

The I-IMDF need to hide the IPTV service provider infrastructure for external domains, but also provide necessary functionality to interconnect to heterogeneous content sources (which can hold a variety of coding, transport, signaling schemas) and convert to content/metadata/signaling to formats supported by Converged NGN based IPTV.

2. Serving - IPTV Media Delivery Function (S-IMDF): this element handles the processing of contents (e.g. encoding, content protection and transcoding), and is also responsible for the storage of contents and metadata as well as the propagation of content information within the IMS based IPTV systems. It is on top of the hierarchy and provides centralized oriented services such as Content on demand for long tail content (less popular content), or recording/storing of user independent content (n-PVR or Time shifted TV or Near CoD).
3. Primary - IPTV Media Delivery Function (P-IMDF): this element is the primary contact point of the users which provides also the streaming and downloading functionalities for all IPTV services according to the required quality, format and type of casting (multi-/uni-/broad-casting) for particular user's end device, and access network. This element could also store the most frequently accessed CoD assets or user specific contents (specific user n-PVRs, user generated content). The P-IMSDF could be responsible for the adaptation of IPTV architecture to other access technologies or distribution network:

   - Preferred way is that the P-IMDFs are located as near as possible to UE for example near to edge of the network. These elements can be combined with specific elements of access network (e.g. in case of using OMA BCAST and 3GPP MBMS also with integrated control elements in case of MBMS, for example, P-IMDFs can contain also Broadcast-Multicast Service Centre (BM-SC). The BM-SC provides functions needed for user service provisioning and content delivery in MBMS capable UMTS).
   - P-IMDF can also support mobility and seamless handover between different technologies.
   - P-IMDF may required to transcode or adapt the content to required bitrate, codec or content encapsulation to specific transporting technologies.
   - Can be used as security elements for the content protection (e.g. digital rights management and content encoding).

The element responsible for the service discovery and selection functions (SDSF) could support multiple formats and mechanisms. But the main enhancement in proposed architecture is the SDSF's potential to aggregate metadata information from multiple sources (e.g. content provider, electronic program guide provider, internet, broadcasted service information), and provide them everywhere to the UE in personalized manner and allows them to integrate them with other relevant information (presence, statistics, recommendations, etc.), too.

Last but not least element is for sure the IPTV converged application function (ICAF) which could provide combinational services and converged services with enhanced service logic and service orchestration. The ICAF can interact with other NGN application servers and subsystems and can used for user's profile information to personalized service behavior based on user's preferences and settings.

Similar to other NGN based IPTV also the converged one need to ensure relevant resource allocation and QoS handling. But we differentiate with the IPTV service provider infrastructure (fixed, mobile or wireless access technologies) from other distribution possibilities, like public internet (without QoS and limited bitrates), terrestrial/satellite distribution (mainly unidirectional for broadcasting with other technologies used for interaction, back channel signaling or unicast services), or the P2P content distribution network (with limited provider capability to control the content but there are already operators oriented to the P2P networks with hierarchical architecture and supernodes as part of the IPTV provider architecture, e.g. co-located with P-IMDF in our case).

Any IPTV architecture could not be complete without other functionalities which we have also included in the proposed concept:

- IPTV supporting function (e.g. content preparation & manipulation)
- IPTV management functions (e.g. content management)
- IPTV security functions (e.g. content protection, IPTV service protection)
- IPTV charging (based on NGN charging for online/offline charging but enhanced for IPTV specific scenarios)
- Interworking or interfacing with other NGN subsystems

6 Conclusion

This chapter has presented the advances in the field of multimedia services with the focus on the latest trends in concepts of NGN based IPTV architecture and services. Presented sections provide an overview about actual trends and standardization effort which will form the IPTV services in following years. Because watching television is still an activity which directly or indirectly affects our lives and perception of the world we can expect also a revolution in this field. The way of how users are consuming television has dramatically changed and this trend will be most probably effected in coming next years by the next generation of IPTV solutions. We can already recognize the movement from passive watching and browsing of channels only available on cable or satellite. End users want to access different content from different sources on the Internet or over user generated content via web communities. When the operator wants to react to these trends he needs to build up a converged architecture which will aggregate various contents from multiple sources and provide this content with quality/format, personalized, secured, reliable manner with much more interactivity and users interactions.

Several NGN based IPTV architectures have been presented with the intent to explain the differences as well as evolution possibilities. Additionally to the existing approaches of NGN based IPTV like NGN integrated IPTV subsystem or IMS based IPTV, we provide an overview about the concept for converged NGN based IPTV.

Acknowledgments

First author is senior designer in Slovak Telekom where he have been actively participating on R&D project ScaleNet and ETSI TISPAN NGN Release 2/3 standardization with more as 50 agreed contributions included in several TISPAN IPTV related specifications (e.g. IPTV related work items WI0005, WI2048, WI2049, WI3127, WI3137, WI7029, WI1059, WI2070, 2074). Some parts of this chapter are based on contribution of author within ETSI TISPAN. We would like thanks to ETSI for permission to reproduce some text and figures from published ETSI TISPAN specifications.

This paper also presents some of the results from participation on various research project at STU such as NGNlab project [NGNlab], European CELTIC EUREKA project Netlab [Netlab], Slovak National research projects: AV project 4/0019/07: Converged technologies for next generation networks (NGN), Slovak National basic research projects VEGA 1/0720/09 and VEGA 1/4084/07.

Web Pages

[DVB] Digital Video Broadcasting web site, http://www.dvb.org/technology/standards/

[ITU-T] International Telecommunication Union web page, http://www.itu.int/ITU-T/

[ITU-T SGI] ITU-T web page of Internet Protocol Television Global Standards Initiative, http://www.itu.int/ITU-T/gsi/iptv/

[ETSI] ETSI web page, http://www.etsi.org

[ETSI TISPAN] ETSI TISPAN web page, http://www.etsi.org/tispan/

[3GPP] 3rd Generation Partnership Project web page, http://www.3gpp.org/

[OMA] Open Mobile Alliance web page, http://www.openmobilealliance.org/

[IETF] Internet Engineering Task Force web page, http://www.ietf.org/

[ATIS] Alliance for Telecommunications Industry web page, http://www.atis.org/

[CableLabs] CableLabs web page, http://www.cablelabs.com

[OIPF] Open IPTV Forum web page, http://www.openiptvforum.org/

[HGI] Home Gateway Initiative web page, http://www.homegatewayinitiative.org/

[NGNlab] NGNlab – NGN laboratory at Slovak University of Technology in Bratislava, project web page http://www.ngnlab.eu

[Netlab] NetLab: Use Cases for Interconnected Testbeds and Living Labs, project web page, http://www.celtic-nitiative.org/Projects/NETLAB/default.asp

References

[1] Draft Recommendation ITU-T Y.1901 (Y.iptv-req), Requirements for the support of IPTV services, ITU-T(September 2008)

[2] ETSI TS 102 034 V1.2.1 (2006-09) Technical Specification, DVB; Transport of MPEG 2 Based DVB Services over IP Based Networks, ETSI (2006)

[3] Recommendation ITU-T Y.1910 (09/2008), IPTV functional architecture, ITU-T (2008)

[4] ETSI TS 181 016 V2.0.0 (2007-11) for Release 2, Draft ETSI TS 181 016 V3.2.3 (2009-02) in Release 3, TISPAN; Service Layer Requirements to Integrate NGN Services and IPTV (2007)

[5] ETSI TS 182 028 V2.0.0 (2008-01) in R2, Draft ETSI TS 182 028 V3.2.3 (2009-04) in R3 TISPAN; IPTV Architecture; Dedicated subsystem for IPTV functions (2008)

[6] ETSI TS 182 027 V2.0.0 (2008-02) R2, Draft ETSI RTS 182 027 V3.1.2 (2008-12) R3, TISPAN; IPTV Architecture; IPTV functions supported by the IMS subsystem (2008)

[7] 3GPP TS 23.228 V7.7.0 (2007-03), IP Multimedia Subsystem (IMS); Stage 2 (2007)

[8] 3GPP Technical Specification, MBMS 3GPP TS 23.246 V8.2.0 (2008-06); Multimedia Broadcast/Multicast Service (MBMS); Architecture and functional description (Release 8), 3rd Generation Partnership Project (2008)

[9] Open Mobile Alliance Document OMA-AD-Service-Environment-V1_0_4-20070201-A, OMA Service Environment 1.0.4, OMA (2007)

[10] OMA draft OMA-AD-BCAST-V1_0-20081209-C: BCAST Mobile Broadcast Services Architecture, Candidate Version 1.0 – 12/2008, Open Mobile Alliance (2008)

[11] RFC 791 - Internet Protocol, IETF (1981)

[12] RFC 2460 - Internet Protocol, Version 6 (IPv6) Specification, IETF (1998)

[13] RFC 3261 - SIP: Session Initiation Protocol, IETF (2002)

[14] RFC 2326 - Real Time Streaming Protocol (RTSP), IETF (1998)

[15] RFC 3376 - Internet Group Management Protocol, Version 3, IETF (2002)

[16] RFC 3810 - Multicast Listener Discovery Version 2 (MLDv2) for IPv6, IETF (2004)

[17] RFC 3550 - RTP: a transport protocol for real-time applications, IETF (2003)

[18] RFC 2616 - Hypertext Transfer Protocol – HTTP/1.1, IETF (1999)

[19] Alliance for Telecommunications Industry Solutions (ATIS) IPTV Architecture Requirements (ATIS-0800002), ATIS IPTV Interoperability Forum, IIF (2006)

[20] Alliance for Telecommunications Industry Solutions (ATIS) IPTV High Level Architecture Standard (ATIS-0800007), ATIS IPTV Interoperability Forum, IIF (2007)

[21] CableLabs document PKT-TR-ARCH-FRM-V05-080425, PacketCable^{TM} 2.0, Architecture Framework Technical Report (2008)

[22] Open IPTV Forum – Service and Platform Requirements – V 2.0, Open IPTV Forum (2008)

[23] Open IPTV Forum – Functional Architecture – V 1.2, Open IPTV Forum (2008)

[24] Home Gateway Technical Requirements: Release 1, Version 1.0, Home Gateway Initiative (2006)

[25] Mikoczy, E., Sivchenko, D., Xu, B., Moreno, J.: IPTV services over IMS - Architecture and standardization. IEEE Communication Magazine 47(5), 128 (2008)

[26] 18bTD221r2_WI02074_NGN_Integrated_IPTV_migration, TISPAN 18 bis contribution, NGN integrated IPTV – architecture & migration scenarios, ETSI (2008)

[27] López, D., Mikoczy, E., Moreno, J., Cuevas, A., Vázquez, E.: IP Multimedia Subsystem (IMS) Handbook. In: Ilyas, M., Ahson, S.A. (eds.) IPTV modeling and architecture over IMS, ch. 22. CRC Press, Boca Raton (2008)

[28] ETSI TS 183 064 V2.1.1 (2008-10), TISPAN; Dedicated IPTV subsystem Stage 3 Specification, Technical Specification Draft (2008)

[29] TS 183 063 V2.1.0 (2008-06), TISPAN; IMS based IPTV Stage 3 Specification (2008)

[30] DVB Document A128, DVB-IP Phase 1.3 in the context of ETSI TISPAN NGN, DVB (September 2008)

[31] Mikóczy, E.: Next Generation of Multimedia Services – NGN based IPTV architecture. In: 15th International Conference on Systems, Signals and Image Processing, IWSSIP 2008, Bratislava, Slovak Republic, June 25-28 (2008)

[32] WG3TD049r3_WI3127_Annex_L_protocols&specs, ETSI TISPANWG3 Interim meetingcontribution Paris, 28 January – 1 February, ETSI (2008)

[33] 16bTD327_WI3137_signalling_flow, ETSI TISPAN#16Bis meeting contribution 16bTD327, Sophia Antipolis, March 3-7, ETSI (2008)

Exploring the Characteristics of High Definition Television Systems

Goran Gvozden, Emil Dumic, and Sonja Grgic

Abstract. The continuous growth of digital video industry driven by constant improvements in digital processing technology has stimulated today's broadcasters to start a new transition. This new transition toward better, high definition television system will introduce and offer the audience a true, realistic experience of a perceived high resolution video content. High Definition Television (HDTV) is under consideration in many countries around the world based on the availability of flat panel displays and increasing production of HDTV content. The main difference between today's Standard Definition Television (SDTV) and HDTV systems includes increased number of scanning lines, progressive scanning capability and increased aspect ratio. Apart from the video format, another HD variation on SD is a slightly different colorimetry and multichannel audio comparable to the sound quality of a compact disc. Due to these parameter modifications larger capacity of media storage and larger bandwidth in transmission systems is needed which, on the other hand, stimulates the utilization of new and more efficient video coding techniques. This chapter examines the aforementioned HDTV characteristics as well as the influence of video coding tools on a high definition video quality. Three different HDTV video formats, 720/50p, 1080/25i and 1080/50p compressed with different sets of H.264/AVC coding tools were tested and evaluated using different picture assessment methods. Based on the results the most suitable HDTV video format is selected and recommended for usage in HDTV environment.

1 Introduction

From the very beginning the world of television has been going through various changes driven by the constant needs to provide the general public with the actual,

Goran Gvozden
RTL Televizija
Zagreb, Croatia
e-mail: ggvozden@ieee.org

Emil Dumic and Sonja Grgic
University of Zagreb
Faculty of Electrical Engineering and Computing
Zagreb, Croatia
e-mail: emil.dumic@fer.hr, sonja.grgic@fer.hr

M. Grgic et al. (Eds.): Rec. Advan. in Mult. Sig. Process. and Commun., SCI 231, pp. 341–373.
springerlink.com

realistic visual experience of a surrounding world. These changes, followed by constant research, numerous ideas, different designs and concepts reinforced by continuous improvements in digital technology resulted in a sophisticated and advanced digital television ready to meet the user demands (Robin and Poulin 1998, Watkinson 2000). Furthermore, introduction of digital technology provided better quality; wider range of services; efficient usage of spectrum; higher transmission and storage capacities; integration with computer networks and at last, a basis for successful introduction and implementation of the high definition television (Jones et al. 2006). Thus, the world of television including broadcasters, equipment manufacturers, regulatory bodies, production companies, network operators and general public as well is once again faced with a new transition that will bring us into an era of high definition television. The successfulness and acceptance of high quality television system will depend on numerous factors such as availability of high resolution, large, flat-panel displays; high quality production and post-production equipment; significantly reduced costs and availability of adequate broadcast services (Ive 2004).

The high definition television can be described as a group of different video signal formats characterized by significantly higher amount of the information in relation to standard definition television. Consequently, increase in amount of information is due to higher number of scanning lines, higher number of samples per line, usage of progressive scanning and higher aspect ratio (Richardson 2006). This huge increase of information requires more careful production, post-production and broadcast chain analysis, design and implementation. Thus, increase of video information requires definition and implementation of advanced digital studio interfaces, more efficient video coding schemes, storage technologies and transmission systems. Additionally, in relation to standard definition television colorimetric parameter values have been changed which additionally complicates the implementation of the production and broadcast chain (Poynton 2007). Above all, due to variety of image formats and scanning frequencies additional conversion is necessary for the video material exchange when source and destination systems use different combination of high or standard definition image format and scanning frequency.

Twenty years ago high definition system was defined in (ITU Recommendation ITU-R BT.801 1990) as: A High Definition System is a system designed to allow viewing at about three times the picture height, such that the system is virtually, or nearly, transparent to the quality of portrayal that would have been perceived in the original scene or performance by a discerning viewer with normal vision acuity. Such factors include improved motion portrayal and improved perception of depth. Consequently, the best visual experience will be reached when a distance between a viewer and a display is about three times the picture height. This distance of three picture heights is only theoretical and based on an individual's visual acuity, viewing conditions, displayed picture quality, used compression method and scanning format (Sugawara et al. 2005, EBU I34 2002, EBU I35 2003).

In this chapter we will describe the characteristics of high definition television systems and the consequences on the production and broadcast chain caused by the selection of certain combination of image format and scanning standard. In the second section after introduction we will give an overview of high definition television systems development; from early analogue systems used throughout the world till today's modern digital systems. Following section includes high definition television standardization process. In the fourth section we will give a detailed description of high definition television characteristics including colorimetric parameter values, optoelectronic conversion, sampling frequencies and aspect ratio followed by the explanation of interdependence among viewing distance, image format and display resolution. The fifth section explains the requirements on high definition television production, post-production and broadcast chain with all the advantages and drawbacks caused by the selection of particular high definition image format and scanning standard combination. Considering the variety of image formats and scanning combinations we will give an overview of a research done in a field of subjective and objective high definition picture assessment in order to find the most suitable combination. Respectively, effects of H.264/AVC compression method, inevitable and necessary for implementation of high definition systems will be presented along with subjective picture assessment methods necessary for selection and efficient utilization of certain high definition video formats.

2 Evolution of HDTV Systems around the World

The term High Definition Television (HDTV) is being mentioned and referred to in many different contexts. It has been frequently used as a representative of something new, something different and something better that will have considerable influence on our lives. Although we can think of it as a new, the term High Definition Television doesn't represent completely new concept; quite contrary, we can say that it is a pretty old concept created decades ago. The work on improving the quality of perceived video signal was continuous from the beginning of the television, constrained, of course by contemporary technology.

The term HDTV was first used in 1937 to describe the television system developed in Europe which had 405 scanning lines alternating at 25 frames per second. In United States, in 1940s, the National Television Standard Committee (NTSC) developed a television standard comprising of 525 scanning lines alternating at 30 frames per second (Udelson 1982). At that time "newly" developed standard was also referred to as a "high definition" standard.

Work on the development of a high definition television standard as we know it today started back in 1960s, in Japan (Ninomiya 1995). The first results presented in Tokyo in 1975 included high quality television signal with 1125 scanning lines alternating at 30 frames per second with aspect ratio of 5:3 where viewing distance was about three times the picture height. Afterwards, the aspect ratio was changed to 16:9 by CCIR Plenary Assembly because it was close to 35-mm film production ratio. Moreover, number of 1125 scanning lines was adopted in order to enable interoperability and simplify conversion, to and from conventional 525/625 scanning line television standards. In early 1980s manufacturing of high quality, analogue signal processing and production equipment, as well as the implementation of a Multi sub-Nyquist Sampling Encoding (MUSE transmission system enabled transmission of native 30 MHz analogue signal using 8.1 MHz frequency modulated channel suitable for terrestrial and satellite distribution systems. In 1985 system was tested and four years afterwards Japanese television started the experimental high quality analogue signal satellite broadcasting. Although system was defined and implemented it failed to become globally accepted concept mainly because of the limitations of the contemporary technology (Hatori and Nakamura 1989).

Meanwhile, United States and Europe lagged behind in a high definition race. Therefore they initiated the development projects for implementation of HDTV systems (Jurgen 1989). In Europe, during the 1960s two systems were in use, PAL and SECAM. An initiative to develop and define the new high definition system was an excellent opportunity to create and standardize unique television system for the whole Europe which could cope with the high definition standard developed in Japan. The goal was to develop high definition television standard having 1250 scanning lines alternating at 50 fields per second. Two projects were started in order to achieve this; Eureka EU 95 and PALplus. In the case of Eureka EU 95 project the main goal was to design a system that could be used for cable and satellite high definition television signal distribution; High Definition – Multiplexed Analogue Components (HD-MAC) was chosen as most appropriate. HD-MAC technology was successful for transmission of high definition television signal; nevertheless, attempts and enormous efforts to replace the SDTV signal didn't give results because of various commercial, technical and program reasons (Fox, 1995). Eventually, project was finished in 1995. The second development project PALplus was imagined as a solution for a terrestrial high definition television distribution. Project started in late 1980s while the complete system started to function in 1995. PALplus didn't use HD-MAC technology therefore it couldn't convey the high definition television signal over the conventional terrestrial television channels (Ellis 1997).

A numerous comities were established in order to study and find out a common standard that could benefit everyone. That was the time of joint action and pursuit after common worldwide HDTV standard. Various forms of future HDTV standards were demonstrated and presented, as well as possible solutions of compatibility among already developed standard television systems. Unfortunately, the noble idea to have the common standard was faced with insuperable obstacles manifesting in painful, long-lasting development and implementation process. With the absence of mutual agreement and finding the compromise the world again, like many times before, failed to find a better solution that could enable more efficient exchange and distribution of video material (Wood D 2007).

The development of digital processing technology along with the emersion of advanced signal modulation and compression methods enabled gradual cessation of analogue television systems and provoked the development of more efficient, more robust and more economical digital television systems designed for production, processing and transmission of SDTV and HDTV signals. Nowadays, there are numerous digital television standards in use for television signal distribution throughout the world; ATSC (Advanced Television Systems Committee) in America; DVB (Digital Video Broadcasting) in Europe; ISDB (Integrated Services Digital Broadcasting) in Asia (Cugnini et al. 2007). In the next section we will explain basic video signal characteristics in digital high definition television systems.

3 Standardization of HDTV Systems

As we mentioned before the development of digital technology enabled the emergence of digital television systems throughout the world. Before we present the high definition television standards currently used worldwide it is of great importance to mention the standard that played crucial role in a transition from analogue to digital television systems. The standard that unified the world of digital television was the ITU-R BT.601 standard (Wood 2007, Baron S and Wood D 2005). Emergence of this standard ensured the interoperability, equipment development and easier international production and program exchange. The successfulness of the newborn standard resided in a decision to abandon the technique of digital composite coding, because of the variety of already existing analogue standards NTSC, PAL and SECAM, and to orientate on digital component coding. Usage of digital component encoding enabled the inclusion of both 525/60 and 625/50 systems due to selection of common sampling frequencies for the luma and chroma signals. Moreover, standard defined the construction of the luma and croma signals as well as the number of their samples per digital line (ITU Recommendation ITU-R BT.601 1995). ITU-R BT.601 standard certainly affected the development of modern digital television systems and can be considered as a basis of today's digital high definition television systems. Hereafter, we will describe digital high definition standards and emphasize the importance of the characteristics described in it.

Today, different picture formats with different scanning standards are being used in digital television systems. In order to maximize the interoperability between different picture formats and scanning standards; reduce the quality loss caused by conversion process; improve the worldwide production and program exchange in high definition environment, different standards have been regulated and issued by different organizations and associations such as International Telecommunication Union (ITU), Society for Motion Pictures and Television Engineers (SMPTE) and European Broadcasting Union (EBU).

SMPTE specified many standards for high definition television systems. We will mention the two most important standards. The first one is SMPTE 296M (1280 x 720 Progressive Image Sample Structure – Analogue and Digital Representation and Analogue Interface) defined in 2001 which describes the image format of 1280 samples per 720 active lines, 16:9 aspect ratio and various scanning frequencies (SMPTE 296M 1998). The second standard is SMPTE 274M (1920 x 1080 Image Sample Structure, Digital Representation and Digital Timing Reference Sequences for Multiple Picture Rates) defined in 2003 describing the image format of 1920 samples per 1080 active lines, 16:9 aspect ratio, square sampling, and colorimetric characteristics (SMPTE 274M 2001).

International Telecommunication Union as a leading international organization for production of standards mandatory for implementation throughout the world, regulated the ITU-R BT.709 standard which describes the parameter values for production and international program exchange for high definition material (ITU Recommendation ITU-R BT.709 2001). With this recommendation ITU-R defined the Common Image Format that has 1920 samples per 1080 active lines. It comprises of two parts where the first part describes the conventional analogue high definition television systems while the second part covers the digital high definition television systems using square pixels. This document considers the benefits of a common display format and recommends the standard aspect ratio and commonality of colorimetry for various display technologies available or under development. Parameter values described in the second part of the ITU-R BT.709 recommendation correspond to the parameter values defined in SMPTE 274M standard.

4 High Definition Characteristics

Television systems differ on various characteristics such as image format resolution; scanning frequency; the way optoelectronic conversion is performed; aspect ratio; signal bandwidth; sampling frequency etc. In next section we will describe main high definition characteristics, differences among the high definition video formats and the influence they have on production and broadcast chain.

4.1 High Definition Video Format

Before we describe the scanning formats in a high definition environment the term video format has to be clarified. The video format is the combination of the image format, formed out by a particular number of pixels and lines, a frame rate at which the image is repeated in time and a scanning method, progressive or interlaced. Nomenclature used to interpret the combination of image format, frame rate and scanning mode is somewhat different around the world, but only for representation of interlaced scanned images. Hence, image formats scanned using interlaced method standardized by SMPTE are represented with field rate number while those standardized by ITU and EBU are represented with frame rate.

Considering the spatial format among all the formats specified for high definition television systems extraction of two major formats can be made, the first one has 1280x720 structure having 0.92 million pixels per image and the other has 1920x1080 structure comprising of 2.07 million of pixels per image. Figure 1 depicts two major high definition formats for 50Hz and 60Hz world.

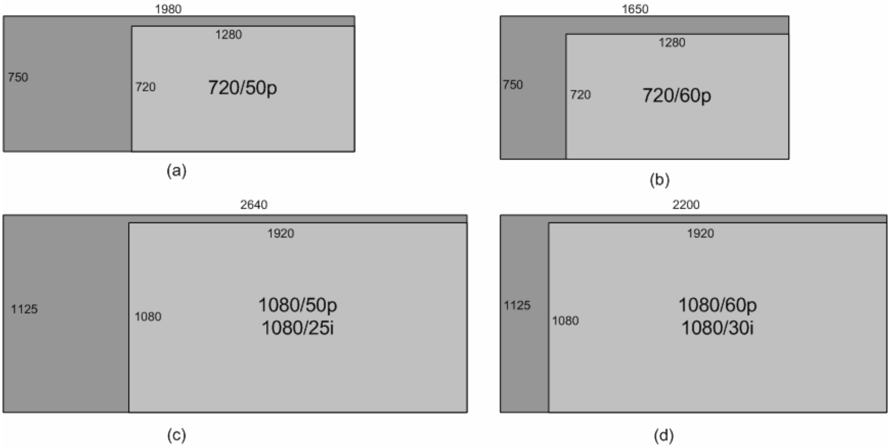


Fig. 1 Two major high definition formats representing total and active pixel image areas (a) 1280x720 active image at 50 frames per second, (b) 1280x720 active image at 60 frames per second, (c) 1920x1080 active image at 50 frames per second, (d) 1920x108050 active image at 60 frames per second

With regard to diversity of methods used for acquisition, production postproduction and displaying in a television and film production area; these two formats inherited different scanning frequencies. Table 1 gives an overview of scanning formats defined for mentioned image formats. Moreover, in order to cope and ensure coexistence of different scanning formats segmented frame has been standardized as well. Segmentation procedure corresponds to progressive capture of the image which is prior to transportation divided into two segments. This process ensures alleviation of transition between the old and new technologies. Practically, segmentation enables smoother accommodation, conversion and transport of material captured with lower scanning frequencies into formats suitable for postproduction, distribution and displaying. Segmented pictures have PsF nomenclature.

This diversity of video formats is very challenging for the television systems interoperability and efficient program exchange. It requires well designed digital systems, additional processing based on complex conversion methods and processing power as well. More details about technical process, types of conversion including format conversion, scaling process, aspect ratio conversion, scanning format conversion and their usage in production, postproduction and broadcast area interested reader can found in (Deame 2007).

Table 1 High definition picture format and scanning standards

Nomenclature	*Y' and R'G'B' samples per active line*	*Active lines per frame*	*Vertical frequency [Hz]*	*Y' and R'G'B' sampling frequency [MHz]*	*Total Y' samples per line*	*Total lines per frame*
720/60p	1280	720	60.00	74.250	1650	750
720/59.94p	1280	720	59.94	74.176	1650	750
720/50p	1280	720	50.00	74.250	1980	750
720/30p	1280	720	30.00	74.250	3300	750
720/29.97p	1280	720	29.97	74.176	3300	750
720/25p	1280	720	25.00	74.250	3960	750
720/24p	1280	720	24.00	74.250	4125	750
720/23.98p	1280	720	23.98	74.176	4125	750
1080/60p	1920	1080	60.00	148.500	2200	1125
1080/59.94p	1920	1080	59.94	148.35	2200	1125
1080/50p	1920	1080	50.00	148.500	2640	1125
1080/30i	1920	1080	30.00	74.250	2200	1125
1080/25i	1920	1080	25.00	74.250	2640	1125
1080/30p	1920	1080	30.00	74.250	2200	1125
1080/29.97p	1920	1080	29.97	74.176	2200	1125
1080/25p	1920	1080	25.00	74.250	2640	1125
1080/24p	1920	1080	24.00	74.250	2750	1125
1080/30psf	1920	1080	30.00	74.250	2200	1125
1080/29.97psf	1920	1080	29.97	74.176	2200	1125
1080/25psf	1920	1080	25.00	74.250	2640	1125
1080/24psf	1920	1080	24.00	74.250	2750	1125

4.2 Colorimetric Characteristics

High definition television standardization process introduced new, amended definition of the colorimetric parameter values. Before we explain the reason of this change we will describe and explain certain fundamental principles of colorimetric science and its application in the video systems (Luther and Inglis 1999, Russ 2006).

During the process of image capturing video equipment mimics the process of visual information acquisition and processing performed by the human visual system. The visual information is conveyed by the light which is then used to create the representation of a surrounding world. The representation of the surrounding world is possible only within the constrained electromagnetic spectrum which comprises of wavelengths capable to stimulate photosensitive cells on the back of an eye in the area called retina. Incident light reaching the retina cells produce an electrochemical reaction which results in visual perception. Photoreceptor cells can be differentiated based upon their response to incident radiation manifested in different spectral curves corresponding to red, green and blue portions of spectrum. Thus, we can say that human vision is trichromatic. Hence, using these three primary components the entire range of human color perception can be covered. This range was defined by the Comission Internationale de l'Eclairge (CIE) in 1931. Overview of the CIE system accompanied by explanation of objects colors measurement and the aspect of color measuring for imaging applications is given in (Ohno 2000).

In video systems acquisition is performed in a similar way. Three primaries, red, green and blue are combined in order to create representation of the visual scene. Nevertheless, the reproducible range of colors in video systems is reduced comparing to human visual systems. The entire range of all possible colors reproducible by the human visual system is referred to as the gamut which is defined by the triangle whose vertices are the location of the primaries on the CIE chart. A combination of three primaries in predetermined ratios according to the principles and CIE standards results in luminance. Since luminance is linearly dependent on light intensity and since it has broad range from lowest to highest value which would require a huge number of bits for representation, the most video systems apply non-linear transformation between brightness and voltage performing perceptual coding. The non-linear pre-correction function is performed on all three RGB components resulting in gamma corrected R'G'B' values. Weighted sum of new gamma corrected R'G'B' components results in luma component denoted as Y'. In high definition television systems luma is defined as Y'=0.2126R'+0.7152G'+0.0722B' with the gamma pre-correction function comparable to a square root function (γ=0.45). Besides reducing the number of bits required to represent brightness sensation the non-linear pre-correction function is needed to provide authentic and reliable representation of color displayed on TV sets and perceived by viewer. Using non linear pre-correction in video capturing systems actually compensates and rectifies the nonlinearity originated in display devices in order to achieve authentic reproduction of luminance.

In (Poynton 1996, Poynton 1997) the principle of constant luminance and the correct gamma process is described and explained. The importance of computing the luminance from tristimulus linear light RGB values is shown as well as the consequences in color reproduction that arise due to misapplied and misunderstood principle of the gamma process. Furthermore, Poynton disputes the need to change the chromaticity parameter values, used in the standard television systems, because of no practical significance. Additionally, these new parameters will definitely be the source of complex up-conversion and down-conversion processes between conventional standard definition and high definition television systems. The complexity could force the manufactures even to omit performing and including the mentioned processes which could, at the end of the chain, result in perceivable errors. ITU-R Recommendation BT.1361 recommendation describes the unified colorimetry parameters appropriate for future television and imagining systems (ITU Recommendation ITU-R BT.1361 1998). This document recommends chromaticity coordinates for standardized primary colors, analogue encoding equations, digital encoding equations, and strategy for extended color gamut and optimized integer coefficients for extended color gamut. In next section we will describe necessary parameters required for achieving the optimum visual experience.

4.3 Optimizing Perception of High Definition Video

In order to determine the viewing experience a different factors such as quality of a displayed video signal, ability of a display to authentically reproduce the image, viewing environment and viewing distance needs to be considered. Hereafter we will focus on the viewing distance and the size of the display to describe the relation between them as well as their influence on optimizing the viewing experience of perceived high definition images.

As mentioned, standardization bodies defined different image formats for high definition systems. The question is how the size of each image format and display device can affect visual experience of the viewer. Though, this is subjective category dependant on characteristics of an individual's visual system it can be defined and put in a relation for a broad viewer public. This relation can be explained introducing the inherent characteristic of the human visual system called visual acuity which describes the ability of the individual to differentiate and identify adjacent pixels, lines or objects on the displayed image. The closer the pixels are or the ticker the lines are the spatial frequency is higher and the ability of the human visual system to make a differentiation is reduced. However, there is a threshold below which a certain individual will not be able to distinguish the pixels or lines. The differentiation is possible when viewing distance is such that a viewing angle subtends more than one minute of an arc. Figure 2 depicts factors involved in determining the optimum visual experience.

Accordingly, calculation of optimum viewing distance can be made if high definition image width and height is known. Herein, we will give mathematical relations defining the dependence of aforementioned factors. The minimum picture object size d that can be discriminated by the viewer is defined with the following relation:

$$d = 2 \cdot D \cdot tg\frac{\alpha}{2} \tag{4.1}$$

where α represents the angle of 1 minute of the arc and D is the distance between the viewer and the object. Based on this, we can calculate a number of possible vertically discriminated objects (lines) N_L or horizontally discriminated object

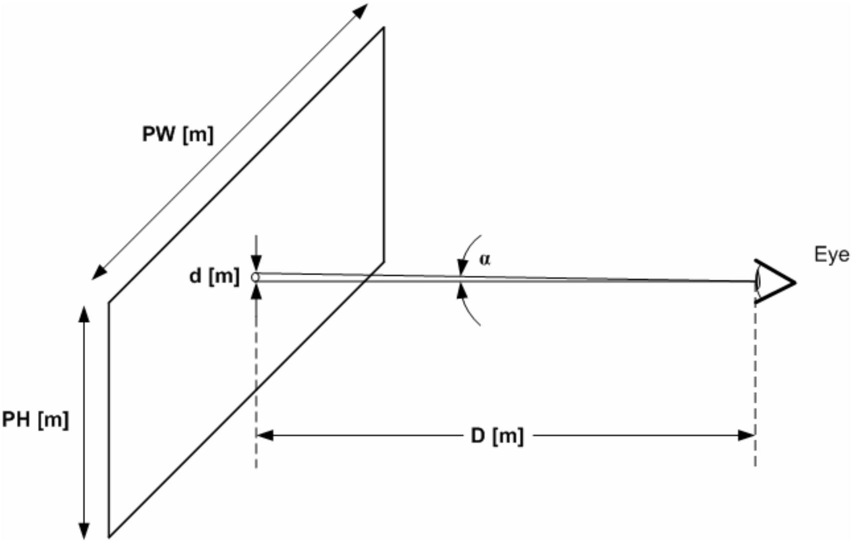


Fig. 2 Factors affecting the visual acuity

(samples) N_S for a particular picture display having picture width and height (P_W and P_H respectively) with ratio of 16:9:

$$\begin{aligned} N_L &= \frac{P_H}{d} \\ N_S &= \frac{P_W}{d} \end{aligned} \tag{4.2}$$

Based on these relations we can determine parameters necessary for optimized viewing experience in the high definition environment. Figure 3 depicts the interdependence among scanning lines, optimized viewing distance and display size. Calculations were made for different viewing distances and display sizes currently present on the market. The aim was to show the limitations of the particular display size ability to show the full resolution of particular source picture format. Therefore, to show the 1280x720 high definition picture format at the viewing distance of 3 meters 50 inch display size is required. Display sizes smaller than 50 inch do not allow viewers to perceive original picture format unless viewing distance is decreased. The same situation refers to 1920x1080 high definition picture formats. At viewing distance of 3 meters 75 inch display is required. Therefore we can see that selection of picture formats having higher resolution affects the display size asking for the larger display sizes as well as for shorter viewing distances.

Furthermore, based on relations we can calculate the optimum viewing distances expressed in picture heights P_H for both high definition video formats and compare them with standard definition video formats. Figure 4 depicts the optimum horizontal

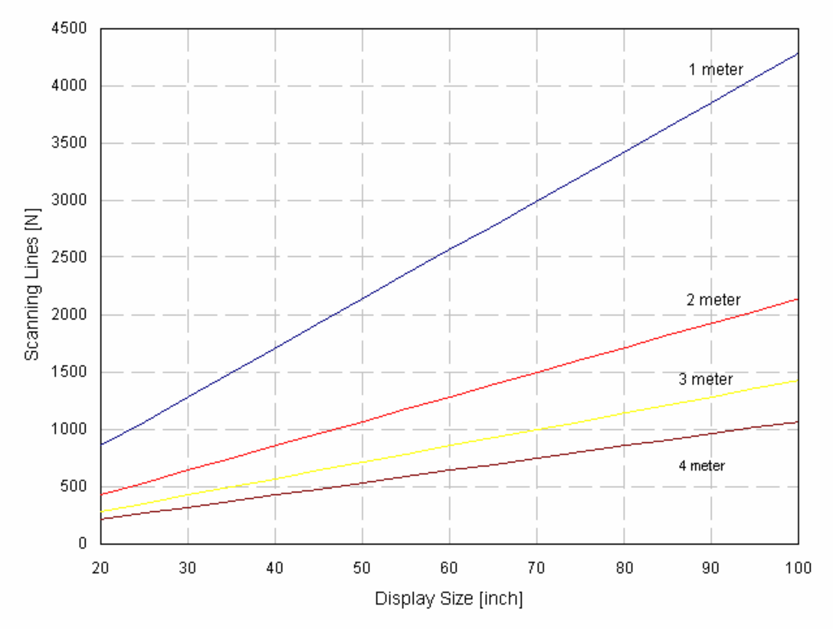


Fig. 3 Interdependence of viewing distance, number of scanning lines and display size

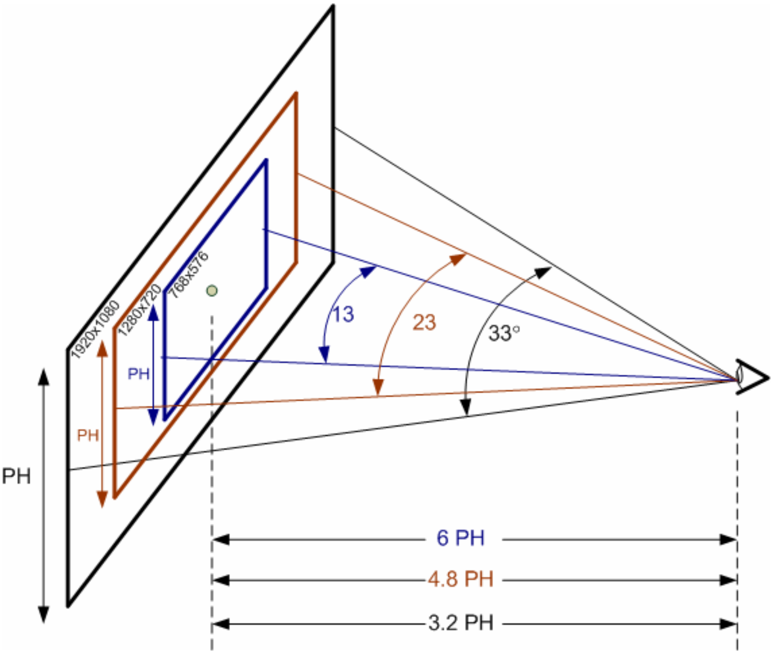


Fig. 4 Optimized viewing distance and viewing angle for standard and high definition video formats

and vertical viewing angle that corresponds to viewing distance measured in picture heights (P_{H}) for standard and high definition video formats.

Nevertheless, besides theoretical methods numerous subject evaluations of optimum viewing distances for different formats have been performed. In order to confirm the authenticity of the concept which defines 1 minute of the arc as the minimum angular resolution experiments have been carried out and described in (Drewery 2004). Obtained results confirmed that the interdependence among the viewing distance, display size and source picture format can be used to select the appropriate television standard that is to be delivered to home environment. Furthermore, results confirmed the theoretical calculation for 720/50p video format, recommended by authors for delivery to home displays in range from 26 to 28 inch in diagonal. However, the recommended video format would, for mentioned display sizes, cause large saturation of human visual system with unnecessary details.

Another survey, made twenty years ago, aimed to investigate the appropriate display size and viewing distance for flat panel displays in the home environment found out that display sizes from 40 to 50 inches in diagonal would be the most practicable and desirable. The results obtained for the optimal viewing distances in the domestic environment were in range from 4 P_{H} – 6 P_{H} while the optimal distance for critical assessment of picture quality was set between 2.5 P_{H} – 3.5 P_{H}. The results have shown differences between laboratory viewing environment and domestic viewing environment determined by practical limits such as room dimensions and furniture arrangement (Tanton and Stone 1989/1990).

Similar tests were performed by the NHK in Japan (Sugawara 2005). The participants of the research were subjected to viewing the high definition pictures under different viewing angles ranging from 15° to 60° and two viewing distances, 2m and 3m. Accordingly, the increase in the viewing angle resulted in higher levels of viewing experience. Moreover, higher levels of visual experience were achieved for longer distance which in this case was set to 3m. The results confirmed the advantages of introducing the high definition formats which enable wider viewing angles corresponding better to effective visual field perceived by the human visual system. Accordingly, authors concluded that among the standardized formats the 1920x1080 picture format is the closest to the psychological characteristic of HVS.

International Telecommunication Union issued a guideline on selection and usage of different image formats in various broadcasting applications (ITU Recommendation ITU-R BT.1845 2008). The guidance is based on optimum viewing distance and optimum viewing angles measured in suitable viewing conditions. Therefore, the appropriate metric of picture rasters needs to be selected for adaptation of program material in broadcast applications. Additional information on viewing distances in conventional television environment can be found in (ITU Recommendation ITU-R BT.654 1986, ITU Recommendation ITU-R BT.1128 1995, ITU Recommendation ITU-R BT.811 1994, ITU Recommendation ITU-R BT.1129 1998, EBU Recommendation R28 1997).

4.4 Aspect Ratio in HDTV Systems

A particular term called Aspect Ratio (*AS*) is used in a television terminology to describe the proportions of the picture intended to be depicted on the picture display. It represents the ratio between picture width (*W*) and picture height (*H*).

In video industry a various aspect ratios are being used for displaying pictures on the screen display. Standard Definition Television Systems have the aspect ratio of 4:3 (or 1.33:1) while the High Definition Television Systems use 16:9 (1.78:1) ratios between picture's width and height. In relation to Standard Definition Television Systems high definition television families of 1920x1080 and 1280x720 picture formats represent increase in picture area of 79% and 56% respectively. Figure 5 depicts difference in effective viewing picture size among SDTV and HDTV systems.

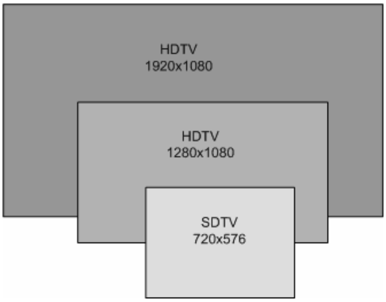


Fig. 5 Comparison of active picture sizes among HDTV and SDTV systems

The ratio of picture width and height defined for HDTV systems enables more comfortable and relaxed viewing of particular program content because of the characteristic of the human visual system which perceives visual information in horizontal better than those in vertical direction. Hence, 16 by 9 ratio corresponds better to effective visual field of a human visual system which subtends an angle around 130° (Darmon 1997).

Standard definition television systems still dominate in the production and transmission applications throughout the world. Accordingly, although more suitable, introducing the new aspect ratio will yield difficulties in the transition period from standard toward high definition television. With the growth of high definition equipment and its availability, more and more material will be produced in high definition. This will lead to the situation where two systems will have to function simultaneously. Thus, produced high definition material will be used in standard definition systems and material produced with standard definition equipment will be used in high definition systems. Obviously, there is a need for aspect ratio and format conversion in different areas such as production, post-production, broadcasting and picture displaying. In the next section we will describe the usage of high definition video formats in different television environments.

5 Choosing the Right Standard

Before providing the viewers with the high level of visual sensation, television signal conveying the visual information, undergoes processes such as capturing, acquisition, editing, mixing, distribution and transmission. In order to preserve visual information convenient for satisfactory displaying a detailed analysis and good system design is essential. Therefore in the transition period towards high definition future different technical issues ought to be analyzed; the transmission platform – terrestrial, satellite or cable; the production and post-production high definition video format; the most appropriate compression method (Wood 2004). Hereafter, we will describe the requirements on production, post-production, storage and transmission of television signal in the high definition television environment.

5.1 Serial Digital Interface in High Definition Studio Environment

In comparison to standard definition television signal, high definition television signal, in its uncompressed form is represented with the huge number of information. The amount of information depends on used scanning standard and picture format. Although transportation of information in a studio environment can be carried via parallel interconnection it is more convenient, simplified and economical to convey the information over single physical media, whether it's optical fiber or coaxial cable. Video data carried via serial digital interface comprises of luma and croma components multiplexed in 4:2:2 ratios, followed by scrambling, error protection

Table 2 Bit rates for different combination of image format and scanning standard

High Definition System	*Y' bandwidth [MHz]*	*Y' sampling frequency [MHz]*	*CR, CB Bandwidth [MHz]*	*CR, CB Sampling frequency [MHz]*	*Bit rate at 4:2:2 using 10 bit quantization [Gbps]*
720/60p	30	74.176	15	37.088	1.484
720/50p	30	74.250	15	37.125	1.485
720/30p	30	74.176	15	37.088	1.484
720/25p	30	74.250	15	37.125	1.485
1080/60p	60	148.352	30	74.176	2.967
1080/50p	60	148.500	30	74.250	2.970
1080/30p	30	74.176	15	37.088	1.484
1080/25p	30	74.250	15	37.125	1.485
1080/30i	30	74.176	15	37.088	1.484
1080/25i	30	74.250	15	37.125	1.485

and conversion to NRZI coding format for the purpose of minimizing the signal's low frequency components, spreading the energy spectrum and enabling the reliable clock recovery at the receiving side. Serial digital interfaces for standard definition studio signals are defined in (ITU-R BT.656 1998, ITU-R BT.799 2007, SMPTE 259M 2006, EBU Tech. 3267 1992) while the high definition studio signal interfaces, based on Common Image Format, operating at 1.485 GHz and 2.95 GHz, are defined in ITU-R BT.1120, SMPTE 424M and SMPTE 292M. Among these high definition serial digital interfaces, SMPTE 424 defines single optical or coaxial link at 3 Gbps, SMPTE 372 defines dual link at 2.95 Gbps, SMPTE 292 defines coaxial link at 1.485 Gbps, SMPTE 297 optical fiber link at 1.485 Gbps while ITU-R BT.1120 defines dual link operations at 1.450 and 2.900 Gbps.

According to Table 2 and depicted calculated bit rates it is obvious that the usage of uncompressed, baseband high definition video signal affects and limits the possibilities of today's storage and transmission technologies used in television systems. Therefore, there is indisputable necessity to reduce the amount of data in order to make it suitable for storage and transmission applications.

5.2 Studio Production Scanning and Picture Standards

The program production chain involving processes of capturing, compression and post-processing is responsible for viewer's visual experience of a perceived program. Therefore it is very important to take great care on parameters such as lighting conditions, scene composition and shot-framing during capturing which has influence on colorimetric, contrast, noise level and presence of high details and movement in a recorded video material. Moreover, all these characteristics are of great importance in compression process required for further manipulation; post-processing of the recorded video material and archiving for future usage as well.

Standard definition environment and high definition even more, requires selection of suitable scanning and picture standards capable not only to fulfill the strict demands and match high criterions of viewers but to do so in a simplified and economical way taking care at the same time about interoperability, implementation stability and most of all, the picture quality preservation. Although various standards have been issued recommending different picture formats in combination with different scanning frequencies all parties involved in production of high definition video material should consider possible advantages and drawbacks of particular picture format and scanning standard combination. This also includes selection of appropriate standard for transferring the film and archive material to high definition studio video format, post-processing, video material storage and studio distribution.

Selection of common picture format having 1920 samples and 1080 lines with 50/60 Hz progressive scanning definitely has advantages due to better representation of motion portrayal, higher vertical and horizontal resolution making the down-conversion and compression techniques more practical and economical. Moreover, video cameras manufactured today already have CCD/CMOS (Charged Coupled Device/Complemetary Metal Oxide Semiconductor) chips able to produce video signals in 50/60 Hz progressive CIF format thus, avoiding the need for usage of interlaced high definition standards. However, there is an issue which compromises and complicates the usage of 1920/50p/60p formats in high definition production; it can be found in absence of suitable and practical high definition serial digital interface in today's studio equipment.

Furthermore, usage of progressive segmented 1080/25psf common picture format was confirmed to be advantage in digitization of film material due to similar frame rate and vertical resolution affecting the better preservation of film details (EBU Tech 3315 2006). Nevertheless, the 1080/50p/60p is desirable standard which usage will depend on market availability.

Due to storage limitations, during the process of picture capturing resulting video signal is being transformed from its uncompressed, baseband form into the compressed form stored in a recording medium such as cassette, optical disc or solid state memory. In a modern, file based, non-linear and non-real time HDTV production the usage of optical discs and memory cards as well as interfaces enabling higher data throughput is definitely an advantage. Considering this new environment, selection of appropriate compression technique is very important since it reduces quality of a recorded material. Bearing in mind the goal to preserve quality and efficiently use storage capacity this is certainly demanding task asking for compromise between compression level and picture quality.

Numerous evaluation of compression format aimed for high definition usage has been conducted (Visca and Hoffman 2008). Results obtained in these evaluations and format comparison showed that for acquisition purposes picture formats having 4:2:2 sampling structure should be used without additional vertical and horizontal sub-sampling. Recommended quantization is 8-bit for mainstream program while 10-bit is preferred for high-end acquisition introducing complex graphics and animations. Furthermore, it was determined that for production and archiving of a mainstream program preferred bit rate should not be less than 100 Mbps if used format is based on I frames only and 50 Mbps if used format is based on Long GOP (Group of Pictures) MPEG-2. Above all, multiple generations of high definition recorded material should be avoided and kept at least up to 4 to 5 generations (EBU Recommendation R124 2008). Although, there is a clear goal to move the production to tapeless, file based environment, tapes are still dominant and widely used in standard as well as in high definition acquisition and production. Accordingly, international telecommunication Union gave an overview of currently available tape recording formats for HDTV signals which can be found in (ITU Recommendation ITU- R BR.1375 2007).

5.3 Broadcast Scanning and Picture Standards

Unlike production and acquisition, appropriate picture format selection for broadcasting environment requires additional factors to be considered. Besides its positive characteristics manifesting in a significantly enhanced television experience, high definition television signal has brought some negative consequences especially in the field of contribution and distribution of a television signal. High definition video signal characterized with huge amount of data requires higher transmission channel capacities. Since frequency spectrum is very expensive and limited, careful analysis, transmission system design and frequency planning is needed as well as the selection of convenient compression technique.

As mentioned, transportation of television signal can be done over cable, satellite and terrestrial platform. Each platform is characterized by its own modulation and channel coding technique determined by the propagation conditions of a transmission environment. Accordingly these characteristic affect the channel capacity which combined with efficient coding technique followed by multiplexing enable transportation of higher number of television signals. Various standards have been issued throughout the world to define the requirements on transmission systems for contribution and distribution of a television signal. Above all, the choice of high definition picture format and scanning standard also plays an important role because each combination of picture format and scanning standard could require additional spectrum resources or channel capacity, additional display processing at receiving equipment or more demanding compression process. Hence, using progressively scanned pictures with higher scanning frequency alleviates the compression process and excludes the need for conversion process in display avoiding accompanying negative effects which result in loss of picture quality (EBU Tech 3312, 2006). On the other hand, usage of picture formats combined with interlaced scanning is characterized by less vertical-temporal information which complicates the compression process and requires additional deinterlacing process in a display device. Based on performed tests, advantages of progressively scanned picture formats have been recognized and confirmed. Accordingly, it was confirmed that interlaced picture formats (1080/25I) require 20% more bit rate in comparison to the progressive picture format (720/50p) in order to achieve the same subjective picture quality. Moreover, relatively new H.264/AVC compression technique showed great efficiency over MPEG-2 compression technique enabling 50% bit rate saving for the same picture quality (EBU Recommendation R 124 2008).

In 2004 EBU Technical committee recommended that distribution and transmitting of high definition television signal should be based on the 720/50p standard although 1080/50p standard was mentioned as a desirable option for future applications. Next section will describe the interdependence among high definition picture format and scanning standards, compression techniques and bit rates in order to confirm the most convenient combination for usage in transmission environment.

6 Picture Quality Assessment in High Definition Environment

Before making a decision what combination of picture format and scanning standard is the most appropriate for usage in particular television field, whether it is production or broadcast a detailed analysis has to be made. This implies taking each high definition video format and submitting it to complete production and broadcast processing chain. In previous sections we described advantages and drawbacks of using particular video format in the production and broadcast area as well as the effects on the viewer's visual experience. Thus, selecting the particular combination is very risky and constrained by many technical and economical factors including production and broadcast equipment availability, the processing quality, system design, spectrum boundedness, availability of receivers and display devices to general public, etc. Hereafter we will describe how appropriate video format and compression method selection affects the picture quality perceived by the viewer in the home environment.

Quality of television signal is exposed to different factors prior to being displayed in the home environment. Considering that signal delivered out of studio camera or some other equipment conforms to standard, and that the environment conditions enable unobstructed propagation, we can say that the picture quality is mainly determined by the applied compression technique and compression ratio. Consequently, in order to select the best compression technique and compression ratio that will not compromise the quality of delivered video signal various methods of picture quality assessment are being used. Therefore, an efficient method of picture quality measurement is required to identify and determine the level of reconstructed picture quality degradation. Bearing this in mind, even a quality correction and improvement of video signal can be easier. Hereafter we will explain different methods for video quality evaluation and the difficulties encountered during such process.

6.1 Picture Quality Assessment Methods

A natural video scene is composed of different objects characterized by the shape, texture, dimension and position. They can be barely noticed or large enough to cover the whole picture. These spatial (objects shape and dimension) and temporal (moving objects, camera movement) characteristics determine the efficiency and complexity of a compression process. To determine the level of picture degradation can be a very difficult and demanding process because it depends on subjective experience of the viewer which is the result of a complex interaction between eye and brain. Methods used for the assessment of the level of picture degradation can be roughly assigned into two groups of methods; subjective and objective quality assessment methods.

Although more accurate than objective methods, subjective quality measurement is more demanding, time consuming and expensive, dependent on many parameters (active or passive viewing, viewing environment). Methodology for subjective quality assessment can be found in (ITU Recommendation ITU-R BT.500-11 2002, ITU-T P.910 1999) with the detailed description of viewing conditions, required source signals, criterions for test sequences selection and results presentations. As opposed to subjective quality assessment methods objective methods are faster and computationally easier because of using the mathematical models which make them more attractive for implementation in video systems (ITU-R Recommendation ITU-R BT.1683 2004, Webster 2004). Therefore the goal is the development of quantitative measures that can automatically predict video quality. The drawback of the objective methods is their inability to reproduce and align to human visual perceptual system. Due to demanding and complex process subjective methods are used mainly during the development and system design but for the operational part objective methods are preferred (ITU Recommendation BT.1683 2004).

Most of the objective assessment methods determine the video quality calculating the difference between original and reconstructed video after compression where the amount of differences represents degradation level of reconstructed video. A thorough analysis and description of different objective methods can be found in (Wang et al. 2003). Wang describes procedures and explains difficulties of designing and developing an objective video quality metric that would correlate well with subjectively perceived video quality. Moreover, for better understanding a brief introduction to the relevant physiological and psychophysical components of the HVS is given. Some of the objective picture quality measures include:

- *PSNR* (Peak Signal to Noise Ratio) (Grgic et al. 2004)
- *SSIM* (Structural Similarity Index) (Wang et al. 2004)
- *VQM* (Video Quality Measure) (Xiao 2000)

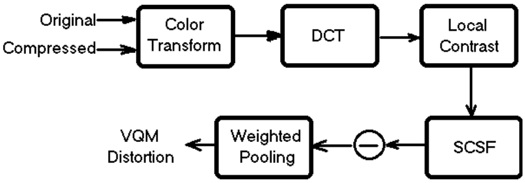
PSNR is the ratio between the maximum possible power of a signal and the power of noise. *PSNR* is usually expressed in terms of the logarithmic decibel. In expression (6.1) $a_{i,j}$ and $b_{i,j}$ are pixels from original and compressed picture. x and y describe height and width of an picture. *MSE* stands for Mean Square Error:

$$PSNR = 10\log_{10}\frac{255^2}{MSE} \qquad MSE = \frac{\sum_i\sum_j (a_{i,j} - b_{i,j})^2}{x \cdot y} \tag{6.1}$$

The Structural Similarity (*SSIM*) is a novel method for measuring the similarity between two pictures (Wang Z et al. 2004). It is computed from 3 picture measurement comparisons: luminance, contrast and structure. Each of these measures is calculated over the 8x8 local square window which moves pixel-by-pixel over the entire picture. At each step, the local statistics and *SSIM* index are calculated within the local window. Because resulting *SSIM* index map often exhibits undesirable "blocking" artifacts, each window is filtered with Gaussian weighting function (11x11 pixels). In practice, one usually requires a single overall quality measure of the entire picture, so Mean *SSIM* (*MSSIM*) index is computed to evaluate the overall picture quality. The *SSIM* can be viewed as a quality measure of one of the pictures being compared, while the other picture is regarded as of perfect quality. It can give results between 0 and 1, where 1 means excellent quality and 0 means poor quality. Figure 6 is an overview of the flowchart of *VQM* (Xiao 2000).

First step is color transform. Both MPEG and H.264/AVC use the YUV color space, so it can use the raw data directly. After that we transform original and compressed picture using DCT transform. Approximately said, this step separates incoming pictures into different spatial frequency components. Third step is converting each DCT coefficient to local contrast (LC). After this step, most values lie between [-1, 1]. Fourth step converts LC to just-noticeable differences (jnds). The DCT coefficients are converted to just-noticeable differences by multiplying

Fig. 6 VQM measuring algorithm



each DCT coefficient by its corresponding entry in the SCSF (Spatial Contrast Sensitivity Function) matrix. For static SCSF matrix, MPEG default quantization is used. For dynamic matrix each entry in static SCSF matrix is raised to a power to account for the temporal property of SCSF. The power is decided by the frame rate of video sequences. Last step is weighted pooling of mean and maximum distortion. The two sequences are subtracted first. At this step *VQM* also incorporates contrast masking into a simple maximum operation and then weights it with the pooling mean distortion. This reflects the facts that a large distortion in one region will suppress sensitivity to other small distortion, because weighted maximum distortion into pooled distortion is much better than pooled distortion alone.

6.2 Work on High Definition Picture Assessment

Hereafter we will present a work done in the area of subjective high definition picture assessment, the implementation suitability as well as the potentials of particular high definition video format in production and broadcasting area. Hence, the goal of the researches was to find the video format that would preserve the video information in the process chain and give the best viewing experience in the home environment.

Research done in (Haglund et al. 2002) set up a number of tests to find out the implications of 576/25i, 720/50p and 1080/25i video formats on broadcasting applications as well as their displaying on contemporary widespread WideVGA (Wide Video Graphics Array) flat panel displays. Test sequences available in all three formats were coded using MPEG-2 method (ISO/IEC Recommendation 13818-2 1998) at different bit rates. Obtained results showed significant advantage of 720/50p video format over other two video formats. Authors concluded and recommended that among tested formats, 720/50p should be used for broadcasting applications, especially terrestrial which are more constrained regarding the channel capacity. The best performance 720/50p format owes to progressive scanning which makes it more suitable for compression process and more appropriate for displaying without the need for demanding in-display processing.

In (Hoffman et al. 2006) the authors made research on different high definition formats in order to find the most suitable for delivery and viewing on the flat panel display having 1920x1080 pixel resolutions. Used test sequence, available in 1080/50p, 1080/25i and 720/50p format, were coded at different bit rates using H.264/AVC coding method while the picture quality assessment was done using Double Stimulus Impairment Method at 3PH and 4PH viewing distance. As expected, picture quality degraded with lower bit rates but more rapidly for 1080i/25 than for other two formats. The best results were obtained for 1080p/50 format followed very closely with 720/50p format despite the up-scaling process needed to show it on 1920x1080 displays. Nevertheless, considering the pixel rate of 1080/50p in relation to other two video formats the results for 1080/50p were remarkable. Authors concluded that this is due to the higher quality of 1080/50p conforming to display resolution and thus avoiding additional processing such as de-interlacing and scaling; usage of progressive scanning is efficiently exploited by the compression process; the characteristics of a test sequence were suitable for the coding and displaying process. Herein, we have to emphasize that this research does not give an answer to the question which format is the best, it only reports on characteristics of each individual high definition format.

In (Hoffman et al. 2008) the authors used a novel subjective assessment method called Triple Stimulus Continuous Evaluation Scale. Three picture formats 1080/50p, 1080/25i and 720/50p, representing potential candidates for usage in broadcasting applications, were compressed using H.264/AVC coding method at different bit rates before being shown on the three displays in a particular vertical arrangement. The method confirmed that the direct comparison of different high definition video format is possible and that progressively scanned pictures enable easier coding and displaying process. Results also showed better performance of 720/50p format, especially for complex and demanding content and at lower bit rates while the 1080/50p format is more appropriate for simple content and for higher bit rates where its high resolution is not masked by compression artifacts. Authors concluded that test results for 720/50p and 1080/25i were better when source material had higher spatial resolution (2160p/50 gained by film digitization in relation to 1080/50p gained from CCD camera). Based on the results authors developed an idealized system chain for contemporary HDTV environment.

6.3 Test Setup and Results for Different Compression Tools

In previous section we mentioned that compression process is responsible for degradation of picture quality. Nevertheless, the selection of efficient compression method can affect transmission and storage capacity enabling better utilization of available storage space or frequency spectrum. Therefore, in order to determine the level of picture quality degradation we performed subjective and objective assessment of compressed high definition video formats. Moreover, influence of certain combination of compression tools on high definition video formats was determined.

Two test sequences representing different levels of program content complexity were selected from (Haglund 2002) where the first sequence CrowdRun was more complex, full of details and motion critical in comparison to the second test sequence InToTree representing simple content. Used sequences were available in all three high definition formats 1080/50p, 1080/50i and 720/50p. The compression process was performed with software application using H.264/AVC method (ITU Recommendation H.264/AVC 2005) to compress the test sequences to required bit rates at 3 Mbps, 6 Mbps, 9 Mbps, 12 Mbps, 15 Mbps and 18 Mbps. In order to examine the influence of certain compression tools on picture quality each test sequence was further compressed omitting particular H.264/AVC tool. Thus, after the coding process we had five different configurations H.264/AVC coding tools for each video format at each bit rate. We denoted them as: A (basic coding tools without the Context Adaptive Binary Arithmetic Coding); B (basic coding tools with reduced motion vector accuracy from quarter to half pixel accuracy); C (basic coding tools using 16x16 block size from motion prediction process); D (basic coding tools without de-blocking functionality); E (basic coding tools including depicted in Table 3).

Table 3 Basic coding setup

Coding Parameter	*Video Format* *720/50p*	*Video Format* *1080/25i*	*Video Format* *1080/50p*
Profile	High Profile	High Profile	High Profile
Level	3.2	4.0	4.2
Scanning	Progressive	Interlaced	Progressive
Aspect Ratio	16:9	16:9	16:9
I picture	33	33	33
Number of B pictures	3	3	3
Reference Pictures	4	4	4
Reference B Picture	Yes	No	Yes
Prediction Block Size	16x16	16x16	16x16
Motion Vector Accuracy	1/4	1/4	1/4
Adaptation Filter	yes	yes	yes
Entropy Coding	CABAC	CABAC	CABAC

Quality measurement and assessment of coded video test sequences was performed with Peak Signal-to-Noise Ratio (*PSNR*). Above all, we examined and compared the time required for coding of video formats.

The results of the video quality assessments were expressed in a bit rate saving form. The bit rate savings between two configurations were calculated by determining the difference between bit rates needed to achieve the same PSNR values. Curve with the poorest PSNR results was used as a reference during calculation of the bit rate savings. Bit rate saving is defined as:

$$S_{bit} = \frac{A(PSNR) - B(PSNR)}{A(PSNR)} \cdot 100 \ [\%] \tag{6.2}$$

where *A* represents the bit rate of an inferior configuration necessary to achieve certain PSNR value while *B* represents the bit rate of the better configuration necessary to achieve the PSNR value of an inferior coder. Besides this interpolation process an easier approximate method for bit rate calculation is given in (Bjontegaard 2001). Anchor combination in our tests was the combination *A* representing basic coding configuration where the CABAC option was omitted. This combination was set as the anchor because the obtained results were the worst.

The results of bit rate savings among used combinations for all three video formats and two test sequences, Crowd Run and InToTree, are depicted in Tables 4 through 9. Looking at average values we can tell that the best results for both test sequences and all three video formats were obtained for configuration E using basic tools. As expected, obtained results for all configurations show tendency of reduction in bit rate savings with higher measured *PSNR* values or with higher bit rates. It can be concluded that at higher bit rates quality of picture is improving and amount of impairments are lower due to lower compression factor. Therefore, whichever configuration is used its influence is fading at higher bit rates. In case of configuration D where de-blocking function is omitted the values for bit rate savings are reduced for less demanding InToTree test sequence. Furthermore, the importance of using de-blocking function at lower bit rates can be seen, especially for more demanding content. De-blocking function is very important at lower bit rates since it smoothes the block edges improving the appearance and reducing the blocking distortion which is more noticeable, particularly for progressively scanned pictures. However, in case of less demanding content, which is easier to compress this is not so prominent. Usage of less accurate motion prediction in configuration B and bigger block size in configuration C degrades the picture quality at higher bit rates and in higher extent for more complex content. Nevertheless, smaller block size and higher motion vector accuracy can positively affect the picture quality but on the other hand they can increase complexity and bit rate. Therefore, a compromise has to be made. Beside this, we have to mention time measured to code the sequences. Hence, 720/50p sequences were coded faster then 1080/50i and 1080/50p; 720/50p was coded two times faster than 1080/50p video format.

Table 4 Bit rate savings for Crowd Run 1080/50p format relative to Configuration A

PSNR Y' [dB]	*Configuration B*	*Configuration C*	*Configuration D*	*Configuration E*
	Bit rate saving [%]	*Bit rate saving [%]*	*Bit rate saving [%]*	*Bit rate saving [%]*
24.16	14.50	14.75	3.00	15.50
26.03	11.67	11.67	6.17	13.33
27.79	12.67	12.11	8.11	15.22
29.06	10.08	8.83	6.33	12.83
30.09	7.73	6.53	5.73	11.00
30.89	7.33	6.17	6.44	10.72
Average	10.66	10.01	5.96	13.10

Table 5 Bit rate savings for Crowd Run 1080/25i format relative to Configuration A

PSNR Y' [dB]	*Configuration B* *Bit rate saving [%]*	*Configuration C* *Bit rate saving [%]*	*Configuration D* *Bit rate saving [%]*	*Configuration E* *Bit rate saving [%]*
25.16	18.50	15.25	9.50	21.00
26.98	11.00	8.83	7.17	14.50
28.74	7.78	6.44	6.78	13.56
29.95	5.58	4.75	6.50	12.00
31.00	3.53	3.27	5.80	10.33
31.86	2.11	2.00	5.22	9.22
Average	8.08	6.76	6.83	13.44

Table 6 Bit rate savings for Crowd Run 720/50p format relative to Configuration A

PSNR Y' [dB]	*Configuration B* *Bit rate saving [%]*	*Configuration C* *Bit rate saving [%]*	*Configuration D* *Bit rate saving [%]*	*Configuration E* *Bit rate saving [%]*
25.54	11.75	11.00	7.75	14.50
27.32	8.17	6.83	6.67	10.67
29.07	7.00	4.67	6.89	10.33
30.40	5.42	3.17	6.42	8.58
31.46	4.73	2.27	6.73	8.60
32.38	3.61	1.44	5.83	7.50
Average	6.78	4.90	6.71	10.03

Table 7 Bit rate savings for InToTree 1080/50p format relative to configuration A

PSNR Y [dB]	*Configuration B* *Bit rate saving [%]*	*Configuration C* *Bit rate saving [%]*	*Configuration D* *Bit rate saving [%]*	*Configuration E* *Bit rate saving [%]*
32.43	16.50	16.25	12.50	18.25
33.58	9.67	10.17	8.17	12.17
34.48	9.56	10.67	9.56	14.33
35.09	6.92	8.08	8.00	12.08
35.53	6.87	8.13	9.07	12.27
35.89	6.22	7.56	9.39	11.67
Average	9.29	10.14	9.45	13.46

Table 8 Bit rate savings for InToTree 1080/25i format relative to configuration A

PSNR Y	*Configuration B*	*Configuration C*	*Configuration D*	*Configuration E*
[dB)	*Bit rate saving [%]*	*Bit rate saving [%]*	*Bit rate saving [%]*	*Bit rate saving [%]*
33.83	12.50	9.00	10.25	17.75
34.97	8.17	4.67	7.83	12.50
35.94	6.56	1.78	8.44	12.67
36.61	5.50	1.00	9.08	11.58
37.13	5.80	1.33	10.73	12.40
37.59	5.33	1.33	10.56	11.44
Average	7.31	3.19	9.48	13.06

Table 9 Bit rate savings for InToTree 720/50p format relative to configuration A

PSNR Y	*Configuration B*	*Configuration C*	*Configuration D*	*Configuration E*
[dB]	*Bit rate saving [%]*	*Bit rate saving [%]*	*Bit rate saving [%]*	*Bit rate saving [%]*
34.11	7.50	7.75	11.25	13.00
35.33	4.67	5.00	9.17	10.17
36.38	4.78	5.22	11.33	12.11
37.13	3.83	4.50	10.50	10.25
37.71	4.87	5.53	11.73	10.87
38.23	3.83	5.67	12.00	10.83
Average	4.91	5.61	11.00	11.20

6.4 Reliability of Objective Picture Quality Measures

The aim of the research presented in this section was to compare the reliability of objective picture quality measures that have been introduced in section 6.1. To be able to compare three of the above mentioned methods, we used subjective quality results from (Hoffmann et al. 2006). Test sequence used in this comparison was Crowd Run sequence, which source material can be downloaded from (Crowdrun source material). Sequence in three different resolutions (1080/50p, 1080/25i and 720/50p) was first converted from .sgi to .yuv format using sgi2yuv program, (sgi2yuv program). Afterwards these uncompressed .yuv sequences were converted to .mpg program stream using H.264/AVC compression (Main Profile).

Basic settings for H.264/AVC encoder were:

- MP@level3.2, 720p; MP@level4.0, 1080i; MP@level4.2, 1080p sequence
- 24 Mbit/s maximum bitrate value
- keyframe interval: 33 frames
- 3 B-pictures, using B-splices as reference
- 4 reference frames
- search shape: 8x8

- CABAC entropy coding
- MPEG Program Stream
- variable bitrate, average bitrate values: 6, 8, 10, 13, 16, 18 and 20 Mbit/s.

Original and compressed sequences were compared using PSNR, SSIM and VQM quality measures. Figure 7 (a)-(c) shows different objective quality measures in relation to bit rates.

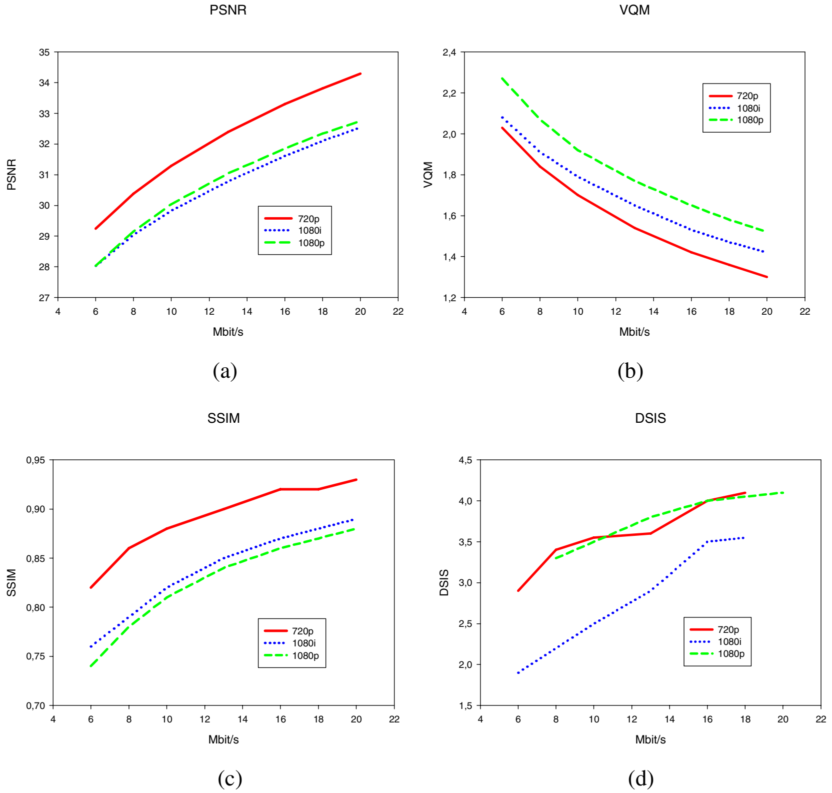
Fourth, subjective quality measure *DSIS* (Double Stimulus Impairment Scale) for sequence Crowdrun has been taken from reference (Hoffmann et al. 2006), for distance "3H" (H is display height). Basically, in this subjective measure 21 non-expert viewers, male and female of average age, were selected as observers after screening for normal vision. Training sequences and an explanation were given before the viewings, and short relaxation breaks between the test series were offered to the observers. *DSIS* method according to ITU-R BT.500-11 (ITU Recommendation BT.500-11 2002) was used in preparation and presentation of the test sequences. During the voting period the observers were asked whether they could observe a difference between the reference and the test signal and to mark their result in the corresponding category according to the five in ITU-R BT.500-11 defined terms (5-imperceptible, 4-perceptible, but not annoying, 3- slightly annoying, 2-annoying, 1-very annoying), Figure 7 (d).

From the objective quality measures it could generally be concluded that resolution 720/50p gives the best results. Also, 1080/50p resolution gives worst results (or similar as 1080i when comparing with *PSNR*). When comparing subjective quality measure, results are different. Resolutions 1080/50p and 720/50p give similar results and 1080/25i gives much worse *DSIS* than other two tested resolutions. So it can be seen that objective measures cannot be used for comparison of different resolution formats. Also, it can be seen from *DSIS* that any progressive resolution format is much better than interlaced format.

Figure 8 shows different objective video quality measures relative to *DSIS*. In Table 10 it is shown Pearson's correlation between objective measures and *DSIS*, which can be calculated as:

$$\begin{gathered} r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)\cdot s_x \cdot s_y}, \mathrm{i} = 1,\ldots,n \\ \bar{x} = \frac{1}{n}\cdot\sum_{i=1}^{n} x_i, \bar{y} = \frac{1}{n}\cdot\sum_{i=1}^{n} y_i \\ s_x = \sqrt{\frac{1}{n-1}\cdot\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ s_y = \sqrt{\frac{1}{n-1}\cdot\sum_{i=1}^{n}(y_i - \bar{y})^2} \end{gathered} \qquad (6.3)$$

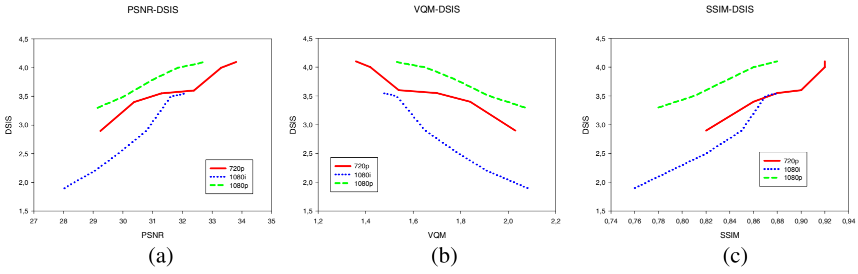
x_i and y_i are sample values, $\bar{x}$ and $\bar{y}$ sample mean, s_x and s_y standard deviation (calculated using $n - 1$ in the denominator) and r_{xy} is calculated Pearson product-moment correlation coefficient. Pearson's correlation reflects the degree of linear relationship between two variables. The coefficient ranges from −1 to 1. A value



(a) (b)

(c) (d)

Fig. 7 Different video quality measures relative to bit rate: (a) PSNR measure, (b) VQM measure, (c) SSIM measure, (d) DSIS measure



(a) (b) (c)

Fig. 8 Different video quality measures relative to DSIS (for the same bit rate when comparing quality measure and DSIS): (a) PSNR measure, (b) VQM measure, (c) SSIM measure, (d) DSIS measure

of 1 shows that a linear equation describes the relationship perfectly and positively, with all data points lying on the same line and with *Y* increasing with *X*. A score of −1 shows that all data points lie on a single line but that *Y* increases

Table 10 Correlation of objective quality measures with subjective quality measure, *DSIS*

	720/50p	*1080/25i*	*1080/50p*
PSNR-DSIS	0.971	0.990	0.992
VQM-DSIS	-0.972	-0.986	-0.993
SSIM-DSIS	0.975	0.981	0.996

as *X* decreases. A value of 0 shows that a linear model is inappropriate – that there is no linear relationship between the variables.

From the Table 10 it can be seen that all objective quality measures (*PSNR*, *VQM* and *SSIM*) give similar correlation with *DSIS*, when resolution is fixed, which means that tested objective quality measures are equally good for testing video quality of the same resolution format.

7 Conclusion

Television is playing a crucial role in the world of modern communications systems, global connectivity and constant availability of all kind of information and services. Yet, its objective to entertain, educate and inform will definitely benefit from implementation of high definition television systems capable to deliver realistic representation of a surrounding world. On this course there are certain obstacles, already described in this chapter, which we will have to confront to.

Transition toward high definition environment did not have the same starting position around the world. This asynchronous development and implementation produced different rules, guidelines and different standards for high definition production and broadcast applications. Nevertheless, with digital technology, differences seem easier to overcome. The important question is still here; how to ensure seamless transition, interoperability and exchange between different systems; what high definition format is the best and most appropriate?

When selecting the high definition format that will be used in certain system a numerous factors have to be considered. The most important task before system design and implementation is to select video format that will enable a preservation of video information throughout complete production, post-production and broadcast chain. The goal is to deliver the information that will induce the highest level of viewer's visual sensation. As we said numerous factors need to be consider and those are: appropriate production and acquisition format enabling easy handling, recording, exchange; spectrum limitations and transmission system capacity needs to be consider as well; storage limitations; and the last in the chain, flat panel displays with their characteristics enabling reproduction of the visual information as good as the original is.

The format that can provide aforementioned is for sure progressively scanned format with 1080 lines and 1920 samples per line. Already standardized, it represents format which can unified the television world of future. Nevertheless, although convenient for production systems due to high spatial resolution, it implies huge amount of data and demands higher channel capacities, which at the end requires new and more efficient compression techniques.

References

Baron, S., Wood, D.: Rec.601. – the origins of the 4:2:2 DTV standard. EBU Tech. R (2005), http://www.uer.biz/en/technical/trev/trev_304-rec601_wood.pdf (accessed January 5, 2009)

Bjontegaard, G.: Calculation of average PSNR differences between RD-curves. Doc.VCEG-M33, ITU-T Q6/16, Austin, TX, USA (2001), http://wftp3.itu.int/av-arch/video-site/0104_Aus/VCEG-M33.doc (accessed March 15, 2008)

Crowdrun source material, http://vqeg.its.bldrdoc.gov/HDTV/SVT_MultiFormat/

Darmon, C.: The 16:9 format – a technical and artistic challenge. EBU,Geneva (1997)

Deame, J.: Format and Standards Conversion. In: Williams, E.A. (ed.) NAB Engineering Handbook, 10th edn. Focal Press, Burlington (2007)

Drewery, J., Salmon, R.: Tests of visual acuity to determine the resolution required of a television transmission system. BBC Research & Development Tech. Rep., WHP 092 (2004), http://www.bbc.co.uk/rd/pubs/whp/whp-pdf-files/WHP092.pdf (accessed January 12, 2009)

Cugnini, A., Wood, D., Asami, H.: Worldvide Standards for Digital Television. In: Williams, E.A. (ed.) NAB Engineering Handbook, 10th edn. Focal Press, Burlington (2007)

EBU-I34. The Potential Impact of Flat Panel Displays on Broadcast Delivery of Television, Geneva (2002), http://www.ebu.ch/CMSimages/en/tec_text_i34-2002_tcm6-4783.pdf (accessed January 16, 2009)

EBU-I35. Further Consideration on the Impact of Flat Panel Home Displays on the Broadcasting Chain, Geneva (2003), http://www.ebu.ch/CMSimages/en/tec_text_i35-2003_tcm6-10879.pdf (accessed January 16, 2009)

EBU Recommendatioin R28. Subjective assessment method to be normally used for 625-line television pictures. EBU, Geneva (1997)

EBU Recommendation R124. Choice of HDTV Compression Algorithm and Bitrate for Acquisition, Production & Distribution. EBU, Geneva (2008)

EBU Tech. 3267. EBU Interfaces for 625–line Digital Video Signals at the 4:2:2 level of CCIR Recommendation. EBU, Geneva (1992)

EBU Tech. 3312. Digital Terrestrial HDTV Broadcasting in Europe. EBU, Geneva (2006)

EBU Tech. 3315. Archiving:Experiences with telecine transfer of film to digital formats. EBU, Geneva (2006)

Ellis, R.J.G.: The PALplus Story. Architects' Publishing Partnership Ltd., Manchester (1997)

Fox, B.: The digital dawn in Europe [HDTV]. IEEE Sp. 32(4), 50–53 (1995)

Grgic, S., Grgic, M., Mrak, M.: Reliability of Objective Picture Quality Measures. Journal of Elect. Engineering 55(1-2), 3–10 (2004)

Haglund, L.: Overall-Quality Assessment When Targeting Wide-XGA Flat Panel Displays, SVT/IRT, Stockholm/Munich (2002), https://up.ebu.ch/CMSimages/en/tec_svt_widexga_final_tcm6-44922.pdf (accessed July 21, 2007)

Haglund, L.: The SVT High Definition Multi Format Test Set. SVT. Stockholm (2006), https://up.ebu.ch/CMSimages/en/tec_svt_multiformat_v10_tcm6-43174.pdf (accessed January 12, 2007)

Hatori, M., Nakamura, Y.: 1125/60 HDTV Studio Standard Intended To Be A World-wide Unified Standard. IEEE Transactions on Broadcasting 35(3), 270–278 (1989)

Hofman, G.: Color Space (2006), `http://www.fhoemden.de/~hoffmann/ciexyz29082000.pdf` (acessed December 22, 2008)

Hoffmann, H., Itagaki, T., Wood, D., Bock, A.: Studies on the Bit Rate Requirements for a HDTV Format With 1920 x 1080 pixel Resolution, Progressive Scanning at 50 Hz Frame Rate Targeting Flat Panel Displays. IEEE Transactions on Broadcasting 52(4), 417–443 (2006)

Hoffmann, H., Itagaki, T., Wood, D., Hinz, T., Wiegand, T.: A Novel Method for Subjective Picture Quality Assessment and Further Studies of HDTV Formats. IEEE Transactions on Broadcasting 54(1), 1–13 (2008)

ISO/IEC Recommendation. Information technology - Generic coding of moving pictures and associated audio information: Video. DIS 13818-2. ITU, Geneva (1998)

Ive, J.: Image formats for HDTV. EBU Technical Review. Geneve, 33–39 (2004)

ITU-R Recommendation BT.654. Subjective quality of television pictures in relation to the main impairments of the analogue composite television signal. ITU, Geneva (1986)

ITU-R BT.801-4. The Present State of High–Definition Television. International Telecommunication Union Tech. Rep. ITU-R Report 801-4. Geneva (1990)

ITU-R Recommendation BT.811. The subjective assessment of enhanced PAL and SECAM systems. ITU, Geneva (1994)

ITU Recommendation ITU-R BT.601. Studio encoding parameters of digital television for standard 4:3 and wide-screen 16:9 aspect ratios. ITU-R, Geneva (1995)

ITU-R Recommendation BT.1128. Subjective assessment of conventional television systems. ITU, Geneva (1997)

ITU-R Recommendation BT.1129. Subjective assessment of standard definition digital television (SDTV) systems. ITU, Geneva (1998)

ITU Recommendation BT.656. Interfaces for digital component video signals in 525 - line and 625 - line television systems operating at the 4:2:2 level. ITU, Geneva (1998)

ITU Recommendation BT.1361. Worldwide unified colorimetry and related characteristics of future television and imaging systems. ITU, Geneva (1998)

ITU-T Recommendation ITU-T P.910. Subjective video quality assessment methods for multimedia applications. ITU, Geneva (1999)

ITU Recommendation BT.500-11. Methodology for the subjective assessment of the quality of television pictures. ITU, Geneva (2002)

ITU-R Recommendation BT.709-5. Parameter Values for the HDTV Standards for Production and International Programme Exchange. Geneva (2002)

ITU Recommendation BT.1683. Objective perceptual video quality measurement techniques for standard definition digital broadcast television in the presence of a full reference. ITU, Geneva (2004)

ITU Recommendation. Advanced Video Coding for Generic Audiovisual Services. H.264 and ISO/IEC 14496-10 (MPEG4-AVC). ITU, Geneva (2005)

ITU Recommendation BR.1375. High-definition television (HDTV) digital re-cording formats. ITU, Geneva (2007)

ITU Recommendation BT.799. Interface for digital component video signals in 525-line and 625 line television systems operating at the 4:4:4 level of Recommendation ITU-R BT.601. ITU, Geneva (2007)

ITU Recommendation BT.1845. Guidelines on metrics to be used when tailoring television programmes to broadcasting applications at various image quality levels and sizes. ITU, Geneva (2008)

Jones, G.A., Defilippis, J.M., Hoffman, H., Williams, E.A.: Digital Television Station and Network Implementation. Proceeding of the IEEE 94(1), 22–36 (2006)

Jurgen, R.K.: Chasing Japan in the HDTV race. IEEE Sp. 26(10), 26–30 (1989)

Luther, A., Inglis, A.: Color Video Fundamentals. In: Poynton (ed.) A Video Engineering, 3rd edn. McGraw-Hill, USA (1999)

Ninomiya, J.: The Japanese Scene. IEEE Sp. 32(4), 54–57 (1995)

Ohno, Y.: CIE Fundamentals for Color Measurements. In: IS&T NIP16 Conference, Vancouver (2000), http://physics.nist.gov/Divisions/Div844/facilities/photo/Publications/OhnoNIP16-2000.pdf (acessed December 22, 2008)

Poynton, C.: Gamma. In: Poynton (ed.) A Technical Introduction to Digital Video, 1st edn. Wiley, New York (1996)

Poynton, C.: The magnitude of nonconstant luminance errors (1997), http://poynton.com/PDFs/Mag_of_nonconst_luminance.pdf (accessed January 10, 2009)

Poynton, C.: Digital Video and HDTV Interfaces. Wiley, New York (2007)

Richardson, M.: HD Basics and Beyond: A Primer for Video Professionals (2006)

Robin, M., Poulin, M.: Digital Television Fundamentals. McGraw Hill, New York (1998)

Russ, J.C.: The Image Processing Book. CRC Press, Boca Raton (2006)

sgi2yuv program, http://www.ldv.ei.tum.de/Members/tobias/videotools/sgi2yuv.zip/

SMPTE 296M. 1280 x 720 Progressive Image Sample Structure – Analogue and Digital Representation and Analogue Interface. Tech. Rep., New York (1998)

SMPTE 274M. 1920 x 1080 Image Sample Structure, Digital Representation and Digital Timing Reference Sequences for Multiple Picture Rates. Tech. Rep., New York (2001)

SMPTE 259M. Television - SDTV Digital Signal/Data - Serial Digital Interface. SMPTE, New York (2006)

Sugawara, M., Mitani, K., Kanazawa, M., Okano, F., Nishida, Y.: Future Prospects of HDTV – Technical Trends Toward, 1080 p. (2005), http://www.nhk.or.jp/digital/en/technical/pdf/02_1_1.pdf (accessed January 15, 2009)

Tanton, N.E.: Results of a survey on television viewing distance. BBC R&D Rep. WHP 090 (2004), http://www.bbc.co.uk/rd/pubs/whp/whp-pdf-files/WHP090.pdf (accessed January 12, 2009)

Tanton, N.E., Stone, M.A.: HDTV DISPLAYS: Subjective effects of scanning standards and domestic picture sizes. BBC Research Department Report (1989/1990)

Udelson, J.: The great Television Race: A History of the American Television Industry, Tuscaloosa, 1925–1941 (1982)

Visca, M., Hoffman, H.: HDTV production codec tests. EBU Technical Review (2008), https://www.ebu.ch/en/technical/trev/trev_2008-Q3_HD-Prod-Codecs.pdf (accessed January 14, 2009)

Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image Quality Assessment: From Error Visibility to Structural Similarity. IEEE Trans. on Image Proc. 13(4), 600–612 (2004)

Wang, Z., Sheikh, H.R., Bovik, A.C.: Objective Video Quality Assessment. In: Furth, Marqes (eds.) The Handbook of Video Databases: Design and Applications, 1st edn. CRC Press, Boca Raton (2003)

Watkinson, J.: The Art of Digital Video. Focal Press, Oxford (2000)

Webster, A.: Objective perceptual. Assessment of video quality: Full reference television. ITU, Geneva (2004), http://www.itu.int/ITU-T/studygroups/com09/docs/tutorial_opavc.pdf (accessed December 5, 2008)

Wood, D.: High Definition for Europe – a progressive approach. EBU Tech. Rew., 24–32 (2004)

Wood, D.: Influential ITU standards marks 25th anniversary – Recommendation 601 drives digital television world wide. ITU, Geneva (2007), http://www.itu.int/itunews/manager/display_pdf.asp?lang=en&year=2007&issue=03 (accessed January 5, 2009)

Wood, D.: The Development of HDTV in Europe – a tale of three cities: Dublin, Dubrovnik and Geneva. EBU Tech. R., Geneva (2007), http://www.ebu.ch/CMSimages/en/311-wood_hdtv_tcm6-52695.pdf (accessed September 14, 2009)

Xiao, F.: DCT-based Video Quality Evaluation. Final Project for EE392J (Winter 2000)

Encryption Based Robust Watermarking in Fractional Wavelet Domain

Gaurav Bhatnagar and Balasubramanian Raman

Abstract. In this chapter, a robust watermarking technique based on encryption in fractional wavelet Domain is presented to improve the protection and authentication of the images. The core idea of the proposed watermarking scheme is to encrypt an image via fractional wavelet transform and then watermark is embedded in encrypted image by modifying the singular values. After embedding, watermarked encrypted image is decrypted with the help fractional wavelet transform to get the watermarked image. First, watermarked image is encrypted by same algorithm at the receiver's end and then watermark is extracted by proposed watermark extraction algorithm. The feasibility of this method and its robustness against different kind of attacks are verified by computer simulations.

1 Introduction

The success and substantial proliferation of the web technologies have created an environment in which some of very crucial issues for digital media such as illegal copying, distribution, editing, copyright protection, authentication etc become very easy. This has led to an increasing need for developing some standard solution to prevent these issues. As the possible solution, Cryptography (or encryption techniques) and steganography come to our help. However these terminologies have their own limitations. Cryptography is the art of secret (crypto) writing (graphy). It is the science of using mathematics to encrypt and decrypt data. Cryptography enables owner to store important information or transmit it across insecure channel so that it cannot be read by anyone except the authorized recipient. Cryptography is an effective solution to the distribution issue. It provides end-to-end security while distributing digital media over a large variety of distributions systems. It is mainly concerned with the secured communication instead of ulterior copyright infraction. It is

Gaurav Bhatnagar and Balasubramanian Raman
Department of Mathematics,
Indian Institute of Technology Roorkee, Roorkee-247 667, India
e-mail: goravdma@gmail.com, balaiitr@ieee.org

M. Grgic et al. (Eds.): Rec. Advan. in Mult. Sig. Process. and Commun., SCI 231, pp. 375–416.
springerlink.com

about protecting the content of the message. As per their existence is concerned, a new terminology named Steganography comes into picture. Steganography is the art of covered (secret) writing (graphy). It typically relates only to covert point to point communication between two parties. Thus, steganographic methods are either not robust or having limited robustness against modification of the digital data, occurred during transmission, storage or file conversion. Digital Watermarking terminology is developed and finding more and more support by the research community as a possible solution to overcome robustness problem. Thus, rather than steganography, watermarking is used wherever the protection of media is concerned. The basic idea behind watermarking is to insert an information (the watermark) into a digital media, which can be later extracted or detected for variety of purposes including identification and authentication purposes. The watermark is called metadata and digital media to be protected is called cover-data. The embedding is done in such a way that it must not cause serious degradation to the original digital media. Cryptography and Watermarking are complementing each other and a complete digital media security depends on both.

2 Encryption Problem Formulation

Cryptography or Encryption techniques are the sciences of using mathematics to encrypt and decrypt data. It allows two people, commonly known as Alice(sender) and Bob(receiver), to communicate with each other securely. The core idea of encryption is to modify the message in such a way that its content can be reconstructed only by a legal recipient. The message is known as the *plaintext* or *cleartext*. The method of disguising the plaintext in such a way as to hide its content is called *encryption* and this encrypted message is called *ciphertext*. The process of getting original message back from the ciphertext is called *decryption*. The process is shown in the figure 1.

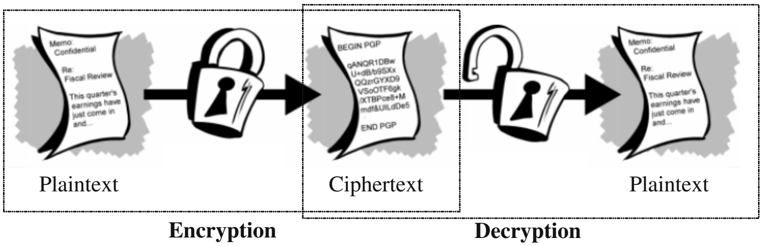


Fig. 1 General Encryption and Decryption System

Encryption techniques have a long and fascinating history. The most complete non-technical history of the encryption techniques can be found in the book *The Codebreakers* by David Kahn. This book points out the history of encryptions techniques from the time of Egyptians to the 20^{th} century where it played a crucial role

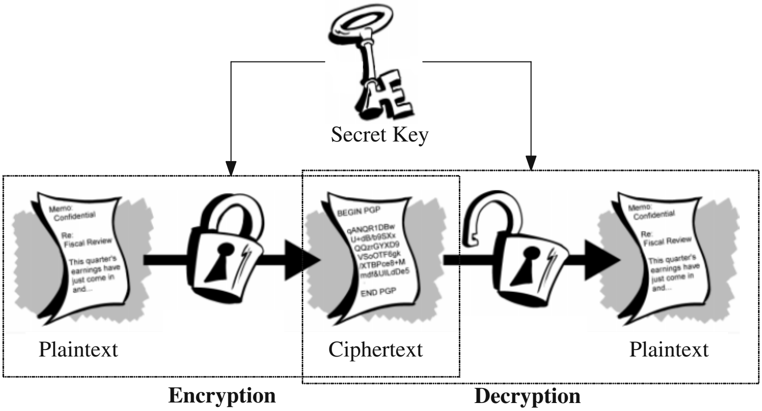


Fig. 2 Secret-key Encryption and Decryption System

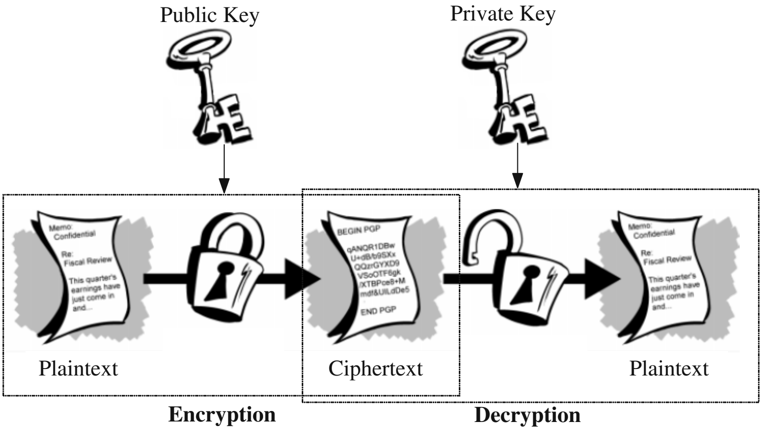


Fig. 3 Public-key Encryption and Decryption System

in the outcome of both world wars. Kahn has written it first in 1963 and has covered all those aspects which were most significant to the development of the subject. In literature two types of encryption techniques exist: *secret-key* (or symmetric key) and *public key* (or asymmetric key) encryption. Figures 2 and 3 show an illustration of the secret-key and public key encryption respectively. In secret-key encryption system one key is used to encrypt and decrypt the data while in public key encryption system uses a pair of keys. One key is publicly available and is used to encrypt the message. Second key is secret and also known as private key and is used for decryption. The benefit of public key encryption is that anyone who has a public key can encrypt the data but cannot decrypt it. Only the authorized person (one who has the private key) can decrypt the message.

The most famous and the simplest example of encryption is Caesar's Cipher. Julius Caesar wants to convey a message to his general but he didn't trust his messengers. So he shifted every alphabet in his message by 3. Only someone who knows the Caesar's key *shift by 3* can decrypt the message and read it. For example, if we want to encrypt the word "GAURAV" using Caesar's key then we have to slide every alphabet up by 3. Hence, starting with

Original: ABCDEFGHIJKLMNOPQRSTUVWXYZ

and sliding every alphabet up by 3, we get

Encrypted: DEFGHIJKLMNOPQRSTUVWXYZABC

Finally, after encryption the word *GAURAV* become *JDXUDY*. Mathematically, the encryption system is defined as:

$$M_c = Encryp(M, K_e) \tag{1}$$

$$M_d = Decryp(M_c, K_d) \tag{2}$$

where M, K_e and K_d are the original message, public and private key respectively. *Encryp* and *Decryp* are the encryption and decryption functions that give encrypted (M_c) and decrypted (M_d) message. Hence, encryption techniques try to protect the content of a message.

In literature, a lot of different encryption algorithms have been proposed and used frequently and widely, such as AES, RSA or IDEA[1, 2, 3, 4]. Most of encryption algorithms are using text or binary data. Generally, it is very difficult to use existing algorithm directly to multimedia data since multimedia data are having high redundancy, large volumes and requiring real-time operations namely displaying, cropping, copying, conversion etc. Hence traditional encryption algorithms are not suitable for multimedia data in the real time applications. As a result, there is a strong need of developing some new efficient multimedia encryption algorithms suitable for real time applications.

As from previous discussion, it is clear that the requirements for the multimedia encryption algorithm[5, 6, 7] is different from traditional encryption algorithm. Unlike traditional encryption algorithm, multimedia encryption requires cryptographic security along with the perceptual security. The meaning of perceptual security is that after seeing the encrypted multimedia, one cannot judge the overview of the original multimedia. For example, figures 4(a,d) show the synthetic and Trui images. The encrypted versions using raster and ZIG-ZAG scan are shown in figures 4(b,e) and 4(c,f) respectively. As can be seen, figures 4(b,e) is give the exact overview whereas from figures 4(c,f) no one can get the overview of original media. This is due to the close relation between the adjacent pixels in an image, which cannot be removed by raster scan while ZIG-ZAG scan removed this relation successfully. Both the requirements and their evaluation process are discussed below.

1. *Cryptographic Security*: Cryptographic security can be viewed as the ability of the encryption algorithm to resist the cryptanalysis process. Cryptanalysis is the

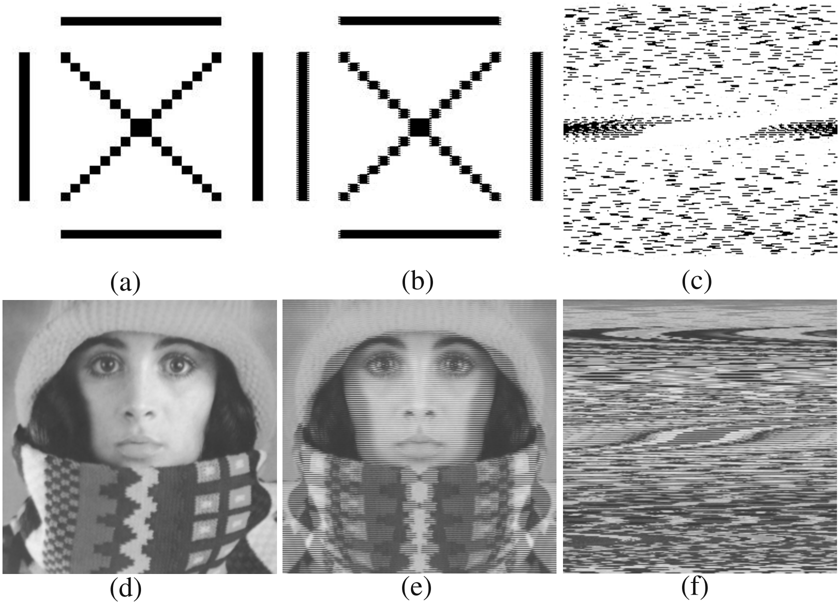


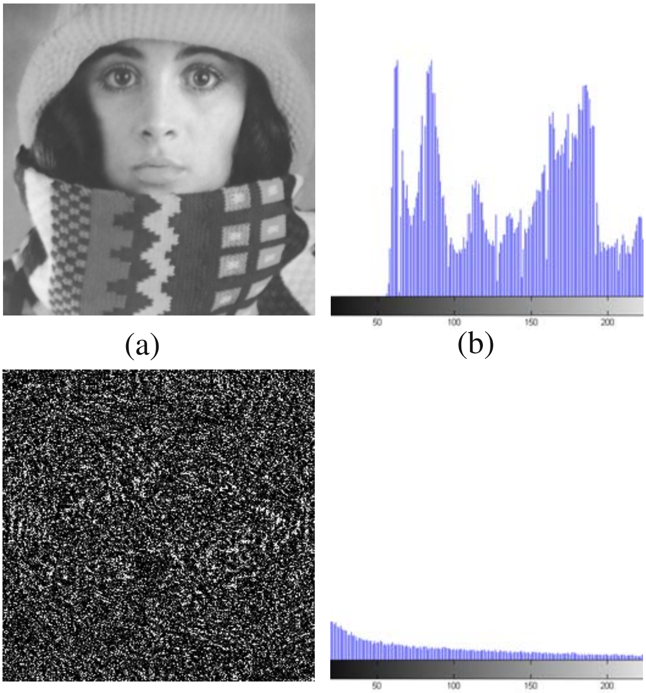
Fig. 4 a,d) Original Images b,e) Encrypted Images using Raster scan c,f) Encrypted Images using ZIG-ZAG scan

science of analyzing and breaking the secure channel. Classical cryptanalysis includes a variety of attacks such as analytical reasoning, mathematical analysis, related-key attack, statistical attack, pattern finding and finally patience, determination and luck. If the encryption algorithm resists against most of the attacks then that the encryption algorithm is said to be of robust against cryptographic attacks.

2. *Perceptual Security*: As it has been pointed out that the perceptual security is to produce that kind of encrypted multimedia which cannot give the overview of the original multimedia. In order to evaluate perceptual security, some subjective and objective metrics are used.

   a. *Subjective Metric*: While giving the communication theory of secrecy system [8], Shannon said “It is possible to evaluate most of the encryption algorithm by statistical analysis” and therefore he suggested two methods based on the histogram and on the correlations of adjacent pixels in the cipher image.

      - *Evaluation via Histograms*: The basic idea is to compare the histograms of the original and encrypted media. It can be seen that the histogram of the encrypted media is fairly uniform and is significantly different from the histogram of original one. One typical example is shown in figure 5.



(a) (b)

(c) (d)

Fig. 5 Histograms of the original and encrypted image

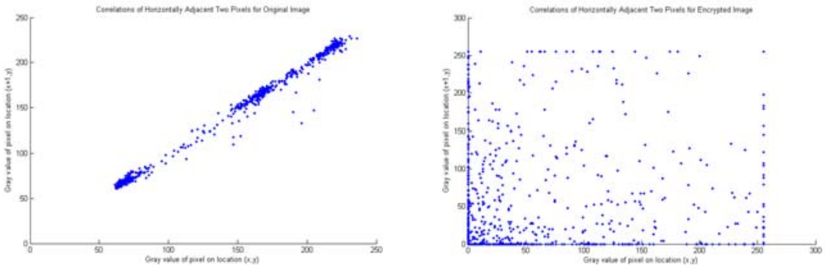
- *Evaluation via Correlations of Adjacent Pixels*: To test the correlation between two adjacent pixels, three ways are there, either take two vertically adjacent pixels or take two horizontally adjacent pixels or take two diagonally adjacent pixels in the encrypted image. First, randomly select P pairs of adjacent pixels and then calculate their correlation coefficient as:

$$cov(x,y) = E(x - E(x))(y - E(y)) \tag{3}$$

$$r_{xy} = \frac{cov(x,y)}{\sqrt{var(x)}\,\sqrt{var(y)}} \tag{4}$$

  where x and y are gray levels of two adjacent pixels in the image. Figure 6 shows the correlations of two horizontally adjacent pixels in the original and encrypted image. The correlation coefficients for the horizontally, diagonal and vertical directions are given in table 1.

b. *Objective Metric*: A metric which provide more efficient test methods and is suitable for computer simulations are called objective metrics. Unfortunately, there is no suitable metric for multimedia content has been explored, except some quality metrics for multimedia. Thus, the quality metrics can be used



(a)

(b)

Fig. 6 The correlations of two horizontally adjacent pixels in a) Original (figure 5(a)) b) Encrypted (figure 5(c)) Image

Table 1 Correlation coefficients of adjacent pixels in Original (figure 5(a)) and Encrypted (figure 5(c)) Images

Adjacent Pixels	Original Image	Encrypted Image
Horizontal	0.9965	0.0266
Vertical	0.9638	0.0171
Diagonal	0.9487	-0.0313

as the objective metric of encrypted media. One of the typical and frequently used quality metric for image quality is signal-to-noise ratio (SNR) and peak signal-to-noise ratio (PSNR). Between these, PSNR is of greater worth and it is defined as:

$$PSNR = 10\ log_{10} \frac{f_{max}^2}{RMSE^2} \tag{5}$$

$$RMSE = \sqrt{\frac{1}{MN} \sum_{x,y} [f(x,y) - f^e(x,y)]^2} \tag{6}$$

where $f(x,y)$ and $f^e(x,y)$ are the original and encrypted media. Generally, the higher the PSNR is, the higher the encrypted-image quality is. Hence, for a good image encryption algorithm, the encrypted-images PSNR should be small enough.

3 Watermarking Problem Formulation

The first model for secret communication was proposed by Simmons[9] in 1984, which is very well-known *The Prisoner's Problem*. According to this problem, two friends Alice and Bob are arrested for some crime and are confined in two different cells of a prison. Both of them want to escape from the prison but unfortunately they want to communicate with each other through a warden named Wendy. However they will not be able to communicate with each-other through encryption because if Wendy notices any suspicious message then she suppresses their communication completely. Finally, both of them decide to communicate invisibly so that they will not come in Wendy's suspicious confine and hence they have to setup a subliminal channel. The most practical way to do this is to hide meaningful information or message in some harmless message. Alice has created a picture of cow lying in a green meadow and send this picture to Bob. She transmits meaningful information via colors of the objects in the picture and Wendy has no idea about it. Unfortunately there are some other problems which may occur during their secret communication like Wendy could change the colors of the objects and destroy all the information or she could destroy original message and send a new message to one of the prisoners via subliminal channel pretending to be the other. This type of model is applicable to many real life situations where invisible communication or (in other words) *steganography* is needed. Generally, Alice and Bob represent two communication parties whereas Wendy represents intruder/attacker who is able to read and probably alter the messages (figure 7). Unlike encryption techniques, steganography goes a bit further. It tries to conceal the existence of the messages. However as far as the robustness against malicious activities of intruder/attacker is concerned *Digital Watermarking* is used rather than steganography, since watermarking is resilience against malicious activities. Watermarking is closely related to steganography but having different philosophies, requirements and applications.

A general watermarking system or technique[10, 11, 12, 13, 14, 15, 16] consists of two main components: 1) Watermark Embedder 2) Watermark Detector or Extractor (shown in figure 8). The watermark embedder combines the cover-data (C_0) with the metadata (W) and creates the watermarked cover (C_W). This embedding operation takes place in two steps or phases. In the first phase, the original metadata (W) is mapped into another form, called reference metadata ($\widehat{W}$). This mapping is done with the help of a watermark key K, which can be used to

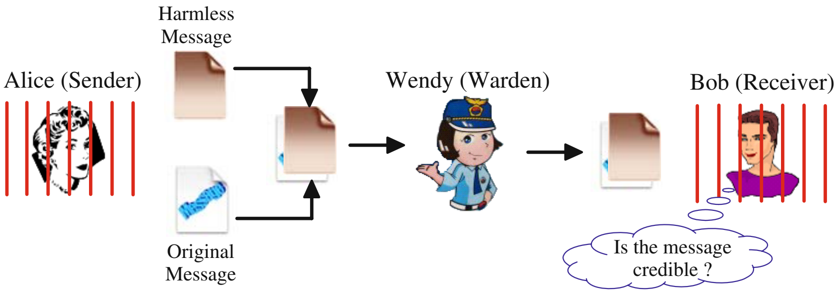
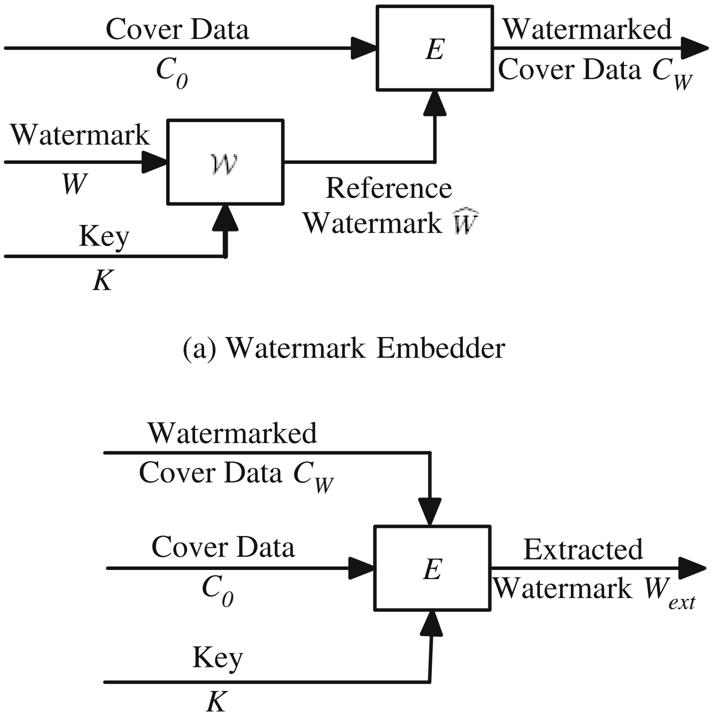


Fig. 7 Illustration of The Prisoner's Problem



(a) Watermark Embedder

(b) Watermark Decorder/Extracter

Fig. 8 General Watermarking System

enforce security. This used key K could be either public or private or combination of these two. In the second phase, $\widehat{W}$ is added to C_0, to produce the watermarked cover C_W. Watermark detector/extractor either extracts the metadata (W) from the watermarked cover C_W, or it produces a estimate of W. The detection/extraction of the metadata is done with help of the same key K, which is used in embedding. Mathematically, the embedding and detecting/extracting process are defined as:

$$C_W = E(C_0, \widehat{W}) \tag{7}$$

$$W_{ext} = D(C_W, C_0, K) \tag{8}$$

where $\widehat{W} = \mathscr{W}(W, K)$, here $\mathscr{W}$ is a function that maps W into $\widehat{W}$ with the help of K. E is the encoder function that takes a cover-data and a watermark, and it generates a watermarked cover-data. Similarly D is the decoder function that takes watermarked cover, original cover-data and a key K to generate the estimate of W. The extracted metadata W_{ext} will then be compared with the owner's metadata W, considering a similarity function $\mathscr{S} = sim(W, W_{ext})$ and is defined as:

$$\mathscr{S} = sim(W, W_{ext}) \begin{cases} = 1, \text{ if } & W = W_{ext} \\ < 1, \text{ if } & W \neq W_{ext} \end{cases} \tag{9}$$

Depending on the requirements and purposes, watermarking schemes are usually characterized by four constraints:

1. *Robustness:* Robustness refers to the ability of the watermark to be preserved even after distortions introduced by standard or malicious data processing, which may be either intensionally or un-intensionally. These distortions are also known as watermarking attacks.
2. *Imperceptibility:* The imperceptibility of the watermark refers to its perceptual transparency. In other words, the human eye should not be able to detect differences between the watermarked and original media.
3. *Capacity:* Capacity refers to the maximum amount of information that can be hidden in the media. This directly affects the robustness and the perceptual transparency.
4. *Security:* Security refers to the fact that un-authorized persons should neither detect nor read the watermark. The best way to improve security is the selection of secret key and those who knows the secret key can detect or read the watermark.

These four constraints are usually opposing each other and hence an application dependent trade-off must be found. A possible tradeoff is given in the figure 9. Generally, the imperceptibility is the most important constraint because the artifacts introduced while watermarking process are not only annoying and un-desirable but also reduce or destroy its commercial value. Unfortunately, improving robustness often implies perceptual degradations, improving imperceptibility leads to less

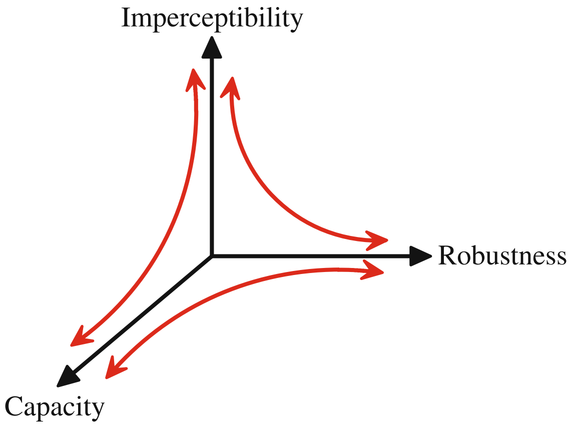


Fig. 9 Tradeoff between these constraints

Table 2 Metrics for Evaluating Imperceptibility

Maximum Difference	$MD = max\lvert f(x,y) - g(x,y)\rvert, \forall x,y$
Average Absolute Difference	$AAD = \frac{1}{MN}\sum_{x,y}\lvert f(x,y) - g(x,y)\rvert$
Norm. Average Absolute Difference	$NAAD = \frac{\sum_{x,y}\lvert f(x,y)-g(x,y)\rvert}{\sum_{x,y}\lvert f(x,y)\rvert}$
Mean Square Error	$MSE = \frac{1}{MN}\sum_{x,y}[f(x,y) - g(x,y)]^2$
Normalized Mean Square Error	$NMSE = \frac{\sum_{x,y}[f(x,y)-g(x,y)]^2}{\sum_{x,y}f^2(x,y)}$
Root Mean Square Error	$RMSE = \sqrt{\frac{1}{MN}\sum_{x,y}[f(x,y) - g(x,y)]^2}$
*Laplacian Mean Square Error	$LMSE = \frac{\sum_{x,y}\left(\nabla^2 f(x,y) - \nabla^2 g(x,y)\right)^2}{\sum_{x,y}\left(\nabla^2 f(x,y)\right)^2}$
L^p-Norm	$L^p = \left(\frac{1}{MN}\sum_{x,y}\lvert f(x,y) - g(x,y)\rvert^p\right)^{1/p}$
Signal to Noise Ratio	$SNR = -10\ log_{10}(NMSE) = -10\ log_{10}\left(\frac{\sum_{x,y}[f(x,y)-g(x,y)]^2}{\sum_{x,y}f^2(x,y)}\right)$
Peak Signal to Noise Ratio	$PSNR = 10\ log_{10}\frac{f_{max}^2}{RMSE^2} = 10\ log_{10}\frac{f_{max}^2}{\frac{1}{MN}\sum_{x,y}[f(x,y)-g(x,y)]^2}$
Image Fidelity	$IF = 1 - NMSE = 1 - \frac{\sum_{x,y}[f(x,y)-g(x,y)]^2}{\sum_{x,y}f^2(x,y)}$

$f(x,y)$ and $g(x,y)$ represent a pixel, whose coordinates are (x,y), in the original and watermarked image. * $\nabla^2 f(x,y) = f(x+1,y) + f(x-1,y) + f(x,y+1) + f(x,y-1) - 4f(x,y)$

robustness, hiding more information (high capacity) leads to less robustness and conversely. There is no universal metric for evaluating imperceptibility, the most popular metrices used as imperceptibility measure are given in table 2. Among these metrices, the most popular imperceptibility metric used in image and video processing is the Peak Signal to Noise Ratio (*PSNR*).

Finally, the detection and the extraction of the watermark do not have the same meaning. The detection of a watermark only intends to check whether the media is watermarked or not. The extraction of the watermark stands for obtaining the watermark pattern, which is embedded in the media with a given secret or public key. The extraction is also classified into three categories:

1. *Informed or Non-Blind Extraction: Non-Blind Extraction:* If the original media is needed to extract the watermark or check whether the suspect media is watermarked or not. The knowledge of a key and of the embedding algorithm are not sufficient at the extraction side.
2. *Semi-Informed Extraction:* In this case, only some information related to original data is needed rather than complete data. It is very close to the blind extraction.
3. *Blind Extraction:* In this case, only the key and the knowledge of the algorithm are needed to extract the watermark without any original data.

Now, to verify the presence of watermark, different measures can be used to show the similarity between the original and the extracted watermark. Some of the most popular similarity measure are given in the table 3. Among these similarity

Table 3 Similarity Measure

Normalized Cross-Correlation	$NC = \frac{\sum_{x,y} w(x,y) \cdot \widetilde{w}(x,y)}{\sum_{x,y} w^2(x,y)}$
Correlation Quality	$CQ = \frac{\sum_{x,y} w(x,y) \cdot \widetilde{w}(x,y)}{\sum_{x,y} w(x,y)}$
Zero Mean Cross-Correlation 1	$ZMCC_1 = \frac{\sum_{x,y} w(x,y) \cdot \widetilde{w}(x,y)}{\sqrt{\sum_{x,y} w^2(x,y)}\sqrt{\sum_{x,y} \widetilde{w}^2(x,y)}}$
Zero Mean Cross-Correlation 2	$ZMCC_2 = \frac{\sum_{x,y} [w(x,y) - mean(w)] \cdot [\widetilde{w}(x,y) - mean(\widetilde{w})]}{\sqrt{\sum_{x,y} [w(x,y) - mean(w)]^2(x,y)}\sqrt{\sum_{x,y} [\widetilde{w}(x,y) - mean(\widetilde{w})]^2(x,y)}}$
Histogram Similarity	$HS = \sum_{i=0}^{255} \lvert F_w(i) - F_{\widetilde{w}}(i) \rvert$, where $F_w(i)$ is the relative frequency of level i in a 256 level image.
Normalized Least Square Error	$NLSE = \frac{\sum_{x,y} \lvert w(x,y) - \widetilde{w}(x,y) \rvert^2}{\sum_{x,y} w^2(x,y)}$
Bit Error Rate	$BER = \frac{\sum_{x,y} p(x,y)}{m \times n}$, where $p(x,y) = \begin{cases} 1, \text{ if } w(x,y) \neq \widetilde{w}(x,y) \\ 0, \text{ if } w(x,y) = \widetilde{w}(x,y) \end{cases}$

$w(x,y)$ and $\widetilde{w}(x,y)$ represent a pixel, whose coordinates are (x,y), in the original and extracted watermark image.

measures, the most popular measure used in image and video processing is the Zero Mean Cross-Correlation ($ZMCC_2$).

4 Classification of Watermarking Techniques

Watermarking techniques can be classified into various categories in various ways. It can be categorized according to either working domain or type of document or human perception or according to applications. Among these, categorizations by working domain and human perception have attracted main interest of the research community. The complete classification of watermarking techniques is shown in the figure 10. Among spatial and frequency domain techniques, spatial domain techniques are less complex but not robust against various attacks. Frequency domain techniques are robust as compared to spatial domain techniques. This is due to the fact that when image is inverse transformed, watermark is distributed irregularly over the image, making the attacker difficult to read or modify. In visible watermarking, the embedded watermark appears visible to the viewers on a careful inspection. In invisible-robust watermarking, watermark is embedded in such a way that it cannot be perceptually noticed and it can be recovered only with appropriate decoding/extracting mechanism.

In invisible-fragile watermarking, watermark is embedded in such a way that any small or big manipulation or modification of the image would destroy the watermark. If watermark detection/extraction process requires original or reference image then it is called an invisible robust private watermarking otherwise it is called invisible robust public watermarking. The invisible robust watermarking schemes that can be

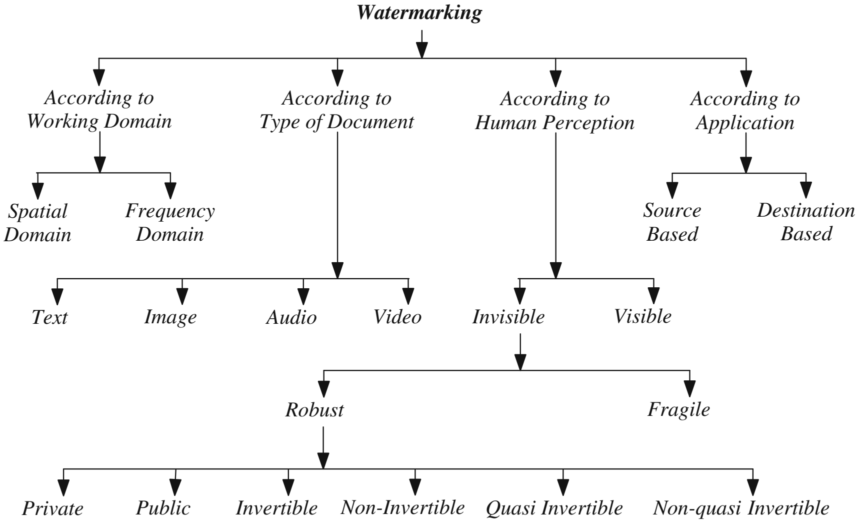
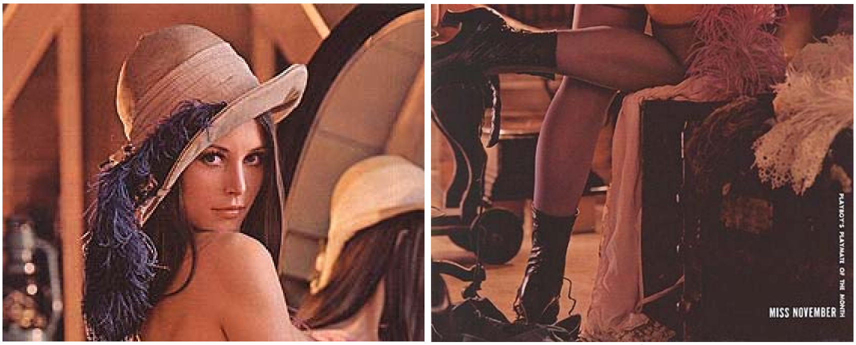


Fig. 10 Classification of Watermarking Techniques

attacked by creating a *counterfeit original image* are called invertible watermarking scheme. If a unique watermark identifies the owner from all the distributed copies of media then technique is called source based watermarking technique and is desirable for ownership identification or authentication. On the other hand, if each distributed copy has a unique watermark identifying the particular buyer then technique is called destination based watermarking technique. Generally, it is used to trace the buyer in the case of illegal reselling.

5 Applications of Watermarking Techniques

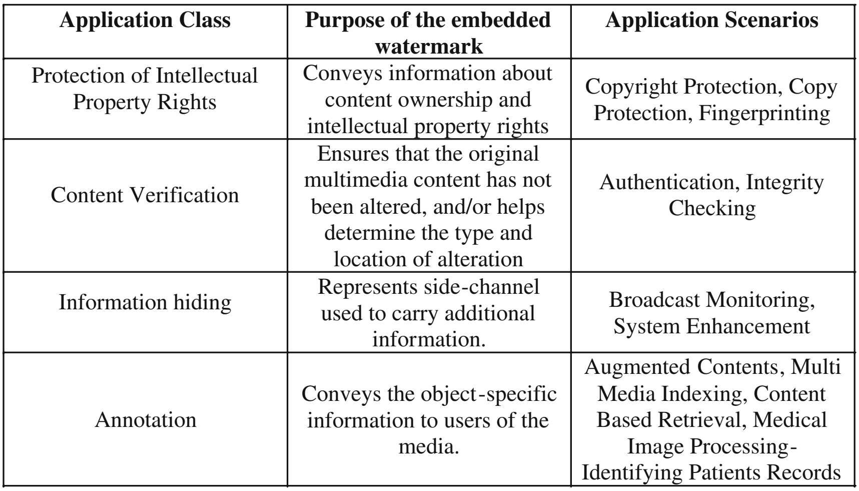
Before exploring the applications, let us explore a scenario, which leads to the birth of the watermarking. In November 1972, Playboy magazine published an image of Lena Soderberg. This image has the copyright notice on the edge of the image. After the image has been scanned for use as the test image by the image processing community, most of persons have been cropped the portion of the image on which copyright notice was printed (see figure 11). Later on Lena image became the most frequently used test image in image processing research, also appeared in a number of journals and books without any reference to its rightful owner, Playboy Enterprises, Inc. There exist a number of different scenarios (like above) in which watermarking plays a key role and according to these scenarios different applications are explored and also classified in a number of different ways. Based on the type of information conveyed by the watermark, applications are given in table 4. Each of the given application has its own specific requirements with respect to robustness, capacity, imperceptibility and security to fulfill the desired goal. It is consequently obvious that there exists no watermarking scheme which would be optimal for every application.



(a) (b)

Fig. 11 a) The Lena Image used as test image by research community b) The cropped part of the original Lena image having copyright notice

Table 4 Applications of Watermarking Techniques



Application Class	Purpose of the embedded watermark	Application Scenarios
Protection of Intellectual Property Rights	Conveys information about content ownership and intellectual property rights	Copyright Protection, Copy Protection, Fingerprinting
Content Verification	Ensures that the original multimedia content has not been altered, and/or helps determine the type and location of alteration	Authentication, Integrity Checking
Information hiding	Represents side-channel used to carry additional information.	Broadcast Monitoring, System Enhancement
Annotation	Conveys the object-specific information to users of the media.	Augmented Contents, Multi Media Indexing, Content Based Retrieval, Medical Image Processing-Identifying Patients Records

6 Attacks on Watermarking Techniques

A watermarking system is said to be robust if watermark survives after the distortions which are introduced by standard or malicious data processing. Mainly, all possible attacks are categorized into four category namely Active Attacks, Passive Attacks, Collusion attacks and Forgery attacks. Among these attacks, active attacks are of main interest since these are very critical for many applications. All these attacks are summarized in brief as follows:

1. *Active Attacks*: Here the hacker/intruder tries to remove the watermark or make it undetectable. This type of attack is critical for many applications, including owner identification, proof of ownership, fingerprinting and copy protection. Among this category a lot of attacks are there. A list of attacks in this category are summarized as follows:

   a. *Lossy Compression*: JPEG is currently one of the most widely used compression algorithms for images. Generally, JPEG compression process consist of two stages: quantization and entropy coding. Between these two stages the most of information loss occurs in quantization stage and hence watermark loss is also occur.
   b. *Geometric Transformations*: Geometric distortion affecting image and video includes flipping, rotation, spatial scaling, translation, skew or shear, perspective transformation, row column deletion and changes in aspect ratio.
   c. *Enhancement Attacks*: This includes low and high pass filtering, sharpening, contrast adjustment, histogram equalization, gamma correction, restoration and colour quantization.

   d. *Noise Addition*: Generally, addition of additive and uncorrelated multiplicative noise is responsible not only for the degradation and distortion in the image but also for degrading the watermark information.
   e. *Printing-scanning*: This process introduces geometrical as well as noise-like distortions.

2. *Passive Attacks*: In this case, the hacker/intruder is not trying to remove the watermark, but is trying to determine whether a watermark is present in the media or not. Most of the scenarios above are not concerned with this type of attack.
3. *Collusion attacks*: These are a special case of active attacks, in which the hacker/intruder tries to construct a copy of media with no watermark. For this they acquire several copies of media and embed different watermarks in those copies.
4. *Forgery attacks*: Here, the hacker/intruder tries to embed a valid watermark rather than removing the embedded one. These attacks are of main concerned because if hacker/intruder is able to embed valid watermarks then in extraction process fake watermark comes out to be extracted one.

7 Literature Survey for Digital Watermarking

The existing literature includes several taxonomies for digital watermarking. Among these, the most common taxonomies are embedding in spatial and frequency domains. Spatial domain methods[17, 18] are less complex and not robust against various attacks as no transform is used in them. The basic idea behind spatial domain methods is the modification of pixel intensities while embedding watermark. Transform domain methods are robust as compared to spatial domain methods. This is due to the fact that when image is inverse transformed, watermark is distributed irregularly over the image, making the attacker difficult to read or modify. The basic idea behind transform domain methods is to transform the media by the means of Fourier Transform(FT)[19], Discrete Cosine Transform(DCT)[20], Fractional Fourier Transform[21, 22, 23], Wavelet Transform[24, 25, 26, 27, 28, 29] etc. Then, the transform domain coefficients are altered to embed the watermark and finally inverse transform is applied to obtain the watermarked digital media.

Schyndel *et al.* [17] have proposed two methods in which first method is based on bit plane manipulation of the LSB whereas second method is based on the linear addition of the watermark to the image data, which is more difficult to decode, offering inherent security. Hwang *et al.* [18] have presented a watermarking scheme employed in spatial domain using hash functions. Cox *et al.* [19] have presented the most popular watermarking schemes based on the Spread Spectrum Communication. The watermark is embedded into the first *k* highest magnitude DFT/DCT coefficients of the image and extraction is done by comparing the DFT/DCT coefficients of the watermarked and the original image. Barni *et al.* [20] have proposed a watermarking algorithm, which operates in the frequency domain, embeds a pseudo-random sequence of real numbers in a selected set of DCT coefficients.

The watermark can be reliably extracted blindly by exploiting the statistical properties of the embedded sequence. Djurovic *et al.* [21] have proposed fractional Fourier transform based watermarking scheme for the multimedia copyright protection. After decomposing image via FRFT, transformation coefficients are reordering in non-increasing sequence and the watermark is embedded in the middle coefficients. Feng *et al.* [22] have proposed a blind watermarking algorithm in which multiple chirps are used as watermark and embedded in the spatial domain directly but detected in the FRFT domain. Yu *et al.* [23] have used the same logic proposed by Feng *et al.* [22], the only difference is that the embedding is done in FRFT domain where watermark position and the transform order are used as the encryption keys.

Xia *et al.* [24] have added a pseudo-random sequence to the largest coefficients of the detail bands where perceptual considerations are taken into account by setting the amount of modification proportional to the strength of the coefficient itself. Watermark detection is achieved through comparison with the original un-watermarked image. Barni *et al.*[25] proposed a method based on the characteristics of the human visual system operating in wavelet domain. Based on the texture and the luminance content of all image sub-bands, a mask is accomplished pixel by pixel. Kundur *et al.*[26] proposed the use of gray scale logo as watermark. They addressed a multiresolution fusion based watermarking method for embedding gray scale logos into wavelet transformed images via salience factor. Wang *et al.*[27] and Zhang *et al.*[28]proposed a new watermarking algorithm based on wavelet tree quantization. The detailed survey on wavelet based watermarking techniques can be found in [29].

Recently, a new transform, singular value decomposition (SVD)-based[30, 31, 32, 33] watermarking technique and its variants have been proposed. These approaches work on the simple concept of finding the SVD of a cover image or the SVD of each block of the cover image, and then modify the singular values to embed the watermark. Gorodestski*et al.* [30] have proposed a scheme in which the host image is first segmented into blocks of size 4×4 and the largest singular value of each block is quantized to embed one bit of data. Liu *et al.*[31] have proposed an algorithm based on SVD. In this algorithm, authors find the singular values of the host image and then modify it by adding the watermark. SVD transform is again applied on the resultant matrix to find the modified singular values. These singular values are combined with the known component to get the watermarked image. Inverse process is used for the extraction of watermark. Chandra *et al.*[32] have described a method for embedding singular values of the watermark into the singular values of entire image. Ganic *et al.*[33] have proposed an optimal watermarking scheme. In which the embedding strength factor depends on the host and watermark image singular values. Recently, some researcher's have presented hybrid watermarking schemes in which they have combined SVD with other existing transforms. SVD based scheme withstands a variety attacks but it is not resistant to geometric attacks like rotation, cropping etc. Hence, for improving the performance hybridization is needed. Ganic[34] have presented hybrid-watermarking scheme based on DWT and SVD. After decomposing the cover image into four bands, SVD is applied on each band, and modify the singular values of each band with the singular values of the visual watermark. Sverdlov[35] have used the same concept taking DCT and SVD. DCT coefficients are mapped into four quadrants via ZIG-ZAG scan and modify the singular values of each quadrant. Li *et al.*[36] have proposed the same hybrid DWT-SVD domain watermarking scheme by exploiting the properties of human visual system. Chang *et al.*[37] have proposed a new technique in which embedding is done in D and U components.

8 Fractional Fourier Transform (FRFT)

The Fourier transform (FT) is undoubtedly one of the most valuable and frequently used tools in signal processing and analysis. However, if the rotation is needed through an angle α, then Fourier transform is not applicable. To overcome this problem, concept of fractional Fourier transform (FRFT) was introduced by Victor Namias in 1980[38, 39]. The fractional Fourier transform[40] is the generalization of Fourier transform. The fractional Fourier transform is also called rotational Fourier transform or angular Fourier transform since it depends on a parameter α and can be interpreted as a rotation by an angle α in the time-frequency plane or decomposition of the signal in terms of chirps.

The 1D FRFT of a function $s(t)$ is defined as

$$F^{\alpha}[s(t)](x) = \int_{-\infty}^{\infty} s(t) K_{\alpha}(t,x)\, dt \tag{10}$$

where α is the transform order (or angle) and $K_{\alpha}(t,x)$ is the transform kernel and is given by:

$$K_{\alpha}(t,x) = \begin{cases} \sqrt{1 - icot\alpha} & \\ \quad e^{i\frac{t^2+x^2}{2} cot\alpha - ixt\ csc\alpha} & \alpha \neq n\pi \\ \delta(t-x), & \alpha = 2n\pi \\ \delta(t+x), & \alpha = 2n\pi \pm \pi \end{cases} \tag{11}$$

where n is a given integer. The FRFT of a signal exists under the same conditions in which its Fourier transform exists. The inverse FRFT can be visualized as the FRFT with transform order $-\alpha$. The main property of FRFT is that the signal obtained is in purely time domain if transform order (α) is 0 and in purely frequency domain if transform order (α) is $\pi/2$.

8.1 Properties of the Fractional Fourier Transform

Let F^{α} denote the FRFT operator with transform order α and s be the input signal. Under these notations, some of the important properties of the FRFT are summarized as follows:

1. *Identity Operator*: F^0 is the identity operator. The FRFT of order $\alpha = 0$ is the input signal itself. The FRFT of order $\alpha = 2\ \pi$ is also act as the identity operator because it can be viewed as the successive application of the ordinary Fourier transform 4 times. Mathematically,

$$F^0[s(t)] = F^{2\ \pi}[s(t)] = s(t) \tag{12}$$

2. *Fourier Transform Operator*: $F^{\pi/2}$ is the Fourier transform operator. The FRFT of order $\alpha = \pi/2$ gives the Fourier transform of the input signal.
3. *Successive applications of FRFT*: Successive applications of FRFT are equivalent to a single transform whose order is equal to the sum of the individual orders. Mathematically,

$$F^{\alpha}(F^{\beta}[s(t)]) = F^{\alpha+\beta}[s(t)] \tag{13}$$

4. *Inverse*: The inverse FRFT to reconstruct the original signal is the FRFT of order $-\alpha$, i.e.

$$F^{-\alpha}(F^{\alpha}[s(t)]) = F^{-\alpha+\alpha}[s(t)] = F^0[s(t)] = s(t) \tag{14}$$

8.2 The Discrete Fractional Fourier Transform

Discrete fractional Fourier transform[41] must obey the rotational properties as the continuous FRFT. These rotation properties can be easily realized by the power law of kernel matrix in discrete case. So, the fractional power of kernel matrix is required for computing the DFRFT. Dickinson *et al.* [42] introduced a commuting matrix S to compute the real eigenvectors of the DFT kernel matrix

$$F : S = \begin{pmatrix} 2 & 1 & 0 & 0 \cdots & 1 \\ 1 & 2\ \cos\omega & 1 & 0 \cdots & 0 \\ 0 & 1 & 2\ \cos 2\omega & 1\ \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots\ \ddots & \vdots \\ 1 & 0 & 0 & 0 \cdots & 2\ \cos(N-1)\omega \end{pmatrix} \tag{15}$$

where $\omega = 2\pi/N$ and N is the size of the DFT kernel matrix. Matrix S commutes with matrix F, and it satisfies the commutative property: FS=SF. The eigenvectors of matrix S are also the eigenvectors of matrix F, but their eigenvalues are distinct. Since the matrix S is real and symmetric, the eigenvalues of S are all real and the eigenvectors of S are orthonormal to each other. The transformation kernel of DFRFT can be easily defined by determining the fractional powers of the eigenvalues. The transform kernel of DFRFT is computed as:

$$K_{\alpha} = F^{2\alpha/\pi} = VD^{2\alpha/\pi}V^T \tag{16}$$

$$= \begin{cases} \sum_{k=1}^{N} \exp^{-ik\alpha} \upsilon_k \upsilon_k^T, & \text{for } N\text{=}4m\text{+}1,4m\text{+}3; \\ \sum_{k=1}^{N-1} \exp^{-ik\alpha} \upsilon_k \upsilon_k^T + \exp^{-iN\alpha} \upsilon_{N-1} \upsilon_{N-1}^T, & \text{for } N\text{=}4m,4m\text{+}2. \end{cases} \tag{17}$$

where υ_k is the k^{th} order DFT Hermite eigenvector, that is, $V = [\upsilon_1, \upsilon_1, \ldots, \upsilon_N]$. Matrix D is a diagonal matrix, in which the diagonal entries have the same eigenvalues corresponding to the column eigenvectors of matrix V in its diagonal entries.

The DFRFT of signal s can be computed as follows:

$$S_\alpha = K_\alpha s = F^{2\alpha/\pi} s = V D^{2\alpha/\pi} V^T s \tag{18}$$

Due to separability of the transform, two dimensional fractional Fourier transform can be obtained by successively taking one dimensional fractional Fourier transforms along both the axis (in both continuous and discrete case). The FRFT of some simple signals are given in the table 5.

Table 5 Fractional Fourier Transform of Some Simple Signals

Signal	FrFT with angle α	Condition
$\delta(t-\tau)$	$\sqrt{\frac{1-icot\alpha}{2\pi}} e^{i\frac{\tau^2+u^2}{2} cot\alpha - iu\tau \; csc\alpha}$	$\alpha \neq n\pi$
1	$\sqrt{1+i\tan\alpha} e^{-i\frac{u^2}{2}\tan\alpha}$	$\alpha - \pi/2 \neq n\pi$
$\exp i\nu t$	$\sqrt{1+i\tan\alpha} e^{-i\frac{\nu^2+u^2}{2}\cot\alpha + iu\nu \; \sec\alpha}$	$\alpha \neq n\pi$
$\exp ict^2/2$	$\sqrt{\frac{1+i\tan\alpha}{1+c\tan\alpha}} e^{i\frac{u^2}{2}(c-\tan\alpha)/(1+c\tan\alpha)}$	$\alpha - \arctan(c) - \pi/2 \neq n\pi$
$\exp -(t^2/2)$	$e^{-(u^2/2)}$	
$H_n(t)\exp -(t^2/2)$	$e^{-in\alpha} H_n(u) e^{-(u^2/2)}$	
$\exp -c(t^2/2)$	$\sqrt{\frac{1-i\cot\alpha}{c-i\cot\alpha}} e^{i\frac{u^2}{2}(c^2-1)\cot\alpha/(c^2+\cot^2\alpha)}$ $e^{-\frac{u^2}{2} \; c \; \csc^2\alpha/(c^2+\cot^2\alpha)}$	

9 Wavelet Transform

In wavelet analysis[43, 44, 45, 46, 47], an orthonormal set of functions (Schauder basis) as in the Fourier techniques, as well as a nonorthogonal, but linearly independent basis (Riesz basis) are employed. The collection of functions may not be linearly independent (frames). Wavelets can approximate discontinuous functions with a fewer number of functions than Fourier techniques. Wavelet Transform has a lot of benefits over sinusoids based transforms. By using wavelets, local properties are detected very easily. Wavelets have a special ability to analyze signal in both time and frequency domain simultaneously. Unlike sinusoids based transforms, wavelet can also used to analyze transient and time varying (non-stationary) signals. The edge effects are reproduced very accurately in wavelet technique as it uses discontinuous functions.

Wavelets are basically functions that are localized in frequency around a central value and that are limited in time (i.e., they are of finite support and hence localized in time around a central value). In this case, wavelets are different from the functions that are used in Fourier analysis in that they do not have a constant waveform (i.e., they have the same envelope but their shape varies with frequency) nor they are of finite support. Wavelets exhibit constant shape because they are generated from only one function. The 1D Mother Wavelet function ψ (in continuous form) defined as:

$$\psi_{s,\tau}(t) = \frac{1}{\sqrt{s}} \psi \left(\tfrac{t-\tau}{s} \right) \tag{19}$$

where $s,\tau \in R$ s.t $s \neq 0$ and satisfy,

$$\int_{-\infty}^{\infty} \psi(t)\, dt = 0 \tag{20}$$

where s and τ are the dilation (scale) and translation (position) parameters respectively. The 1D continuous wavelet transform denoted by $W_f(a,b)$ of a 1D function $f(t)$ is defined by:

$$W_f(s,\tau) = \int_{-\infty}^{\infty} f(t)\ \psi_{s,\tau}(t)\ dt \tag{21}$$

where ψ is the mother wavelet function as mention above. Now as we have inverse Fourier transform to reconstruct the original signal, we also have inverse continuous wavelet transform which is defined as:

$$f(t) = \frac{1}{C_\psi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} W_f(s,\tau)\ \psi_{s,\tau}(t)\ \frac{ds\, d\tau}{s^2} \tag{22}$$

Where C_ψ is defined as follows:

$$C_\psi = \int_{-\infty}^{\infty} \frac{|\widehat{\psi}(u)|^2}{|u|} du \tag{23}$$

where $\widehat{\psi}(u)$ is the Fourier Transform of $\psi(t)$.

9.1 The Discrete Wavelet Transform

If the parameters s and τ are chosen to be $s = s_0^j$ and $\tau = k\tau_0 s_0^j$, the 1D mother wavelet in discrete form as another type of wavelet is derived as:

$$\psi_{j,k}(t) = s_0^{-j/2} \psi(s_0^{-j} t - k\tau_0) \tag{24}$$

In general, $j,k \in Z$ and s_0=2, τ_0=1. The 1D discrete wavelet transform of a signal $f(t)$ is constitute of two parts, first part is the approximation part which is just approximation of original signal and the other one is called detail part. Approximation part is the low frequency whereas detail part is the high frequency of the signal along the axis. These parts are expressed as:

$$W_\phi(j_0,k) = \frac{1}{\sqrt{M}} \sum_t f(t)\ \phi_{j_0,k}(t) \tag{25}$$

$$W_\psi(j,k) = \frac{1}{\sqrt{M}} \sum_t f(t)\ \phi_{j,k}(t), \forall j \geq j_0 \tag{26}$$

Where ϕ is defined same as ψ and is called *Scaling Function* and it gives the *Approximation Part* of the signal and ψ gives the *Detail Part* of the signal. To reconstruct original signal inverse formula is expressed as follows:

$$f(t) = \frac{1}{\sqrt{M}} \sum_j \sum_k W_\psi(j,k)\ \psi_{j,k}(t) + \frac{1}{\sqrt{M}} \sum_k W_\phi(j_0,k)\ \phi_{j_0,k}(t) \tag{27}$$

Normally, we let $j_0 = 0$ and select M to be power of 2 i.e. $M = 2^J$. Hence t = 0,1,2,3,...,M-1; j = 0,1,2,3,...,J-1; $k = 0, 2^0, 2^1, 2^2, ..., 2^j$. The shift integer k is chosen in such a way that $\psi(s_0^{-j}t - k\tau_0)$ covers the whole line for all values of t. Thus the wavelet transform separates the objects into different components in its transform domain and studies each component with a resolution matched to its scale.

Due to separability of the transform, two dimensional wavelet transform can be obtained by successively taking one dimensional wavelet transforms along both the axis (in both continuous and discrete case).

10 Fractional Wavelet Transform (FRWT)

The continuous fractional wavelet transform (FRWT) was first defined by Mendlovic and Zalevsky *et al.*[48] in 1997. The continuous fractional wavelet transform (FRWT) of 1D function $f(t)$ is written as

$$W^\alpha(u,s,\tau) = \int_{-\infty}^{\infty} F^\alpha[f(t)](x)\ \psi_{s,\tau}(x)\ dx \tag{28}$$

where s and τ are the dilation (scale) and translation (position) parameters respectively. In other words, we can say that FRWT is the combination of FRFT and WT. Hence, all the properties of FRFT and WT are available in FRWT. Obviously, the FRWT domain is also the combination of time and frequency domains. The properties of FRWT are summarized as follows:

1. *Single Frequency/Wavelet transform operator:* W^0 is the Wavelet Packet transform operator. The FRWT of order $\alpha = 0$ is the wavelet packet transform of the input signal.
2. *Dual frequency operator:* The meaning of dual frequency operator is that the input signal is transformed by two different transforms in succession. $W^{\frac{\pi}{2}}$ is the

Dual frequency operator i.e. the FRWT of order $\alpha = \frac{\pi}{2}$ gives the dual frequency (Fourier-wavelet) transformed signal.

3. *Successive applications of FRWT:* Successive applications of FRWT are equivalent to a single transform whose order is equal to the sum of the individual orders.

$$W^{\alpha}(W^{\beta}) = W^{\alpha+\beta}$$

To reconstruct the original signal back from transformed signal, the inverse fractional wavelet transform is defined as:

$$f(t) = \frac{1}{C_{\psi}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F^{-\alpha}[W^{\alpha}(u,s,\tau)](x)\ \psi_{s,\tau}(t)\ \frac{ds\, d\tau\, dx}{s^2} \tag{29}$$

Where C_{ψ} is defined as follows:

$$C_{\psi} = \int_{-\infty}^{\infty} \frac{|\widehat{\psi}(u)|^2}{|u|} du \tag{30}$$

where $\widehat{\psi}(u)$ is the Fourier Transform of $\psi(t)$.

The computation of FRWT is correspond to the steps shown in figure 12, as explained in[48].

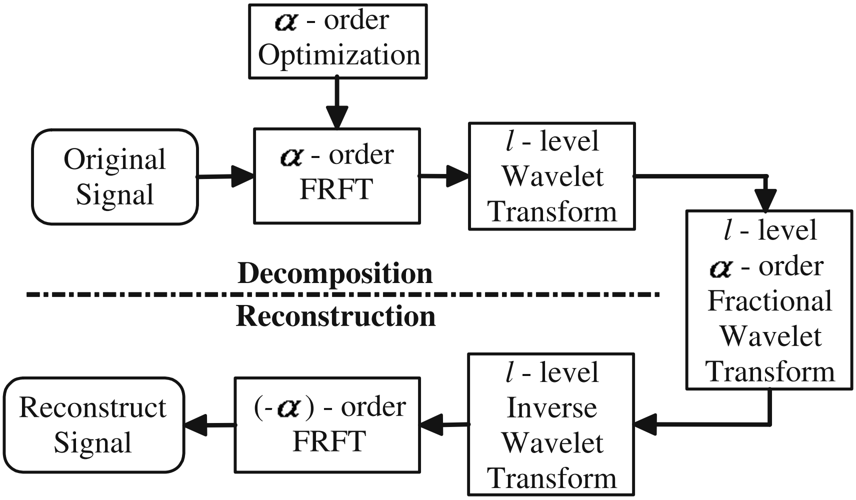

α - order Optimization
Original Signal
α - order FRFT
l - level Wavelet Transform
l - level α - order Fractional Wavelet Transform
Decomposition
Reconstruction
Reconstruct Signal
(-α) - order FRFT
l - level Inverse Wavelet Transform

Fig. 12 Computation of Fractional Wavelet Transform

In α-order optimization, the transform order is optimized for FRWT. To achieve desired goal, a Trial-and-error algorithm is used. In this algorithm, original signal is transformed via FRWT for arbitrary value of α and then original signal is reconstruct via inverse FRWT. After reconstruction, error is calculated according to Eq. 31.

$$\varepsilon = \int_{-\infty}^{\infty} |f(t) - \widetilde{f}(t)|^2 \, dt \tag{31}$$

where $f(t)$ and $\widetilde{f}(t)$ are original and reconstructed signals respectively. Choose that value of α as optimized transform order, for which the minimum error is obtained between original and reconstructed signal. Hence, the value of α is determined in such a way that the mean-square error between original and reconstructed signal is minimal. This trial-and-error algorithm may be long and followed by many calculations. However, for a given signal this process should be done only once.

For the discrete case, the same procedure is followed. The only difference is the use of discrete FRFT and WT instead of continuous one. Further, due to separability of the transform, two dimensional FRWT can be obtained by successively taking one dimensional FRWT along both the axis(in both continuous and discrete case). Mathematically, the FRWT of 2D function $f(t_x,t_y)$ is written as

$$W^{\alpha_x,\alpha_y}(u,v,s_1,\tau_1,s_2,\tau_2) = FRWT^{\alpha_y}_{t_y\rightarrow v}\{FRWT^{\alpha_x}_{t_x\rightarrow u}\{f(t_x,t_y)\}\} \tag{32}$$

where s_1, s_2 and τ_1, τ_2 are the dilation (scale) and translation (position) parameters along x and y direction respectively.

11 Singular Value Decomposition

The SVD is the novel technique for analyzing spectral information of the signals. This transform was introduced for square matrices by Beltrami[49] in 1873 and Jordan [50] in 1874 independently, and was extended for rectangular matrices by Eckart and Young[51] in 1930.

Let X be a general real(complex) matrix of order $m \times n$. The singular value decomposition (SVD) (figure 13) of X is the factorization

$$X = U * S * V^T \tag{33}$$

where U and V are *orthogonal(unitary)* and $S = diag(\sigma_1, \sigma_2, \ldots, \sigma_r)$, where σ_i, $i = 1(1)r$ are the singular values of the matrix A with $r = min(m,n)$ and satisfying

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r \tag{34}$$

The first r columns of V are the *right singular vectors* and the first r columns of U are the *left singular vectors*.

Use of SVD in digital image processing has some advantages. First, the size of the matrices from SVD transformation is not fixed. It can be a square or rectangular. Secondly, singular values in a digital image are less affected if general image processing is performed. Finally, singular values contain intrinsic algebraic image properties. All the properties of SVD are summarized as follows:

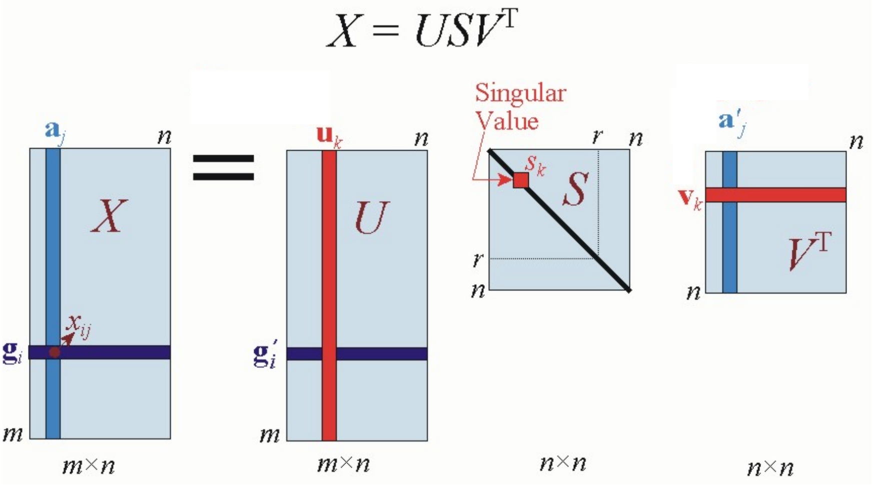


Fig. 13 Illustration of the Singular Value Decomposition (SVD)

- Stability: When a small perturbation is added to the matrix, large variance of its singular values does not occur.
- Singular values represent algebraic properties of an image.
- To some extent, singular values possess algebraic and geometric invariance.
- Rotation: given an image I and its rotated (with arbitrary angle) I^r, both have the same singular values.
- Translation: given an image I and its translated I^t, both have the same singular values.
- Scaling: given an image I and its scale I^s, if I has the singular values σ_i, then I^s has the singular values $\sigma_i * \sqrt{L_R L_C}$ where L_R and L_C are the scaling factor of rows and columns respectively. If rows (columns) are mutually scaled, I^s has the singular values $\sigma_i * \sqrt{L_R}(\sigma_i * \sqrt{L_C})$.
- Transpose: given an image I and its transposed I^T, both have the same singular values.
- Flip: given an image I and its row and column filled I_{rf} and I_{cf},transposed I^T, all have the same singular values.

12 Proposed Watermarking Scheme

In this section, we discuss some motivating factors in design of our approach to watermarking based on encryption technique which uses a meaningful gray scale logo/image as watermark instead of randomly generated Gaussian noise type watermark. First, host image is encrypted with the help fractional wavelet transform and then encrypted image is used for embedding. For embedding, encrypted image is partitioned into four sub-images via sub-sampling. Finally, embedding is done by modifying singular values of sub-images with the watermarks singular values. Then decryption is performed to get the watermarked image. This watermarked image is forwarded to the insecure communication channel. At the receiver end, first possibly attacked image is encrypted by same encryption algorithm and then extract the watermark using proposed extraction algorithm. Encryption, decryption, embedding and extraction algorithm are discussed below in detail.

With out loss of generality, assume that F represents the host image of size $M \times N$, W represents the watermark of size $m \times n$ and the watermark image is smaller than the host image by a factor 2^{Q_1} and 2^{Q_2} along both the direction, where Q_1 and Q_2 are any integers greater than or equal to 1. Block diagram of proposed algorithm is shown in figure 14.

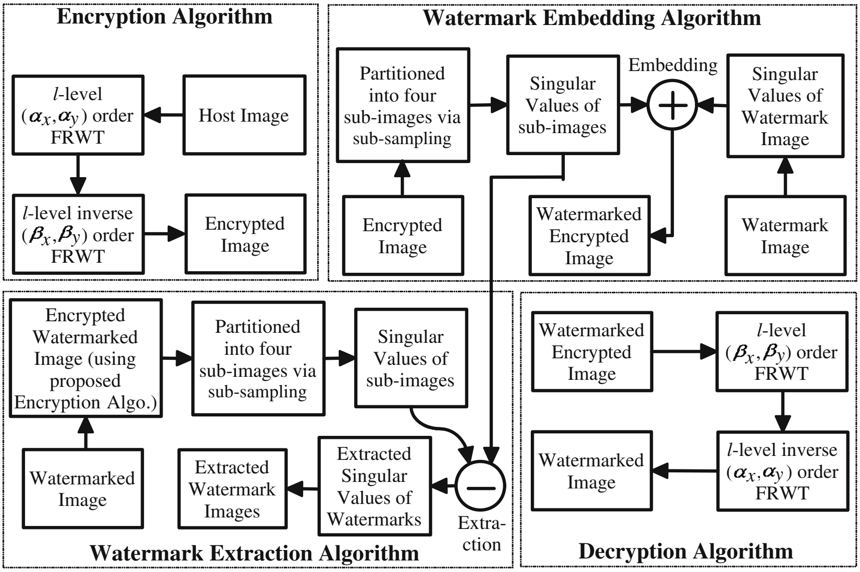

Encryption Algorithm
Host Image
l-level (α_x,α_y) order FRWT
l-level inverse (β_x,β_y) order FRWT
Encrypted Image
Watermark Embedding Algorithm
Partitioned into four sub-images via sub-sampling
Encrypted Image
Singular Values of sub-images
Embedding
Singular Values of Watermark Image
Watermark Image
Watermarked Encrypted Image
Encrypted Watermarked Image (using proposed Encryption Algo.)
Watermarked Image
Partitioned into four sub-images via sub-sampling
Singular Values of sub-images
Extracted Singular Values of Watermarks
Extracted Watermark Images
Extra-ction
Watermark Extraction Algorithm
Watermarked Encrypted Image
l-level (β_x,β_y) order FRWT
l-level inverse (α_x,α_y) order FRWT
Watermarked Image
Decryption Algorithm

Fig. 14 Block Diagram of Proposed Encryption and Watermark Algorithm

Encryption Algorithm. The goal of encryption is to obtain an encrypted image F_e with the help of original host image F. The process is given as follows:

Step 1: Perform l-level *fractional wavelet transform* with transform orders $(\alpha_x, \ \alpha_y)$ on the host image, which is denoted by f_l^θ, where $\theta \in \{$ LL, LH, HL, HH $\}$ and $l \in [1, L]$.

Step 2: Perform l-level *inverse fractional wavelet transform* with transform orders $(\beta_x, \ \beta_y)$ on f_l^θ, which is denoted by F_e.

Embedding Algorithm. The goal of embedding algorithm is to embed watermark W in the encrypted image F_e. The process is formulated as follows:

Step 1: The encrypted image F_e is partitioned into four sub-images via sub-sampling, denoted by F_e^{ν}, where $\nu \in \{$ TL, TR, BL, BR $\}$ (TL=Top-Left, TR=Top-Right, BL=Bottom-Left and BR=Bottom-Right).

$$\begin{aligned} F_e^{TL}(i,j) &= F(2i,2j) & F_e^{TR}(i,j) &= F(2i,2j+1) \\ F_e^{BL}(i,j) &= F(2i+1,2j) & F_e^{BR}(i,j) &= F(2i+1,2j+1) \end{aligned} \tag{35}$$

where $i = 1(1)M/2$, $j = 1(1)N/2$.

Step 2: Perform SVD on all sub-images F_e^{ν} and watermark image W,

$$F_e^{\nu} = U_{F_e^{\nu}}\ S_{F_e^{\nu}}\ V_{F_e^{\nu}}^T \tag{36}$$

$$W = U_W\ S_W\ V_W^T \tag{37}$$

Step 3: Modify the singular values of all sub-images with the singular values of the watermark as:

$$(\sigma_{F_e^{\nu}})^{new} = \sigma_{F_e^{\nu}} + \gamma\ \sigma_W \tag{38}$$

where γ gives the watermark strength for each sub-image.

Step 4: Perform inverse SVD to construct all modified sub-images,

$$(F_e^{\nu})^{new} = U_{F_e^{\nu}}\ (S_{F_e^{\nu}})^{new}\ V_{F_e^{\nu}}^T \tag{39}$$

Step 5: After embedding, construct watermarked encrypted image via inverse sub-sampling, which is denoted by F_e^{water}.

Decryption Algorithm. The goal of decryption algorithm is to obtain watermarked image G from F_e^{water}. The process is given as follows:

Step 1: Perform l-level *fractional wavelet transform* with transform orders $(\beta_x,\ \beta_y)$ on F_e^{water}, which is denoted by $f_{l,e}^{\theta}$, where $\theta \in \{$ LL, LH, HL, HH $\}$ and $l \in [1,L]$.

Step 2: Perform l-level *inverse fractional wavelet transform* with transform orders $(\alpha_x,\ \alpha_y)$ on $f_{l,e}^{\theta}$ to get watermarked (decrypted) image G.

Extraction Algorithm. The goal of extraction algorithm is to extract watermark W from watermarked image G. We have used encrypted image F_e, U_W and V_W as the keys to extraction and hence saved for this purpose. The process is formulated as follows:

Step 1: The watermarked image (G) is encrypted via encryption algorithm, which is denoted by G_e.

Step 2: The encrypted images F_e and G_e are partitioned into sub-images via sub-sampling, denoted by F_e^{ν} and G_e^{ν}, where $\nu \in \{$ TL, TR, BL, BR $\}$.

Step 3: Perform SVD on all sub-images F_e^{ν} and G_e^{ν},

$$F_e^{\nu} = U_{F_e^{\nu}}\ S_{F_e^{\nu}}\ V_{F_e^{\nu}}^T \tag{40}$$

$$G_e^v = U_{G_e^v}\ S_{G_e^v}\ V_{G_e^v}^T \tag{41}$$

Step 4: Extract the watermark singular values from all sub-images as:

$$\sigma_{ext}^v = \frac{\sigma_{G_e^v} - \sigma_{F_e^v}}{\gamma} \tag{42}$$

Step 5: Perform inverse SVD to construct all patterns of extracted watermarks,

$$W_{ext}^v = U_W\ S_{ext}^v\ V_W^T \tag{43}$$

13 Results and Discussions

The performance of the proposed watermarking algorithm is explored using MATLAB platform and a number of experiments are performed on different images of size 256 × 256, namely Goldhill, Lena, Clock and Trui (shown in figure 15(a)). Four different gray scale logos/images of size 128 × 128, namely Springer logo, Ducky, Work in Progress logo, and Cup are used as watermark images (shown in figure 16). Springer logo is embedded into Goldhill image, Ducky is embedded into Lena image, Work in Progress logo is embedded into Clock image and finally Cup is embedded into Trui image. The watermarked image quality is measured using PSNR (Peak Signal to Noise Ratio). No perceptual degradation is observed between the original and watermarked image according to Human Visual System (figure 15). The similarity between the original and the extracted watermark singular values is measured by Zero Mean Cross Correlation ($ZMCC_2$). The values of PSNR and

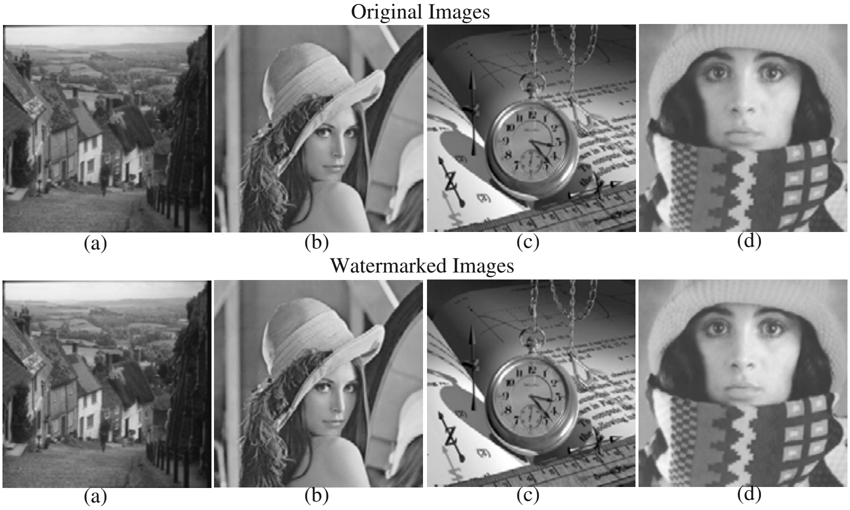


Fig. 15 a,b,c,d) Original e,f,g,h) Watermarked Images

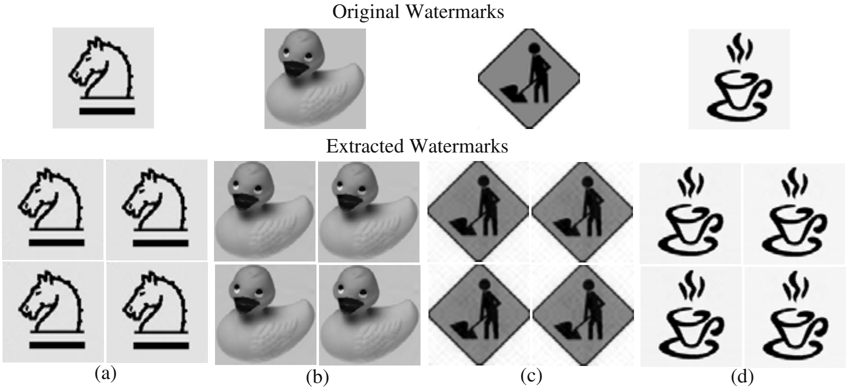


Fig. 16 Original and extracted watermarked images from a) Goldhill b) Lena c) Clock d) Trui images

correlation coefficients for all test images are shown in the tables 6 and 7 respectively. The value of transform orders and encryption algorithm are used as the keys

Table 6 PSNR (in Db) for all Images

	PSNR (in Db) between Original and		
Images	Encrypted Image	Watermarked Encrypted Image	Watermarked (Decrypted) Image
Goldhill	10.3133	10.9906	35.3124
Lena	9.3466	9.5464	37.5968
Clock	10.1820	10.4105	36.9715
Trui	7.7439	8.0321	37.6952

Table 7 Correlation coefficients for all Extracted Watermark Images

Images	ρ			
	TL	TR	BL	BR
Goldhill	1.0000	1.0000	1.0000	1.0000
Lena	0.9999	0.9999	0.9999	0.9999
Clock	0.9994	0.9992	0.9993	0.9993
Trui	1.0000	1.0000	1.0000	1.0000

in the proposed scheme because without knowing transform order and encryption algorithm, one cannot extract the watermark image correctly. Hence, the security of proposed scheme is increased by using both encryption and watermarking terminologies. To make the algorithm less complex, we can opt the value of $(\alpha_x,\ \alpha_y)$, which is calculated from transform order optimization step (see section 10) and the value of $(\beta_x,\ \beta_y)$ is chosen randomly and used as the key.

The robustness of the proposed watermarking algorithm is carried out by attacking the watermarked image by a variety of active attacks namely Average and Mean Filtering, Gaussian noise addition, JPEG Compression, Row-Column Deletion, Flipping, Cropping, Resize, Shearing, Histogram Equalization, Wrapping, Sharpen and Contrast Adjustment. After these attacks on the watermarked image, the extracted logo is compared with the original one. For further analysis, Trui image is used, since Trui image is having higher PSNR value among all the test images. In figure 17, original, encrypted, watermarked encrypted and watermarked images and the corresponding histogram are shown. The encrypted image (figure 17(b)) appears to be a degraded image, from which it is very difficult to ascertain the original image without knowing the algorithm. The change in the histogram and correlation plot of adjacent pixels of the original (figure 17(a)) and encrypted image (figure 17(b)) verify that all the properties of encryption are satisfied by the proposed algorithm. After embedding and decrypting the watermarked encrypted image, watermarked image is retrieved and the corresponding histograms are depicted in 17(c,d)). The correlation plot of two horizontally adjacent pixels in original, encrypted, watermarked encrypted and watermarked (decrypted) images are depicted in figure 18 and the correlation coefficients for the horizontally, diagonal and vertical directions are given in the table 8.

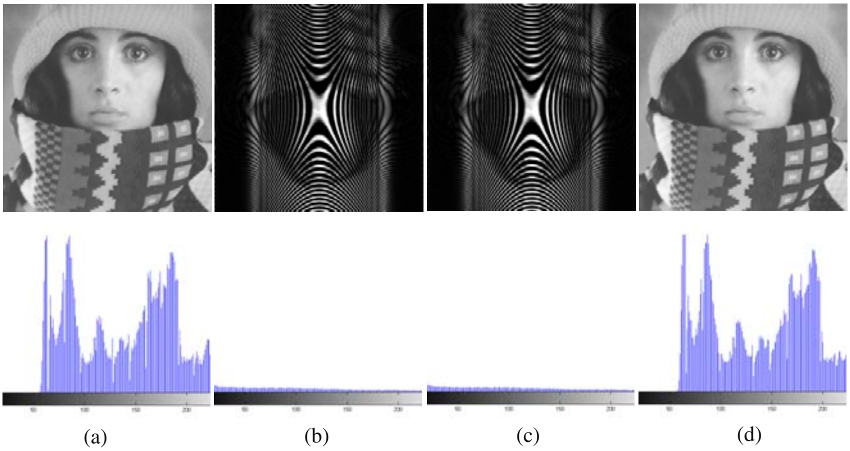
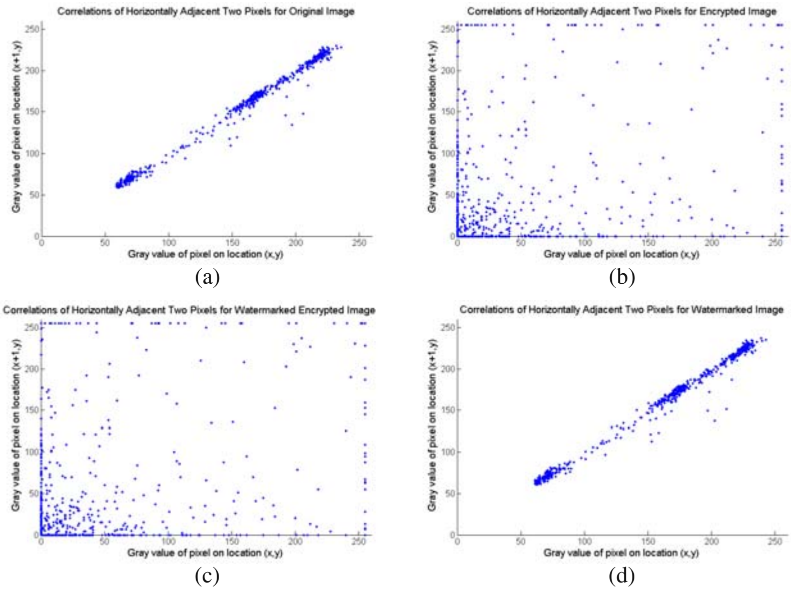


Fig. 17 a) Original b) Encrypted c) Watermarked Encrypted d) Watermarked (Decrypted) Images



(a) (b)

(c) (d)

Fig. 18 The correlation plot of two horizontally adjacent pixels in a) Original b) Encrypted c) Watermarked Encrypted d) Watermarked (Decrypted) Images

Table 8 Correlation coefficients of adjacent pixels in Original, Encrypted, Watermarked Encrypted and Watermarked (Decrypted) Images

Adjacent	Image			
Pixels	Original	Encrypted	Watermarked Encrypted	Watermarked (Decrypted)
Horizontal	0.9962	0.0453	0.0445	0.9961
Vertical	0.9834	0.1424	0.1398	0.9832
Diagonal	0.9871	0.0081	0.0079	0.9867

The most common manipulation in digital image is filtering. Watermark is extracted after applying 7 × 7 averaging and median filtering. The results are shown in the figures 19 and 20. To verify the robustness of the watermarking scheme, another measure is noise addition. In our experiments, *P*% additive Gaussian noise is added in the watermarked image. In figure 21, extracted watermarks from 50% Gaussian noise attacked watermarked image is shown. Storage and transmission of digital data is the most common operation for this purpose a lossy coding operation is often performed on the data to reduce the memory and increase efficiency. Hence, we have also tested our algorithm for JPEG compression(90:1) and the results are shown in figure 22. The proposed algorithm has also been tested for row-column deletion and resizing attacks. In row-column deletion, some rows and columns of the watermarked image are deleted randomly and then extract the watermarks. Figure 23 shows the result of randomly deleted 5 rows and 5 columns attack. The proposed algorithm is also tested for flipping (both horizontal and vertical) attack. The results for horizontal and vertical flipping are shown in figures 24 and 25 respectively. Another frequently used action on images is cropping. In figure 26, results for cropping are shown. For resizing, the size of the image is reduced to 64 × 64 and again carried back to the original size 256 × 256 (figure 27). Results for shearing, histogram equalization, wrapping, sharpen and contrast adjustment are shown in figures 28, 29, 30, 31 and 32 respectively. For shearing attack, watermarked image is sheared along x-axis and filled in the area (figure 28). Wrapping is the process of giving 3D effect to an object by wrap a selection around a shape. Figure 30 shows the extracted watermarks when object is wrapped around spherical shape. For sharpening attack, the sharpness of the watermarked image is increased by a factor of 80 and then watermark is extracted (figure 31). For Contrast Adjustment, the contrast of the watermarked host image is increased by 50%(figure 32). The correlation coefficients of all extracted watermarks after all attacks are given in table 9.

Table 9 Correlation coefficients of adjacent pixels in Original, Encrypted, Watermarked Encrypted and Watermarked (Decrypted) Images

Adjacent	ρ of Watermark extracted from			
Pixels	TL	TR	BL	BR
Average Filtering (7 × 7)	0.8589	0.8631	0.8615	0.8618
Median Filtering (7 × 7)	0.8186	0.8166	0.8210	0.8152
Gaussian Noise Addition (50%)	0.4428	0.4198	0.4419	0.4366
JPEG Compression (90 : 1)	0.9924	0.9925	0.9930	0.9927
Row-Column Deletion (5 Rows-5 Columns)	0.8586	0.9904	0.8840	0.9910
Horizontal Flipping	0.9987	0.9986	0.9984	0.9986
Vertical Flipping	0.9994	0.9995	0.9995	0.9995
Symmetric Cropping	-0.9842	0.9932	0.8983	0.8132
Resizing (256 → 64 → 256)	0.9814	0.9822	0.9829	0.9824
Shearing (along *x* axis)	0.7271	0.6706	0.7356	0.6728
Histogram Equalization	0.9901	0.9909	0.9898	0.9907
Wrapping (around spherical shape)	0.7533	0.7469	0.7476	0.7413
Sharpen (increased by 80%)	0.4299	0.4282	0.4333	0.4263
Contrast Adjustment (increased by 50%)	0.7283	0.7364	0.7273	0.7353

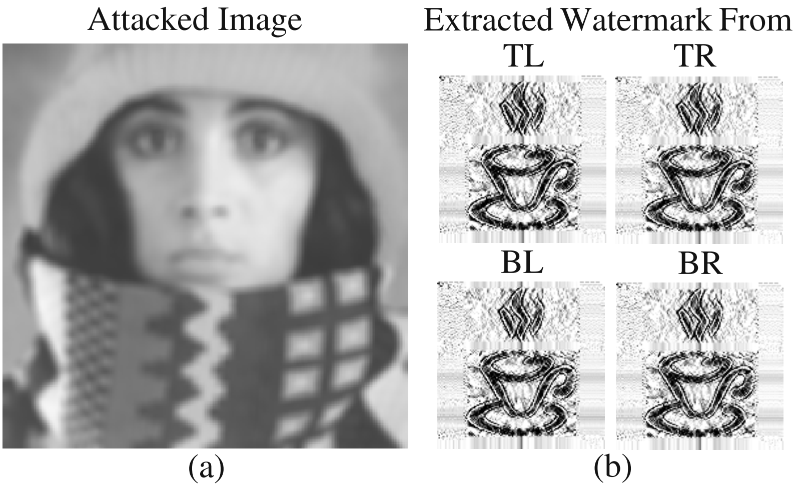


Fig. 19 Results for Average Filtering (7 × 7)

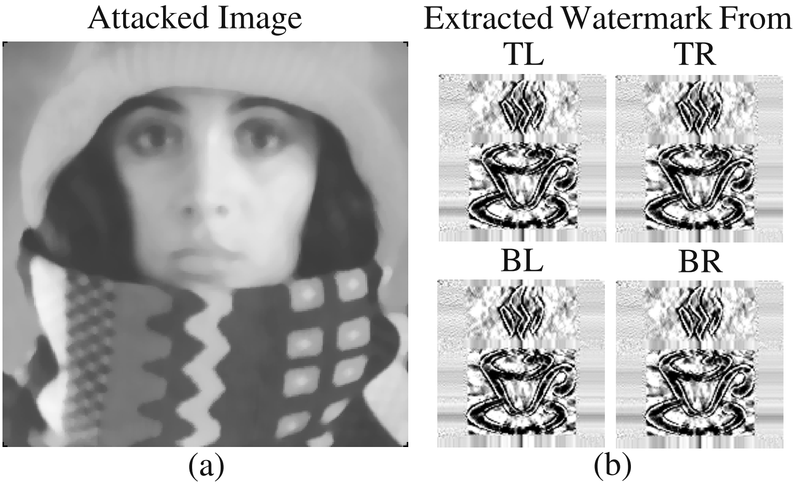


Fig. 20 Results for Median Filtering (7 × 7)

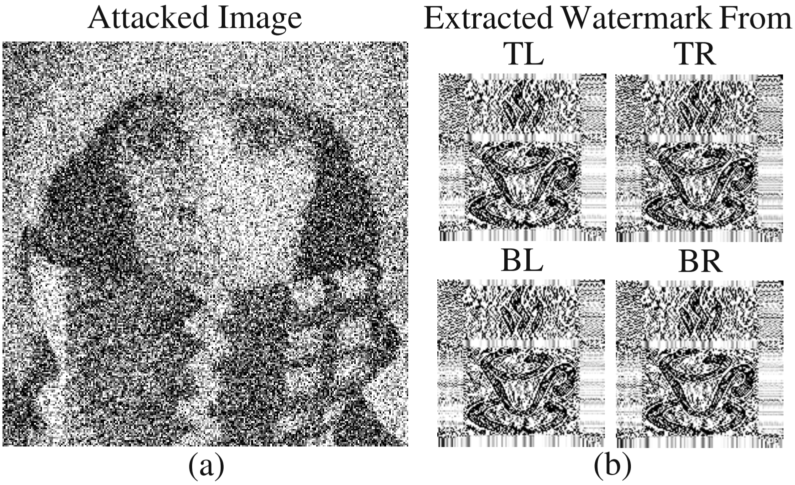


Fig. 21 Results for Gaussian Noise Addition (50%)

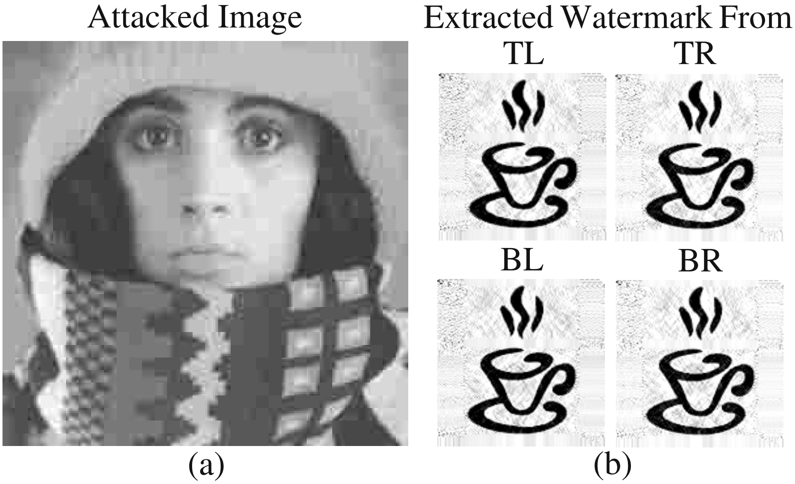


Fig. 22 Results for JPEG Compression (90 : 1)

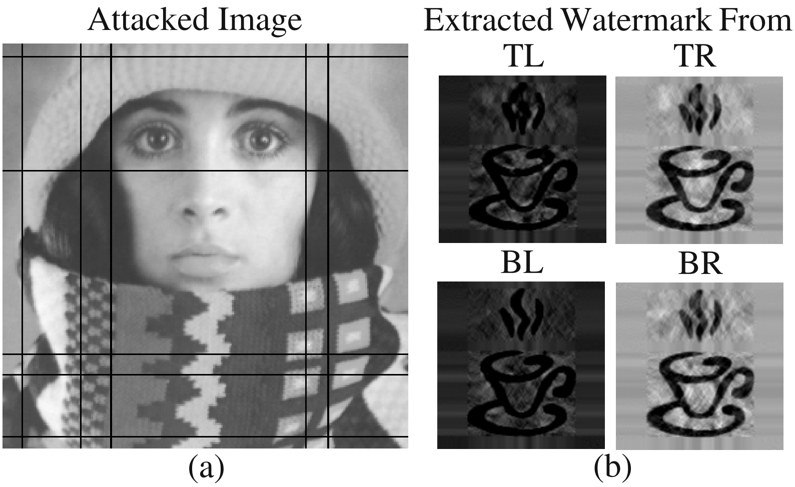


Fig. 23 Results for Row-Column Deletion (Randomly deleted 5 Rows and Columns)

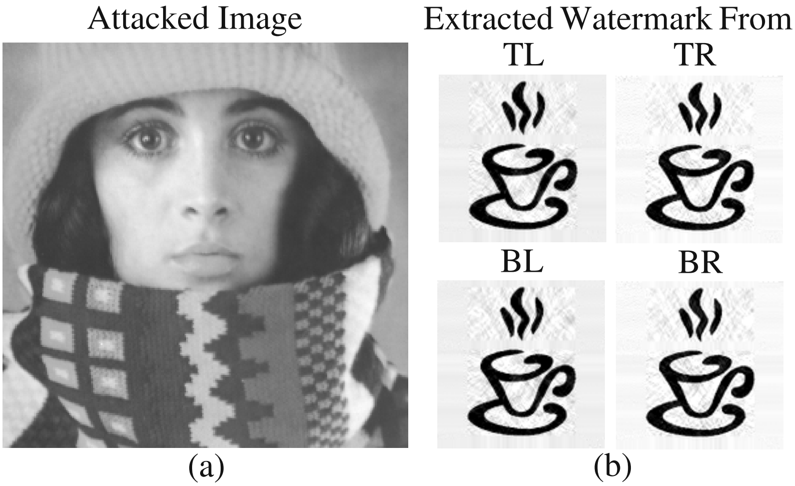


Fig. 24 Results for Horizontal Flipping

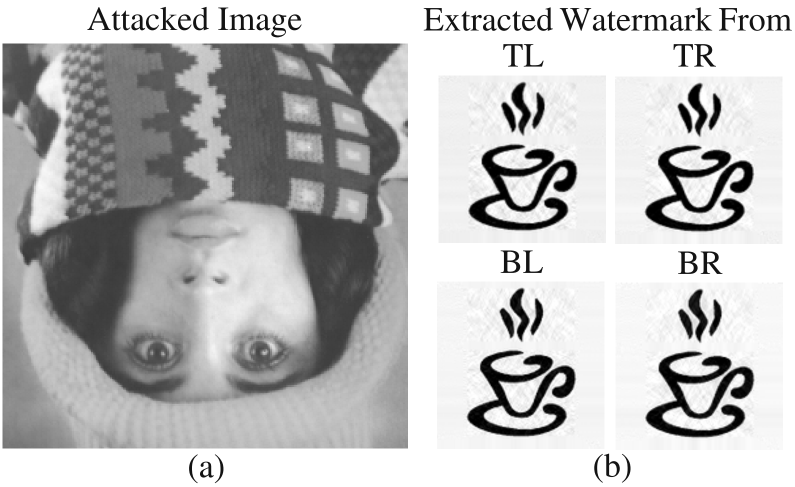


Fig. 25 Results for Vertical Flipping

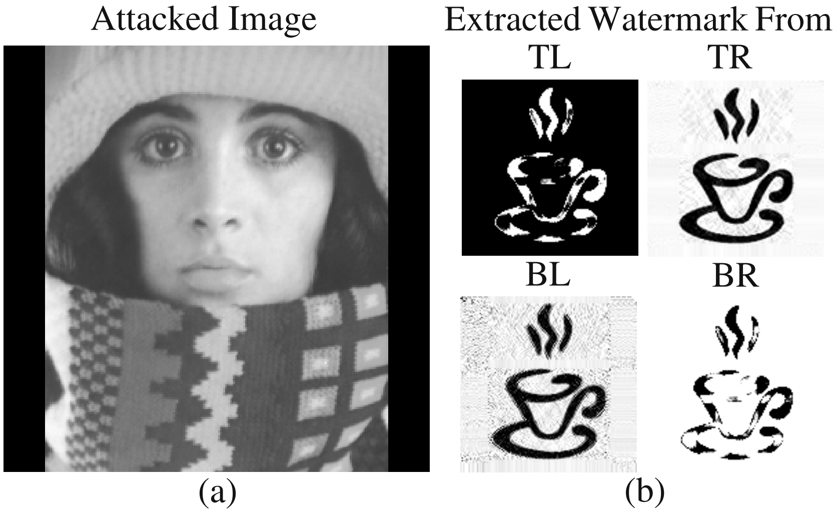


Fig. 26 Results for Symmetric Cropping (Deleted 20% area)

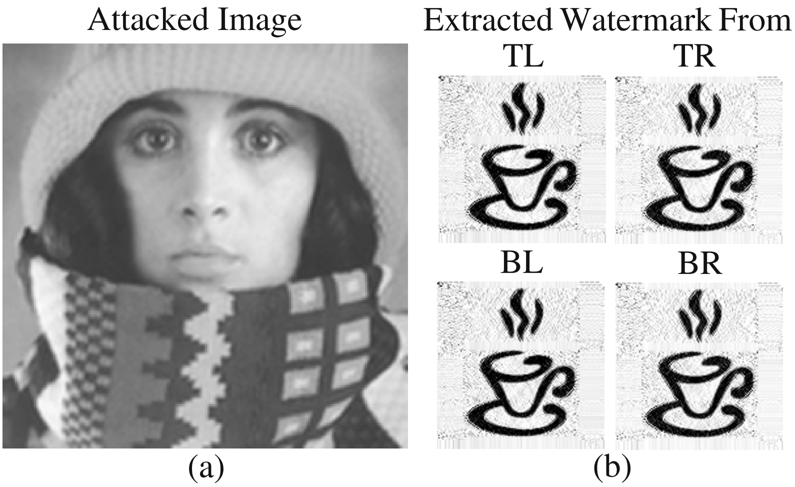


Fig. 27 Results for Resizing (256 → 64 → 256)

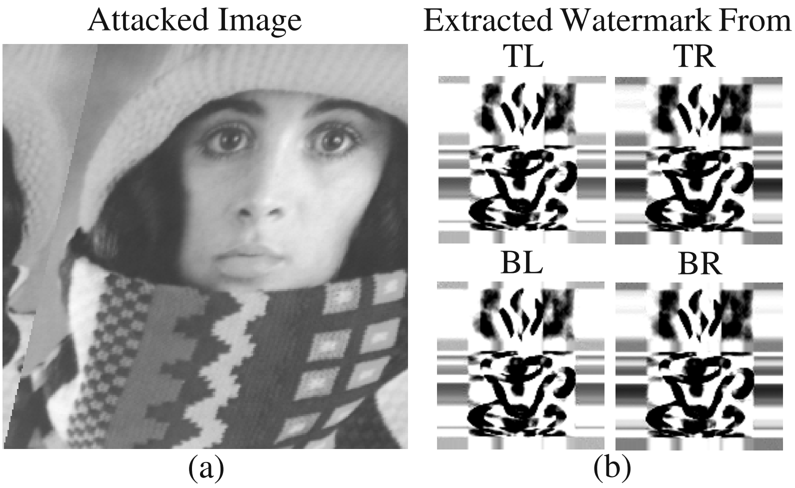


Fig. 28 Results for Shearing along *x*-axis

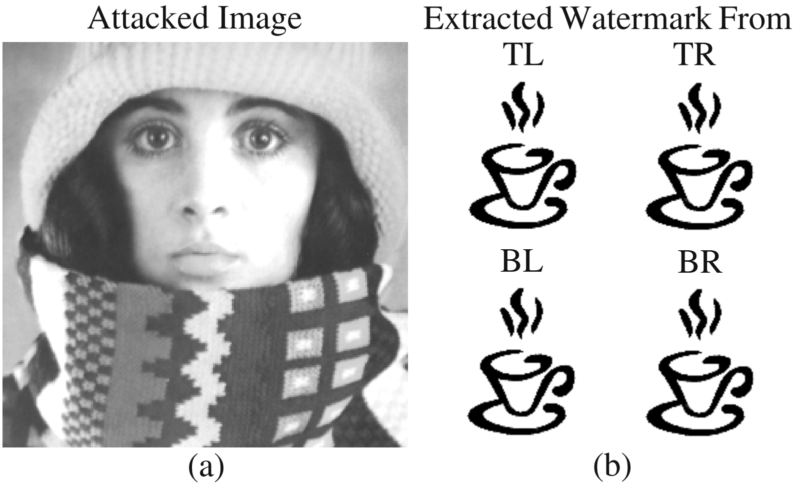


Fig. 29 Results for Histogram Equalization

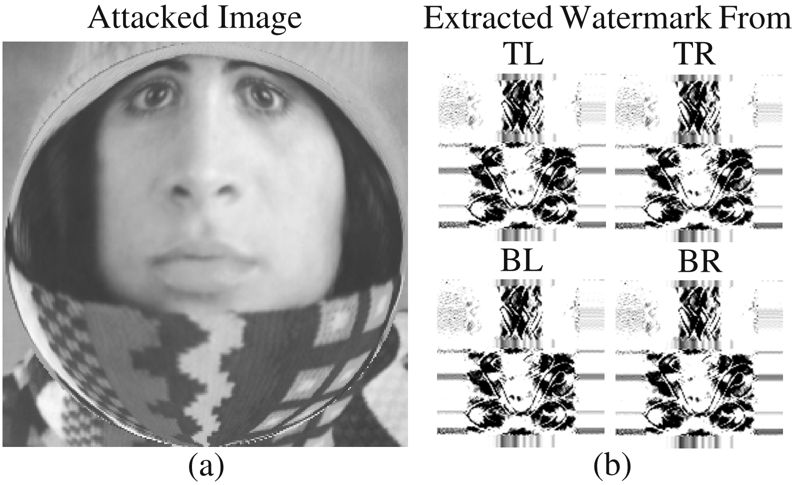


Fig. 30 Results for Wrapping (*when wrapped around a spherical shape*)

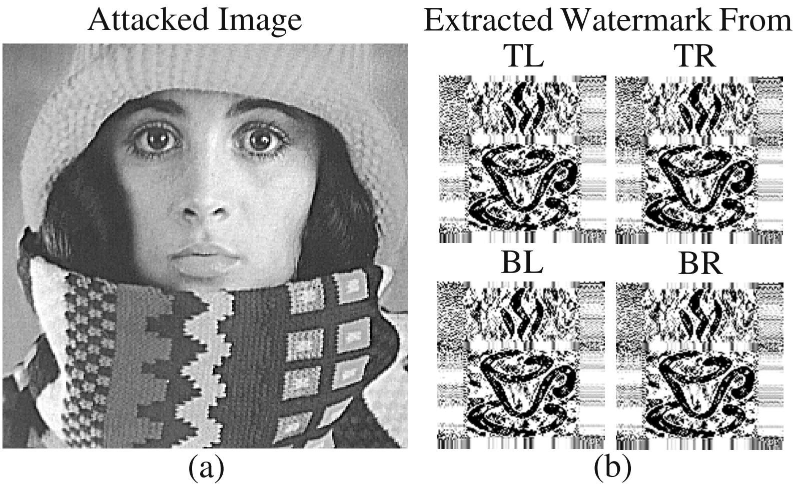


Fig. 31 Results for Sharpen (increased by 80%)

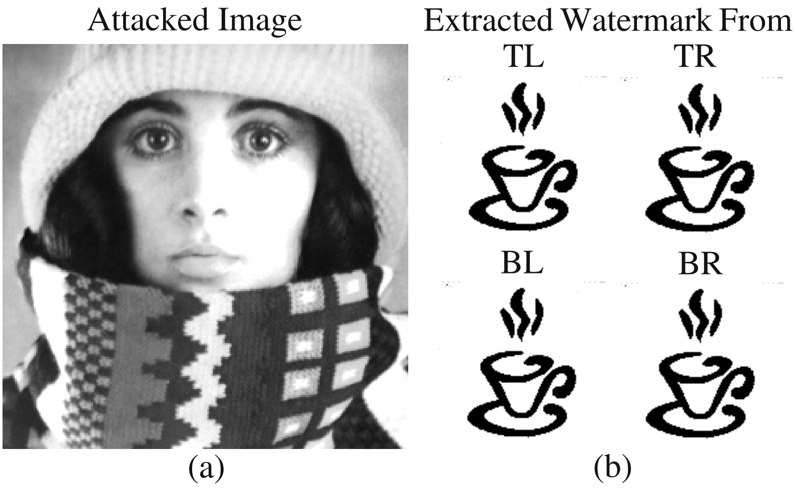


Fig. 32 Results for Contrast Adjustment (increased by 50%)

14 Conclusions

In this chapter, the overview of multimedia encryption and watermarking techniques are given. Both techniques realize on different functionalities and have been developed independently. Both the techniques are complementing each other and they do not provide complete protection either. For instance, in case of encryption, if an intruder is able to decrypt the data successfully, the original data is again vulnerable to duplication. While in the second case if the digital watermark is removed, then also a copy of data can be distributed without any problem. Hence, for a complete multimedia security, the hybridization of both the techniques is needed. The possible way of hybridization is multimedia data can be first watermarked and then encrypted or vice versa. Moreover, a new robust hybrid technique is proposed in this chapter to enhance the security level of the multimedia data. The image is encrypted using the fractional wavelet transform and this encrypted image is then watermarked using singular value decomposition. The proposed technique satisfied all the properties of encryption and watermarking, which has also been empirically verified by computer simulations.

References

[1] Menezes, A.J., Van Oorschot, P.C., Vanstone, S.A.: Handbook of Applied Cryptography. CRC Press, Boca Raton (1996)

[2] Stallings, W.: Cryptography and Network Security: Principles and Practice. Prentice-Hall, Upper Saddle River (1999)

[3] Stinson, D.R.: Cryptography: Theory and Practice, 2nd edn. Chapman and Hall/CRC, Boca Raton (2002)

[4] Mollin, R.A.: An Introduction to Cryptography. CRC Press, Boca Raton (2006)

[5] Furht, B.: Handbook of Internet and Multimedia Systems and Applications. CRC Press, Boca Raton (1999)

[6] Furht, B., Socek, D.: Multimedia Security: Encryption Techniques. In: IEC Comprehensive Report on Information Security, International Engineering Consortium, Chicago, IL (2003)

[7] Lian, S.: MultiMedia Content Encryption: Techniques and Applications. CRC Press, Boca Raton (2008)

[8] Shannon, C.E.: Communication theory of secrecy system. Bell System Technical Journal 28, 656–715 (1949)

[9] Simmons, G.J.: The Prisoner's Problem and the Subliminal Channel. In: Advance in Cryptology, Proc. of CRYPTO 1983, pp. 51–67. Plenum Press (1984)

[10] Cox, I.J., Miller, M.L., Bloom, J.A.: Digital watermarking. Morgan Kaufmann, San Francisco (2001)

[11] Katzenseisser, S., Petitcolas, F.A.P.: Information Hiding Techniques for Steganography and Digital Watermarking. Artech House, Boston (2000)

[12] Arnold, M., Schmucker, M., Wolthusen, S.D.: Techniques and Applications of Digital Watermarking and Content Protection. Artech House (2003)

[13] Muharemagic, E., Furht, B.: Multimedia Security: Watermarking Techniques. In: Comprehensive Report on Information Security, International Engineering Consortium, Chicago, IL (2004)

[14] Cox, I.J., Miller, M.L., Bloom, J.A.: Watermarking applications and their properties. In: Proc. of Int. Conf. on Information Technology 2000, Las Vegas, pp. 1–5 (2000)

[15] Kutter, M., Petitcolas, F.A.P.: A fair benchmark for image watermarking systems. In: Proc. of SPIE, Electronic Imaging 1999, Security and Watermarking of Multimedia Contents, Sans Jose, CA, USA, vol. 3657, pp. 25–27 (1999)

[16] Mohanty, S.P.: Digital Watermarking: A Tutorial Review. In: Report on Dept. of Electrical Engineering, Indian Institute of Science, Bangalore, India (1999)

[17] Schyndle, R.G.V., Tirkel, A.Z., Osbrone, C.F.: A Digital Watermark. In: Proc. of IEEE Int. Conf. on Image processing, vol. 2, pp. 86–90 (1994)

[18] Hwang, M.S., Chang, C.C., Hwang, K.F.: A watermarking technique based on one-way hash functions. IEEE Transcations on Consumer Electronics 45(2), 286–294 (1999)

[19] Cox, I.J., Killian, J., Leighton, F.T., Shamoon, T.: Secure Spread Spectrum Watermarking for Multimedia. IEEE Transcation on Image Processing 6(12), 1673–1687 (1997)

[20] Barni, M., Bartiloni, F., Cappellini, V., Piva, A.: A DCT Domain System for Robust Image watermarking. Signal Processing 66(3), 357–372 (1998)

[21] Djurovic, I., Stankovic, S., Pitas, I.: Digital watermarking in the fractional fourier transformation domain. Journal of Network and Computer Applications 24(4), 167–173 (2001)

[22] Feng, Z., Xiaomin, M., Shouyi, Y.: Multiple-chirp typed blind watermarking algorithm based on fractional Fourier transform. In: Proc. of Int. Sym. on Intelligent Signal Processing and Communication Systems, pp. 141–144 (2005)

[23] Yu, F.Q., Zhangi, Z.K., Xu, M.H.: A Digital Watermarking Algorithm for Image Based on Fractional Fourier Transform. In: Proc. of First IEEE Conf. on Industrial Electronics and Applications, pp. 1–5 (2006)

[24] Xia, X., Boncelet, C.G., Arce, G.R.: A multiresolution watermark for digital images. In: Proc. Fourth IEEE Int. Conf. on Image Processing, Santa Barbara, CA, vol. 3, pp. 548–551 (1997)

[25] Barni, M., Bartiloni, F., Piva, A.: Improved wavelet based watermarking through pixel wise masking. IEEE Transcations on Image Processing 10, 783–791 (2001)

[26] Kundur, D., Hatzinakos, D.: Towards robust logo watermarking using meltiresolution image fusion. IEEE Transcations on Multimedia 6, 185–197 (2004)

[27] Wang, S.H., Lin, Y.P.: Wavelet tree quantization for copyright protection watermarking. IEEE Transcations on Image Processing 13(2), 154–165 (2004)

[28] Zhang, X.D., Feng, J., Lo, K.T.: Image watermarking using tree-based spatial-frequency feature of wavelet transform. J. Visual Comm. Image Representation 14, 474–491 (2003)

[29] Meerwald, P., Uhl, A.: A survey of Wavelet-Domain Watermarking Algorithms. In: Proc. of SPIE, Electronic Imaging, Security and Watermarking of Multimedia Contents III, San Jose, CA, USA, vol. 4314 (2001)

[30] Gorodetski, V.I., Popyack, L.J., Samoilov, V., Skormin, V.A.: SVD-based approach to transparent embedding data into digital images. In: Gorodetski, V.I., Skormin, V.A., Popyack, L.J. (eds.) MMM-ACNS 2001. LNCS, vol. 2052, p. 263. Springer, Heidelberg (2001)

[31] Liu, R., Tan, T.: An SVD-Based Watermarking Scheme for Protecting Rightful Ownership. IEEE Transactions on Multimedia 4(1), 121–128 (2002)

[32] Chandra, D.V.S.: Digital Image Watermarking Using Singular Value Decomposition. In: Proc. of 45th IEEE Midwest. Sym. on Circuits and Systems, pp. 264–267 (2002)

[33] Ganic, E., Zubair, N., Eskicioglu, A.M.: An Optimal Watermarking Scheme Based on Singular Value Decomposition. In: Proc. of the IASTED Int. Conf. on Communication, Network, and Information Security (CNIS 2003), Uniondale, NY, pp. 85–90 (2003)

[34] Ganic, E., Eskicioglu, A.M.: Robust Embedding of Visual Watermarks Using DWT-SVD. Journal of Electronic Imaging (2005)

[35] Sverldov, A., Dexter, S., Eskicioglu, A.M.: Robust DCT-SVD Domain Image Watermarking for Copyright Protection: Embedding Data in All Frequencies. In: Proc. of European Signal Processing Conference (2005)

[36] Li, Q., Yuan, C., Zong, Y.Z.: Adaptive DWT-SVD Domain Image Watermarking Using Human Visual Model. In: ICACT 2007, pp. 1947–1951 (2007)

[37] Chang, C.C., Tsai, P., Lin, C.C.: SVD-based digital image watermarking scheme. Pattern Recognition Letters 26, 577–1586 (2005)

[38] Namias, V.: The fractional order Fourier transform and its application to quantum mechanics. Journal of Inst. Math. Appl. 25, 241–265 (1980)

[39] McBride, A.C., Kerr, F.H., Namias, V.: Fractional Fourier transforms. IMA Journal of Appl. Math. 39, 159–175 (1987)

[40] Almeida, L.B.: The fractional Fourier transform and time-frequency representations. IEEE Transaction on Signal Processing 42, 3084–3091 (1994)

[41] Pei, S.-C., Yeh, M.-H.: Two dimensional discrete fractional Fourier transform. Signal Processing 67, 99–108 (1998)

[42] Dickinson, B.W., Steiglitz, K.: Eigenvectors and functions of the discrete Fourier transform. IEEE Trans. Acoust. Speech Signal Process 30, 25–31 (1982)

[43] Mallat, S.G.: A theory for Multiresolution Signal Decomposition: The Wavelet Representation. IEEE Tran. Pattern Analysis and Machine Intelligence 11, 674–693 (1989)

[44] Meyer, Y.: Orthonormal Wavelets in Wavelets: Time-Frequency methods and Phase Spaces, pp. 21–37. Springer, Heidelberg (1989)

[45] Chui, C.K.: Wavelets: A Tutorial in Theory and Applications. Academic Press, London (1992)

[46] Daubechies, I.: Ten Lectures notes on Wavelet, vol. 61. SIAM, Philadelphia (1992)

[47] Daubechies, I.: The wavelet transform time frequency localization and signal analysis. IEEE Transcations on Information Theory 36, 961–1005 (1990)

[48] Mendlovic, D., Zalevsky, Z., Mas, D., García, J., Ferreira, C.: Fractional wavelet transform. Appl. Optics. 36, 4801–4806 (1997)

[49] Beltrami, E.: Sulle funzioni bilineari (On Bilinear Functions). Giornale di Matematiche ad Uso degli Studenti Delle Universita 11, 98–106 (1873); An English translation by-Boley, D. Tech. Report, Dept. of Computer Science, Univ. of Minnesota, Minneapolis, pp. 90–37 (1990).

[50] Jordan, C.: Memoire sur les formes bilineaires (Memoir on Bilinear Forms). Journal de Mathematiques Pures et Appliquees, Deuxieme Serie 19, 35–54 (1874)

[51] Eckart, C., Young, G.: The approximation of one matrix by another of lower rank. Psychometrika 1, 211–218 (1936)

Multimedia Encryption: A Brief Overview

Nidhi S. Kulkarni, Balasubramanian Raman, and Indra Gupta

Abstract. The augmentation in the field of communication technology has lead to an extensive use of multimedia applications. Multimedia applications, especially, over wireless networks, can easily be intercepted, thus, making its security an essential and challenging issue. Multimedia encryption is the core enabling technology that provides confidentiality and prevents unauthorized access of the content. Real time constraints, large amount, and unique characteristics of multimedia data inhibits the use of traditional cryptographic algorithms over multimedia data. Recent years have witnessed an astounding development in the direction of format compliant, perceptual, and scalable encryption techniques that support advanced functionalities. This chapter gives a snapshot of the conventional encryption, and an up-to-date treatise of the principles, techniques, attacks, and advancements of multimedia encryption techniques developed to meet desired goals in specific applications.

1 Introduction

Advancement in the field of computers and communications has led to a phenomenal growth in transmission, distribution, and processing of digital multimedia data over wired/wireless channel. The open nature of these wired/wireless channels makes data transfer over them vulnerable to various kind of attacks and hence, media content protection has become an essential requirement. Techniques like encryption, watermarking, steganography, fingerprinting, etc. have been developed to protect valuable multimedia assets from unauthorized access and consumption.

Various techniques developed for content protection serve different purposes; watermarking or fingerprinting embeds owner's mark/fingerprint in original data for

Nidhi S. Kulkarni and Indra Gupta
Department of Electrical Engineering,
Indian Institute of Technology Roorkee, Roorkee-247667, India
e-mail: nskindee@iitr.ernet.in, indrafee@iitr.ernet.in

Balasubramanian Raman
Department of Mathematics,
Indian Institute of Technology Roorkee, Roorkee-247667, India
e-mail: balarfma@iitr.ernet.in

M. Grgic et al. (Eds.): Rec. Advan. in Mult. Sig. Process. and Commun., SCI 231, pp. 417–449.
springerlink.com

authentication, steganography hides important information in the host image and encryption provides complete confidentiality to the desired content. Thus, multimedia encryption is applied to the data before transmission or distribution to protect confidentiality of the content, prevent unauthorized access, provide persistent access control, and rights management of the content.

Varied applications have different requirements; military applications stress on complete confidentiality; pay per view applications entails perceptual degradation rather than high end security while some real time applications tend to decrease consumption of computational resources, thus, providing an acceptable level of security. This chapter gives an up-to-date treatise of multimedia encryption principles and techniques developed to meet the desired goals in specific applications. Section 2 discusses the preliminaries of conventional encryption (for beginners, to facilitate the description of multimedia encryption), can be skipped by the reader if desired.

2 Fundamentals of Conventional Encryption

The unique structural and statistical properties of multimedia data makes its encryption as an extension of conventional encryption. A snapshot of the fundamentals of conventional encryption system is given in this section to facilitate the discussion on multimedia encryption. However, a detailed description on conventional cryptography can be accessed in [1, 2, 3].

Encryption can be applied on bits/bytes or on blocks. To avoid ambiguity bits/bytes are referred as symbols in the current context and encrypted bitstream is referred as codestream.

- M denotes a set called a message space and consists of a string of symbols, text file etc.. A message from M is known as a plaintext, which can be read or understood without any special measures.
- The ciphertext space, C is a set consisting of string of symbols, different than the symbols in plaintext space. The string of symbols in ciphertext space is generally in unreadable gibberish form.
- The elements of keyspace K, i.e. keys are used for converting plaintext into ciphertext in conjunction with the encryption algorithm.
- Each element $e\ \varepsilon\ K$ uniquely determines a bijection from M to C and is denoted by E_e (i.e. $E_e : M \rightarrow C$). E_e is the encryption function or transformation. The process of applying the transformation E_e to a message $m\ \varepsilon\ M$ is usually referred as encrypting m or encryption of M.
- For each $d\ \varepsilon\ K$, D_d denotes a bijection from C to M (i.e. $D_d : C \rightarrow M$). D_d is the decryption function or decryption transformation. The process of applying the transformation D_d to a ciphertext $c\ \varepsilon\ C$ is usually referred as decrypting c or decryption of C.
- An encryption scheme, also known as a cipher consists of a set $\{E_e : e\ \varepsilon\ K\}$ of encryption transformations and a corresponding set $\{D_d : d\ \varepsilon\ K\}$ of decryption

transformations with the property that, for each $e \varepsilon K$, there is a unique key $d \varepsilon K$ such that $D_d = E_e^{-1}$ i.e. $D_d(E_e(m)) = m$ for all $m \varepsilon M$.

- The keys e and d in the preceding definition are referred as a key pair and sometimes denoted by (e,d), where e and d can be same.
- To construct an encryption scheme, one requires to select a message space M, ciphertext space C, key space K, set of encryption transformations $E_e : e \varepsilon K$, and corresponding set of decryption transformations $D_d : d \varepsilon K$. These five parameters together form a cryptosystem represented by $\{M,C,K,E_e,D_d\}$. A fundamental cryptosystem is shown in fig. 1.
- Cryptanalysis involves the understanding of mathematical techniques with an intention to defeat cryptographic techniques, and more generally, information security services.

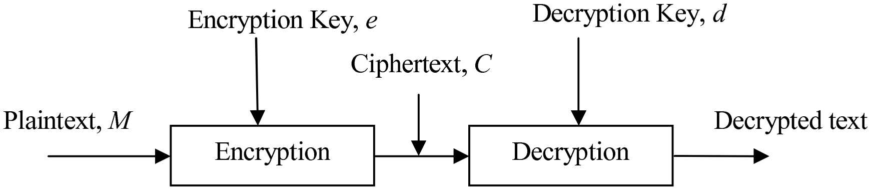


Fig. 1 A practical cryptosystem

- A cryptanalyst is someone who engages in cryptanalysis.
- Cryptology enfolds the area of cryptography and cryptanalysis.

A fundamental premise in cryptography is that the sets forming a cryptosystem, $\{M,C,K,E_e : e \varepsilon K,D_d : d \varepsilon K\}$ are known publicly. When two parties wish to communicate securely using an encryption scheme, the only thing which is kept secret is the particular key pair (e,d), selected mutually by the communicating parties.

There are two general types of cryptosystem : symmetric cryptosystem and public key cryptosystem. In symmetric key cryptosystem, either the encryption and decryption key are same or decryption key can be calculated from the encryption key. Symmetric key cryptosystems are fast, easy to implement in hardware, and are widely used. However, it suffers from key distribution problem wherein it requires the communicating parties to agree upon, and exchange keys securely.

A public key cryptosystem, also known as an asymmetric key cryptosystem, has different encryption and decryption keys. Furthermore, the decryption key cannot be calculated from the encryption key, which is public by nature. A stranger can use the encryption key to encrypt the message, but only the genuine receiver having decryption key will be able to decrypt the message. The receiver does not get any information regarding the sender. In other words, the system provides secrecy at the expense of authentication. The encryption key and decryption key are often called the public key and private key respectively. Public key cryptography is computationally intensive, difficult for hardware implementation, and not very efficient for high speed applications. Hence, we shall focus on symmetric key encryption and its applications for multimedia in this chapter.

2.1 Symmetric Key Encryption

An encryption system is said to be symmetric if for each associated encryption/decryption key pair (e,d), it is computationally easy to determine d knowing only e, and to determine e from d. Based on the operation of a cipher, symmetric key encryption can be distinguished into two types: block ciphers and stream ciphers.

2.1.1 Block Cipher

A block cipher is an encryption scheme which first breaks up the plaintext messages into blocks of fixed size, padded if necessary and then encrypts these blocks, one at a time. Identical plaintext blocks will be converted to a similar ciphertext blocks using the same key. Data Encryption Standard (DES) and Advanced Encryption Standard (AES) are two widely used block ciphers; DES operates on 64 bit block with a key of 56 bits while AES operates on 128 bit block with key sizes of 128, 192 or 256 bits.

Block ciphers can be used as substitution cipher or transposition ciphers. In substitution block ciphers, ciphertext is formed by replacing a block of plaintext with another block. Transposition block cipher produces ciphertext block by simply permuting the symbols in a plaintext block. A simple transposition cipher preserves the number and type of symbols within a block, and thus, is easily cryptanalyzed.

A number of modes having different features and properties can be used in conjunction with block cipher to provide confidentiality for messages of arbitrary length. Most commonly used modes are the Electronic Codebook (ECB) mode, the cipher block chaining (CBC) mode, the Cipher Feedback (CFB) mode and the Output Feedback (OFB) mode.

In the ECB mode, each plaintext block is encrypted into a ciphertext block independently. Ciphertext blocks can be considered as codebooks of plaintext blocks, as identical plaintext blocks will always result in identical ciphertext blocks. In the CBC mode, current plaintext block is XORed with previous ciphertext block, or with an initialization vector, before encryption is performed. The initialization vector is a random block and need not to be secret but, it should be unique for every message encrypted with the same key. In CBC, same plaintext results in different ciphertext block, but still, it has a disadvantage of being sequential in nature.

In the CFB mode, a queue of the size of block is formed by filling it with initialization vectors. This queue is encrypted repetitively with the block cipher and after each encryption, the queue is shifted left by n bits, where, n is same or smaller than the size of queue. The leftmost n bits are shifted out of the queue and XORed with the first n bits of the plaintext to form first n bits of ciphertext. The same n bits are also fed back to the right side of the queue. Both CBC and CFB modes encrypt the data in sequential manner and have same error propagation characteristics. The OFB operates in the same way as that of CFB except that the left shift is a circular shift i.e. the leftmost n bits are shifted back to the right side of the queue.

2.1.2 Stream Cipher

Stream cipher is an important class of symmetric key encryption scheme which operates on streams of plaintext or ciphertext, with one symbol at a time using time varying transformation. Stream ciphers execute at a higher speed than block ciphers, and have lower hardware complexity. In stream ciphers, same plaintext symbol will always be encrypted to a distinct ciphertext symbol due to different encryption transformation for each plaintext symbol. A stream cipher applies simple encryption transformation according to randomly generated keystream using a seed value or chaos based maps.

Stream ciphers are further classified into synchronous and self synchronizing stream ciphers. In synchronous stream ciphers, the next state of cryptosystem is independent of both plaintext and ciphertext, as, each plaintext symbol is encrypted independent of others, thus, no error is propagated from one ciphertext symbol to another. This makes synchronous stream ciphers extremely useful when encrypted multimedia is transmitted over error prone wireless networks. OFB mode in block cipher, RC4 and SEAL are most widely used synchronous stream ciphers.

In self synchronizing stream ciphers, the next state of a cryptosystem is dependent on the previously generated ciphertexts, thus, resulting in limited error propagation. A block cipher in CFB mode can be interpreted as self synchronizing stream cipher. Exclusive-OR (XOR) operation is the most commonly used encryption transformation in a stream cipher.

2.2 Cryptanalysis

While cryptography is the science of securing data, cryptanalysis is the science of analyzing and breaking secure communication. As per the Kerckhoff's principle, it is assumed that except the key, a cryptanalysts has full access to the description of algorithms and the insecure channel over which, a message is transmitted. An attempted cryptanalysis is called an attack and a secure encryption algorithm must withstand the following type of attacks:

1. **Ciphertext only attack:** A cryptanalyst tries to recover the corresponding plaintext or the encryption keys by observing, ciphertext of several plaintexts encrypted with the same key. A cryptosystem vulnerable to this type of attack is considered to be completely insecure.
2. **Known plaintext attack:** A cryptanalyst has access to the ciphertext and associated plaintext for several messages and tries to deduce the key, used to encrypt the messages or, to develop an algorithm to decrypt any new messages encrypted with the same key.
3. **Chosen plaintext attack:** In this case, a cryptanalyst is allowed to choose the plaintext, that is encrypted and, observe the corresponding ciphertext. The cryptanalyst's goal is same as that in a known plaintext attack.

4. **Adaptive chosen plaintext attack:** This is a special case of chosen plaintext attack where a cryptanalyst not only chooses a plaintext that is encrypted but he can also modify his choice based on the results of previous encryption.
5. **Chosen ciphertext attack:** A cryptanalyst can choose ciphertexts to be decrypted and, has access to corresponding plaintexts through the knowledge of decryption equipment and not the decryption key. The objective is then to deduce plaintext from different ciphertexts without having access to the decryption equipment.
6. **Adaptive chosen ciphertext attack:** It is a chosen ciphertext attack where the choice of ciphertext may be modified based on the results of previous decryption.
7. **Exhaustive key search:** It is also known as brute force attack and, has a very high complexity. In this kind of attack, a cryptanalyst tests each of the possible keys one at a time until, the correct plaintext is recognized. This attack can be combined with any of the previous attacks to reduce the number of possible keys.

3 Prototype for Multimedia Encryption

Multimedia encryption has number of issues that are not present in the text encryption. It is a special application of general encryption, in which multimedia data is transformed into an unintelligible or perceptually degraded form. The simplest way to encrypt a 2 or 3 dimensional multimedia data is, to consider it as a 1-D data stream and encrypt with any available cipher like DES, AES, IDEA etc. It is generally referred as naive encryption which provides less security, and sacrifices many desirable features that multimedia applications may require. Moreover, direct application of some encryption algorithm to multimedia data requires high computational power and, introduces delay in real time communication. The desirable requirements and characteristics of multimedia encryption techniques, are discussed hence.

3.1 Desirable Requirements and Characteristics of Multimedia Encryption

Multimedia data unlike traditional textual data, has high data rate, possess components of different importance, are more loss tolerant, and highly adaptable. These unique properties of multimedia data have posed significant challenges to conventional encryption techniques, that were initially designed for textual data. Direct application of conventional encryption techniques to multimedia data may not provide adequate security, and may render performance of some of the advanced technologies. Certain unique requirements are related to each other, while others are mutually competitive. Design of a practical multimedia system may involve tradeoff, and careful balance of conflicting requirements according to specific application. The unique desirable features for a multimedia cryptosystem can be described as:

1. **Complexity:** Multimedia encryption and decryption techniques require a lot of computational resources, time, and power due to processing of quantum of data. In general, multimedia data is efficiently compressed before transmission or distribution to save storage space and bandwidth. The whole process i.e. compression and encryption is performed to save computational resources and make the data secure. On the contrary, if compression and encryption are not applied properly, they can consume a lot of computational resources, battery power, bandwidth, and time. Hence, the complexity of encryption and decryption process becomes an important consideration while designing a practical cryptosystem.
2. **Compression efficiency:** An efficient compression technique is required to store, transmit or distribute bulky multimedia data in resource constrained environment. Compression of multimedia data generates a lot of processing overhead, which may manifest in several ways. Encryption before compression may lower the compression efficiency by modifying well designed compression parameters, or by modifying the statistical properties of multimedia data. Compression efficiency overhead incurred during encryption due to additional headers in compressed codestream for decryption parameters, boundary indicators of encrypted segments, etc. should be minimized.
3. **Perceptibility:** Perceptibility means encryption of multimedia data in a way to make encrypted content partially perceptible without access to the decryption key i.e. some content is allowed to leak out even after encryption. Different applications like pay-TV, video on demand (VoD) services, military or financial applications require varied level of perceptibility for the protected content. The main purpose of multimedia encryption for pay after trial services is to provide content degradation rather than complete secrecy whereas military or financial applications require highest protection level with zero perceptibility. Different level of secrecy requires different multimedia encryption scheme with varied complexity and cost. Multimedia encryption should be designed to meet, the desired perceptibility level with minimal complexity.
4. **Format-compliant:** Many multimedia systems were designed without much consideration of encryption. So, a later add-on encryption technique may not be recognized or supported by existing infrastructure and installed device. It is desirable that the encrypted codestream be compliant to the specific syntax of multimedia data to solve this backward compatibility problem. This will help in performing various content and network adaptations directly on the protected bitstreams, without performing cryptographic operations. This format compliant encryption inherits many carefully designed and desirable properties of unprotected compressed bitstream, such as error resiliency, scalability and protocol friendliness. The newly found JPEG2000 standard tries to retain the format compliant property while protecting the multimedia data [4]. Many format compliant content encryption techniques have been developed recently that address various network friendly end to end security issues for multimedia delivery.
5. **Error resilience:** The avalanche property of encryption scheme propagates a single bit error in ciphertext to almost entire decrypted plaintext, especially, if the cipher employed is synchronous stream cipher or block cipher. It is highly

undesirable that the encrypted stream cannot be decoded when bit errors are introduced, which frequently occur in multimedia applications over wireless networks due to congestion, buffer overflow and other network imperfections. A well designed multimedia cryptosystem should; confine the encryption incurred error propagation to minimize perceptual degradation, enable quick recovery from bit errors, and fast resynchronization from packet losses. Earlier encryption schemes did not consider this error resilience property, however, the concept of error resilient encryption is investigated in [5].

6. **Adaptability & Scalability:** Inspite of the varying characteristics and processing capabilities of different devices catering to multimedia data, it is desirable that the encrypted data adapts to the targeted device. This may require transparency between the encryption technique and the adaptation process, which becomes more difficult in case of fluctuating transmission bandwidth. Scalable coding offers a solution for this by partitioning and organizing the codestream in hierarchical structures, according to some scalable parameter like quality, rate, spatial or temporal features. Based on the scalabilities offered, a user can extract a part of the codestream that best fits his/her application without requiring the receiver to decode the entire compressed and encrypted codestream. The Moving Picture Experts Group (MPEG) and the Joint Photographic Experts Group (JPEG) have recently adopted MPEG-4 Fine Granularity Scalability (FGS) [6] and JPEG 2000 [7] as their respective scalable coding standards. Encryption of scalable multimedia should preserve the scalabilities offered by the underlying scalable codec so that the desirable feature of easy adaptation is not impaired.
7. **Multi-level encryption:** Multimedia content can be encrypted into a single cipher codestream to support simultaneous access; to multiple types based on required quality, resolution, frame size and rate, and to multiple layers for each access type. A desirable feature of multi-level multimedia encryption is that a single encrypted codestream should support multiple accesses, so that different applications can extract a best fit representation as per the processing capabilities and characteristics of the device, i.e. Multi access encryption allows different users to obtain different version of multimedia data from a single encrypted codestream. A user can access only those types and levels that he or she is authorized to. Authorization to access a higher privileged level would allow access to all the lower privileged levels of the same type. Multi level encryption is an elegant enabling technology to support the business model of ”what you see is what you pay” with a single encrypted codestream.
8. **Content agnostic:** Different codecs have been designed to efficiently compress different multimedia data i.e. image, audio or video, and each codec generates its own codestream. Content agnostic desired in some applications is a direct contradiction to the requirement of syntax compliance. Content agnostic multimedia encryption implies that the encryption do not depend on the content type or specific technology used in compression. That is, a single encryption or decryption module can be used to process a wide variety of multimedia types and encoded bitstreams. Microsoft’s Advanced Systems Format (ASF) [8] is a general multimedia format that supports many multimedia types and codecs. Multimedia

encryption in Microsoft's Windows Media Rights Manager (WMRM) [9] and the Open Mobile Alliance (OMA)'s DRM [10] also adopt the content-agnostic approach.

9. **Bandwidth expansion:** Many encryption techniques increases the size of encrypted data as compared to the original multimedia data, which is unsolicited for the storage, transmission, or distribution of multimedia data in a resource constrained environment (limited bandwidth and small storage space). It is highly desirable that the underlying encryption technique should not lead to bandwidth expansion.

3.2 Attacks on Multimedia Encryption

A multimedia cryptosystem should be able to withstand various attacks discussed in section 2.2, especially, the known plaintext attack due to two reasons. Firstly, most of the commercial videos start with a set sequence of frames and secondly, it is easy to predict a local portion in multimedia data; silence in audio stream, smooth portion in image or static portion in a video. In addition to the attacks discussed in section 2.2, various other attacks that a cryptanalyst can explore are:

1. **Approximation Attack [11]:** Traditional all-or-nothing situation in generic data security is not always appropriate for measuring the security of multimedia encryption. It is necessary that intelligible perceptual information is not leaked out from the ciphertext, even if exact recovery is not possible. However, a cryptanalyst can approximately recover the encrypted multimedia content due to high spatial and temporal correlation in image, video or audio data if the syntax, context or statistical information is known a priori. Since, the perceptual quality of multimedia data is closely tied with the value of multimedia content, it is desirable that strongly protected multimedia data be robust against approximation attack. Luminance Similarity Score (LSS) and Edge Similarity Score (ESS) [11] depicts the security of multimedia encrytion technique against approximation attack.
2. **Error concealment Attack [12]:** In this attack, statistical information and knowledge of media format are exploited by a cryptanalyst to achieve a perceptual break in the multimedia cryptosystem. This attack is similar to the approximation attack except the amount of degradation produced. The leftover redundancy after compression process is exploited by a cryptanalyst to conceal perceptual quality degradation caused by bit errors or lost data. Both the approximation and error concealment attacks are applicable to all selective encryption techniques, producing different amount of degradation. Error concealment attack is considered successful, when it gives an inferior quality version of the original data, while approximation attack generally reveals the edges and contours of the images.
3. **Statistical Attack [12]:** A cryptanalyst exploits the predictability of a particular element, or predictable relationships between data segments of original bitstream and cipher codestream. The relationships are exploited to either determine the plaintext without the knowledge of decryption key, or to substantially reduce the

search space such that a brute force attack is feasible. The correlation among different portions of multimedia data is explored to launch this statistical attack and can be easily applied on any selective encryption technique as compared to total encryption technique. A cryptanalyst can use the correlation existing between the unencrypted portion and the encrypted portion to either reverse the encryption, or to deduce information about the encrypted portion. Thus, an efficient compression technique which reduces the correlation existing in multimedia data can be used to defend this attack.

4 Multimedia Encryption Techniques

Due to transmission and distribution of multimedia data over bandwidth limited open natured wired/wireless channel, compression and encryption has become an integral part of multimedia communication. Compression exploits the statistical properties of multimedia data to remove redundancy and to reduce the size of data whereas encryption introduces randomness in data to make it unintelligible. Encryption can be combined with compression at any of the three stages: before compression (A,B,C in fig. 2), during compression (position D in fig. 2) or after compression (position E in fig. 2).

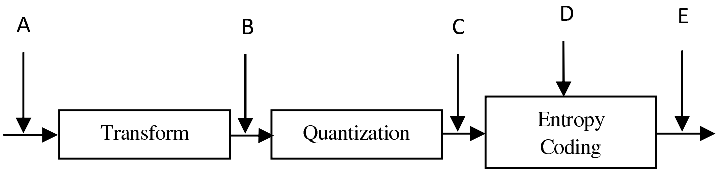


Fig. 2 Encryption domains

Encryption before compression disturbs the structural and statistical properties of multimedia data, leading to much reduced compressibility. However, Johnson et al. [13] has proposed an approach to compress the encrypted data by employing distributed source coding theory. This gives the same compression gain as compressing the unencrypted data in the case of ideal Gaussian source, but the compression gain reduces for commonly used source. Moreover, encryption before compression cannot easily support many other forms of delegate processing and hence, it is not much pursued by researchers.

The second case encryption after compression do not affect the compression ratio but introduces little overhead, which is acceptable for most practical cryptosystems. However, if generic encryption is applied then, the structure and syntax readily available in the unencrypted bitstream may get destroyed. One solution to this can be to encrypt the content carrying fields of the compressed multimedia bitstream such as, fields of motion vectors & DCT coefficients in MPEG video, and to keep the structure & header/marker unchanged.

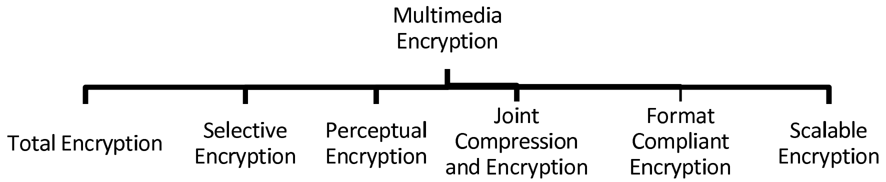


Fig. 3 Categorization for multimedia encryption techniques

The third case i.e. encryption during compression is a joint signal processing and cryptographic approach, which provides encryption and compression during entropy coding stage by modifying the compression parameters. This approach provides sufficient security level without sacrificing compression efficiency or increasing computational load. Most of the practical cryptosystems are developed on the second or third type of combination due to their capability to support advanced functionalities and delegate processing.

Furthermore, multimedia encryption techniques can be applied either in spatial domain or in transform domain. However, the spatial domain based techniques cannot withstand various attacks present in the communication channel. Moreover, important parts cannot be identified in spatial domain based techniques, which are required to achieve advanced functionalities like conditional access, format compliant, support for FGS (Fine Granular Scalability) coding etc. Due to these limitations of spatial domain, encryption in transform domain is widely implemented, and hence discussed here. However, spatial domain encryption is mentioned intermittently to give a complete overview of the concerned topic.

A principal classification of multimedia encryption techniques is given in fig. 3, whereas an informative and precise explanation for these classified techniques is covered in subsequent subsections.

4.1 Total Encryption

In multimedia data, some portions carry more importance as compared to rest of the content; smoothness, silence or still objects carry less information as compared to edges, voice or moving objects in image, audio and video respectively. Total encryption is a simple and straightforward, but it is a time consuming process where entire multimedia stream is encrypted without contemplating the importance of various parts of multimedia data. Since entire multimedia data is encrypted, total encryption tends to achieve highest level of security.

Encryption is usually applied after compression so as not to disturb the compression efficiency. However, Choo et al. [14] has explored full encryption on an uncompressed MPEG-4 video format using XOR and transposition operation, which is not desirable considering the storage and bandwidth constraints.

In the same context, relationship between chaos and cryptography has been exploited by several researchers from the past two decades; a brief overview can be found in [15, 16]. These chaos based methods have been used both, as stream and block ciphers to provide full encryption to the uncompressed multimedia content [17]. However, security provided by these chaos based total encryption techniques is not competitive to the security provided by standard cryptographic algorithms, and hence, loses its robustness against transmission errors [18, 19].

Total encryption in transform domain is usually performed by partitioning and then packetizing the compressed codestream into structured packets. Each packet has a data field which is encrypted using some cryptographic algorithm, and a header field which is generally left unencrypted. Decryption information can be inserted into headers. Microsoft's WMRM [9] and DMA's DRM [10] adopt such an approach. The format to support total encryption is ASF [8] for Microsoft and the DRM Content Format [20] for OMA.

However, the approach of encrypting only data field and keeping header field unencrypted is not completely secure. Since the data field is totally encrypted, it does not leak out any information about multimedia content without the knowledge of decryption key. But, unencrypted headers can be exploited by an adversary to extract basic information of the protected content. The traffic pattern revealed from unencrypted header fields can also help the adversary to speculate general information. To deal with the problem, Dang & Paul [21] proposed end to end encryption of data field and link encryption of whole packet.

Various other encryption techniques are also proposed in literature which provides high security level with support for advanced functionalities like conditional access, FGS, format compliance etc. Grangetto et al. [22] proposed format compliant total encryption for JPEG2000 images at entropy coding stage using randomized MQ coder. The proposed encryption scheme can also be used for selective encryption and conditional access. A multi-access total encryption for encryption of MPEG-4 FGS has been proposed in [23], where data in each video packet is independently encrypted with a cipher.

4.2 Selective Encryption

Unlike total encryption schemes, selective encryption techniques encrypt only a part of data to make the entire multimedia content incomprehensible. The subset of data that needs to be encrypted is generally the crucial/important data from either the final bitstream, or from the intermediate steps of compression process. Encrypting this small amount of crucial data consumes less computational resources as compared to the encryption of a large amount of unimportant data, for achieving same level of degradation.

Although same principle is applied in all the schemes, they still differ from each other, based on data encrypted, criteria chosen to select the crucial data, ways to encrypt the data, the domain used, and the bitstream used (generated by exploiting the inherent properties of the codec). [24, 25, 26, 27] gives a partial review of performance and security aspect of various selective encryption techniques. In this section various multimedia techniques are categorized on the basis of data i.e. image, video or audio data which are discussed in respective subsections.

4.2.1 Selective Encryption for Images

Selective encryption of images can be done in spatial, DCT or wavelet domain. Each domain has its own limitations and properties; spatial domain techniques are not robust to various cryptanalytic attacks, whereas DCT domain based techniques suffer from a problem of blocking artifacts, which arises during compression. Of the three domains, wavelet domain is better, and is widely explored by researchers due to its multiresolution capability, good compression efficiency and support for scalable coding. Representative selective encryption techniques based on these three domains are described hence.

Selective Encryption of Images in Spatial Domain: A simple approach for selective encryption of images in spatial domain to encrypt bitplanes before compression. It is observed that most significant bitplanes of any grayscale image carries more perceptual information than the least significant bitplanes. Podesser et al. [28] and Droogebroeck et al. [29] has exploited this property to propose selective bitplane encryption in spatial domain. Fig. 4 shows the visual results from [28] for the encryption of different bitplanes. It is observed that encryption of only most significant bitplane leaves the structural information visible. Hence, encryption of two most significant bitplanes to hide the structures, and encryption of four significant bitplanes to achieve high level of confidentiality was proposed [28]. Whereas, Droogenbroeck & Benedett [29] has proposed the encryption of bitplanes in reverse order i.e. from the least significant bitplane to most significant bitplane to achieve perceptual degradation rather than complete confidentiality. Since both the encryption schemes are proposed on uncompressed images; so compression of these encrypted images generate a lot of overhead and reduces the compression ratio due to disturbed statistical properties of the image.

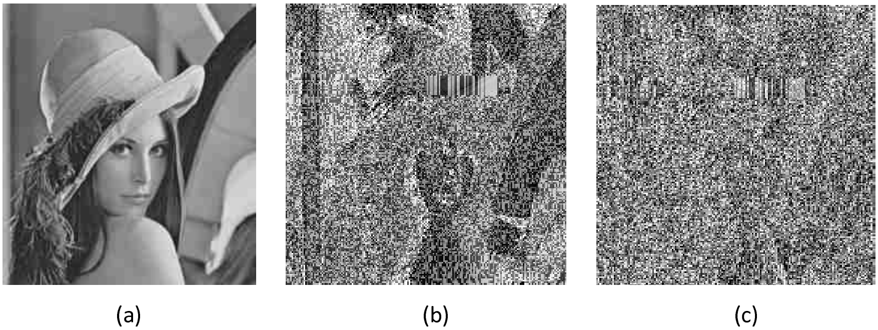


Fig. 4 (a) Original lena image (b) MSB plane encrypted [28] (c) two MSB planes encrypted [28] ©2002 IEEE

Other representative spatial domain techniques implement quadtree image compression before encryption of crucial information. Quadtree compression is computationally efficient and the structure is formed by partitioning an image block in a recursive manner with original image as initial block. Parameters attached to leaf node describe the corresponding block. Cheng & Li [30, 31] proposed to encrypt only the quadtree structure and leaves the parameter attached to leaf nodes unencrypted for quadtree image compression, where only one parameter is associated with a leaf node to describe the average intensity of corresponding block. The unencrypted leaf node values have to be transmitted in some order. In-order traversal of the quadtree make the encryption technique susceptible to cryptanalytic attacks, thus, the ordering which encodes leaf values one level at a time from the highest level to lowest level is recommended.

Chang et al. [32] also proposed selective encryption of image during quadtree compression. In this approach, each internal node of the quadtree structure follows a scan sequence, different from the scan sequence of all other nodes. 24 different scan sequences are defined for accessing four quadrants, thus making $24^n \times 4^{n(n-1)/2}$ possibilities to encode an image of resolution $2^n \times 2^n$, making it robust against brute force attack.

Selective Encryption of Images in DCT Domain: Selective encryption in spatial domain do not support advanced functionalities, which otherwise can be obtained in transform domain with much added security. Transforming an image/video frame from spatial to Discrete Cosine Transform domain [33] in blockwise form (of size 8×8) yields 64 coefficients. The coefficient with zero frequency has excellent energy compaction for highly correlated data and is called DC coefficient, while remaining 63 coefficients are recognized as AC coefficients. Selective encryption techniques in DCT domain exploit the importance of these DC and AC coefficients to obtain varying levels of security.

Encrypting the bitstream of leading DCT coefficients in each DCT block has been proposed by Kunkelmann & Reinema [34], but Cheng and Li [35] reported that even when 50% of the total JPEG compressed file is encrypted, then also confidentiality cannot be achieved as contours remain visible. Tang [36] proposed few permutation based encryption schemes for commercial applications which includes (i) permutation of all AC coefficients while keeping DC coefficient unencrypted, (ii) random permutation of DC coefficients, (ii) making DC coefficient zero and random permutation of AC coefficients, (iii) splitting DC coefficient and then permutation of AC coefficients, (iv) permuting all coefficients and making last AC coefficient as zero. It is observed that simple permutation of AC coefficients, with unencrypted DC introduces only perceptual degradation while permutation of all DC coefficients makes the image incomprehensible.

However, permutation based techniques can be broken easily. Moreover, Uehara and co-authors [37] showed that it is possible to recover DC coefficients from AC coefficients with reasonable quality. Thus, encryption technique proposed by Tang do not provide adequate security and are weak against known plaintext and chosen ciphertext attacks.

Contrary to the encryption of DC coefficients, a technique has been proposed by Droogenbroeck & Benedett [29] to encrypt all DCT coefficients except the DC coefficient, or DC coefficient plus some AC coefficients of low frequency. This is achieved by encrypting the sign and magnitude of non-zero AC coefficients, and leaving the portion encoded from Huffman table, unencrypted. This scheme does not provide confidentiality but just introduces perceptual degradation in the image. This scheme was further extended by Droogenbroeck [38] to propose the concept of multiple selective encryption, where a part of original image is encrypted by one owner while the other part is encrypted by the second owner. This scheme offer advantages like flexibility, multiplicity, spatial selectively and format compliance, but requires a trade-off between processing power and speed.

Apart from these grayscale selective encryption technique, Shiguo et al. [39] proposed selective encryption of colored JPEG images by introducing AC coefficient confusion between subsections, sign encryption of DCT coefficients using chaotic stream cipher or permutation of blocks in luminance and chrominance plane using Space Filling Curves (SFC). The encryption scheme is secure against known plaintext attacks, supports direct bit-rate control but the use of SFC affects the compression ratio slightly.

Selective Encryption of Images in Wavelet Domain: The wavelet transform forms a pyramid decomposition having coefficients at different hierarchy level. These coefficients often have correlation among themselves, and can be grouped together to form a zerotree, indicating significant and insignificant coefficients/sets. Selective encryption in wavelet domain is mainly done using this significant or insignificant information obtained during wavelet compression.

A partial encryption scheme for Set partitioning in Hierarchical Trees (SPIHT) was proposed by Cheng & Li [31], which encrypts only the bits related to significant pixels and sets in two highest pyramid levels, as well as the parameter n that determines the initial threshold to form List of Significant Pixels (LSP), List of Insignificant sets (LIS) and List of Insignificant Pixels (LIP). Compression performance of SPIHT is not affected by this encryption scheme but it requires a deep parsing in the compressed bitstream during encryption and decryption. This encryption scheme was further extended for selective encryption of videos. However, Said [40] measured the security of partial encryption technique proposed by Cheng & Li [31] and reported that it may not be easy to get the decryption key or the original information but significant improvement in quality can be achieved with relatively low complexity by using information present in non-encrypted data.

On the similar concept, AES encryption of complete tree structure formed during wavelet packet decomposition was proposed by Pommer et al. [41]. The amount of data encrypted is very less as compared to other selective encryption techniques due to encryption of only header information instead of any visual data. Use of decomposition tree, formed by best basis algorithm share common features for many images which would consequently facilitate an attack leading to potential security weakness. Thus, PRNG approach is used to generate the tree structure for encryption instead of following the best basis algorithm. Moreover, the proposed approach is not secure when uniform quantizer is used to encode the coefficients. But, if zerotree coder is used instead of uniform quantizer then the difficulty for attack becomes higher, complexity of launching brute force attack increases and compression rates for a given image quality improves.

Secret permutation of wavelet coefficients in each subband of wavelet compressed image was proposed by Uehara et al. [42]. Lian et al. [43] also proposed an alternative scheme where coefficients among child nodes sharing the same parent are permuted for the quadtree structure of the wavelet decomposition of an image. This can be enhanced by encryption of lowest subband with a cipher.

First selective encryption technique for JPEG 2000 compressed images has been presented by Grosbois et al. [44] to achieve conditional access. The authors proposed to pseudo-randomly invert some bits in the coding passes of last layers, i.e., those that contribute detail to the image. A decoder knowing the seed of the random sequence (i.e., the key) can undo the scrambling and correctly decode such layers; otherwise, attempts to decode the protected layers will impair the obtained visual quality. Another scheme for JPEG2000 was proposed in [45], which performs encryption during entropy coding stage by replacing the default MQ coder with private initial table, generated using a secret key and a mapping function. The proposed scheme substitutes each initial index of 19 context labels with one of the 47 states, with the help of a mapping function. The scheme does not affect the compressibility of JPEG2000 coder, and can achieve perceptual encryption as well as region of interest encryption. It qualifies for both format compliant and joint compression & encryption scheme.

Other than the mentioned selective encryption schemes for grayscale images, Martin [46] proposed an efficient encryption scheme for colored images using C-SPIHT compression algorithm. The algorithm encrypts only the significant bits of the individual coefficients encountered during the first K sorting passes of C-SPIHT algorithm, instead of encrypting the entire tree or part of the tree. The parameter K can be controlled to make a trade-off between confidentiality and processing overhead.

4.2.2 Selective Encryption for Video

Due to bulky size of video data, it is usually transmitted in compressed format such as MPEG-1, MPEG-2, H.263, MPEG-4 or Motion JPEG2000. Thus, encryption algorithms for digital video usually work in the compressed domain. The similarity between compression technologies used for images and video do not make any image encryption technique applicable to video encryption, without modifications.

In selective encryption, crucial information is encrypted and not so important information is kept unchanged. As video encryption evolved, different researchers perceived different portions of video format as important for encryption purpose, that can yield desired security in less computational complexity and time. Video encryption schemes are thus categorized based on the content encrypted such as headers, motion vectors, frames, macro-blocks, DCT coefficients etc. or the levels of security provided instead of the domain, as in the case of image encryption.

- **Header and prediction based encryption:** According to MPEG, Intracoded frames (I-frame) are considered as a reference frames, and the reconstruction of predicted coded frames (P-frames) and bidirectional coded frames (B- frames) are dependent on the availability of the preceding I-frames. It is assumed that if attackers are not able to retrieve original I frame from its encrypted counterpart, then it will be difficult to reconstruct the remaining P and B frames even if they are unencrypted.

  Maples & Spanos [47] and Li et al. [48] has used this idea to encrypt only I-frames in the video data. Large amount of data needs to be encrypted in this case because I-frames are not predicted and hence, requires a large number of bits per picture as compared to P- or B-frames. The amount of data to be encrypted can be reduced by reducing the frequency of I-frames, but that will lead to long delay in switching channels or prolonged perceptual distortion when packet loss occurs.

  Agi and Gong [49] argued that the basic idea of encrypting only I-frames is not correct. They showed that a large portion of encrypted video is still visible even in the absence of I-frames, especially, for a video sequence with high degree of motion. This content leakage is due to the unencrypted I-blocks in P- & B-frames and partly because of the inter frame correlation. This may suit to certain applications like pay-per view but not in the applications requiring better security. To increase the security level, Agi and Gong [49] proposed (i) to increase the

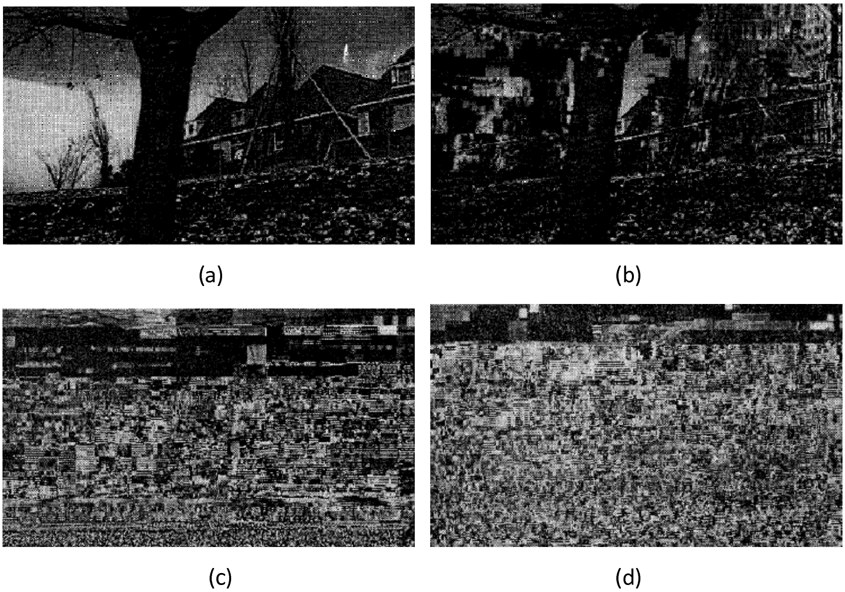


Fig. 5 (a) Original frame [48] (b) encryption of only I-frames [48] (c) Encryption of only I-frames with higher I-frame frequency [49] (d) encryption of I-blocks in all frames [49] ©1996 IEEE

frequency of I-frames, or (ii) to encrypt the I-blocks in P- and B-frames along with the I-frame encryption. Results from [49] are shown in fig. 5 and depicts that both the proposed techniques are better than only I-frame encryption, proposed in [48]. However, a trade-off is required between desired security and computational load or complexity. Spanos and Maples later suggested encrypting not only I-frames, but also encryption of all headers as improvements. Header encryption has an advantage of providing security in low complexity.

Another method to increase security and to reduce the amount of data to be encrypted was proposed by Qiao and Nahrstedt [50, 51]. The bytes of each I-frame are divided into two streams, one consisting of bytes at odd indices and the other consisting of bytes at even indices. The first half is replaced by XOR of the two streams and the second half is encrypted using any standard cryptographic algorithm like DES. Low correlation between bytes in the MPEG bitstream makes this approach quite secure.

- **Hierarchical encryption:** Several hierarchical selective video encryption algorithms are proposed by researchers which provide varied levels of security. The user can choose the amount and content of data to be encrypted as per the desired level of security. Meyer and Gadegast [52] proposed SECMPEG having four levels of implementation with high speed software implementation. The algorithm for lower level is always a subset for the algorithm level immediately above it. The first level includes only header encryption while in the second level the algorithm encrypts parts of the I-block in addition to the implementation in the first level. In the third level, all the I-frames and all the I-blocks of P- and B-frames are encrypted while in the fourth level, the algorithm works in the same manner as that of a naive algorithm.

  Another hierarchical encryption algorithm was proposed by Alattar et al. [53] having three levels of security. In the first level, encryption is applied to data associated with every n^{th} I-macroblock. In the second level, headers of all predicted macroblocks are encrypted along with the data associated with every n^{th} I-macroblock. To reduce the computation load, third level encrypts the headers of every n^{th} predicted macroblock in addition to the data associated with every n^{th} I-macroblock. Similarly, Li et al. [54] proposed three levels of security, first level considered encrypting only I-frames of MPEG video stream with standard PGP (pretty good privacy) encryption algorithm. Second level included encryption of I-frames and P-frames while the highest level included encryption of all the frames.

- **Encryption of DCT coefficients and motion vectors:** Another approach for selective encryption for videos is to selectively encrypt the DCT coefficients or motion vectors of MPEG video. Shi and Bhargava [55] proposed Video Encryption Algorithm (VEA) which uses a secret key to randomly flip the sign bit of all the coefficients. The sign bits are flipped by XORing sign bits with a keystream constructed by repeating a pseudo-randomly generated bitstream of length m.

Repetition of generated pseudorandom bitstream makes this scheme vulnerable to known plaintext and chosen plaintext attacks.

A modification in VEA (MVEA) was suggested by the same authors [56], in which sign bits of DC coefficients in I-frames and sign bits of motion vectors in B- and P-frames are flipped by XORing with a secret key. However, the security of MVEA is weaker than VEA. Security improvement was then achieved by real-time VEA (RVEA) [57], which is a combined version of VEA and MVEA. In RVEA, the XOR operation was replaced by some standard block cipher. For each macroblock, atmost 64 coefficients are encrypted from low frequency to high frequency. In these techniques, search space is not large, thus making them vulnerable to brute force attacks. Moreover, these schemes only introduces perceptual degradation rather than providing confidentiality. Computational time is also more due to deep parsing into the compressed bitstream for extraction of selected sign bits only, both during encryption and decryption.

- **Permutation based encryption:** Several schemes have been proposed which encrypt video data by permuting randomly selected blocks, macroblocks, coefficients or motion vectors. An early work in this direction was proposed by Tang [36] to use a random permutation list for replacing the zig-zag scan sequence in mapping the 8 × 8 DCT block into a 1 × 64 vector before run length encoding for each block. The DC coefficient is split into two halves, with the highest AC coefficient of the block set to the higher half of this splitted DC coefficient. However, the splitting and merging both during encryption and decryption adds computational overhead to the approach and the use of a non zig-zag scanning results in reduction of compression efficicncy [58].

  Zeng and Li [59] also proposed a scheme in which sign encryption of DC coefficients is combined with the permutation of DCT coefficients at the same frequency location within each segment consisting of several 8 x 8 blocks or macroblocks. Same authors proposed selective encryption of sign bits along with the permutation of blocks (formed by partitioning each subband into same sized non-overlapping blocks) for wavelet based video compression. Permutation is also used by Wen et al. [60] which also qualifies for a format compliant encryption and will be discussed in the next section. But, security of only permutation based schemes is not very high, as these schemes are vulnerable to various kinds of attacks such as known plaintext and chosen plaintext attacks, if permutation table is not changed frequently [51].

- **Miscellaneous Schemes:** Another popular encryption scheme called CVES was proposed in [61] which employs multiple chaotic maps for encryption of video data. A simple chaotic stream cipher and a block cipher are combined to construct a fast and secure product cipher. In this technique, each plaintext block is first XORed by a chaotic signal pseudo randomly selected from the Encryption Chaotic System (ECS) and then substituted by a pseudo-random S-box generated from all chaotic orbits of ECS. This scheme is competitive against most traditional ciphers and is little slower than AES reference code. However, to increase the encryption speed, avalanche property is sacrificed in the process.

Cheng et al. [31] proposed the encryption of both residual error and motion vectors, which represent the leaf values of quadtree decomposed structure. The proposed use of Multiple Huffman Trees (MHT) [62] during compression of video data can also be classified under joint compression and encryption techniques and hence discussed in the later sections.

An approach of randomly corrupting MPEG video information before distribution was proposed by Griwodz et al. [63, 64]. The correct bytes are sent to an authorized user at the time of content decoding. The author observed that encryption of only 1% of total data could degrade the quality of data. A similar approach SURSLE was described by Wen et al. [60] as a part of MPEG-4 IPMP (Intellectual Property Management and Protection Extension) standard. In this approach, X consecutive bits are encrypted followed by Y unencrypted bits which are then followed by Z encrypted bits and so on. But this approach is format incompatible and it is not necessary that the encrypted bits are always the crucial bits like in other selective encryption techniques.

A Region of Interest (ROI) based encryption is also proposed by Dufaux et al. [65] for MPEG-4 or Motion JPEG2000 videostream. It randomly flips the signs of wavelet coefficients belonging to AC subbands and ROI. But this technique requires the ROI to be transmitted to receiver as metadata (in case of MPEG-4) either as a part of codestream or on a separate channel, which leads to increase in computational overhead, whereas the shape can be implicitly embedded in JPEG2000 using ROI mechanism.

4.2.3 Selective Encryption of Audio

Speech and audio signals are an essential component of most multimedia applications. It is an important part of the huge telephone industry and a component of advanced audiovisual services such as videoconferencing and news broadcasting. Encryption on audio signals is performed after compression to save bandwidth during transmission. MP3 is the most commonly used compression format that outputs frames consisting of frame header, side information, and main data. Main data further contains the sclaefactor and Huffman codeword values for each subband.

Thorwirth et al. [66] proposed selective encryption for MP3 bitstreams in which equal sized blocks are formed, by taking portions of compressed main data associated with the most critical frequency. These equal sized blocks are then encrypted using standard block ciphers and encrypted bitstream is then reassembled in order to preserve the MP3 format. Since encryption is applied after compression, this scheme has very little compression efficiency overhead, but it requires a deep parsing into the MP3 bitstream to identify crucial data.

Torrubia and Mora [67, 68] exploited the least significant bits of scale-factor to produce the ciphered stream. In this scheme, j least significant bits of the total scale-factor bits are encrypted by XORing them with a pseudo-random number generator (PRNG), the value of j is decided on the basis of quality loss parameter. In addition to this, a PRNG can also be used to select a new codeword from the same codebook and replace any original Huffman codeword with this newly selected codeword.

It is well known that most of the spectral energy of audio signal is concentrated from 20 Hz to 14 kHz, and MP3 encoder usually maps this segment into a big value region. Moreover, Modified Discrete Cosine Transform (MDCT) coefficients are also partitioned into several frequency regions during Huffman encoding. Based on this, a different perceptual encryption for MP3 audio is described by Servetti et al. [69]. The big value region is divided into three sub-regions, called region 0, region 1 and region 2 with increasing frequencies, and encoded with different huffman tables. Encryption of one bit out of 20 in region 1 is sufficient to produce perceptual degradation, if region 0 is left unencrypted [69]. This is just 1.1% of the total bitstream encrypted for a 128kbps stereo bitstream at 44.1 kHz. However, if both region 0 and region 1 are kept unencrypted then 70 to 100 bits of region 2 have to be encrypted, which is 6.3-8.3% for a 128kbps stereo stream at 44.1 kHz. By making the index of encrypted streams as "not used" in Huffman table and setting the big value number to its maximum, MP3 player would be able to play this encrypted bitstream at a degraded quality.

4.2.4 Issues with Selective Encryption

The main aim of selective encryption is to reduce computation complexity and load while securing the multimedia data. Certain issues that needs to be considered before performing any selective encryption technique are:

- Selective encryption techniques do not strive for maximum security but trade security for computational complexity.
- Encryption techniques based on only permutation of multimedia content are not secure; known plaintext and chosen plaintext attacks can easily be launched.
- Inappropriate choice of selective encryption may lead to same computational overhead as that of total encryption, which gives more security and high level of confidentiality. This is due to the processes involved in selecting and then encrypting the crucial part of the multimedia content which requires deep parsing into the compressed bitstream.
- The statistical properties of multimedia data gets altered, if selective encryption is performed before compression or during compression, thus, resulting in lower compression efficiency.
- It is not necessary for all selective encryption schemes to be format complaint. With the lack of format compliance property of encrypted multimedia, encryption unaware players or standard-compliant players are not able to give even the degraded version of multimedia data.

4.3 Perceptual Encryption

The main purpose of perceptual encryption is to produce a cipher codestream that is degraded, yet recognizable or playable version of the original multimedia content without decryption. It aims at decreasing the perceivable quality of data by encryption of smallest subset of multimedia content, and requires to fulfill the syntax-compliance property to be playable on a standard player. The desired amount of degradation can be obtained by controlling two parameters: zone of encryption and quality factor. Quality factor specifies the level of degradation to be produced in multimedia content while zone of encryption (not in audio data) specifies the visual regions where encryption should be applied. Various perceptual encryption techniques for image, video and audio data have been developed, some of which are described in previous sections [28, 29, 45, 67, 69].

Yekkala et al. [70] proposed lightweight encryption of images in bit domain by encrypting only those blocks that contain edges. In the proposed scheme, image is divided into non-overlapping fixed size blocks and the blocks having number of bits greater than the predefined threshold are encrypted. The perceptual degradation can be controlled by adjusting the threshold value.

Torrubia and Mora [71] proposed another perceptual encryption scheme for JPEG images in which Huffman codewords, used to encode AC coefficient are replaced by alternative codewords. Sign bit encryption and permutation of four wavelet coefficients belonging to the same parent node in the quadtree structure of wavelet decomposition is used to provide perceptual encryption.

Similarly, Lian et al. [72] proposed sign bit encryption, bit plane permutation and inter block permutation to have perceptual encryption in JPEG2000 codestreams. Lian et al. [73] also proposed perceptual encryption for MPEG-4 bitstream. Video Object Layers (VOLs) from the highest layer to the lowest layer for each video object, code blocks in each video object plane from the highest subband to the lowest subband, and bitplanes in each code block from the least significant bitplane to the most significant bitplane are encrypted. In both the schemes mentioned above, quality factor controls the data to be permutated or encrypted.

4.4 Joint Compression and Encryption

Encryption can be carried out at different levels inside the multimedia encoder: the image level, transform level, quantization/bitplane level, entropy coding level, and the codestream level [74]. Encryption at the image, transform, and quantization/bitplane levels (described in previous sections) may reduce coding efficiency due to alteration in the statistics of data that is input to entropy coder. In entropy coding level encryption, new entropy coders try to change the compression parameters or procedures, to achieve the purpose of compression and encryption in a single step, instead of encoding the data in a fixed and public manner. These entropy coders provide encryption by using secret keys to encode the data in such a way, that an adversary should not be able to decode the data without the secret key. This approach is often combined with selective encryption for greater efficiency.

Joint compression and encryption will only be successful if the resulting system 1) does not considerably reduce compression rate 2) require less processing time than compression followed by encryption and 3) can provide demonstrable security. Due to inherent advantages of arithmetic coder and Huffman encoder, these are widely used in joint compression and encryption techniques, and hence, encryption techniques can be broadly categorized on the basis of coder used.

4.4.1 Arithmetic Coder Based Techniques

Due to high coding efficiency and the capability to employ adaptive coding strategy, Arithmetic Coding (AC) is being adopted by plenty of recent standards like H.264/AVC and JPEG2000. Thus, many schemes of joint compression and encryption are based on arithmetic coding.

Barbir [75] proposed an encryption scheme where an arithmetic coder with random adaptation instants is employed, thus, making the decoder unable to track the source statistics computation. The proposed technique though solves the problem of attacks to AC employing non-random adaptation, but still exhibits a performance loss (reduced coding efficiency) as compared to non-encrypted arithmetic coder. Moreover, the implementation is dependent on selected adaptation strategy as, encryption is performed in statistical modeler and not in AC.

Another encryption scheme based on AC was proposed by Grangetto et al. [22] by introducing randomization during the arithmetic coding stage i.e. encryption using Randomized Arithmetic Coding (RAC) for JPEG2000 coded images. In conventional arithmetic coding, the first occurrence of Least Probable Symbol (LPS) or Most Probable Symbol (MPS) is decided in advance between the encoder and decoder. Whereas, in the proposed technique, first occurrence of LPS or MPS changes with coding of every bit, and is decided by a random number (either 0 or 1) generated using a seed value. This random number acts as a key to encryption algorithm and is required for synchronization between encoder and decoder for exact decryption of multimedia data. This scheme can be used to support total encryption, selective encryption and conditional access. However, Jakimoski et al. [76] has argued that encryption using RAC approach do not have any advantages over standard approach in terms of security and compression efficiency, but the author has not shown any cryptanalytic attack.

4.4.2 Huffman Encoder Based Techniques

Another type of joint compression and encryption schemes concentrate on Huffman encoder, as decoding Huffman coded bitstream without the knowledge of coding table is very difficult. Gillman et al. [77] kept the used Huffman table as confidential to achieve encryption, but this approach is vulnerable to known plaintext and chosen plaintext attacks. Ciphertext-only attacks are also feasible due to the limited number of available Huffman tables [78].

Security can be improved by increasing randomness in using *m* statistical models instead of one statistical model [79]. A random sequence is used to select one of the *m* statistical model to encode an incoming symbol. Wu et al. [79] suggested to generate large number of Huffman tables by using Huffman tree mutation technique. Though, it is difficult to manage large number of Huffman tables but this scheme is secure to ciphertext only & known plaintext attacks, and vulnerable to chosen plaintext attacks.

4.5 Format-Compliant Encryption

End to end security requires that content adaptation be performed directly on the protected content. To achieve this goal, content protection solution must be designed to ensure that the encrypted bitstream is still compliant to the syntax specifications of a specific format such that an encryption unaware format compliant player can play the encrypted bitstream directly without crash, although, the rendered content may be unintelligible.

In format compliant encryption, encryption operations are executed intelligently on the compressed bitstream, and full bit-level compliance to the compression syntax can be maintained. It would support many carefully designed and desirable properties of the unprotected compressed bitstream, such as error resiliency, scalability and protocol friendliness. In addition, many random access, network bandwidth adaptation, and error control techniques that were developed for unprotected bitstream, would still work with the protected bitstreams.

Some of the schemes explained earlier, also qualifies for format compliant encryption [36, 45]. Pazarci and Dipcin [80] proposed a scheme to scramble multimedia data before compression to achieve syntax compliance by partitioning a video frame into scrambling blocks and applying a linear transformation to pixel values in each scrambling block consisting of one or more MPEG macroblocks (MB). MPEG compression applied on this encrypted video adversely affects the compression efficiency, and the scheme is vulnerable to known plaintext attacks.

Recently, a format compliant selective encryption and scrambling framework was developed in which spatial shuffling of basic shuffling unit is done in compressed bitstream [60]. This basic shuffling unit can be fixed length coded (FLC) 8×8 block or Variable Length Coded (VLC) run level codeword. FLC encryption is done on FLC-coded DCT sign, dquant (difference of quantization step size between current and previous MB) and intra DC information while VLC encryption is performed on motion vectors. Results from [60] are shown in fig. 6 which shows that both FLC and VLC fields should be encrypted to provide maximum scrambling. This scheme is vulnerable to error concealment attack and has relatively small bitrate overhead, as it is simply a cryptographic key based reorganization of already compressed bitstream.

Wen et al. [81] also proposed a format compliant based encryption in which a fixed length index is associated with each variable length codeword to encrypt the indexes and to map them back to codewords. This approach works well with Huffman and Golomb codes but still has problem of emulated markers and reduced

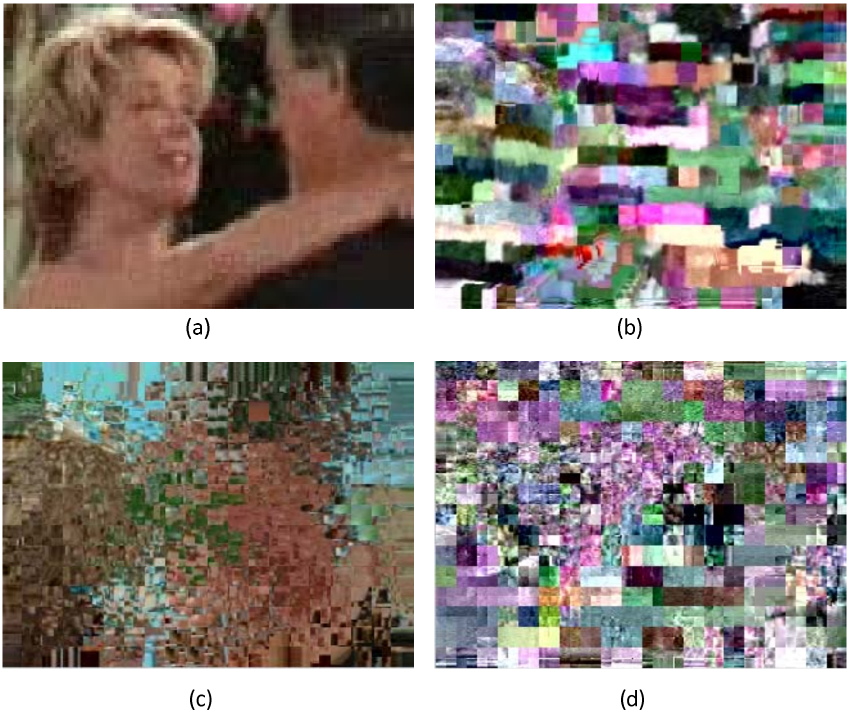


Fig. 6 Results for flower sequence(a) Original sequence [60] (b) only Intra DCs and DCT signs are encrypted [60] (c) only MVs are encrypted [60] (d) Intra DC's, DCT signs and MV's are encrypted [60] ©2002 IEEE

coding efficiency. Also, the concept do not apply to entropy coders with fractional codeword length such as arithmetic coders.

Several other format compliant encryption schemes have been proposed for JPSEC, the security part of latest international still image compression standard JPEG2000. These format compliant techniques are developed to ensure backward compatibility with JPEG2000 part 1 to maintain important properties such as scalability and error resilience. Random flipping signs of wavelet coefficients in high frequency subbands was proposed in [44] to maintain format compliance. Encryption parameters can be inserted after the last termination marker of a code block to exploit the fact that bits appearing after a termination marker will not be read by a compliant JPEG2000 entropy decoder [44, 82]. But, this scheme has a drawback that if data of last coding pass is lost alongwith the encryption parameters during transmission, then received data may not be decrypted, thus, resulting in error propagation.

The parts of JPEG2000 that can be exploited for format compliant encryption include encrypting the data in each individual packet [44, 83, 84], code block contribution to a packet or codeword segments [83, 85], with mechanisms to prevent marker emulation i.e. to prevent any two consecutive bytes of the encrypted bit-stream to assume a value in the interval of [0xFF90,0xFFFF], which are reserved for use by JPEG2000 markers.

4.6 Scalable Encryption

Scalable coding encodes a signal into a single codestream which is partitioned and organized in a hierarchical structure according to certain scalable parameters like quality, resolution etc. The best representation that fits the specific application can be extracted from the codestream based on the scalabilities offered. JPEG and MPEG has recently adopted wavelet based scalable image coding format, JPEG2000 and scalable video coding format called FGS for MPEG-4 standard respectively. All these scalable codecs offer FGS.

In MPEG-4 FGS, a video sequence is compressed into a single stream consisting of two layers: the base layer and the enhancement layer. The base layer is a non-scalable coding of video sequence at the lower bound of a bitrate range. The enhancement layer encodes difference between original sequence and reconstructed sequence from the base layer in a scalable manner, to offer a range of bitrates for the sequence. Scalable encryption can be applied as full encryption as well as selective encryption.

Early scalable encryption techniques were based on layered approach like, Tosun and Feng [86] partitioned the DCT coefficients into three layers: base, middle and enhancement layers. Encrypting only base layer provides minimum protection and, the level of protection increases with the encryption of middle and enhancement layers. Encryption of different layers can be carried out independently either with the same key or a different key.

Various scalable encryption techniques are also proposed for JPEG2000 such as Norcen and Uhl [82] has proposed to encrypt data in each independent packet using standard ciphers like AES or DES in cipher feedback mode without any size expansion while Zhu et al. [85] proposed the encryption of bitstream of each codeword segment in each code block obtained from the coding passes of most significant bitplane to those of least significant bitplane. The encrypted bitstreams are then partitioned and distributed to different layers. But cipher bitstream generated by this method may cause marker emulation problem, which may lead to erroneous parsing or synchronization, especially under error prone environments. A simple solution to avoid this problem is to encrypt the data in each packet, CCP or codeword segment resulting in a syntax compliant cipher codestream [87].

Several scalable encryption techniques are proposed for MPEG-4 bitstream as well. A lightweight syntax compliant selective encryption scheme is proposed in [88] to encrypt MPEG-4 FGS codestream. Different encryption schemes are used to encrypt the base layer and the enhancement layer so that each scheme can be designed to fully exploit the features of each layer. The base layer is fully or selectively encrypted and enhancement layers are encrypted by sign bit flipping of DCT coefficients and scrambling of motion vector (MV) sign bits, or MV residues in FGST-VOP (Video Object Plane). However, results from [88] shown in fig. 7

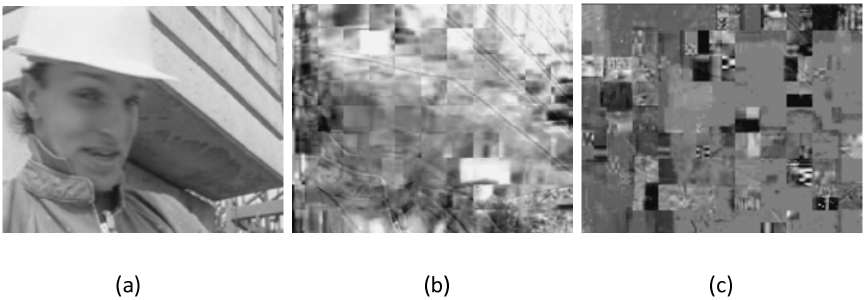


Fig. 7 (a) original foreman image [88] (b) selective encryption mode [88] (c) total encryption mode [88] ©2005 IEEE

depicts that there is a problem of content leakage if selective encryption is given to the base layer.

A full encryption scheme for MPEG-4 FGS is proposed by Yuan et al. [23] which applies Chain & Sum (C & S) cipher to encrypt the video data in each Video Packet (VP) independently. The scalable granularity is reduced to a VP level after encryption. In other words, an entire VP is either dropped or maintained in an adaptation manipulation on an encrypted codestream. Any ciphertext error in an encrypted VP renders the whole decrypted VP unusable, no matter where the error occurs in a VP. This is due to the dependency on the whole ciphertext in decryption with the C & S cipher.

An improved version with scalable granularity smaller than a VP and better error resilience was proposed in [89], which operates in two modes: the Video Packet Encryption mode (VPE) and the Block Encryption mode (BE). In the VPE mode, compressed data in each VP are independently encrypted with a syntax-compliant cipher which produces ciphertext that neither emulates the VP delimiters nor increases the length. In the BE mode, the compressed enhancement data of each 8 x 8 block or macroblock (MB) are independently encrypted with a stream cipher or a block cipher operating in the OFB mode bitplane-wise from the Most Significant Bit (MSB) to the Least Significant Bit (LSB). The scheme provides a decent solution to typically competitive requirements of fine scalable granularity and minimal impact on compression efficiency.

Most scalable encryption schemes use a single key to encrypt an entire scalable codestream, while DRM protected scalable codestream implies that multiple keys are needed to encrypt a scalable codestream to support ”what you see is what you pay”. It requires efficient key generation and management system to allow different users to access different access types and levels of a single encrypted codestream.

4.7 Miscellaneous Techniques

In recent times, multimedia encryption techniques for varied applications has been explored in various domains and combinations.

- Lian et al. [90] proposed a commutative watermarking and encryption scheme based on H.264/AVC compressed codestream which uses 128-bit AES cipher to selectively encrypt the Motion Vector Difference (MVD), intra-prediction mode (IPM) and residue data. Stream cipher is used to encrypt the watermark implementing quantization embedding. The distortion introduced in the video data makes replacement attack and Said's attack, difficult to launch.
- Sun et al. [91] discussed the security issues and proposed a joint fingerprinting and encryption method, incorporated with advanced access control system and traitor tracing. Solutions like multimedia encryption at certain points using 128 bit block based AES cipher and rewritable fingerprint embedding are suggested to deal with some multi-collusion attacks.
- Chen et al. [92] introduced a new Fractional Wavelet Packet Transform (FWPT) to selectively encrypt the images. The fractional orders and wavelet packet filters serve the purpose of encryption keys. The authors reported that FWPT is more effective than Wavelet Packet Transform (WPT) coding, can realize multilevel decomposition and can secure images more flexibly than the encryptions based on wavelet transform or fractional wavelet transform.

5 Conclusion

The increased use of multimedia data transfer over open natured wired/wireless communication channel has made encryption as its integral part, to make the data confidential, and prevent its unauthorized access. Fundamentals of conventional encryption system are explained to facilitate the discussion of multimedia encryption. This chapter gives a brief overview of the concept, desirable features and possible attacks on multimedia encryption. These techniques are then categorized and explained with the support of extensive literature survey. It can be concluded that multimedia encryption has progressed a lot and still has potential, especially in the direction of scalable, format-compliant and joint compression/cryptographic techniques.

References

[1] Schneier, B.: Applied cryptography: protocols, algorithms, and source code in C. John Wiley & Sons, New York (1996)

[2] Menezes, A.J., van Oorschot, P.C., Vanstone, S.A.: Handbook of applied cryptography. CRC Press, FL (1996)

[3] Mao, W.: Modern cryptography: theory and practice. Prentice Hall PTR, Upper Saddle River (2003)

[4] Wu, M.D.: Efficient and secure encryption schemes for JPEG2000. In: IEEE Int. Conf. Acoustics, Speech, and Signal Processing, Montreal, Quebec, Canada, vol. 5, pp. 869–872 (2004)

[5] Ali, S.T., Wu-chi, F.: Lightweight security mechanisms for wireless video transmission. Proc. IEEE Int. Conf. Information Technology: Coding and Computing, 157–161 (2001)

[6] ISO/IEC, Coding of Audio-Visual Objects, Part-2 Visual, Amendment 4: Streaming Video Profile. ISO/IEC 14496-2/FPDAM4 (2000)

[7] ISO/IEC, Information Technology-JPEG, Image Coding System, Part 1: Core Coding System. ISO/IEC 15444-1:2000 (ISO/IEC JTC/SC 29/WG 1 N 1646R) (2000)

[8] Microsoft. Advanced Systems Format (ASF) Specifications, `http://www.microsoft.com/windows/windowsmedia/format/asfspec.aspx`

[9] Microsoft. Architecture of Windows Media Rights Manager, `http://www.microsoft.com/windows/windowsmedia/howto/articles/drmarchitecture.aspx`

[10] Open Mobile Alliance (OMA), OMA DRM Specification v2.0 (2004), `http://www.Openmobilealliance.org`

[11] Yinian, M., Min, W.: A joint signal processing and cryptographic approach to multimedia encryption. IEEE Trans. Image Processing 15(7), 2061–2075 (2006)

[12] Lookabaugh, T., Sicker, D.C., Keaton, D.M., Guo, W.Y., Vedula, I.: Security analysis of selectively encrypted MPEG-2 streams. In: SPIE. Conf. Multimedia Systems and Applications VI, Orlando, vol. 5241, pp. 10–21 (2003)

[13] Johnson, M., Ishwar, P., Prabhakaran, V., Schonberg, D., Ramchandran, K.: On compressing encrypted data. IEEE Trans. Signal Process 52(10), 2992–3006 (2004)

[14] Euijin, C., Jehyun, L., Heejo, L., Giwon, N.: SRMT: A lightweight encryption scheme for secure real-time multimedia transmission. In: Proc. IEEE Int. Conf. Multimedia and Ubiquitous Engineering (MUE 2007), pp. 1–6 (2007)

[15] Frank, D., Wolfgang, S.: Chaos and cryptography. IEEE Trans. Circuits and Systems-I: Fundamental Theory and Applications 48(12), 1498–1508 (2001)

[16] Ljupco, K.: Chaos based cryptography: a brief overview. IEEE Circuits and Systems Magazine 1(3), 6–21 (2001)

[17] Goce, J., Ljupco, K.: Chaos and cryptography: block encryption ciphers based on chaotic maps. IEEE Trans. Circuits and Systems-I: Fundamental Theory and Applications 48(2), 163–169 (2001)

[18] Jakimoski, G., Kocarev, L.: Analysis of some recently proposed chaos-based encryption algorithms. Phy. Letters 291(6), 381–384 (2001)

[19] Michael, G., Andreas, U., Wild, P.: Transmission error and compression robustness of 2D chaotic map image encryption schemes. EURASIP J. Inf. Security, 1–16 (2007)

[20] Open Mobile Alliance (OMA), OMA DRM content Format v2.0 (2004), `http://www.Openmobilealliance.org`

[21] Dang Philip, P., Chau Paul, M.: Image encryption for secure internet multimedia applications. IEEE Trans. Consumer Electronics 46(3), 395–403 (2000)

[22] Grangetto, M., Magli, E., Olmo, G.: Multimedia selective encryption by means of randomized arithmetic coding. IEEE Trans. Multimedia 8(5), 905–917 (2006)

[23] Yuan, C., Zhu, B.B., Su, M., Wang, X., Li, S., Zhong, Y.: Layered access control for MPEG-4 FGS video. IEEE Int. Conf: Image Processing 1, 517–520 (2003)

[24] Kunkelmann, T., Reinema, R., Steinmetz, R., Blecher, T.: Evaluation of different video encryption methods for a secure multimedia conferencing gateway. In: Danthine, A. (ed.) COST-237 1997. LNCS, vol. 1356, pp. 75–89. Springer, Heidelberg (1997)

[25] Liu, X., Eskicioglu, A.M.: Selective encryption of multimedia content in distribution networks: Challenges and new directions. In: 2nd Int. Conf. Communications. Internet and Information Technology, Scottsdale, AZ (2003)

[26] Lookabaugh, T., Sicker, D.C.: Selective encryption for consumer applications. In: Proc. First IEEE Consumer Communications and Networking Conference, pp. 516–521. Las Vegas (2004)

[27] Shujun, L., Guanrong, C.: Chaos based encryption for digital images and videos. In: Multimedia Security Handbook, pp. 1–26. CRC press, Boca Raton (2004)

[28] Martina, P., Hans-Peter, S., Andreas, U.: Selective bitplane encryption for secure transmission of image data in mobile environments. In: Proc. 5th Nordic Signal Processing Symposium (2002)

[29] Van Droogenbroeck, M., Benedett, R.: Techniques for a selective encryption of uncompressed and compressed images. In: Proc. Advanced Concepts for Intelligent Vision Systems (ACIVS), Ghent, Belgium (2002)

[30] Xiaobo, L., Jason, K., Howard, C.: Image compression and encryption using tree structures. Pattern Recognition Letters 18, 1253–1259 (1997)

[31] Howard, C., Xiaobo, L.: Partial encryption of compressed images and videos. IEEE Trans. Signal Processing 48(8), 2439–2451 (2000)

[32] Chang, H.K.C., Liu, J.L.: A linear quadtree compression scheme for image encryption. Signal Processing, Image Communication 10, 279–290 (1997)

[33] Ahmed, N., Natarajan, T., Rao, K.R.: Discrete cosine transform. IEEE Trans. Comput. C-23(1), 90–93 (1974)

[34] Kunkelmann, T., Reinema, R.: A scalable security architecture for multimedia communication standards. In: Proc. IEEE Int. Conf. Multimedia Computing and Systems, Ottawa, Canada, pp. 660–661 (1997)

[35] Cheng, H., Li, X.: On the application of image decomposition to image compression and encryption. In: Proc. Int. Conf. Communications and Multimedia Security II, Essen, Germany, pp. 116–127 (1996)

[36] Tang, L.: Methods for encrypting and decrypting mpeg video data efficiently. In: Proceedings of the 4th ACM International Multimedia Conference, Boston, MA, pp. 219–230 (1996)

[37] Takeyuki, U., Safavi, N.R., Philip, O.: Recovering DC coefficients in block based DCT. IEEE Trans. Image Processing 15(11), 3592–3596 (2006)

[38] Van Droogenbroeck, M.: Partial encryption of images for real time applications. In: Fourth IEEE Benelux Signal Processing, The Netherlands, Hilvarenbeek, pp. 11–15 (2004)

[39] Shiguo, L., Jinsheng, S., Zhiquan, W.: A novel image encryption scheme based on JPEG encoding. In: Proc. Eighth International Conference on Information Visualisation, pp. 217–220 (2004)

[40] Amir, S.: Measuring the strength of partial encryption schemes. In: IEEE Int. Conf. Image Processing, vol. 2, pp. 1126–1129 (2005)

[41] Andreas, P., Andreas, U.: Selective encryption of wavelet-packet encoded image data: efficiency and security. Multimedia Systems 9, 279–287 (2003)

[42] Uehara, T., Safavi, N.R., Ogunbona, P.: Securing wavelet compression with random permutations. In: Proc. IEEE Pacific-Rim Conf.: Multimedia, Syndney, Australia, pp. 332–335 (2000)

[43] Lian, S., Wang, Z.: Comparison of several wavelet coefficients confusion methods applied in multimedia encryption. In: Proc. Int. Conf. Computer Networks and Mobile Computing (ICCNMC 2003), Shanghai, China, pp. 372–376 (2003)

[44] Grosbois, R., Gerbelot, P., Ebrahimi, T.: Authentication and access control in the JPEG 2000 compressed domain. In: Proc. SPIE 46th Annu. Meeting, vol. 4472, pp. 95–104 (2001)

[45] Jiang-Lung, L.: Effective selective encryption for jpeg2000 images using private initial table. Pattern Recognition 39, 1509–1517 (2006)

[46] Karl, M., Rastislav, L., Konstantinos, N.P.: Efficient encryption of wavelet based color images. Pattern Recognition 38, 1111–1115 (2005)

[47] Maples, T.B., Spanos, G.A.: Performance study of a selective encryption scheme for the security of networked, real-time video. In: Proc. Int. Conf. Computer Communications and Networks, Las Vegas (1995)

[48] Li, Y., Chen, Z., Tan, S.M., Campbell, R.H.: Security enhanced MPEG player. In: Proc. IEEE Int. Workshop Multimedia Software Development, Berlin, Germany, pp. 169–175 (1996)

[49] Agi, I., Gong, L.: An empirical study of secure MPEG video transmissions. In: Proc. Internet Soc. Symp. Network Distributed System Security, San Diego, CA, pp. 137–144 (1996)

[50] Qiao, L., Nahrstedt, K.: A new algorithm for MPEG video encryption. In: Proc. 1st Int. Conf. Imaging Science, Systems and Technology, Las Vegas, pp. 21–29 (1997)

[51] Qiao, L., Nahrstedt, K.: Comparison of MPEG encryption algorithms. Int. J. Computers & Graphics 22(4), 437–448 (1998)

[52] Meyer, J., Gadegast, F.: Security mechanisms for multimedia data with the example MPEG-1 video. Project Description of SECMPEG, Technical University of Berlin, Germany (1995)

[53] Alattar, A.M., Al-Regib, G.I., Al-Semari, S.A.: Improved selective encryption techniques for secure transmission of MPEG video bit-streams. In: Proc. Int. Conf. Image Processing, Kobe, Japan, vol. 4, pp. 256–260 (1999)

[54] Yongcheng, L., Zhigang, C., See-IMong, T., Campbell Roy, H.: Security enhanced MPEG player. In: Proc. First Int. Workshop Multimedia Software Development (1996)

[55] Changgui, S., Bharat, B.: An efficient MPEG video encryption algorithm. In: Proc. IEEE Symp. Reliable Distrubited Systems, pp. 381–386 (1998)

[56] Shi, C., Bhargava, B.: A Fast MPEG video encryption algorithm. In: Proc. 6th Int. Multimedia Conf., Bristol, UK (1998)

[57] Shi, C., Wang, S.Y., Bhargava, B.: MPEG video encryption in real-time using secret key cryptography. In: Int. Conf. Parallel and Distributed Processing Techniques and Applications, Las Vegas, pp. 191–201 (1999)

[58] Qiao, L., Nahrstedt, K.: Is MPEG encryption by using random list instead of zigzag order secure? In: Proc. IEEE Int. Symp. Consumer Electronics, Singapore, pp. 226–229 (1997)

[59] Zeng, W., Lei, S.: Efficient frequency domain selective scrambling of digital video. IEEE Trans. Multimedia 5(1), 118–129 (2003)

[60] Wen, J., Severa, M., Zeng, W., Luttrell, M.H., Jin, W.: A format-compliant configurable encryption framework for access control of video. IEEE Trans. Circuits and Systems for Video Technology 12(6), 545–557 (2002)

[61] Shujun, L., Xuan, Z., Xuanqin, M., Yuanlong, C.: Chaotic encryption scheme for real-time digital video. In: Proc. SPIE Real-Time Imaging VI, vol. 4666, pp. 149–160 (2002)

[62] Wu, C.P., Kuo, C.C.: Efficient multimedia encryption via entropy codec design. In: Proc. SPIE Security & Watermarking of Multimedia Contents III, San Jose, CA, vol. 4314 (2001)

[63] Griwodz, C.: Video protection by partial content corruption. In: ACM Workshop on Multimedia and Security, Bristol, UK, pp. 37–40 (1998)

[64] Griwodz, C., Merkel, O., Dittmann, J., Steinmetz, R.: Protecting VoD the easier way. In: ACM Int. Conf. Multimedia, Bristol, UK, pp. 21–28 (1998)

[65] Frederic, D., Touradj, E.: Region based transform domain video scrambling. In: Proc. SPIE Visual Communications and Image Processing (2006)

[66] Thorwirth, N.J., Horvatic, P., Weis, R., Zhao, J.: Security methods for MP3 music delivery. In: Proc. Thirty-Fourth Asilomar Conference on Signals, Systems, and Computers, Asilomar, CA, vol. 2, pp. 1831–1835 (2000)

[67] Torrubia, A., Mora, F.: Perceptual cryptography on MPEG-1 Layer III bit-streams. In: Int. Conf. Consumer Electronics, Los Angeles, CA, pp. 324–325 (2002)

[68] Torrubia, A., Mora, F.: Perceptual cryptography on MPEG Layer III bit-streams. IEEE Trans. Consumer Electronics 48(4), 1046–1050 (2002)

[69] Servetti, A., Testa, C., De Martin, J.C.: Frequency-selective partial encryption of compressed audio. In: Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, Hong Kong, vol. 5, pp. 668–671 (2003)

[70] Yekkala Anil, K., et al.: Lightweight encryption for images. In: IEEE Int. Conf. Consumer Electronics, pp. 1–2 (2007)

[71] Torrubia, A., Francisco, M.: Perceptual cryptography of JPEG compressed images on the JFIF bitstream domain. In: Proc. IEEE Int. Conf. Consumer Electronics, pp. 58–59 (2003)

[72] Lian, S., Sun, J., Wang, Z.: Perceptual cryptography on JPEG2000 compressed images or videos. In: Fourth Int. Conf. Computer and Information Technology, Wuhon, China, pp. 78–83 (2004)

[73] Lian, S., Sun, J., Wang, Z.: Perceptual cryptography on SPIHT compressed images or videos. In: Proc. IEEE Int. Conf. Multimedia and Expo., Taipei, Taiwan, vol. 3, pp. 2195–2198 (2004)

[74] Wu, M., Mao, Y.: Communication-friendly encryption of multimedia. In: Proc. IEEE Int. Workshop on Multimedia Signal Processing (2002)

[75] Barbir, A.: A methodology for performing secure data compression. In: Proc. Twenty-Ninth South-eastern Symp. System Theory (1997)

[76] Goce, J., Subbalakshmi, K.P.: Cryptanalysis of some multimedia encryption schemes. IEEE Trans. Multimedia 10(3), 330–338 (2008)

[77] Gillman, D., Mohtashemi, M., Rivest, R.: On breaking a Huffman code. IEEE Trans. Information Theory 42(3), 972–976 (1996)

[78] Shi, C., Bhargava, B.: A fast MPEG video encryption algorithm. In: Proc. ACM Int. Conf. Multimedia, Bristol, UK, pp. 81–88 (1998)

[79] Chung-Ping, W., Jay Kuo, C.-C.: Design of integrated multimedia compression and encryption systems. IEEE Trans. Multimedia 7(5), 828–839 (2005)

[80] Pazarci, M., Dipcin, V.: A MPEG2-transparent scrambling technique. IEEE Trans. Consumer Electronics 48(2), 345–355 (2002)

[81] Wen, J., Muttrell, M., Severa, M.: Access control of standard video bitstreams. In: Int. Conf. Media Future, Florence, Italy (2001)

[82] Norcen, R., Uhl, A.: Selective encryption of the JPEG2000 bitstream. In: Lioy, A., Mazzocchi, D. (eds.) CMS 2003. LNCS, vol. 2828, pp. 194–204. Springer, Heidelberg (2003)

[83] Zhu, B.B., Yang, Y., Shipeng, L.: JPEG 2000 syntax-compliant encryption preserving full scalability. In: Proc. IEEE Int. Conf. Image Processing, vol. 3, pp. 636–639 (2005)

[84] Sadourny, Y., Conan, V.: A proposal for supporting selective encryption in JPSEC. IEEE Trans. Consumer Electronics 49(4), 846–849 (2003)

[85] Zhu, B.B., Yang, Y., Li, S.: JPEG 2000 Encryption enabling fine granularity scalability without decryption. In: IEEE Int. Symp. Circuits and Systems, Kobe, Japan, vol. 6, pp. 6304–6307 (2005)

[86] Tosun, A.S., Feng, W.C.: Efficient multi-layer coding and encryption of MPEG video streams. In: IEEE Int. Conf. Multimedia and Expo., New York, vol. 1, pp. 119–122 (2000)

[87] Wu, Y., Deng, R.H.: Compliant encryption of JPEG 2000 codestreams. In: Proc. IEEE Int. Conf. Image Processing, Singapore, pp. 3447–3450 (2004)

[88] Zhu, B.B., Yuan, C., Wang, Y., Li, S.: Scalable protection for MPEG-4 fine granularity scalability. IEEE Trans. Multimedia 7(2), 222–233 (2005)

[89] Zhu, B.B., Yang, Y., Chen, C.W., Li, S.: Fine granularity scalability encryption of MPEG-4 FGS bitstreams. In: IEEE Int. Workshop Multimedia Signal Processing, Shanghai, China (2005)

[90] Shiguo, L., Zhongxuan, L., Zhen, R., Haila, W.: Commutative encryption and watermarking in video compression. IEEE Trans. Circuits and Systems for Video Technology 17(6), 774–778 (2007)

[91] Shih-Wei, S., Chun-Shien, L., Pao-Chi, C.: Joint multimedia fingerprinting and encryption: security issues and some solutions. In: IEEE Int. Conf. Multimedia and Expo., pp. 1531–1534 (2007)

[92] Linfei, C., Daomu, Z.: Image encryption with fractional wavelet packet method. Elsevier, Optik 119, 286–291 (2008)

The Method for Image Copy Detection Robust to Basic Image Processing Techniques

Karol Wnukowicz, Grzegorz Galiński, and Władysław Skarbek

Abstract. The chapter presents a method for content based image replica detection. In this method a compact image signature is extracted. The signature depends on image content, carries distinctive image information, and is invariant to many widely used image processing techniques, which do not lead to significant loss of information, such as lossy compression, resizing, color enhancements and simple rotations. The detection of unmodified and modified image replicas is performed by matching signatures of query images to signatures of original images. The signature is designed to be usable in big image databases and even in video databases. It can be characterized by the following properties: small size (a few dozen bytes), fast extraction and matching, high detection rate for basic image processing techniques. A few millions of signatures per second can be compared on a modern PC.

1 Introduction

In today's networked audiovisual systems and applications a great deal of audiovisual data is gathered and distributed. The data is often modified during the distribution to meet various requirements of networks, storage capacity, or capabilities of user equipment. Different multimedia devices and networks often need data in specific formats, bitrates, and resolutions – so the data must be transformed according to these needs (e.g. different format is preferred by high resolution printers, HD displays, PC monitors, and mobile phones). This makes the possibility that the same material is stored and distributed in modified versions and various formats. The examples of modifications which may be applied to visual data are rescaling, lossy compression, changing of color depth, and image enhancements such as changing brightness, contrast, saturation, blurring or sharpening.

The goal of a copy detection system is to allow the detection and localization of all variants of the same multimedia material, including the modified versions. In contemporary public networks the image and video material can be easily created, shared, and accessed. The easiness of dissemination of such material in digital form creates a new challenge for applications such as copyright protection, content management, media usage monitoring. A good solution for copy detection system

Karol Wnukowicz, Grzegorz Galiński, and Władysław Skarbek
Warsaw University of Technology, Institute of Radioelectronics, 00-665 Warszawa, Nowowiejska 15/19, Poland
e-mail: {K.Wnukowicz,G.Galinski,W.Skarbek}@ire.pw.edu.pl

M. Grgic et al. (Eds.): Rec. Advan. in Mult. Sig. Process. and Commun., SCI 231, pp. 451–466.
springerlink.com

would be the algorithm which extracts unique and distinctive information based on the content of images or videos. Designing a reliable tool for content based replica detection of digital content would significantly help to protect the rights of copyright owners and help to manage the usage of the content. The detection of images and video duplicates should be fast, reliable, and robust to widely used techniques of adaptation or enhancement of the material. Two general solutions for copy detection can be distinguished: watermarking and fingerprinting. Watermarking [10, 13] represents methods which add some information to the visual content. This additional information is hard to modify, remove and detect by unauthorised users or applications, but can be detected and identified by authorised ones. When images are duplicated the hidden information is also duplicated. The disadvantage of this method is the necessity of inserting watermarks to digital content before dissemination. On the other hand, the fingerprinting consists in extraction of media description which is regarded to be the identifier or "fingerprint" of the media. The identifier should represent visual content in a unique way, and should be robust to many common image processing techniques. This chapter presents a solution for fingerprinting copy detection system.

A practical application of the automatic replica detection technique may be, for example, an image copyright protection system, which uses image copy detection module for searching copies of copyrighted images on the Web. Another example is an application of digital photo management. Many users of digital cameras often process their photos in popular image processing applications and multiple (modified) copies of the same photo may be stored in their photo galleries. Copy detection tool would allow users to find photos having different versions in their private photo galleries.

The number of images available on the Web is very high. The application for image copy detection on the Web should be very fast and efficient, which imposes the following requirements: fast extraction and detection, small size of any description data, and good performance of the detection (very high precision and recall). The fulfilment of all these requirements is very hard, so the algorithm needs to accept a trade-off among high detection speed, high detection rates, and robustness to heavy modifications. To achieve high detection speed and high detection rate a compromise in the designing of the algorithm needs to be made, such as restriction of supported image processing techniques which can be applied to images to be detected as image copies.

In this chapter a method for content based image copy detection is presented. This method is based on distinctive signature extracted from visual content. The signature can be regarded as an image fingerprint, and it is robust to many image processing techniques. The signature was designed to be used in big image databases. To achieve this goal, the list of modifications supported by the signature is restricted to common image processing techniques (basic 'attacks') that do not lead to significant loss of information, such as lossy compression, resizing, color conversion and enhancement, blurring, sharpening, noising. The robustness of the signature to strong 'attacks' such as heavy cropping of image is limited. The detection of image replicas is performed by signature matching. The result of matching is the distance between two signatures, and the distance calculated during the matching is used to assess if two images are replicas of each other or not. In other words, the image copy detection system is a simple decision system of which the elementary function is to state whether a given ‘suspected’ image is a copy of the reference original image or not. The decision is based on a distance threshold which decides if two signatures represent copies of each other or not. The threshold value is chosen to obtain a required balance between false detection error and false rejection error.

The small size of the signature on the one hand, and fast extraction and matching on the other hand allows the usage of the signature for the detection of duplicated video segments. The image signature method can be easily extended to detect video duplicates because videos consist of sequences of images for which the signatures can be computed and matched.

The outline of this chapter is the following: Section 2 presents a survey of related works, Section 3 contains the description of signature design, extraction, and matching, Section 4 contains the result of experiments using the proposed image replica detection method, Section 5 presents the initial work on video copy detection as an extension of the proposed still image method, and finally the conclusions are drawn in Section 6.

2 Related Works

A number of papers have been published on content based image replica detection. Two general groups of methods can be distinguished: methods that use global visual features [7, 8], and methods that use local visual features [1, 2, 4, 5, 11, 17]. In the global methods the image content is described using a set of features such as texture, shape, or color. Replicated images would have similar features provided that the features are invariant to modifications that were applied to the images. The detection success rates of the global methods are generally not very high because much of the unique image information is comprised in local distribution of features. To improve the detection performance of global feature methods, in [7] the classifier for replica detection system is presented. This classification system uses selected visual features to build classification scheme which assigns input images to two classes: replicas and non-replicas of a given reference image. The following features were used in the classification system: color histogram, grey level histogram, and Gabor transform statistics. The classification system was based on support vector machines and a single classifier is build for each reference image. The classification system is trained to assign query images to be replicas or not replicas of each of the trained images for a broad range of possible image modifications.

The second group of methods takes into account local distribution of features. In [2] the feature vector is build from coefficients of Discrete Wavelet Transform (DWT) applied to image with normalized size. Due to the nature of DWT, the feature vector depends on global as well as local image characteristics. In [5] the feature vector is based on AC coefficients from DCT transform for image resized to the size 8x8 pixels. The extraction and matching of these signatures is fast, but the detection rate is not suitable for big image databases. Another method was presented in [4]. It uses the concept of key or interest points which gives good results in detecting replicas for a wide range of image modifications. The detection rate of this method is good, but the drawback is high computational cost and the size of image features is relatively large, thus it is not practical for big image databases.

Recently, the possibility of designing a visual identifier, which can be used for image replica detection, was also investigated by standardization activities of MPEG group [9]. The group defined the requirements for image signature (image identifier) to be used for content-based image identification. These requirements contain experimental conditions for two different scenarios: fast algorithm with high success rate for replicas obtained by basic modifications and possibly slower algorithm which will detect image replicas obtained by heavy modifications. The basic image modifications include the following: lossy JPEG compression, image scaling, blurring, noising, color to monochrome conversion, brightness change, color reduction, histogram equalisation, auto-levels, flip, simple rotation (by 90°, 180°, 270°). Moreover, the size of the basic identifier was restricted to 1 kilobit, and the matching speed should be at least 2 million signature pairs per second. Two contributions which fulfil the requirements of basic conditions have been submitted in [1] and [11]. In [1] a signature based on Radon transform is presented. This signature is build from rotation invariant and scaling invariant features obtained by Radon Transform, which also depend on local feature characteristics. The signature proposed in [11] is computed using concentric circle partition. Both signatures were designed to be invariant to the basic 'attacks' defined in the MPEG requirements for image signature.

The algorithm proposed here uses local image features. It can be characterized by fast extraction and detection, small size, and very good detection rate for broad range of popular image modifications. The main effort was dedicated to meet the MPEG requirements defined in [9] for basic identifier (robustness to basic image modifications and fast detection). The algorithm presented in this chapter is compared to the methods proposed in [1, 11] using the same dataset and experimental conditions.

3 Image Signature

The proposed signature is invariant to many common image processing techniques such as color conversion/enhancement, resizing, simple rotation, lossy compression and applying filters. The modifications which cause significant loss of information such as cropping are not supported by the signature – such heavy modifications need more complex solutions, for example the method presented in [4], which require much more computation power for extraction and matching, and thus their applicability to big multimedia databases is limited.

The proposed method is based on our previous work presented in [14], which uses trajectory of features in overlapping image blocks. The blocks were obtained by partitioning an image into fixed number of blocks. In each block a local feature was computed and the successive blocks formed a vector of features. Experiments have shown that the spatial distribution of features is highly correlated for images which are modified copies of each other. Some of the modifications can change the feature value of a single block, but the correlation of vectors formed by the values of consecutive blocks is very high for image replicas.

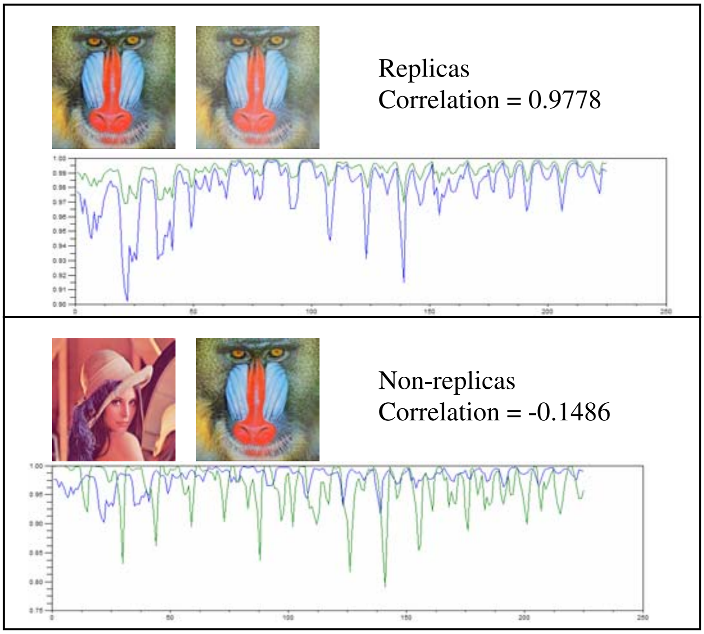


Fig. 1 1-D diagrams of feature vectors and their correlations for modified replicas and non-replicas

Figure 1 shows two example diagrams of feature vectors for image replicas and non-replicas. The top diagram shows block's values of original (baboon) image and image modified by brightening and JPEG compression with quality factor *q*=70. The plotted lines representing the two images are displaced but the correlation is high, it equals 0.9778. The bottom diagram shows block values of images which are non-replicas, and here the correlation is close to 0, which means the signal is different. The correlation of vectors which represent features in local blocks was used to decide if two images are copies of each other. Two images were replicas of each other if the value of correlation was above some threshold, which was set experimentally.

The new signature presented in this chapter is also based on features computed in the same image blocks, but additional processing is added to achieve smaller size of the signature, and better robustness to basic image modifications including simple rotations [15]. The signature use 3 features computed for each block: mean luminance level, energy, and singular energy. These features cover different aspects of local block characteristics, so using them together increases the uniqueness of signature and thus the detection performance. The blocks are grouped into rotation-invariant clusters – this makes the signature invariant to simple rotations. Figure 2 shows the simple rotations which are supported by the signature. Moreover, the signature is build as a bit-string, with one bit for each feature in each block group, where bits represent local change of features from block group to block group, that is, they represent the distribution of local features. Image modifications do not change this distribution of local features. As a result, the signature is smaller with good discriminative properties for replicas and non-replicas, and the signature matching is very fast. The following two subsections present the extraction algorithm of the signature in more details and the method for signature matching.

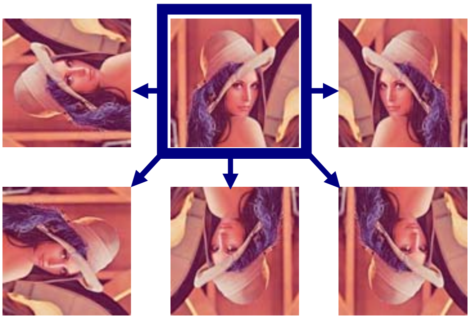


Fig. 2 Simple rotations: left, right, vertical flip, 180°, horizontal flip

3.1 Extraction

The signature extraction algorithm consists of the following steps. These steps are depicted in Figures 3 and 4:

1. Image pre-processing:

- resize image in such a way that the shorter edge has L pixels (L=128 was used);
- convert pixel values of the image to greyscale;
- crop the central part of the image to a size L×L pixels;
- apply blur operation to the image using 3×3 filter to get slightly blurred image (*Img*), and generate second version of the cropped image by histogram equalization (Img_{histeq}).

2. Partition the two image versions *Img* and Img_{histeq} into overlapping blocks B(x,y) of M×M pixels. The blocks are obtained by moving M×M window of pixels across the image from left to right by N pixels and from top to bottom also by N pixels (e.g. M=8; N=4; x, y = [0, 4, 8, …, 120], where x and y identify each block by its upper-left pixel position).
3. Compute features in each image block – 3 features are used: mean luminance level of a block $Y(x,y)$, energy of a block $E(x,y)$, and singular energy of a block $S(x,y)$. The features are described below. Mean levels of luminance Y and singular energies S are computed using the blurred version of image *Img*. Energy E is computed using the image with equalized histogram Img_{histeq}. Using two versions of images increases the detection performance for some image modifications which are supported by the signature.

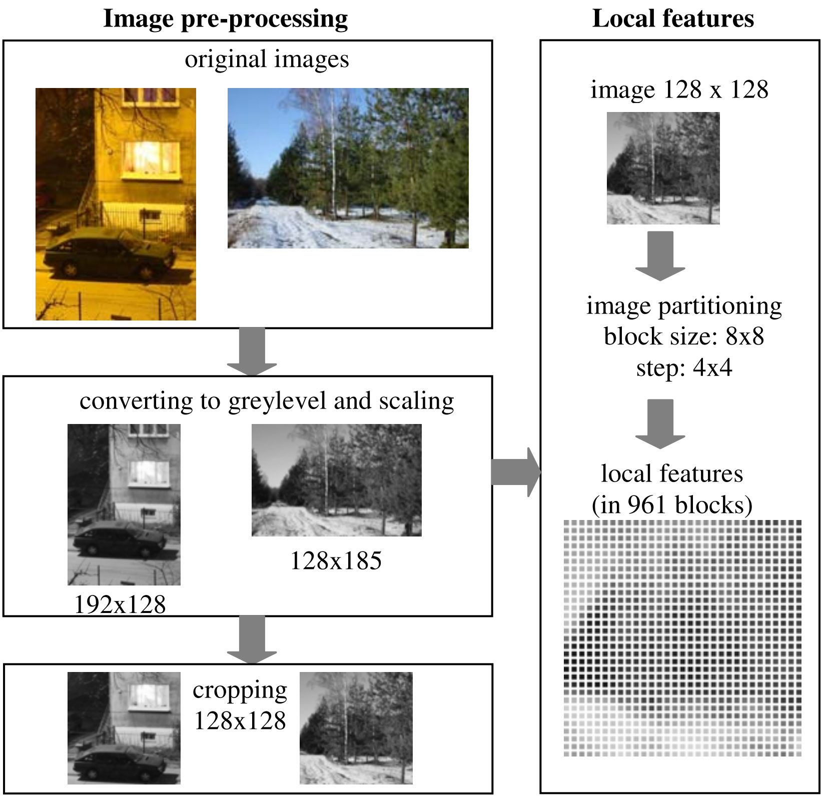


Fig. 3 Image pre-processing and partitioning into overlapping blocks

4. Cluster image blocks $B(x,y)$ in order to obtain block groups $BG(i)$ which are rotation-invariant. Rotation-invariant means that when the image is rotated (using simple rotation) the group is not changed (see Figure 4). The block clustering procedure assigns group number i to each block $B(x,y)$. The number of blocks in a single group is 8, or 4 on the symmetry axes. The number of groups for the chosen set of partition parameters is 120, which means $0 \le i < 120$.
5. Compute features for each group of blocks: mean and standard deviation of feature values in the groups: $Y_{mean}(i)$, $Y_{dev}(i)$, $E_{mean}(i)$, $E_{dev}(i)$, $S_{mean}(i)$, $S_{dev}(i)$. These values of block groups are invariant to basic rotations.
6. Build image signature: the signature is a bit-string built from the feature values in groups of blocks obtained in step 5. The signature bits are computed for each feature F using the following pseudo-code:

```
if $(F(i+1)) > F(i))$        { $Signature\_bit(i) = 1$ }
else                         { $Signature\_bit(i) = 0$ },
```

where F is one of the features: Y_{mean}, Y_{dev}, E_{mean}, E_{dev}, S_{mean}, S_{dev}, and $0 \le i < 119$. Moreover, for energy and singular energy the following rule was applied:

if the block energies of $F(i)$ and $F(i+1)$ are below a predefined threshold set *Signature_bit*(i)=0. The motivation for this rule was that when energies of two consecutive blocks are very low, the signature bits may reflect noise rather than real signal. The signature is extended by 2 bytes – one byte represents the percentage of low-energy blocks in normal image *Img*, the other represents the percentage of low energy blocks in image Img_{hist}.

Mean luminance, block energy, and singular energy for block $B(x,y)$ of the size $M\times M$ pixels are computed in the following way:

Mean Luminance
Mean luminance is the mean value of greylevel pixels in the block:

$$Y(x,y)=\frac{\sum_{m=x}^{x+M}\sum_{n=y}^{y+M} I(m,n)}{M\times M} \tag{1}$$

where $I(m,n)$ is the greylevel value of pixel at position (m,n), M is the size of the squared block with upper left pixel at position x,y.

Block Energy:
More specifically, block energy is the energy of pixel value deviation from mean luminance:

$$E(x,y)=\frac{\sum_{m=x}^{x+M}\sum_{n=y}^{y+M} \left(I(m,n)-Y(x,y)\right)^2}{M\times M} \tag{2}$$

Singular energy:
The usage of singular energy for image replica detection was introduced in [12], where the concept of singular energy trajectory is defined to describe images by vectors of singular energy channels computed for image blocks. Singular energy is defined as fractional distribution of the energy in the first singular channel (defined by singular directions of image blocks). Singular energy is invariant to many image processing techniques including basic rotations and hence it is a good candidate for a block feature of the signature. Singular energy is obtained by singular value decomposition (SVD). Performing the singular decomposition of the matrix represented by block $B(x,y)$, we get k singular values $\sigma_1(x,y),\ldots,\sigma_k(x,y)$. It is well known that the energy of pixels in a block $B(x,y)$ is decomposed into the sum of all squared singular values of $B(x,y)$:

$$\left\|B(x,y)\right\|_F^2=\sum_k \sigma_k^2(x,y) \tag{3}$$

$$\text{where } \left\|B(x,y)\right\|_F^2=\sum_{m=x}^{x+M}\sum_{n=y}^{y+M} I(m,n)^2 \tag{4}$$

$$\frac{\sum_k \sigma_k^2(x,y)}{\left\|B(x,y)\right\|_F^2}=1 \tag{5}$$

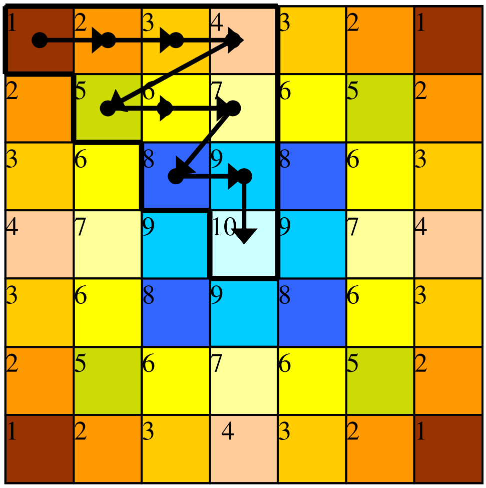
The singular energy value of block $B(x,y)$ of rank $r<k$ is defined as the point in r dimensional unit cube $[0,1]^r$:

$$\left(\frac{\sigma_1^2(x,y)}{\left\|B(x,y)\right\|_F^2},\dots,\frac{\sigma_r^2(x,y)}{\left\|B(x,y)\right\|_F^2}\right) \tag{6}$$

For building the signature only one singular energy component was used. Experiments have shown that the best result was achieved when the second singular energy component is used, so the singular energy $S(x,y)$ is defined as:

$$S(x,y)=\frac{\sigma_2^2(x,y)}{\left\|B(x,y)\right\|_F^2} \tag{7}$$

Figure 4 illustrates the idea of rotation-invariant block clustering (step 4 of the algorithm) and block ordering for building the signature (step 6 of the algorithm). The blocks are marked with squares. The squares depicted by the same number and colour represent blocks belonging to the same rotation-invariant group. The arrows show the scanning order for computing the signature bits using group-to-group differences of the features. The features for each scan point are the means and standard deviations of all the features in blocks of the related group (i.e. having the same numbers in Figure 4). After any simple rotation the groups do not change, and thus the feature values of the scan points are invariant to such rotations.



1	2	3	4	3	2	1
2	5	6	7	6	5	2
3	6	8	9	8	6	3
4	7	9	10	9	7	4
3	6	8	9	8	6	3
2	5	6	7	6	5	2
1	2	3	4	3	2	1

Fig. 4 Rotation-invariant block clustering, and the direction of building the bit-string of image signature

Invariance to various image processing techniques is obtained by the following operations and designing rules:

- conversion to greyscale (invariance to color conversion);
- scaling to fixed size and cropping (invariance to resizing)
- processing of scaled images: blurring and histogram equalization – improved robustness to modifications such as blurring, noising, histogram equalization, etc.;
- block features invariant to all basic modifications;
- rotation-invariant groups of blocks (invariance to basic rotations);
- distance function which represent local distribution of features;

3.2 Matching Image Signatures

The detection of image replicas is performed by matching image signatures. The copy detection system gives a binary answer for single matching: an image is a replica or not. The signatures of original images in image database can be pre-computed and stored. To check if a given query image is a replica of any of the images in the database, the signature of the query image may be extracted and matched to all the signatures of original images stored in the database.

The matching algorithm calculates Hamming distance D between the bit-streams of two signatures; it means the distance is represented by the number of bits which differ between the two signatures. Hamming distance can be implemented very efficiently using XOR operation and a lookup table for each compared byte pairs – a few millions of comparisons per second can be performed on a typical desktop computer. The final decision to classify two images as replicas is made by comparing the distance D with a predefined threshold. The threshold is determined experimentally and depends on required precision and recall rates. Typically lower thresholds will give better precision rate (less falsely detected copies) but the recall rate will decrease (more copies will be missed). On the other hand, higher threshold will give better recall, but the precision will decrease.

In order to achieve high detection speed and good detection performance, an efficient algorithm for comparison of signatures was developed and implemented. The algorithm uses partial matching of signatures (matching of sub-bitstreams of signatures) and 5 thresholds. The signature comparison algorithm is as follows:

1. Calculate Hamming distance D_{15} of the first 15 bytes of the signatures (120 bits corresponding to the feature Y_{dev}): if ($D_{15} > TH_{15}$) the test image is not a replica; else:
2. Calculate, in incremental way, Hamming distance D_{30} of the first 30 bytes (240 bits corresponding to the features Y_{dev} and Y_{mean}): if ($D_{30} > TH_{30}$) the test image is not a replica; else:
3. Calculate Hamming distances D_Y (equal to D_{30}), D_S, D_E of the 3 features: if ($D_Y < TH_Y$ or $D_S < TH_S$ or $D_E < TH_E$) the test image is a replica; else: the test image is not a replica.

In the steps 1 and 2 of the signature matching algorithm, the signatures are partially matched. If non-replica is detected in steps 1 or 2 (i.e., test image is not a replica of any database image) the matching algorithm returns “false” and the detection process is finished. As a result, many of non-replicas are discarded: in the first step 89.32%, and after the second step 99.7%. This solution significantly reduces the matching time since the full distance was computed only for 0.3% of total matching cases during the experiment. The thresholds TH_{15} and TH_{30} are set experimentally in such a way that the number of discarded non-replicas is maximized, and the true replicas are not discarded. The third step of the matching algorithm is performed only when the first two steps do not return “false”. All the thresholds are obtained experimentally using a big image database to get false alarm ratio equal to 0.05 parts per million (ppm), which means that one error of falsely accepted non-replica can occur in 20,000,000 tested non-replicas.

4 Experiments and Results

The detection rate of the image signature has been evaluated using the evaluation conditions and image database defined in the requirements for image signature robust to basic ‘attacks’ defined by MPEG group [9]. The experiments consist of two steps. In the first step, the operational point for the detection is assigned using a database of independent images (non-replicas of each other). The database contains 135,609 images and the distances of all independent image pairs in the database were computed. The number of independent image pairs is: 135,609 * (135,609 – 1)/2 = 9,194,832,636. The operational point is set to obtain false positive rate of the detection equal to 0.05 ppm. This means that a non-replica may be falsely detected as a replica once for every 20,000,000 image comparisons. The operational point is used to set the thresholds for the matching algorithm which give the required false positive rate. These thresholds are then used in the second part of the evaluation: testing the robustness of the signature.

To test the robustness of the image signature, the success rate of the detection is computed corresponding to the operational point 0.05 ppm false alarm ratio obtained in the first part of the experiments. A second image database, which consists of 10,000 original images, is used. These original images are used to generate modified images with the software provided in [9]. The modifications presented in Table 1 are performed on each original image. Up to three modification sets are generated for each modification group using various modification parameters. ’Light’, ‘Medium’ and ‘Heavy’ show the ‘strength’ of the modification, e.g., for jpeg compression the quality factor is (light = 80, medium = 60, heavy = 30). The success rate is measured for each modification independently, and the results are presented in Table 1. Success rate is defined as the ratio of the number of successfully detected image copies to the total number of image copies generated from the originals.

Table 1 Success rate at 0.05 ppm false alarm

Modification	Heavy	Medium	Light	Mean
Brightness change	99.41	99.55	99.91	99.62
Color to monochrome conversion			99.87	99.87
JPEG compression	98.94	99.76	99.91	99.54
Color reduction		98.83	99.3	99.06
Gaussian noise	98.01	99.24	99.71	98.99
Histogram equalization			98.52	98.52
Auto levels			99.6	99.6
Flip (left-right)			99.99	99.99
Blur	99.82	99.91	99.95	99.89
Scaling (preserving aspect ratio)	99.91	99.97	99.96	99.95
Rotation (90º, 180º, 270º)	99.92	99.94	99.93	99.93
Average				99.54

The signature is conformant to the requirements of MPEG Call for Proposal on Image & Video Signature Tools (i.e., basic modifications, fast algorithm). Table 2 contains the comparison of the proposed method (labeled LFT) with other methods. The results for methods submitted by VIL [1] and ETRI [11] are taken from their responses to Call for Proposal on Image & Video Signature Tools, submitted at the 82nd MPEG meeting.

Table 2 Comparison of image signatures

	Success rate	Success rate for	Size	Extraction time
	[%]	histogram equalisation	Bits (Bytes)	(Sign. per second)
LFT	99.54	98.52	730 (91.25)	16 - 27
VIL	99.59	96.62	512 (64)	4.35
ETRI	99.63	98.32	392 (49)	5

Although the success rate of the proposed method is a bit worse, the difference is minimal. The advantage of the proposed method is fast extraction and fast matching. The extraction time of the proposed method given in the table depends on image size and includes the decompression of input images stored in JPEG format. The time performance of matching was measured on 2 GHz Intel Centrino processor with 1 GB RAM. The obtained number of comparisons was above 7 million per second (3-step Hamming distance) in the test of independent images. The methods of VIL and ETRI also use Hamming distance, but they compare the full signature each time.

5 Video Signature

The proposed image signature can be extended to detect duplicated video segments. Video copy detection can be useful in applications such as monitoring of the usage of video material in TV programs (e.g. advertisement), the detection of illegal copies of copyrighted video clips in public networks, or finding the original source of a given video. The description and evaluation of existing algorithms for video copy detection can be found in [6, 16]. They use spatial and/or temporal characteristics of videos to compute various features which are used to match replicated video segments. Some of the algorithms use global features and some use local features including color, texture and motion. Some of the video copy detection algorithms operate in pixel domain, and some others operate in compressed domain of specific video format using features based on DC/AC components of DCT transform, or motion vectors. The algorithms operating in compressed domain have usually lower computation needs, but their usage is restricted to specific coding formats. Another solution for video copy detection is to detect and track video objects in video shots, and the detection of duplicated video segments is performed by matching the objects and their spatio-temporal trajectories.

The proposed video signature is a straightforward extension of image signature presented in Section 3. It consists of concatenated signatures of video frames, with a header containing additional information about the video such as the frame rate. The detection of video segments is performed by matching sequences of signatures of consecutive video frames in two videos. In case of video the signature was modified by using blocks of the size 16×16 moved with step 8 instead of blocks 8×8 moved with step 4, and the matching is done by computing Hamming distance of the signature's bit-strings. Using bigger blocks in case of video signature produces signatures of smaller size than image signature and the matching is faster. In case of video the signature of single frame can be smaller and thus less distinctive than in case of still image because we match not a single frame but a segment of a few video frames, so the amount of information is higher. The searching is performed by comparing frame signatures (one signature every *MD* seconds) from query clip to all signatures extracted from original videos. If a frame match, the longest segment of consecutive frames before and after this frame is being matched. And finally, if the duration of matched segment exceeds minimum segment duration (e.g. *MD* = 5 seconds) the segment is returned as a duplicated video segment.

The segments of arbitrary number of frames can be matched which allows to detect edited video material. The advantage of this method is that the size of matched video segments can vary from single frame to the whole video. Figure 5 shows the general idea of matching video segments of consecutive video frames. The minimum duration *MD* of a video segment to be detected is the parameter of the matching algorithm. The reasonable minimum duration is 5 seconds, which means that the video copy detection application can match any segment in original video to any segment in tested video which is equal or longer than 5 seconds.

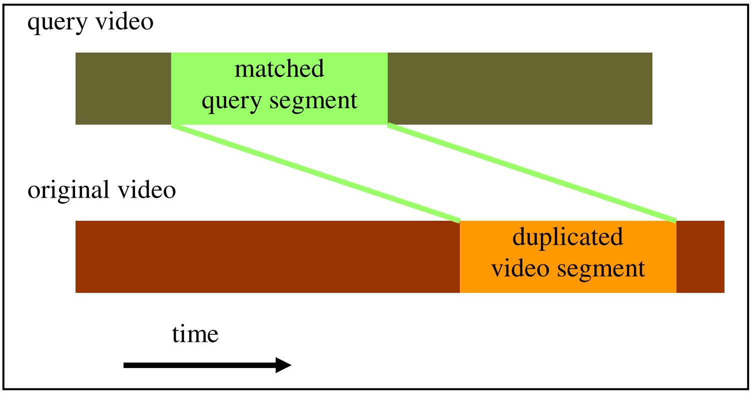


Fig. 5 Partial matching of video segments

Initial experiments have been carried out and they confirmed the suitability of the proposed method for video copy detection applications. The signatures were computed and stored for a few MPEG-2 videos. The total duration of the videos was 7 hours and 40 minutes. Next, a number of short query clips were generated from the original videos using AviSynth software and encoded with free MPEG-2 encoder HC Encoder (available on http://www.bitburners.com/hc_encoder). AviSynth is free software which allows performing various modifications of videos and also cutting and merging video segments. The query clips consisted of randomly selected video segments generated from original videos of the following durations: 10 seconds, 1 minute, and 10 minutes. Some of the clips where modified using the following processing: reduction to CIF format (352×288), brightness change, compression (reduction of the bitrate). Two video segments were regarded to match each other if all consecutive video frames in one segment matched to corresponding frames in the other segment. All query clips and their positions in the original videos were successfully detected.

The processing speed was evaluated on a PC computer with Core 2 Duo 3.16 GHz processor. The results are the following:

- signature extraction: from 67 to 72 frames per second for video in PAL format, including decoding of video frames and signature extraction;
- matching video signature of 10 second video clip to signatures of all original videos (the total duration of original videos: 7 hours, and 40 minutes): 46.8 milliseconds;
- matching video signature of 1 minute video clip to signatures of all original videos: 286 milliseconds;
- matching video signature of 10 minute clip to signatures all original videos: 2 seconds and 925 milliseconds;

The above result show that, the matching time depends on the duration of query clips. For big video databases the matching time may be significant.

6 Conclusions

In this chapter a signature for content-based image copy detection is described. The extraction and matching is very fast and the false alarm is low, so the signature is suitable for big image databases (e.g., images on the Internet) or even video databases. Also the precision and recall rates are high for many commonly used modifications of images. These properties allow the signature to be applied for the detection of image copies in big image databases and on the Internet, e.g. in the Digital Rights Infringements Detection system [15]. The limitation of the signature is that it cannot detect all possible image modifications, especially heavy cropped images.

The proposed algorithm can also be adapted for video copy detection applications using signatures of consecutive video frames. This method allows for the detection of one or more duplicated video segments of arbitrary duration contained in original videos. Initial experiments have been carried out, and the detection speed has been evaluated.

Acknowledgments. This work has been supported by the European Network of Excellence VISNET-II (IST-2005-2.41.6), funded under the European Commission IST 6th Framework Program (http://www.visnet-noe.org).

References

[1] Brasnett, P., Bober, M.: Proposed Improvements to Image Signature XM 31.0. ISO/IEC JTC1/SC29/WG11 M14983 (2007)

[2] Chang, E., Wang, J., Li, C., Wilderhold, G.: Rime: A Replicated Image Detector for the World Wide Web. In: Proc. SPIE Multimed Storage Arch. Syst. III, vol. 3527, pp. 58–67 (1998)

[3] Hampapur, A., Bolle, R.: Comparison of sequence matching techniques for video copy detection. In: Conf. Storage Retr. Media Databases, pp. 194–201 (2002)

[4] Ke, Y., Sukthankar, R., Huston, L.: An Efficient Parts-based Near-duplicate and Sub-image Retrieval System. In: Proc. 12th ACM Int. Conf. Multimed., pp. 869–876 (2004)

[5] Kim, C.: Content-based image copy detection. Signal Process. Image Commun. 18, 169–184 (2003)

[6] LawTo, J., Chen, L., Joly, A., et al.: Video Copy Detection: a Comparative Study. In: Proc. 6th ACM Int. Conf. Image Video Retr., pp. 371–378 (2007)

[7] Maret, Y., Dufaux, F., Ebrahimi, T.: Adaptive Image Replica Detection Based on Support Vector Classifiers. Signal Process. Image Commun. 21(8), 688–703 (2006)

[8] Meng, Y., Chang, E., Li, B.: Enhancing DPF for Near-replica Image Recognition. IEEE Comput. Vis. Pattern Recognit. 2, 416–423 (2003)

[9] MPEG Video Sub-Group, Call for Proposals on Image and Video Signature Tools. MPEG Doc. No. N9216 (2007)

[10] Nikolaidis, N., Pitas, I.: Digital Image Watermarking: An Overview. In: Int. Conf. Multimed. Comput. Syst., vol. 1, pp. 1–6 (1999)

[11] Oh, W.G., Cho, A., Cho, I.H., et al.: Concentric Circle Partition-based Image Signature. ISO/IEC JTC1/SC29/WG11 M14956 (2007)

[12] Skarbek, W.: Singular and principal subspace of signal information system by BROM algorithm. In: Yao, J., Lingras, P., Wu, W.-Z., Szczuka, M.S., Cercone, N.J., Ślęzak, D. (eds.) RSKT 2007. LNCS (LNAI), vol. 4481, pp. 157–165. Springer, Heidelberg (2007)

[13] Swanson, M., Kobayashi, M., Tewfik, A.: Multimedia data-embedding and watermarking technologies. Proc. IEEE 86, 1064–1087 (1998)

[14] Wnukowicz, K., Skarbek, W., Galinski, G.: Trajectory of Singular Energies for Image Replica Detection. In: Int. Conf. Signal Process. Multimed. Appl. SIGMAP, Barcelona, Spain (2007)

[15] Wnukowicz, K., Galinski, G., Tous, R.: Still Image Copy Detection Algorithm Robust to Basic Image Modifications. In: Int. Symp. ELMAR, Zadar, Croatia (2008)

[16] Willems, G., Tuytelaars, T., Van Gool, L.: Spatio-temporal features for robust content-based video copy detection. In: Proc. 1st ACM Int. Conf. Multimed. Inf. Retr., pp. 283–290 (2008)

[17] Wu, M., Lin, C., Chang, C.: Image Copy Detection with Rotating Tolerance. In: Hao, Y., Liu, J., Wang, Y.-P., Cheung, Y.-m., Yin, H., Jiao, L., Ma, J., Jiao, Y.-C. (eds.) CIS 2005. LNCS (LNAI), vol. 3801, pp. 464–469. Springer, Heidelberg (2005)

An Optimally Robust Digital Watermarking Algorithm for Stereo Image Coding

S. Kumar and R. Balasubramanian

Abstract. In this chapter, a robust image watermarking algorithm in discrete wavelet transform (DWT) domain for stereo image coding is presented. First, a disparity-image is computed from the pair of stereo images using a frequency domain based matching criteria. Later, this disparity-image is used as a watermark and embedded into the left stereo image based on a modifying singular values concept. The strength of watermark is optimized using a real coded genetic algorithm to achieve the task of invisibility and robustness. The proposed scheme can achieve the following three main advantages. Any illegal user can not extract any information from the watermarked image since the host image is degraded using the ZIG-ZAG sequence. The second is that a legal user can retrieve the embedded watermark (disparity-image) and so able to recover 3-D information and right image of the stereo-pair. The third advantage is its robustness to the various attacks. Experimental results are presented to evaluate the performance of proposed algorithm in terms of accuracy and robustness.

1 Introduction

In the recent years, digital multimedia technology has shown a significant progress and growing up day by day. This technology offers so many advantages in an intelligent way compared to the old analog counterpart. Some of these advantages are transmission of data, easy editing of any part of the digital content, capability to copy a digital content without any loss in the quality of the content and many other advantages in digital signal processing and other communication applications when

S. Kumar
Dept. of Mathematics and Computer Science, University of Udine, Udine-33100, Italy
e-mail: sanjeev.kumar@dimi.uniud.it

R. Balasubramanian
Dept. of Mathematics, IIT Roorkee, Roorkee-247667, India
e-mail: balarfma@iitr.ernet.in

M. Grgic et al. (Eds.): Rec. Advan. in Mult. Sig. Process. and Commun., SCI 231, pp. 467–493.
springerlink.com

compared to analog systems. The great success of internet makes easy to use and distribute the multimedia data. However, this success creates some disadvantages also like the copyright protection [43] that arises due to great facility in copying a digital content rapidly, perfectly and without limitations on the number of copies. To avoid copyright problem, digital watermarking is proposed as a solution to prove the ownership of digital data. Digital watermarking is based on the science of steganography [36] or data hiding. It aims to embed secret information called watermark into images in a robust manner.

On the other hand, stereo vision is used in many applications of 3-D video applications and machines vision. Typically, the transmission or the storage of a stereo image sequence requires twice as much data volume as a monocular vision system. In this context, digital watermarking can be helpful to transmit the information of stereo images in a secure and robust manner with an equal requirements of monocular image data storage and transmission rate. In the other words, digital watermarking shows a great applicability in the process of secure and robust stereo image coding.

1.1 Digital Watermarks

A digital watermark is embedded inside a host image in visible or invisible manner. This watermark may be in the form of an image, digital signature, tag or label for proving the ownership or authenticity of image. Generation of these watermarks is one of main steps important steps of the process since the information containing by watermarks must be unique [17]. Usually, the generation of watermarks are achieved by involving a third party that should be the trusted as well as watermark specialist and would store a database of all its clients. There are many properties of an ideal digital watermark [10, 45, 51] and some of them are stated as follow:

- Watermark should be robust to filtering, addition of noise such as Gausian noise, compression, rotation and other image operations.
- A digital watermark should be statically and perceptually invisible to prevent security and obstruction of original image.
- Watermark extraction should be easy, fast and computationally intensive.
- The watermark should be able to determine the truth owner of the image.
- The detection of a non-marked image and the non-detection of a marked image should be less.

1.2 Digital Watermarking

In the digital watermarking algorithms, a watermark is embedded into the original data in such a way that it remains present as long as the perceptible quality of the content is at an acceptable level. The owner of the original data proves his/her ownership by extracting the watermark from the watermarked content in case of multiple ownership claims.

The multimedia object may be an image or audio or video. In particular any watermarking scheme contains three parts:

- The encoder.
- The decoder.
- The Comparator.

In general, an encoder function E takes an image I and watermark W as input and generates a watermarked image I'. Mathematically,

$$E(I,W) = I' \tag{1}$$

Here, W may not necessarily be strongly dependent on the host image I. Commonly used embedding techniques can be classified into additive [9], multiplicative [9] and quantization-based schemes [7, 16]. In additive schemes, there are usually very weak dependencies between W and I. In multiplicative schemes, W and I are dependent on each other. Strong local dependencies between W and I exist in quantization based watermarking schemes. However, these dependencies are such that statistically I and W appear independent. In some techniques, host image may be degraded using some procedure like ZIG-ZAG sequence [33], Arnold transform [59] before embedding the watermark for security purpose. Fig. 1(a) represents a general encoding process.

A decoder function D takes a watermarked image J that is possibly corrupted by attacks and decodes a watermark W'. In this process an additional image I can also be included which is often the un-watermarked version of image J. This is due to the

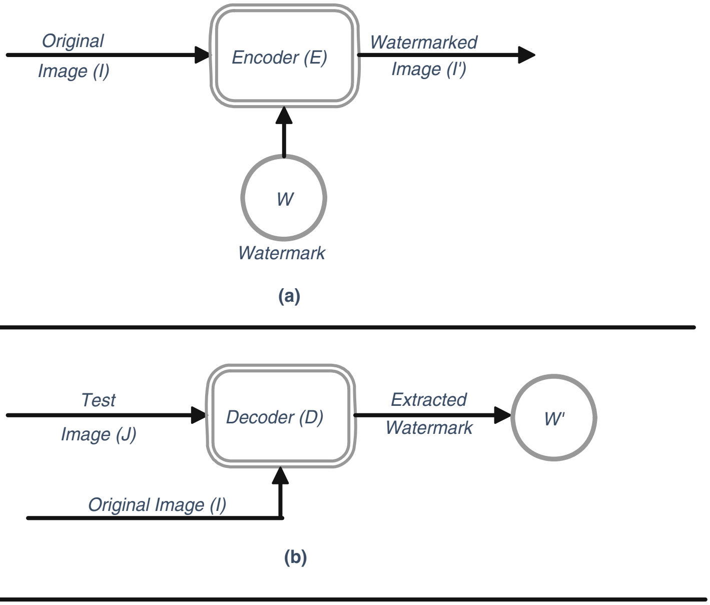


Fig. 1 A simple encoder and decoder in digital image watermarking

fact that some encoding schemes may make use of host images in the watermarking process to provide extra robustness against intentional and unintentional corruption of pixels. Figure 1(b) illustrates a simple decoder for watermarking scheme. Mathematically,

$$D(I,J) = W' \tag{2}$$

The extracted watermark W' is compared with the original watermark W using a comparison criterion. The decision of ownership is made based on the comparison results. If we talk in terms of binary decision, and assume that C_t is the decision function, then

$$C_t(W,W') = \begin{cases} 1, & c \leq t \\ 0, & otherwise \end{cases} \tag{3}$$

where c is the matching result of two watermarks and t is the certain threshold [38].

1.3 Types of Digital Watermarking

Watermarking techniques can be divided into various categories based on different criterions. The watermarks can be embedded in spatial domain. An alternative to spatial domain watermarking is frequency domain watermarking. It has been pointed out that the frequency domain methods are more robust than the spatial domain techniques. A classification of different types of watermarking scheme is shown in the Fig. 2.

In spatial domain based watermarking algorithms, a watermark is embedded directly in the spatial coordinates of host image. However, not much information can be embedded in the images without detecting the flat featureless regions [43]. Some algorithms attempt to incorporate most of the information into textured or definite edges but care must be taken to maintain the originality of host image. A common method of watermarking has been given by altering the least significant bit of each pixel in a pseudo-random manner [3]. However, a poor robustness has been offered by this algorithm and required the original image for extraction of the watermark. An improve method based on the statical detection theory has been given by overlapping the image with a binary mask having same as image [45]. This technique has been found robust against signal processing attacks. Instead of using a mask the same size as of the image, binary patterns forming 2×2 or 3×3 blocks have been used in [42]. These altered methods have been found much more resilient to low pass and median pass filtering techniques. It has been found that a combination of these two methods also proved robust against a medium level of JPEG compression.

Later, few other methods have been developed in the spatial domain using 8×8 blocks as zones and the luminance averaged over each zone. Unfortunately, these old techniques are not effective against rotation, cropping or scaling known as geometric attacks. To overcome this weakness, a method based on the amplitude modulation in color images has been introduced [31]. This proved to be a good reliable method but required more improvement in terms of robustness. In [6], a highly robust approach has been introduced for spatial domain watermarking by changing the sizes and

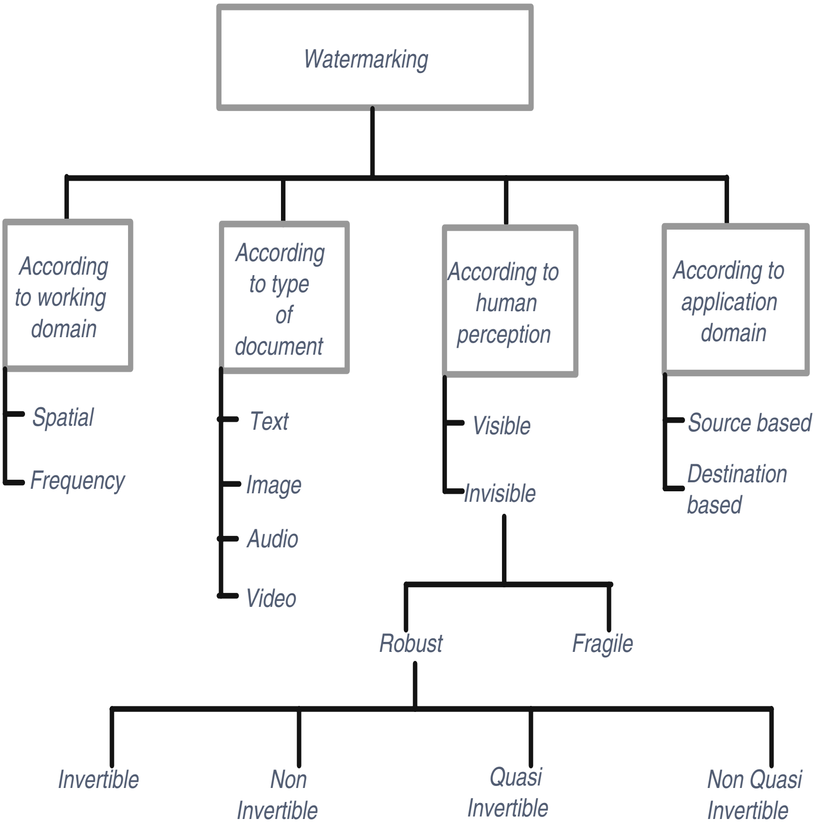


Fig. 2 Types of watermarking scheme

positions of blocks in a random manner. Some methods based on the human vision system have been introduced in [24, 53]. These techniques have been found robust against many attacks excluding a simple low pass filtering or lossy compression.

In recent years, some artificial intelligence based techniques have been developed in spatial domains. A lossless digital image watermarking algorithm based on neural network has been given in [46] by using a combination of the neural network the exclusive-or (XOR) operation to model the relationships among some randomly selected pixels with their neighborhoods. A new rotation, scaling, and translation invariant digital image watermarking scheme based on log-polar mapping and a new phase-only filtering method has been proposed in [58]. In this algorithm, watermark is embedded in the spatial domain imperceptibly based on the human perceptual model. Because of the log-polar mapping, the rotation and scaling in the spatial domain result in shifts in the log-polar mapping domain. Based on the new phase-only filtering method, the shift parameters in the log-polar mapping domain can be computed, and then the rotation degree and scaling ratio in the spatial domain can be successfully obtained through inverse log-polar mapping computation. In

[30, 41], few other spatial domain methods have been given with detail description and experimental results.

The frequency domain-based digital watermarking methods are more popular due to their more robustness when compared to spatial domain-based algorithms. Most of these algorithms are based on the Discrete Cosine Transform (DCT), Fast Fourier Transform (FFT), fractional Fourier transform (FrFT) or Discrete Wavelet Transform (DWT). In [28], host image has been divided into different blocks and then DCT of each block is found. Then these blocks are classified into six different classes in the increasing order of noise sensitivity, such as edge block, uniform with moderate intensity, uniform with high or low intensity, moderate busy, busy and very busy. Each block are assigned different L and M values. The watermark is embedded in each blocks of host image as:

$$I'_{ij} = \alpha_1 I_{ij} + \alpha_2 W_{ij} \tag{4}$$

where I'_{ij} represents the DCT co-efficient of the watermarked image, I_{ij} is the corresponding DCT co-efficient of the original image and W_{ij} is the corresponding DCT co-efficient of the watermark image.

Some other frequency domain-based watermarking schemes have been given in [9, 10]. In these schemes, the watermark has been inserted into the spectral components of the image using the technique analogous to spread spectrum communication. The argument is that the watermark must be inserted in the perceptually significant components of a signal if it is robust to common signal distortions and malicious attacks. However, the modification of these components may lead to perceptual degradation of the signal. A robust strategic invisible approach for insertion-extraction of a digital color image watermark into a host color image has been given in [39]. The beauty of this scheme is the determination of a perceptually important sub-image from the host image in such a way that slight tampering of the sub-image will affect the aesthetic of the host image significantly.

Some similar approaches to [9, 10] for invisible robust watermarking have been presented in [61, 62]. The only differences has been found the use of DWT instead of FFT/DCT and the watermark has been embedded to small number of significant coefficients instead of every high-pass wavelet coefficient. A wavelet-based fragile watermarking scheme for secure image authentication has been given in [23]. The watermark has been generated in DWT domain and then this watermark has been embedded into the least significant bit of the host image. A combination of visible/invisible watermarking called ‘dual watermarking’ has been proposed in [40]. This technique can be applied in the two different ways. It establishes the owners right to the image and second, it detects the intentional and unintentional tampering of the image. This dual watermarking technique works for both gray and color images. In [47], an other dual watermarking scheme has been proposed.

1.4 Attacks in Digital Watermarking

In most watermarking applications, the watermarked image is likely to be processed in some unsecured channel before it reaches the watermark receiver. During this processing, the watermarked image can be affected by various attacks. In watermarking terminology, an attack is any processing that may impair detection of the watermark or communication of the information conveyed by the watermark [52]. There are mainly two popular categories of watermark attacks: removal attacks [50] and geometrical attacks [32]. Removal attacks contain de-noising, compression and collusion attacks, while translation, rotation, pixel-shifting come under the second category.

Robustness can be achieved if significant modifications are made to the host image either in spatial or transform domain. However, such modifications are distinguishable and thus do not satisfy the requirement of transparency (invisibility). The design of an optimal watermarking for a given application always involves a tradeoff between these requirements. Therefore, image watermarking can be considered as an optimization problem. This optimal problem has been solved by several techniques like genetic algorithm [1, 49, 54], fuzzy logic [34], neural networks [56, 57] and support vector machine [19, 48] in spatial as well as transform domain.

1.5 Stereo Image Watermarking

Vision is a very popular tool to recover 3-D depth map information of an object (or scene) from its 2-D perspective images. So far, stereo vision has been used in many applications such as robot vision, aerial mapping, autonomous navigation, visual surveillance, 3-D television, 3-D video applications, virtual machines, medical surgery and so on. Typically, the transmission or the storage of a stereo image sequence requires twice as much data volume as a monocular vision system. However, once the stereo sequence has been compressed, its content is no more visible. Someone should first decompress the stereo image sequence in order to visualize its content, which is time consuming and not desirable in some cases, as in a stereo video data library. Therefore, there are applications where the capacity of a stereo sequences must be reduced without losing the ability to identify its content. To avoid these difficulties, many stereo coding schemes are developed [15, 18, 27]. In this context, stereo watermarking can be very helpful to perform the task of stereo sequence coding.

Recently stereo vision has been applied for image coding and security with the help of disparity map. In [2], a region-based stereo image coding algorithm has been proposed. Three types of regions: occlusion, edge and smooth have been considered for coding. The non-occluded region has been segmented into edge and smooth regions. Each region has been composed of fixed size blocks. The disparity for each block in a non-occluded region has been estimated using a block-based approach. The estimated disparity field is encoded by employing a lossy residual uniform scalar quantizer and an adaptive arithmetic coder based on segmentation. In [27], a wavelet based stereo image coding algorithm has been proposed. The wavelet transform has been used to decompose the image into an approximation and detail images. A new disparity estimation technique is developed for the estimation of the disparity field using both approximation and edge images. To improve the accuracy of estimation of wavelet images produced by the disparity compensation technique, Wavelet based subspace projection technique is developed.

Two different stereo image watermarking scheme using DCT and disparity map is proposed in [25]. A watermark image is embedded into the right image of a stereo image pair in the frequency domain through the conventional DCT operation and the disparity information between the watermarked right image and the left image is extracted. Then disparity data and the left image are simultaneously transmitted to the recipient through the communication channel. At the receiver's end, the watermarked right image is reconstructed from the received left image and the disparity data through the adaptive matching algorithm. The watermark image is finally extracted from this reconstructed right image through the decoding algorithm. Later, this technique has been proposed using DWT instead of DCT [26]. In [8], a stereo watermarking technique has been given using reversible watermarking concept. This scheme investigates the storage and bandwidth requirements reduction for stereo images. The main advantage of their scheme is that the contents of images remain available without additional manipulations.

An object-oriented method for stereo images watermarking has been proposed in [5]. Since, stereo images are characterized by the perception of depth, Therefore this scheme has been relied on the extraction of a depth map from the stereo pairs to embed the watermark. The watermark embedding has been performed in the wavelet domain using the quantization index modulation method.

1.6 Overview on Proposed Approach

This chapter focuses on the problem of embedding the watermark generated using a pair of stereo images in left stereo image in a invisible manner. Our proposal is devoted to make this watermarking scheme as robust as possible. Given a pair of stereo images, disparity-image is estimated in transform domain for reducing the computation cost. Later, this disparity-image is embedded in left stereo image in DWT domain using singular value decomposition (SVD) concept. The ZIG-ZAG sequence is performed on host image before embedding process for providing the extra security. The watermark strength is optimized using a real coded genetic algorithm for robustness of the algorithm together with maintaining the invisibility of watermark in watermarked image. On the receiver's end, disparity-image is extracted and can be used to recover right image of stereo pair as well as recovering the 3-D information of image-scene. Experimental results show that the proposed algorithm is efficient to fulfill the requirements of stereo image watermarking and coding in a secure and intelligent manner.

1.7 Organization of Chapter

The rest of the chapter is organized as follows: a brief description on the theoretical background of DWT, Wavelet decomposition, SVD and GA are given in Sec. 2.

The proposed transform domain based disparity estimation method is described in Sec. 3. In Sec. 4, proposed watermarking embedding and extracting algorithms are presented. GA based optimization of watermark strength is given in 5. The experimental results are presented in Sec. 6 and finally Sec 7 contains few concluding remarks on proposed algorithm.

2 Theoretical Background

In this section, a brief description are given on wavelet transform, wavelet decomposition, singular value decomposition and genetic algorithm.

2.1 Wavelet Transform

Generally, Fourier transform is widely used for many scientific purposes, but it is most suitable for the study of stationary signals where all frequencies have an infinite coherence time. The Fourier analysis brings only global information which is not sufficient to detect compact patterns. In [20], an introduced to local Fourier analysis has been given by taking into account a sliding window, leading to a time frequency-analysis. This method is only applicable to situations where the coherence time is independent of the frequency. On the other hand, the wavelet transform is a better tool for providing a coherence time that is proportional to the period [37].

A 1-D continuous Wavelet transform [22] can be defined as:

$$W(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} f(x) \psi^* (\frac{x-b}{a}) dx \tag{5}$$

where $\psi^*(x)$ is the analyzing wavelet, ψ^* denotes the complex conjugate of ψ and a (> 0) and b are the scale and position parameters, respectively. The transform is linear, co-invariant under translations and dilations. The wavelet transform is very suitable for analyzing hierarchical structures because of its co-invariant under dilations property.

2.2 Wavelet Decomposition

The wavelet decomposition is a multi-resolution representation of a signal using a set of basis functions generated by the dilation and translation of a unique wavelet function. Let $\phi(t)$ be a low pass scaling function and $\psi(t)$ be an associated band pass wavelet function. Using separable products of the scaling function $\phi(t)$ and wavelet function $\psi(t)$, a two dimensional wavelet decomposition can be constructed. With the two dimensional wavelet decomposition, an image function $f(x,y)$ can be decomposed into a set of independent, spatially oriented frequency channels [35]. For one level decomposition, the discrete two-dimensional wavelet transform of the image function $f(x,y)$ can be written as

$$Af = [(f(x,y) * \phi(-x)\phi(-y))(2n,2m)]_{(n,m)\in z^2} \quad (6)$$

$$D^1 f = [(f(x,y) * \phi(-x)\psi(-y))(2n,2m)]_{(n,m)\in z^2} \quad (7)$$

$$D^2 f = [(f(x,y) * \psi(-x)\phi(-y))(2n,2m)]_{(n,m)\in z^2} \quad (8)$$

$$D^3 f = [(f(x,y) * \psi(-x)\psi(-y))(2n,2m)]_{(n,m)\in z^2} \quad (9)$$

A diagram representation of above 1-level wavelet decomposition of 2-D signal $f(x,y)$ is shown in Fig. 3. The image function $f(x,y)$ is decomposed into four components. The details of these components are given in Table 1.

Wavelet decomposition provides both image intensity patterns in the approximation image and image edge patterns in the detail images within a multi-resolution

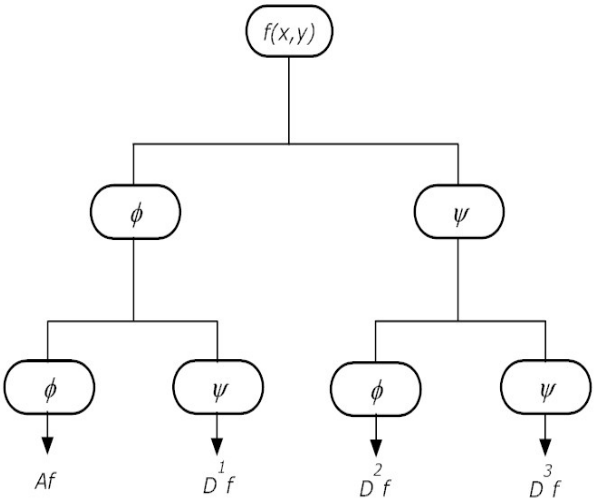


Fig. 3 1-level wavelet decomposition of a 2-D signal $f(x,y)$

Table 1 Details of decomposition components

Components	Channel	Type	frequencies
Af	LL	Approx. part	lowest
$D^1 f$	LH	detail part	vertical-high (horizontal edges)
$D^2 f$	HL	detail part	horizontal-high (vertical edges)
$D^1 f$	HH	detail part	high (diagonal edges)

representation. This is a good property for data compression since the purpose of image data compression is to compactly represent image data.

2.3 Singular Value Decomposition

Let A be a general real(complex) matrix of order $m \times n$. The singular value decomposition (SVD) of A is the factorization

$$A = U * S * V^T \tag{10}$$

where U and V are orthogonal(unitary) and $S = diag(\sigma_1, \sigma_2, ..., \sigma_r)$, where σ_i, $i = 1(1)r$ are the singular values of the matrix A with $r = \min\ (m, n)$ and satisfying

$$\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r \tag{11}$$

The first r columns of V are the ‘right singular vectors’ and the first r columns of U are the ‘left singular vectors’ of A.

2.4 Genetic Algorithm

Genetic algorithm (GA) has been introduced by John Holland in 1975, based on a method for studying natural adaptive systems and designing artificial adaptive systems, with roots in Darwinian natural selection and Mendelian genetics [21]. GAs search a problem representation space of artificial adaptive systems, eliminating weak elements by favoring retention of optimal and near optimal individuals (survival of the fittest), and recombining features of good individuals to perhaps make better individuals.

The elements of search space represent the possible solutions to the problem and are coded as strings (chromosomes), derived from an alphabet. The optimization is performed by manipulating the population of chromosomes, during a number of generations, in each of which the GA creates a set of new individual by crossover and mutation operations. GAs are particularly suitable for applications that require adaptive problem-solving strategies. A GA typically consists of the following components:

- A population of strings or coded possible solutions (chromosomes)
- A mechanism to encode a possible solution (binary or real)
- Objective function and associated fitness evaluation techniques
- Selection/reproduction procedure
- Genetic operators (crossover and mutation)
- Probabilities to perform genetic operations

The chromosomes can be encoded either in binary pattern or in form of real vectors.

3 Disparity Estimation in Wavelet Domain

Most of the energy of an image $f(x,y)$ is concentrated in its approximation part Af that is only one fourth of the original image $f(x,y)$ in size. In this context, the wavelet image Af can be very useful for image analysis and image estimation. In stereo vision, the main problem is the correspondence between different pixels of image-pairs [4]. Some researchers have solved the stereo matching problem in frequency domain [11, 29, 44, 60]. For stereo image coding, the disparity estimation is important for using disparity compensation to eliminate the cross-image redundancy between the two images of a stereo pair. The following are some advantages to estimate the disparity field using the wavelet representation of the stereo images.

- Edge information such as the length and the orientation of edges that are available in the wavelet domain may be used to improve the estimation of the disparity field.
- The size and the dynamic range of the disparity field are reduced in the wavelet domain because of the down sampling in the wavelet decomposition. This may help to save some bits to transmit the disparity field.
- Using a coarse to fine estimation strategy within multi-resolution wavelet representation, computationally, the disparity field can be more efficiently estimated in the wavelet domain.

In the wavelet domain, since the approximation component represents the same scene and contains most of the energy information of original image. Moreover, It is almost similar to the original disparity. The only difference is of the size of original disparity and the disparity estimated between approximation parts of images. Our main aim is to generate a watermark image by estimating the disparity-image. In this way, we can use this disparity-image as the watermark image. The main advantage is that the computational cost reduces to one fourth in stereo matching.

Let I_l and I_r denote the left and right stereo images and $\{AI_l, D^iI_l,\ i=1,2,3\}$ and $\{AI_r, D^iI_r,\ i=1,2,3\}$ denote their 1-level wavelet decomposition. For each pixel in the left image AI_l, similarity scores are computed by comparing a fixed, small window of size 3×3 centered on the pixel to a window in the right image AI_r, shifting along the corresponding horizontal scan line. Windows are compared through the normalized SSD measure, which quantifies the difference between the intensity patterns:

$$C = \frac{\sum\limits_{(\xi,\eta)} [AI_l(x+\xi, y+\eta) - AI_r(x+d+\xi, y+\eta)]}{\sqrt{\sum\limits_{(\xi,\eta)} AI_l(x+\xi, y+\eta)^2 \sum\limits_{(\xi,\eta)} AI_r(x+\xi, y+\eta)^2}} \tag{12}$$

The disparity estimate for pixel (x,y) is the one that minimizes the SSD error:

$$d_0(x,y) = \arg\ \min\ C(x,y,d) \tag{13}$$

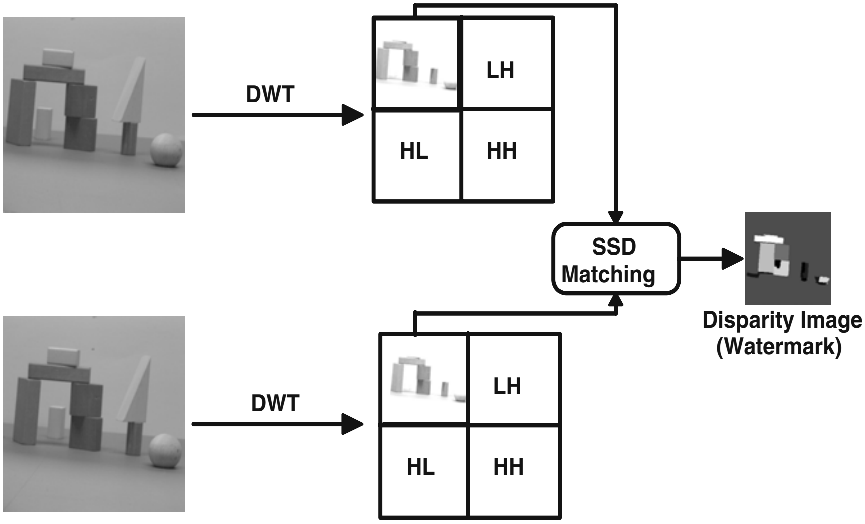


Fig. 4 Proposed stereo disparity algorithm

However we can observe that squared differences need to be computed only once for each disparity, and the sum over the window need not be recomputed from scratch when the window moves by one pixel. The disparity estimation process is shown in Fig. 4.

4 Proposed Stereo Watermarking Algorithm

Discrete wavelet transform and singular value decomposition are used for embedding and extracting of the watermark image. The left image from the stereo-pair is used as the host image. The estimated disparity-image from the wavelet images of stereo-pair is used as the watermark image. Let us consider I_l be the left image of stereo-pair, which is gray scale images of size $m \times n$. The disparity-image estimated by the algorithm given in Sect. 3 is considered as watermark. Therefore, the size of watermark image is $\frac{m}{2} \times \frac{n}{2}$ and it is one fourth of the host image I_l. In this section, the proposed algorithms for watermark embedding and extracting are given in detail.

4.1 Watermark Embedding

Without loss of generality, let the size of watermark image (disparity-image) W is $m_1 \times n_1$, where $m_1 = \frac{m}{2}$ and $n_1 = \frac{n}{2}$. The embedding process cam be accomplished in following steps:

1. First, host image I_l is degraded with the help of ZIG-ZAG sequence. let the degraded image is denoted by I_l^d.
2. Perform DWT on the degraded image I_l^d, let its wavelet approximation component be I_l^{d*}.
3. Perform SVD on I_l^{d*}.

$$I_l^{d*} = U_{I_l^{d*}}\ S_{I_l^{d*}}\ V_{I_l^{d*}}^T \tag{14}$$

4. Perform SVD on watermark image W.

$$W = U_W\ S_W\ V_W^T \tag{15}$$

5. Modify the singular values of degraded image with the singular values of the watermark as

$$S_{new} = S_{I_l^{d*}} + k\ S_W \tag{16}$$

   where k is the watermark strength, calculated using a genetic algorithm approach in this work.
6. Compute J_l^{d*} as follows

$$J_l^{d*} = U_{I_l^{d*}}\ S_{new}\ V_{I_l^{d*}}^T \tag{17}$$

7. Perform inverse DWT to construct the modified degraded image, denoted by J_l^d.

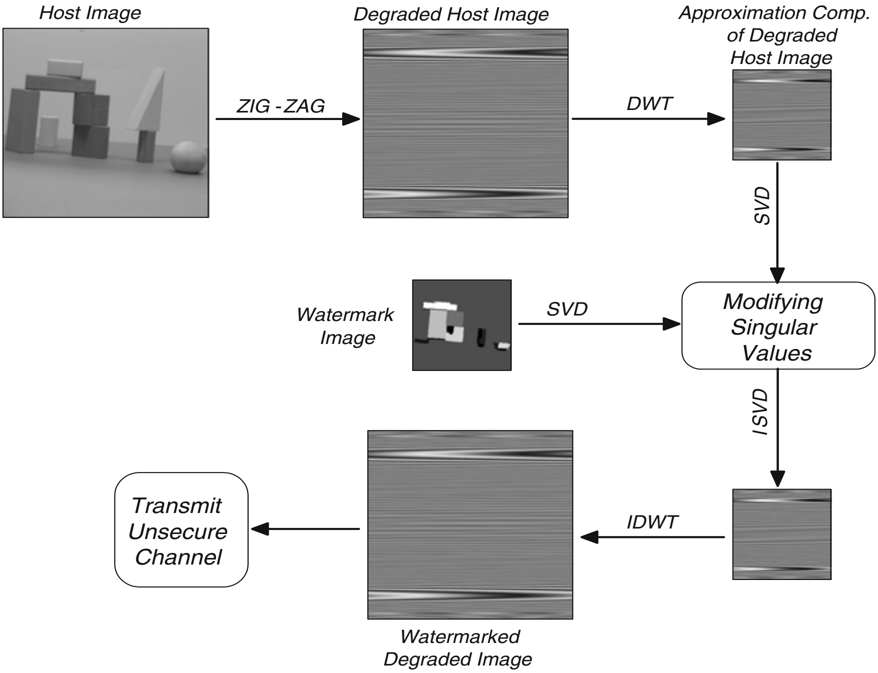

Host Image
ZIG -ZAG
Degraded Host Image
DWT
Approximation Comp. of Degraded Host Image
SVD
Watermark Image
SVD
Modifying Singular Values
ISVD
IDWT
Watermarked Degraded Image
Transmit Unsecure Channel

Fig. 5 Embedding algorithm for watermark

The ZIG-ZAG sequence is used for the propose of extra security. Once, the watermarked degraded image is transmitted into the unsecured channel, it is not possible to know any information about host image also by any hacker.

4.2 Watermark Extracting

The objective of the image decoding process is the extraction of watermark from the watermarked image. At the receiver's end, watermark can be extracted out with the help of following decoding algorithm.

1. First, perform d-ZIG-ZAG operation on the image I_l^d to find its original version I_l.
2. Perform DWT on J_l^d and to find J_l^{d*}.
3. Perform DWT on I_l and let its wavelet approximated component be I_l^*.
4. Perform SVD on J_l^*.

$$J_l^{d*} = U_{J_l^{d*}}\ S_{J_l^{d*}}\ V_{J_l^{d*}}^T \tag{18}$$

5. Perform SVD on host image I_l^*.

$$W = U_{I_l^*}\ S_{I_l^*}\ V_{I_l^*}^T \tag{19}$$

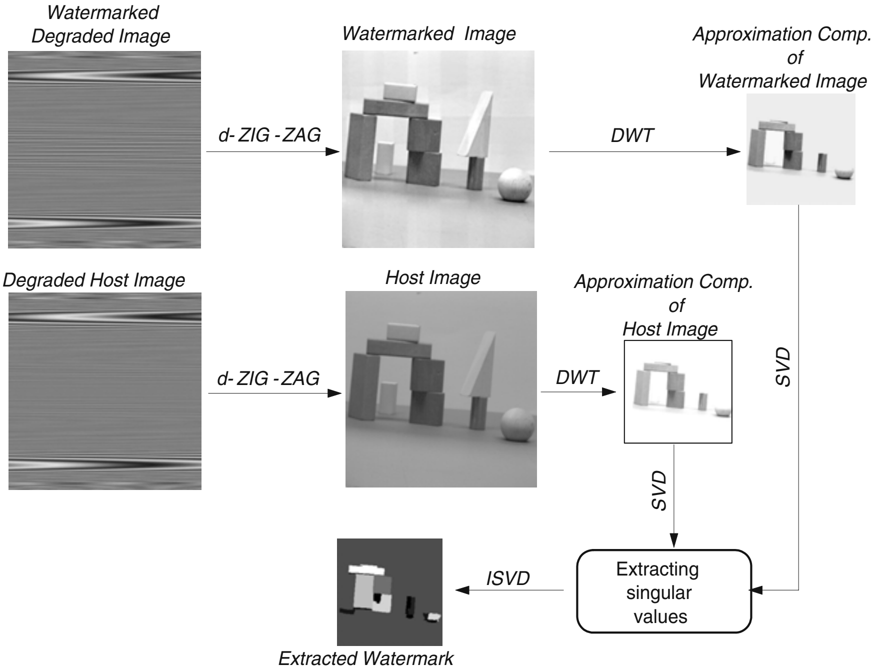

Watermarked Degraded Image
Watermarked Image
Approximation Comp. of Watermarked Image
d- ZIG -ZAG
DWT
Degraded Host Image
Host Image
Approximation Comp. of Host Image
d- ZIG -ZAG
DWT
SVD
SVD
Extracting singular values
ISVD
Extracted Watermark

Fig. 6 Extraction algorithm for watermark

6. Extract the singular values of watermark image

$$S' = \frac{(S_{J_l^{d*}} - S_{I_l^*})}{k} \tag{20}$$

7. Extract the watermark image as

$$W' = U_{I_l^*}\ S'\ V_{I_l^*}^T \tag{21}$$

A comparison between extracted watermark image W' and original watermark image W can be done by computing the cross-correlation between these two images.

5 Optimization of Watermark's Strength

It is very difficult to find the maximum power (strength) of watermark signal while it will remain invisible to the human eye. To accomplish such kind of aim, watermarked image is generated for a given value of watermark strength and allowing one or more persons to judge it. This process is repeated by increasing the value until the human deem the watermark visible. This process is complicated and human eye is not able to correctly decide the visibility of watermark from watermarked image. In this way, researchers have used neural network to compute the power of watermarking strength [12]. However, again neural network needs a large number of images for training process. The other problem with neural network is that one can not decide the correct architecture of neural network to balance the speed, reliability, convergence and memory requirements of neural network.

An important aspect of any watermarking scheme is its robustness against various type of attacks. Watermark impairment can be measured by criteria such as missing probability, or channel capacity. Our aim is to obtain the highest possible robustness without losing the transparency. In this way, robustness can be evaluated by simultaneously considering watermark impairment and the distortion of the attacked data. An attack succeeds in defeating a watermarking scheme if it impairs the watermark beyond acceptable limits while maintaining the perceptual quality of the attacked data. The job for developing a robust algorithm can be accomplished by optimizing the strength of watermark in embedding process. Once, the watermarks are extracted from the attacked watermarked images by the proposed extraction algorithm. The following steps are used to optimize the watermarking strength:

5.1 Chromosome Encoding

The success of any GA depends on the encoding of its chromosomes. GA chromosomes are encoded as real numbers instead of binary string. Total number of 20 chromosomes are used and each chromosome consists of parameter k to be searched.

5.2 Fitness Function

The fitness function is selected by considering two criterions, i.e., invisibility of watermark in watermarked image and robustness of the algorithm. Peak signal-to-noise ratio (PSNR) is used as a measure for invisibility in watermarked image. The higher value of PSNR indicates the better invisibility of watermark. The normalized cross-correlation (NC) between the embedded and extracted watermark images is used to measure the robustness of algorithm. The higher value of NC means that the more robustness in watermarking algorithm. In the GA process, following fitness function is used:

$$f = \frac{1}{((v * PSNR(k)) + \frac{1}{n}\sum_{i=1}^{n} w_i * NC_i(k))} \tag{22}$$

where n is the number of watermark attacks, NC is the normalized cross-correlation between original and extracted watermark image. The factors v and w_i are the weights for $PSNR$ and NC, respectively and having with the following relationship:

$$v + \sum_{i=1}^{n} w_i = 1.0 \tag{23}$$

Each weighting factor represents how important each index is during the searching process of GA. In order to gain the optimal performance of the watermarking system, we have to minimize f using GA process.

5.3 Selection

Selection is the stage of a genetic algorithm in which individual genomes are chosen from a population for later operations like crossover and mutation. Here, we have used roulette-wheel selection. The normalized fitness value is evaluated for each individual based on the fitness function. The population is sorted by descending fitness values. Accumulated normalized fitness values are computed that is the sum of its own fitness value plus the fitness values of all the previous individuals. The selected individual is the first one whose accumulated normalized value is greater than a randomly chosen threshold value. Sometimes, selection is performed by considering only those individuals which have a fitness value higher than a given arbitrary constant.

5.4 Crossover and Mutation

Crossover operation is used to generate the new population from the parent's population. If the crossover operation can provide a good offspring, a higher fitness value can be reached in less number of iterations. A Laplace crossover operator [13] is used for generating the new individuals. Mutation is mainly used for maintaining the diversity in a generation. Here, we have used a power mutation operator [14] for this task. This operator works based on the power distribution function. After the operation of crossover and mutation, a new population is generated. GA process is run until the most fitness chromosomes α is found.

6 Experimental Results

In order to evaluate the performance of the proposed algorithm, MATLAB platform is used for watermark embedding and extracting algorithms. The 'mex' function is used to call 'C++ code of GA' in MATLAB platform. The experimental study has been performed on three different gray scale stereo images namely Fruit, Tsukuba and Arch. These test images are downloaded from the Vision and Autonomous Systems Center's Image Database of Carnegie Mellon University [55]. All these image-pairs are of the size 512×512. First, we find out the disparity-image from stereo images using the proposed transform domain based algorithm. Disparity-image is used as watermark and left image from the stereo pair as the host image for all three pairs.

The parameters for GA should be selected carefully for obtaining the optimal results. Mainly, these parameters are crossover and mutation probabilities, number of generations and population size. To find an optimal setting for these parameters, several experiments has been conducted and selected the best possible value for each parameters (see Table 2). The number of generation is decided by taking the fitness values of independent runs. Maximum number of generation is decided when no significant improvement is noticed in the average of fitness value.

Table 2 Details of GA parameters setting

Parameters	Value
Population size	20.0
Crossover probability	0.85
Mutation probability	0.05
Number of generations	250.0

Four types of attacks are considered in order to analyze the potential of our proposed algorithm. These attacks are average filtering (AF) using a window of size 5×5, rotation (RT) of 30^o in anti clockwise direction, resizing (RS) in bilinear : $512 \rightarrow 256 \rightarrow 512$ and Gaussian noise addition (GN) with mean=0.0 and variance=0.01. The PSNR value is used to evaluate the invisibility of watermark in watermarked image. The PSNR value between host image I and watermarked image I' is calculated as

$$\text{PSNR} = 20\log_{10}\left(\frac{255}{\text{RMSE}}\right) \tag{24}$$

where

$$\text{RMSE} = \sqrt{\frac{1}{m*n}\sum_{i=1}^{m}\sum_{j=1}^{n}[I(i,j)^2 - I'(i,j)^2]} \tag{25}$$

Here, the higher value of PSNR represents the better invisibility of watermark in watermarked image.

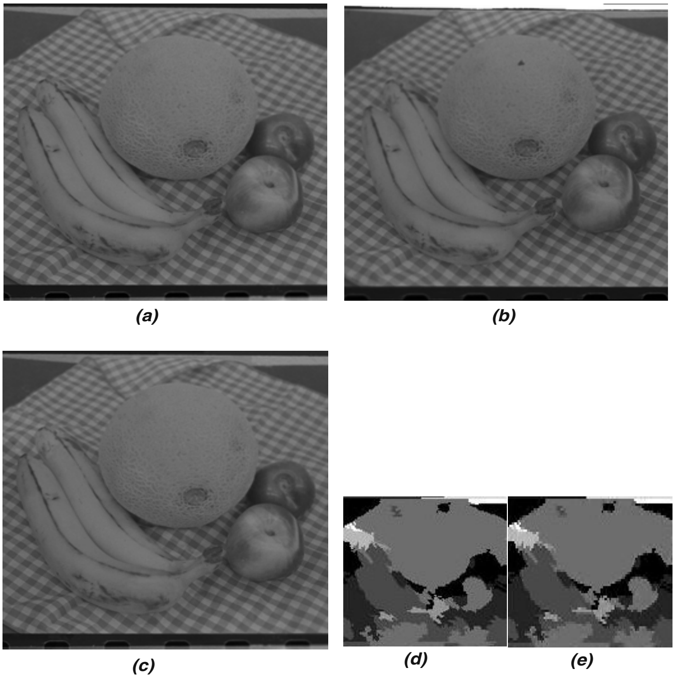


Fig. 7 Results for Fruit image-pairs: (a & b) left and right stereo images, (c) watermarked image, (d) watermark image and (e) extracted watermark

Table 3 PSNR values obtained using various strength of watermarking

Image-pairs	k= 0.9	k=0.7	k=0.5	k=0.3	k=0.1	Proposed k
Fruit	28.9770	34.4505	39.8200	42.3118	45.1218	40.5072
Tsukuba	28.0532	33.1322	38.9432	41.8712	44.5312	40.0840
Arch	29.1132	36.6400	42.6453	44.9008	47.0072	41.3744

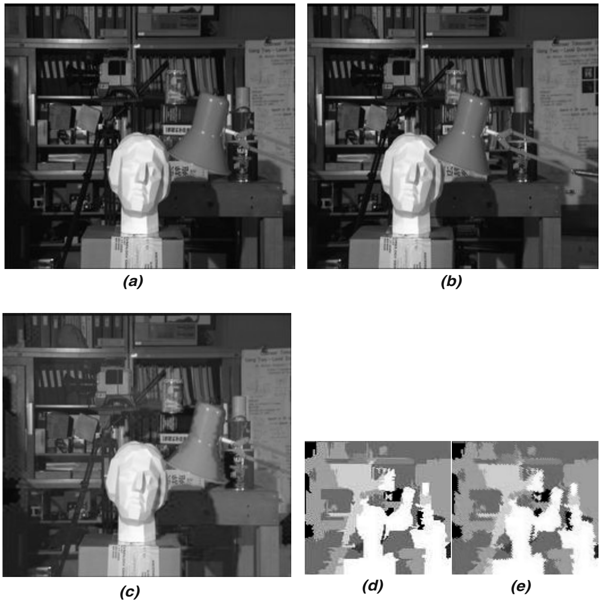


Fig. 8 Results for Tsukuba image-pairs: (a & b) left and right stereo images, (c) watermarked image, (d) watermark image and (e) extracted watermark

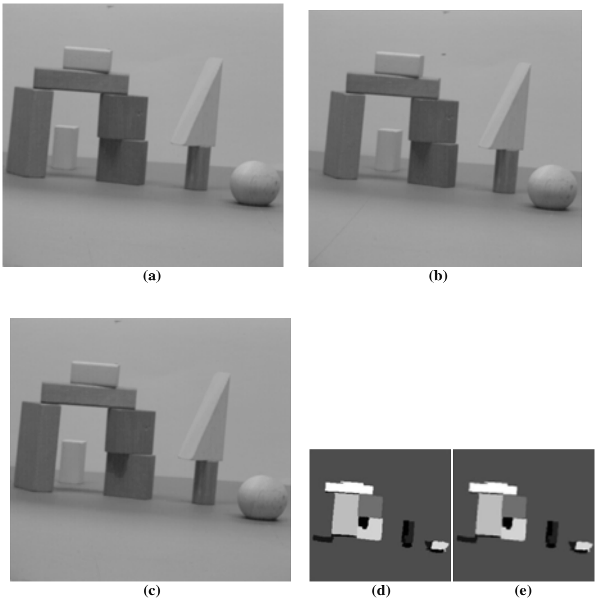


Fig. 9 Results for Arch image-pairs: (a & b) left and right stereo images, (c) watermarked image, (d) watermark image and (e) extracted watermark

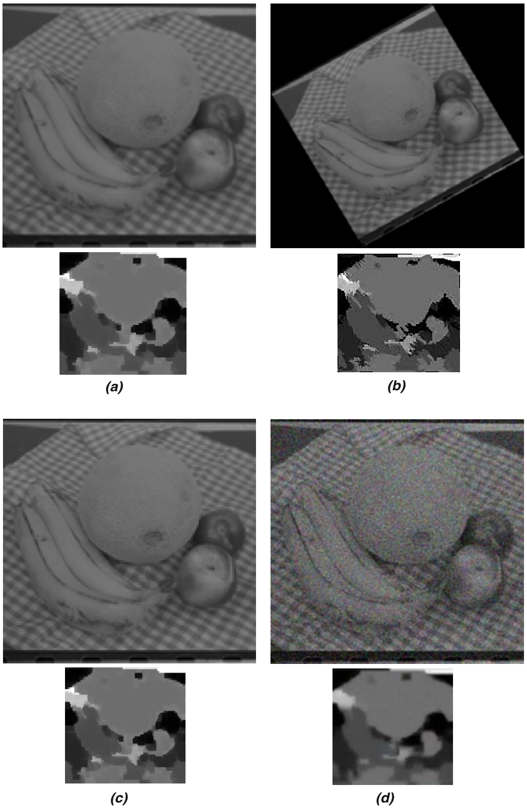


Fig. 10 Attacked watermarked image and extracted watermark for Fruit image-pairs: (a) Average filtering, (b) rotation, (c) resizing and (d) adding Gaussian noise

Table 4 *NC* values between original and extracted watermark images for various attacks

Image-pairs	AF	RT	RS	GN
Fruit	0.9803	0.9841	0.9730	0.9721
Tsukuba	0.9861	0.9874	0.9791	0.9742
Arch	0.9757	0.9801	0.9672	0.9700

Figs. 7, 8 and 9 show the experimental results for Fruit,Tsukuba and Arch image-pairs, respectively. The watermarked image quality is measured using PSNR (see Tab. 3). The robustness of proposed algorithm is represented by the results given in Figs. 10, 11 and 12. The comparison between original watermark W and extracted

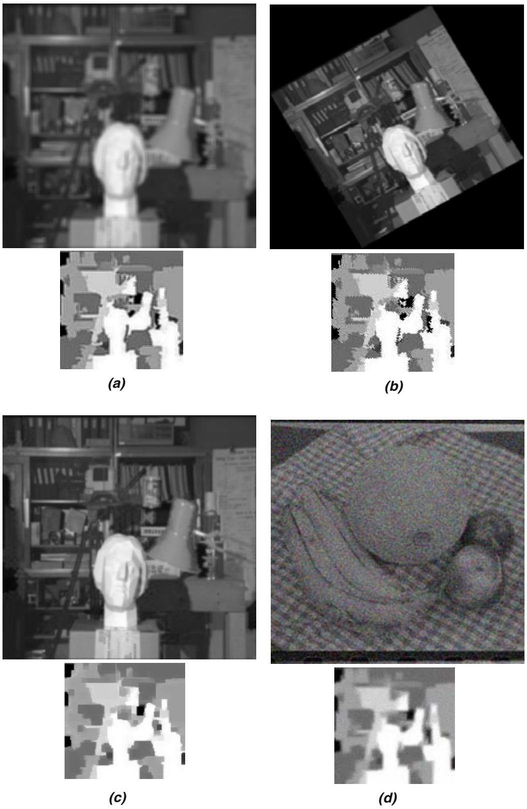


Fig. 11 Attacked watermarked image and extracted watermark for Tsukuba image-pairs: (a) Average filtering, (b) rotation, (c) resizing and (d) adding Gaussian noise

watermark W' from attacked watermarked image is made on the basis of normalized cross-correlation coefficient. The NC between two images is calculated as

$$NC(W,W') = \frac{\sum\limits_{i=1}^{r} w_i * w'_i}{\sqrt{\sum\limits_{i=1}^{r} w_i^2}\sqrt{\sum\limits_{i=1}^{r} w'^2_i}} \tag{26}$$

where w_i are the singular values of the original watermark, w'_i are the extracted singular values and $r = \max \ (m,n)$. Here, the higher value of NC represents the better robustness os proposed algorithm (see Table 4). The presented results are sufficient to prove the feasibility and its superiority over the methods in which strength of watermark is chosen randomly a constant value.

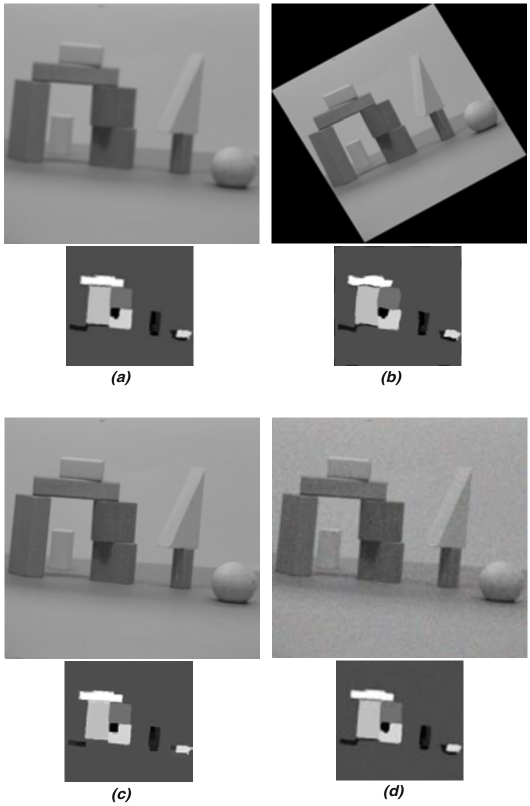


Fig. 12 Attacked watermarked image and extracted watermark for Arch image-pairs: (a) Average filtering, (b) rotation, (c) resizing and (d) adding Gaussian noise

7 Conclusions

In this chapter, an optimally robust digital watermarking algorithm has been presented for stereo image coding. A fitness function has been defined based on the two different requirements, i.e., invisibility and robustness. A real coded GA has been used to minimize the fitness function. The singular values of the host image has been modified for embedding the watermark image in DWT domain. Experimental results have been presented to show the efficiency and applicability of proposed algorithm for the task of stereo image coding. Further this algorithm can be extend in many ways like replacing DCT or FFT in place of DWT and using some other criterion instead of modifying singular values concept. The other type of attacks can be added or replaced in proposed system easily.

References

[1] Aslantas, V.: A singular-value decomposition-based image watermarking using genetic algorithm. Int. Journal on Electronics Communications 62, 386–394 (2008)

[2] Aydinoglu, H., Kossentini, F., Jiang, Q., Hayes, M.H.: Region-based stereo image coding. In: Proc. of IEEE Int. Conference on Image Processing, vol. 2, pp. 57–61 (1995)

[3] Barni, M., Bartolini, F., Cappellini, V., Piva, A.: Copyright protection of digital images by embedded unpercievable marks. Image and Vision Computing 16(12), 897–906 (1998)

[4] Brown, M., Burschka, D., Hager, G.D.: Advances in Computational Stereo. IEEE Tras. on Pattern Analysis and Machine Intelligence 25(8), 993–1008 (2003)

[5] Campisi, P.: Object-oriented stereo-image digital watermarking. Journal of Electron. Imaging 17(4), 043024 (2008)

[6] Caronni, G.: Assuring ownership rights for digital images, Reliable IT Systems VIS95. Viewreg Publishing Company, Germany (1995)

[7] Chen, B., Wornell, G.W.: Dither modulation: a new approach to digital watermarking and information embedding. In: SPIE Proc. on Security and Watermarking of Multimedia Contents, San Jose, CA, USA, vol. 3657 (1999)

[8] Coltuc, D.: On stereo embedding by reversible watermarking. Int. Symposium on Signals, Circuits and Systems 2, 1–4 (2007)

[9] Cox, I., Kilian, J., Leighton, F., Shamoon, T.: Secure Spread Spectrum Watermarking for Multimedia. IEEE Trans. on Image Processing 6(12), 1673–1687 (1997)

[10] Cox, I.J., Kilian, J., Leighton, F.T., Sjamoon, T.: A Secure Robust Watermarking for Multimedia. In: Anderson, R. (ed.) IH 1996. LNCS, vol. 1174, pp. 185–206. Springer, Heidelberg (1996)

[11] Cun, X., Jiasheng, H.: Stereo matching algorithm using wavelet transform. In: SPIE proc. on the Int. Society for Optical Engineering, vol. 4221, pp. 225–229 (2000)

[12] Davis, K.J., Najarian, K.: Maximizing strength of digital watermarks using neural networks. In: Proc. of the IEEE Int. Joint Conference on Neural networks, Washington, DC, USA, vol. 4, pp. 2893–2898 (2001)

[13] Deep, K., Thakur, M.: A new crossover operator for real coded genetic algorithms. Applied Mathematics and Computation 188(1), 895–911 (2007)

[14] Deep, K., Thakur, M.: A new mutation operator for real coded genetic algorithms. Applied Mathematics and Computation 193(1), 211–230 (2007)

[15] Duarte, M.H.V., Carvalho, M.B., Silva, E.A., et al.: Stereo image coding using multiscale recurrent patterns. In: Proc. of IEEE Int. Conference on Image Processing, Rochester, New York, vol. 2, pp. 661–664 (2002)

[16] Eggers, J.J., Su, J.K., Girod, B.: A blind watermarking scheme based on structured codebooks. In: IEE Colloquium: Secure Images and Image Authentication, London, UK (2000)

[17] Fornaro, C., Sanna, A.: Public key watermarking for authentication of CSG models. Computer-Aided Design 32(12), 727–735 (2000)

[18] Frajka, T., Zeger, K.: Residual image coding for stereo image compression. Optical Engineering 42(1), 182–189 (2003)

[19] Fu, Y.G., Shen, R.M., Lu, H.T.: Watermarking scheme based on support vector machine for colour images. Electron Letters 40, 986–993 (2004)

[20] Gabor, D.: Theory of communication. Journal of I.E.E. 93, 429–441 (1946)

[21] Goldberg, D.E.: Genetic Algorithm in search Optimization and Machine learning. Addison-Wesley, Reading (1989)

[22] Grossmann, A., Morlet, J.: Decomposition of Hardy functions into square integrable wavelets of constant shape. SIAM Journal of Mathematical Analysis 15, 723–736 (1984)

[23] He, H.J., Zhang, J., Tai, M.: A wavelet-based fragile watermarking scheme for secure image authentication. In: Shi, Y.Q., Jeon, B. (eds.) IWDW 2006. LNCS, vol. 4283, pp. 422–432. Springer, Heidelberg (2006)

[24] Hsu, C.T., Wu, J.L.: Hidden digital watermarks in images. IEEE Trans. on Image Processing 8(1), 58–68 (1999)

[25] Hwang, D.C., Bae, K.H., Lee, M.H., Kim, E.S.: Real-time stereo image watermarking using discrete cosine transform and adaptive disparity maps. In: SPIE Proc. on Multimedia Systems and Applications VI, Orlando, Florida, USA, vol. 5241, pp. 233–242 (2003)

[26] wang, D.C., Bae, K.H., Lee, M.H., Kim, E.S.: Stereo image watermarking scheme based on discrete wavelet transform and adaptive disparity estimation. In: SPIE Proc. on Mathematics of Data/Image Coding, Compression, and Encryption VI, with Applications, vol. 5208, pp. 196–205 (2004)

[27] Jiang, Q., Lee, J.J., Hayes, M.H.: A wavelet based stereo image coding algorithm. In: Proc. of IEEE Int. Conference on Acoustics, Speech, and Signal Processing, Civic Plaza, Hyatt Regebcy, Phoenix, Arizona, vol. 6, pp. 3157–3160 (1999)

[28] Kankanhalli, M.S., Rajmohan, K.R.: Adaptive Visible Watermarking of Images. In: Proc. of IEEE Int. Conf. on Multimedia Computing Systems, ICMCS 1999, Cento Affari, Florence, Italy, vol. 1, pp. 568–573 (1999)

[29] Kim, Y.S., Lee, J.J., Ha, Y.H.: Stereo matching algorithm based modefied wavelet decomposition process. Pattern Recognition 30(6), 929–952 (1997)

[30] Kim, K.S., Lee, H.Y., Lee, H.K., et al.: Practical, Real-Time, and Robust Watermarking on the Spatial Domain for High-Definition Video Contents. IEICE Trans. on Information and Systems: Special Section on Information and Communication System Security, 1359–1368 (2008)

[31] Kutter, M., Jordan, F., Bossen, F.: Digital signature of color images using amplitude modulation. In: SPIE proc. on Software and Retrieval for Image and Video Databases, pp. 518–526 (1997)

[32] Kutter, M., Petitcolas, F.: A fair benchmark for image watermarking systems. In: SPIE proc. of Electronic Imaging: Security and Watermarking of Multimedia Content, San Jose, California USA, vol. 3657 (1999)

[33] Lim, H., Lee, M.E., Park, S.Y., et al.: Robust watermarking using a block operator for secure copyright protection of digital images. Applied Informatics, 351–265 (2002)

[34] Liu, J.L., Lou, D.C., Chang, M.C., et al.: A robust watermarking scheme using self-reference image. Comput. Standards Interfaces 28, 356–367 (2006)

[35] Mallat, S.G.: A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans. on Pattern Analysis and Machine Intelligence 11, 674–693 (1989)

[36] Marnel, L.M., Boncelet, C.G., Retter, C.T.: Spread spectrum image steganography. IEEE Trans. on Image Processing 8(8), 1075–1083 (1999)

[37] Meyer, Y.: Wavelets. In: Combes, J.M., et al (eds.). Springer, Berlin (1989)

[38] Mohanty, S.P.: Digital watermarking: A tutorial review. Technical Reprot (1999), `http://www.cse.unt.edu/~smohanty/research/OtherPublications/MohantyWatermarkingSurvey1999`

[39] Mohanty, S.P., Guturu, P., Kougianos, E., Pati, N.: A Novel Invisible Color Image Watermarking Scheme Using Image Adaptive Watermark Creation and Robust Insertion-Extraction. In: Proc. of Eighth IEEE Int. Symposium on Multimedia (ISM 2006), pp. 153–160 (2006)

[40] Mohanty, S.P., Ramakrishnan, K.R., Kankanhalli, M.: A Dual Watermarking Technique for Images. In: Proc. of 7th ACM Int. Multimedia Conference, ACM-MM 1999, Orlando, USA, vol. 2, pp. 49–51 (1999)

[41] Nasir, I., Weng, Y., Jiang, J.: A New Robust Watermarking Scheme for Color Image in Spatial Domain. In: Proc. of third IEEE Int. Conference on Signal-Image Technologies and Internet-Based System, Shanghai, China, pp. 942–947 (2007)

[42] Nikolaidis, N., Pitas, I.: Copyright protection of images using robust digital signatures. In: Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 2168–2171 (1996)

[43] O-Ruanaidh, J., Petersen, H., Herrigel, A., Pereira, S., Pun, T.: Cryptographic copyright protection for digital images based on watermarking techniques. Theoretical Computer Science 226(1), 117–142

[44] Pen, H.P.: General stereo image matching using symmetric complex wavelets. In: SPIE proc. on Wavelet Applications in Signal and Image Processing IV, vol. 2825, pp. 697–720 (1996)

[45] Pitas, I.: A Method for signature casting on digital images. In: Proc. of IEEE Int. Conference on Image Processing, vol. 3, pp. 215–218 (1996)

[46] Sang, J., Alam, M.S.: A Neural Network Based Lossless Digital Image Watermarking in the Spatial Domain. In: Wang, J., Liao, X.-F., Yi, Z. (eds.) ISNN 2005. LNCS, vol. 3497, pp. 772–776. Springer, Heidelberg (2005)

[47] Sharkas, M., El-Shafie, D., Hamdy, N.: A Dual Digital-Image Watermarking Technique. Proc. of World Academy of science, Engineering and Technology 5, 136–139 (2005)

[48] Shen, R.M., Fu, Y.G., Lu, H.T.: A novel image watermarking scheme based on support vector regression. Journal of Systems Software 78, 1–8 (2005)

[49] Shih, F.Y., Wu, Y.T.: Enhancement of image watermark retrieval based on genetic algorithm. Jornal on Vision Communication and Image Representation 16, 115–133 (2005)

[50] Su, J.K., Eggers, J.J., Girod, B.: Analysis of digital watermarks subjected to optimum linear filtering and additive noise. Signal Processing 81(6), 1141–1175 (2001)

[51] Swanson, M., Zhu, B., Tewfik, A.: Transparent robust image watermarking. In: Proc. IEEE Int. Conf. on Image Processing, vol. 3, pp. 211–214 (1996)

[52] Voloshynovskiy, S., Pereira, S., Pun, T.: Attacks on digital watermarks: classification, estimation-based attacks and benchmarks. IEEE Magzine on Communications 39(8), 118–126 (2001)

[53] Wang, Z., Bovik, A.C.: Embedded foveation image coding. IEEE Trans. on Image Processing 10(10), 1397–1410 (2001)

[54] Wu, Y.T., Shih, F.Y.: Genetic algorithm based methodology for breaking the steganalytic systems. IEEE Trans. Systems Man Cybernet. Part B: Cybernet. 36, 24–31 (2006)

[55] `http://vasc.ri.cmu.edu/idb/html/stereo/index.html`

[56] Yu, P.T., Tsai, H.H., Lin, J.S.: Digital watermarking based on neural networks for color images. Signal Processing 81, 663–671 (2001)

[57] Zhang, F., Zhang, H.: Applications of a neural network to watermarking capacity of digital image. Neurocomputing 67, 345–349 (2005)

[58] Zheng, D., Liu, Y., Zhao, J.: RST Invariant digital image watermarking based on a new phase-only filtering method. Signal Processing 85(12), 2354–2370 (2005)

[59] Zhong, N., He, Z., Kuang, J., Zhou, Z.: An optimal wavelet-based image watermarking via genetic algorithm. In: Proc. of third IEEE Int. Conference on Natural Computation (ICNC), vol. 3, pp. 103–107 (2007)

[60] Zhoua, J., Xu, Y., Yanga, X.: Quaternion wavelet phase based stereo matching for uncalibrated images. Pattern Recogination Letters 28(12), 1509–1522 (2007)

[61] Zhu, W., Xiong, Z., Zhang, Y.Q.: Multiresolution Watermarking for Images and Video: A Unified Approach. In: Proc. of IEEE Int. Conf. on Image Processing, vol. 1, pp. 465–468 (1998)

[62] Zhu, W., Xiong, Z., Zhang, Y.Q.: Multiresolution Watermarking for Images and Video. IEEE Tran. on Circuits and Systems for Video Technology 9(4), 545–550 (1999)

Reversible Watermarking for 3D Cameras: Hiding Depth Maps

Asad Ali and Asifullah Khan

Abstract. This chapter presents a reversible watermarking approach based on integer wavelet transform and adaptive threshold for a novel application of watermarking. The proposed technique exploits the multi-resolution representation capability of wavelet transform for achieving high payload with low imperceptibility. Depth maps of objects obtained from sequence of 2D images and 3D Camera are secretly embedded for subsequent 3D analysis. Additionally, for efficient generation of the depth map from 2D images, we use a focus measure based on Discrete Cosine Transform and Principal Component Analysis. The approach is able not only in extracting the depth map, but also recovers the cover image. Experimental results conducted on real images acquired using the microscopic control system and 3D camera validates the concept. 3D cameras equipped with self embedding capability could be helpful in medical, military, and law enforcement image processing. Further the technique has minimal computational requirements thus enabling the visualization of embedded implementation in future 3D devices.

1 Introduction

The last decade has seen an exponential increase in digital content generation because of the ease of creation, transmission and storage of such data. It is because of this information explosion that digital watermarking has gained sizable attention from the research and academic communities, mainly due to the problems arising in securing the now easily generated, copied, and transmitted digital content. As compared to the last decade, applications of watermarking are now quite

Asad Ali
Center of Excellence in Science and Applied Technologies (CESAT), Islamabad, Pakistan
email: m.aliasad@yahoo.com

Asifullah Khan
Pakistan Institute of Engineering and Applied Sciences (PIEAS), Islamabad, Pakistan
email: asif@pieas.edu.pk

M. Grgic et al. (Eds.): Rec. Advan. in Mult. Sig. Process. and Commun., SCI 231, pp. 495–521.
springerlink.com

diverse and still increasing. This is because of the consistent emergence of the complex issues related to the security of digital content [1].

A very recent example is the patent filed by Canon, where the embedding of photographer's iris information is performed in the same image being captured. Camera would be able to perform scanning of the iris as the eye is put to the view-finder when the shot is taken. This is thus a combination of digital watermarking and iris recognition systems producing images that can be linked back to the photographer [2].

Similarly, very recently MPEG4 video codec adds watermarking capabilities and thus MPEG4000WA is introduced, which is very attractive being able to perform both authentication and integrity check. It first guarantees that the received data originated from an authentic source and secondly, that the data have not been tampered afterward [3]. Likewise, owing to the success of watermarking technologies, Philips is launching a VTrack digital watermarking solution that may deter the illicit replication of high definition movies in hotels and enable hoteliers to guarantee that the content remains available to their guests only [4].

On the other hand, one of the fundamental objectives of computer vision is to reconstruct 3D structure of objects from 2D images. Image focus analysis is one of the many approaches used for developing 3D shape recovery or depth maps of objects. The basic idea is to estimate best focus for every pixel by taking a series of images. After that, computational approaches are employed for selecting the best focused frame for each pixel [17]. In this context, shape from focus (SFF) is a cheap and fast approach. SFF could be an effective 3D shape recovery tool that may be exploited in machine vision, consumer video cameras and video microscopy [6]. For example, the depth perception is a prominent low-level task that helps a mobile robot understand the three dimensional relationship of the real world objects. Depth map can also help in achieving better surface understanding of electroformed sieves or meshes, photo-etched components, printed circuit boards, silicon engineering, etc [7]. Similarly, other examples where depth maps can be exploited are: 3D features extraction, range segmentation, estimation of object distances from camera in image sequences, and examination of the 3D shape of the microbiological species, etc.

The aim of this work is to securely hide the depth map of an object in one of its corresponding 2D images with applications to 3D cameras. Depth maps used in our experiments are generated either using a 3D camera or are estimated from a sequence of 2D images using a focus measure.

We use a new focus measure proposed in [25] based on Discrete Cosine Transform (DCT). DCT is applied on a small window around a pixel and the focus value is calculated by accumulating energies of the modified AC coefficients. The AC coefficients are modified by subtracting the DC component. The magnitude of this difference, rather than the ratio of AC and DC coefficient is considered to be more valuable in measuring the focused value in transformed domain. To further efficiently learn the variation in DCT domain energy, we employ Principal Component Analysis (PCA). The depth map is then computed by maximizing this new measure based on the absolute difference of AC and DC component and PCA.

Embedding is performed with the intention that the depth map can be extracted accurately as and when needed, but only by an authorized person. Additionally, after the extraction, the proposed method should be able to restore the 2D image to its original state so that any information represented by the image may not be lost. In order to perform this novel task with less distortion being generated at embedding stage, we introduce the concept of adaptive threshold based reversible watermarking. The proposed reversible watermarking approach could be considered as an improvement of the technique proposed by Xuan et al. [8] and our previous work [25].

The sections ahead are organized as: section 2 describes some scenarios where depth maps need to be secretly transmitted, section 3 surveys the literature on reversible watermarking, section 4 describes the depth map generation procedure for images acquired using the microscopic control system, in section 5 we present reversible data hiding technique using adaptive threshold for depth map hiding in its cover image, in section 6 we describe the implementation details for generating depth maps, in section 7 we present the results of reversible watermarking technique for images generated using the microscopic control system and 3D camera, in section 8 we present the potential applications were self embedding can be useful and finally section 9 concludes the topic.

2 Scenarios Where Depth Maps May Be Secretly Transmitted

Prospective applications of hiding depth maps as watermarks could be envisioned in mobile industry, such as, downloading of movies and 3D games on mobile devices [9]. This may either be performed for security reasons or to reduce the bandwidth requirement. Similarly, it may be helpful in extracting depth maps from printed 2D images through mobile devices for web-linking. Additionally, this secret hiding of depth maps could be supportive in military and medical image processing.

In case of medical applications, secure and fast transmission of highly valuable information between two working units is now highly desirable. One example is the safe communication of medical images and videos between island and mainland hospitals for online discussion or telesurgery [10].

Another example is the secure communication of classified 3D shape information related to injuries and skin conditions for the purpose of online negotiating interpretation and legal significance [11]. This type of secure transmission of depth information is priceless for forensic pathologists. Likewise, depth map information is very valuable for microsurgery and DNA studies [12].

As far as military applications are concerned, depth information corresponding to the 2D image could be highly confidential and would certainly require secure transmission. However, if it is intelligently embedded and secured through secret keys, no unauthorized person could extract it. After receiving the 2D image, depth map can be extracted and the cover image restored by an authorized person without compromising on its quality.

3 Reversible Watermarking

Watermark is normally embedded in cover work and is supposed to be extracted, whenever needed. However, the embedding of a watermark introduces distortion in the cover work, which may not be desirable in important applications such as medical, military, and law-enforcement related image processing. For this purpose, the concept of distortion-less or reversible watermarking has been introduced [5].

In digital watermarking applications related to multimedia archives, military and medical image processing; only reversible degradation of the original data is favorable. In multimedia archives, it is not desirable to store both the original and the watermarked versions. However, a content provider mostly wants the original content to be preserved besides the fact that distortion due to watermarking is imperceptible to most users. In other applications like military image processing and crime scene investigations, images are gathered at a high cost. Additionally, they are usually subjected to further processing steps and rigorous analysis. In such scenarios, any loss of original information may result in inaccurate analysis and thus lead to a significant error. The limitations posed by conventional watermarking approaches in applications as listed above, can be eradicated by using a reversible or lossless watermarking scheme [13].

Thus, reversible watermarking deals with the ability of a watermarking scheme to reconstruct the original data from the watermarked version. In addition, it can provide controlled access to the original content. An authorized person can access the original content by removing the watermark, while the watermarked content is still available to everyone else. This ability is not offered by the conventional watermarking approaches, where the distortions induced by watermark embedding are not reversible and thus no one has access to the original content. Regular cryptographic algorithms can also be used to achieve the reversibility property. Nonetheless, the problem with cryptographic approaches is that they cannot maintain the semantic understanding of the cover work.

Among the initial works in reversible watermarking is the one proposed by Fridrich et al. [5]. Vleeschouwer et al. [14] used circular interpretation of bijective transformations to propose a lossless watermarking scheme. Celik et al. [15] achieved high capacity by using a prediction based entropy coder in order to generalize a well-known LSB-substitution technique. Similarly, the work by Xuan et al. [8] is based on reversible embedding of the watermark bits into the middle and high frequency integer wavelet coefficients. Tian et al. [16] embeds data using the difference expansion technique. Recently, Lee et al. [13], introduced a reversible image watermarking using integer-to-integer wavelet transform and exploits the high frequency coefficients of non overlapping blocks for embedding watermark. Our use of block based embedding using adaptive threshold later is motivated by [13].

4 Depth Map Generation Using PCA in DCT Domain

Shape From Focus (SFF) is one of the passive methods for 3D shape reconstruction. A sequence of images is taken by either relocating object in the optical axis direction or by changing the focus of the camera lens. The best-focused pixel among the sequence provides depth information about corresponding object point. Once such information is collected for all points of the object, the 3D shape can be easily recovered. The first step in SFF algorithms is to apply a focus measure operator. Focus measure is defined as a quantity that locally evaluates the sharpness of a pixel. The value of the focus measure increases as the image sharpness increases and attains the maximum for the best focused image. In literature, many focus measures have been reported in the spatial as well as in the frequency domain. Modified Laplacian, Sum Modified Laplacian (SML), Tenenbaum Focus Measure, and Gray Level Variance are commonly used [17]. These methods locally compute the sharpness by considering a small 2D window around each pixel, the size of which affects the depth map accuracy and computational complexity. On the other hand, Bilal et al. [6] suggested that the accuracy of depth maps can be improved by using a 3D window around each pixel.

In this work we have used SFF for obtaining a depth map. The proposed technique, however, is general and able to embed depth maps generated through other advanced approaches [18]. An image sequence $I_k(x, y)$ consisting of k images of an object, each having X×Y pixels, is obtained by moving the image detector in small steps in the optical axis direction. For each pixel in the image volume a small window of size $N \times N$, is transformed by applying DCT . Recently, a new SFF method has been introduced based on DCT and PCA [19]. We use the same idea of employing PCA for efficiently exploiting the variations in energies in transform domain. However, instead of directly applying PCA on the AC energy part corresponding to a pixel position in question, we first compute the absolute difference of AC and DC energies. That is, to better exploit the variation in the AC energy, the DC component is first subtracted from each AC coefficient:

$$F_{(i,j)}^{k} = \sum_{u=1}^{N-1} \sum_{v=1}^{N-1} \left| F(u,v) - \quad F(0,0) \right| \tag{1}$$

where $F(u,v)$ are DCT coefficients of an image block and $F(0,0)$ represents its DC component. The energies of the modified AC coefficients for the sequence of pixels are collected into matrix $\mathbf{M}_{(i,j)} = [m_{kl}]$ where $1 \le k \le Z$ and $1 \le l \le N-1$. The eigenvalues $\boldsymbol{\lambda}$ and their corresponding eigenvectors $\mathbf{E}$ are computed from the covariance matrix $\mathbf{M}_{(i,j)}$. The transformed data $\mathbf{T}$ in eigen-space is then obtained by multiplying matrix $\mathbf{E}$ with the mean $\boldsymbol{\mu}_l$ subtracted data.

$$\mathbf{T} = \mathbf{E} \times (\mathbf{m}_{kl} - \boldsymbol{\mu}_l) \tag{2}$$

The columns of the matrix **T** are known as the principal components or features in eigenspace. The first feature is employed to calculate the depth by using formula (3). The algorithm iterates XY times to compute the complete depth map for the object. Finally, median filter is applied on the depth map to reduce the impulse noise.

$$Depth_{(i,j)} = \arg\max_k |\mathbf{t}_{k1}| \tag{3}$$

5 Adaptive Threshold Based Lossless Data Hiding

Having obtained the depth map using the above procedure or from a 3D camera the major question that arises is how to embed it in one of the cover images with minimum distortion. To answer this question in this section we present the reversible data hiding scheme that utilizes integer wavelet transform and adaptive thresholding for selective embedding and does not require a preprocessing step like histogram modification to avoid possible overflow as used by Xuan [8].

Analysis of ordinary grayscale images shows that binary 0's and 1's are almost equally distributed in the first several 'lower' bit-planes [8]. However, the bias between 0's and 1's gradually increases in the 'higher' bit-planes. In this regard, transformation of the image to frequency domain is expected to be more deliverable for obtaining a large bias between 0's and 1's. For this purpose and to avoid round-off error, we use the second generation wavelet transform, such as IDWT [20]. This wavelet transform maps integer to integer and has been adopted by JPEG2000 as well.

5.1 Watermark Embedding

Now, besides capacity, imperceptibility of a watermarking system is also highly desirable in reversible watermarking. Therefore, embedding is performed only in LH, HL, and HH, subbands which comprises of middle and high frequency coefficients. Further, data is embedded in level 1 and level 2 wavelet coefficients and header information is embedded in level 3 coefficients. In order to achieve security, secret key based permutation is employed to keep the secrecy of hidden information even after the algorithm is made public. Usually, in reversible watermarking approaches, pre-processing has to be performed before embedding to avoid possible overflow but this is not required in our case. The detailed procedure to distribute payload among the wavelet subbands during embedding phase is described below:

1. Initialize T_{MAP} and C_{ERROR} to an (X/N) × (Y/N) zero matrices, where N specifies the block size along X and Y dimensions of the image *I*.
2. The thresholds T_1, T_2 and T_3 are initialized to zero, which correspond to LH, HL and HH sub-bands of the wavelet transform.
3. Divide the input image *I* into blocks of size N × N.

   Iteration
4. Increment $T_1 = T_1 + 2$, $T_2 = T_1 + 1$, $T_3 = T_1 + 2$.
5. Compute the 2D integer wavelet transform (IWT) of *block(i,j)* of image *I* using Cohen-Daubechies-Fauraue (CDF) filters, performing decomposition upto level 3 to obtain middle and high frequency wavelet sub-bands.
6. Simulate watermark embedding in LH, HL and HH sub-bands at the three decomposition levels for all coefficients that satisfy equation (4).

$$X' = \left\{ \begin{array}{ll} 2 \cdot X + b, & if\ |X| < T_q \\ X + T_q, & if\ X \geq T_q \\ X - (T_q - 1), & if\ X \leq -T_q \end{array} \right\} \tag{4}$$

   where $T_q = \{T_1, T_2, T_3\}$ for there respective sub-bands and *X* denotes the frequency domain coefficient in question and *b* is the bit to be embedded in simulation and X' represents the modified coefficient.
7. Compute inverse integer wavelet transform to obtain the modified image block *b(i,j)'*.
8. In order to check that the embedding did not cause overflow or underflow, minimum and maximum values are found for *b(i,j)'*. Besides, Mean Square Error (MSE) is also computed.
9. If gray scale values in *b(i,j)'* are found to be within bounds for an 8-bit image we compute Change In Error (CIE) as:

$$\text{CIE} = \mid \text{MSE} - C_{ERROR}(i,j) \mid \tag{5}$$

   If CIE is found to be less than the global maximum allowed change in error (MAX_{CIE}) the threshold T_1 and MSE are recorded in T_{MAP} and C_{ERROR} respectively and the embedding capacity is incremented accordingly.
10. The iteration continues till T_1 equals T_{MAX}, at which we obtain the matrix T_{MAP} containing threshold values for each block of the input image *I* depending upon its properties that were determined adaptively.
11. Actual embedding of the depth map is then performed in level 1 and level 2 of the wavelet coefficients corresponding to LH, HL and HH sub-bands of each block using equation (4).
12. The threshold map adaptively determined here and used for watermark embedding needs to be embedded in the image to facilitate recovery as embedding has been performed in each block with a different threshold, hence

the T_{MAP} can be compressed using RLE or arithmetic coding to reduce its size significantly.

13. Finally, compressed T_{MAP} is embedded irrespective of the blocks in level 3 wavelet coefficients of LH, HL and HH sub-bands using $T_z = \{T_1, T_2, T_3\}$ to be {2, 3, 3} for each sub-band respectively. Other information embedded as part of the header includes the size of the block i.e. N.
14. The watermarked image I' is obtained by taking the inverse integer wavelet transform.

It can be observed from the embedding procedure that we use different thresholds for LH, HL and HH sub-bands. It follows from our observation in our previous work [25] that the distribution of HL and HH sub-band encompasses more high frequency content as compared to the LH sub-bands. The sub-band, where there is large high-frequency content, embedding of the bits is encouraged. Therefore, we set slightly high thresholds for sub-bands having large high-frequency content as is the case with HL and HH.

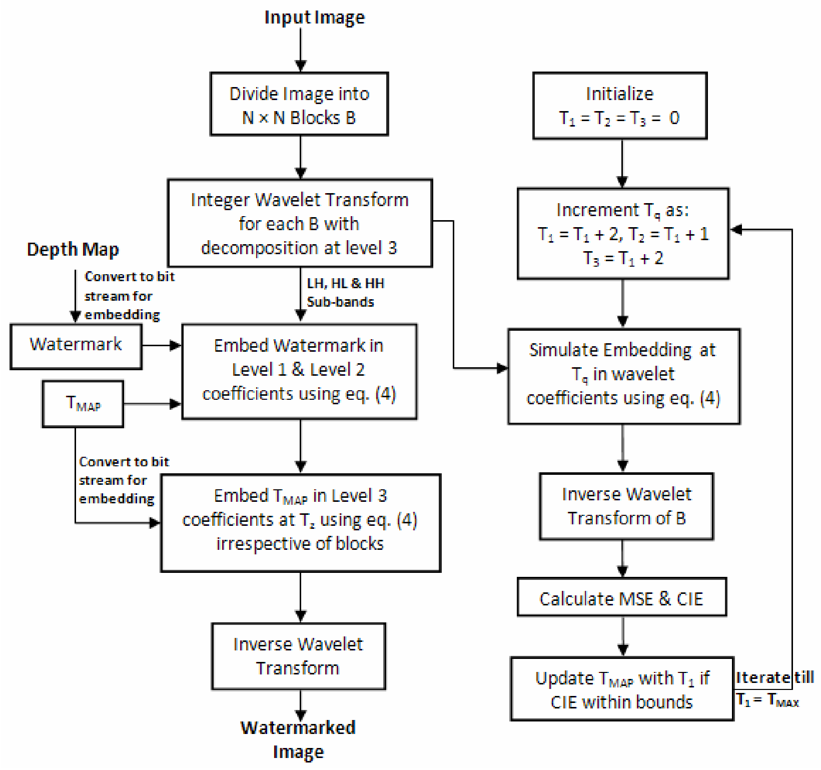

Input Image
Divide Image into N × N Blocks B
Integer Wavelet Transform for each B with decomposition at level 3
Depth Map
Convert to bit stream for embedding
LH, HL & HH Sub-bands
Watermark
Embed Watermark in Level 1 & Level 2 coefficients using eq. (4)
T_{MAP}
Convert to bit stream for embedding
Embed T_{MAP} in Level 3 coefficients at T_z using eq. (4) irrespective of blocks
Inverse Wavelet Transform
Watermarked Image
Initialize $T_1 = T_2 = T_3 = 0$
Increment T_q as: $T_1 = T_1 + 2$, $T_2 = T_1 + 1$ $T_3 = T_1 + 2$
Simulate Embedding at T_q in wavelet coefficients using eq. (4)
Inverse Wavelet Transform of B
Calculate MSE & CIE
Update T_{MAP} with T_1 if CIE within bounds
Iterate till $T_1 = T_{MAX}$

Fig. 1 Flowchart of the watermark embedding algorithm

Besides, embedding is performed in a coefficient if its magnitude is less than the threshold value set for its respective sub-band. Else, embedding is not performed; however, we do need to push the non selected coefficient away from the selected ones in terms of magnitude. So in any case, the coefficient has to be modified. This helps us in elegantly extracting the watermark at the extraction stage. Also the T_{MAP} records values of threshold T_1 only as the other two thresholds T_2 and T_3 can be obtained from it. This is the essence of our technique for improving the quality of the marked image. Fig. 1 shows the steps involved in watermark embedding in block form.

5.2 Watermark Extraction

In order to recover the depth map and restore the cover image, secret key needs to be communicated to the intender to undo the permutations on the embedded data. The steps involved in watermark extraction are described below:

1. Compute the 2D integer wavelet transform (IWT) of image *I'* using CDF filters, performing decomposition up-to level 3 to obtain LH, HL and HH wavelet sub-bands.
2. Using known threshold values of $T_z = \{T_1, T_2, T_3\}$, extract header information comprising of compressed T_{MAP} from level 3 wavelet coefficients and uncompress it to obtain actual T_{MAP}. Also extract the block size and initialize N with it.
3. Divide the input image *I'* into blocks of size N × N.
4. Compute the 2D integer wavelet transform (IWT) of *block(i,j)'* of image *I'* using CDF filters, performing decomposition up-to level 2 to obtain LH, HL and HH wavelet sub-bands.
5. For each block initialize T_1, T_2 and T_3 using values from T_{MAP}. Next, the watermark is extracted from the coefficients and original value is restored. The restoration of the coefficients based on the fact that if marked value is greater than or equal to twice the threshold of that sub-band then threshold value is subtracted from the coefficient. Else if the coefficient value is less than $(-2T_q + 1)$, we add $(T_q - 1)$ to the coefficient to restore the original value. Else the watermark bit is extracted from the LSB of the coefficient and its value is restored using division by 2 and taking the floor of the result. The extraction stage could be mathematically expressed as:

$$X = \left\{ \begin{array}{ll} \lfloor X'/2 \rfloor, & if\ -2T_q + 1 < X' < 2T_q \\ X' - T_q, & if\ X' \geq 2T_q \\ X' + (T_q - 1), & if\ X' \leq -2T_q + 1 \end{array} \right\} \qquad (6)$$

where symbol $\lfloor p \rfloor$ provides the largest integer value smaller than p. By applying equation (6), we can restore the frequency coefficients to their original values.

6. LSB's collected above group to form the original depth map where as the original image I is restored by taking the inverse integer wavelet transform of the restored coefficients at level 1 and level 2 of wavelet sub-bands.

Fig. 2 shows the detailed flowchart for watermark (depth map) extraction algorithm.

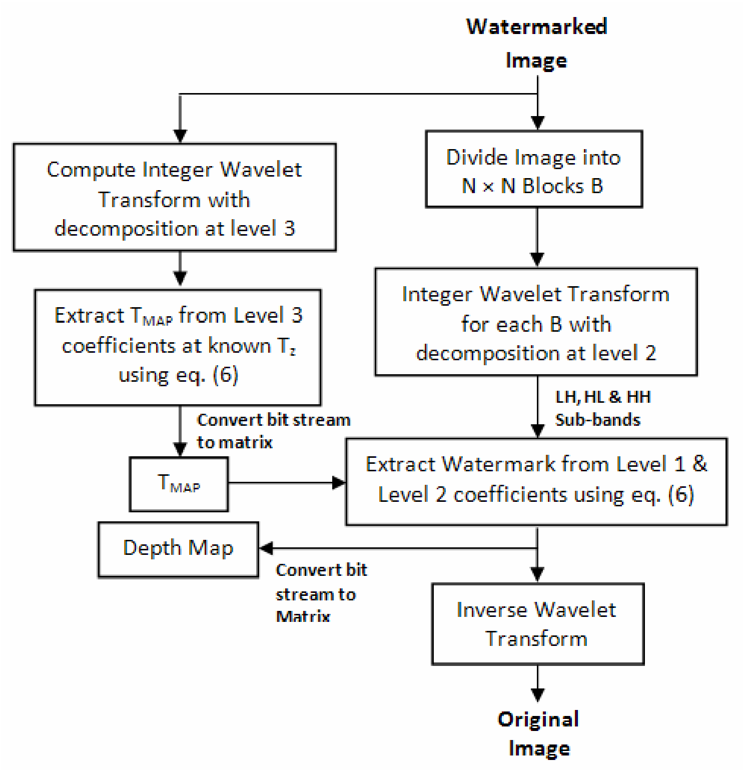

Watermarked Image
Compute Integer Wavelet Transform with decomposition at level 3
Divide Image into N × N Blocks B
Extract T_{MAP} from Level 3 coefficients at known T_z using eq. (6)
Integer Wavelet Transform for each B with decomposition at level 2
LH, HL & HH Sub-bands
Convert bit stream to matrix
T_{MAP}
Extract Watermark from Level 1 & Level 2 coefficients using eq. (6)
Depth Map
Convert bit stream to Matrix
Inverse Wavelet Transform
Original Image

Fig. 2 Flowchart of the watermark extraction algorithm for recovery of the depth map and restoration of the cover image

Keeping in view the intended application in 3D cameras, the proposed technique is simple and independent of complex floating point manipulations. Hence an embedded implementation in FPGA or DSP chip can be easily visualized provided the DSP chip possess a combination of fast memory access operations to the bit-map pixel memory and processing horsepower to handle the volume of matrix arithmetic. Some of the DSP chips designed to handle this sort of processing are already available in open market. These devices satisfy many, or all, of the needs of advanced graphics and imaging applications including a data throughput rate of more than 50 MFLOPS. Also, are capable of accessing large banks of inexpensive memory and of applying multiprocessor resources. Several features of recently developed DSP chips make them ideally suited to imaging and graphics applications.

6 Depth Map Generation: Implementation Details

Matlab is used for the simulations related to both depth map generation and its distortionless embedding. To generate depth map of an object, we first use an optical system for obtaining a sequence of frames. Computational approaches are then applied to generate the depth map. The camera and its corresponding optical system used for obtaining sequences of the above images for SFF analysis is the same as employed in [6]. We have used simulated Cone (no. of frames=97) and TFT-LCD color filter (no. of frames=60) objects for depth map analysis. One of the frames is used as a cover image. The depth is then generated by applying the SFF method based on DCT and PCA described in section 4. Size of the non-overlapping mask for computing depth map is set to 3×3 and 5×5. Generally, overlapping windows are used in SFF based approaches [17]. However, in order to reduce the size of the depth map for subsequent embedding, we have used non-overlapping windows with the assumption that the depth remains the same for all the pixels in that window. Size of each frame of simulated Cone and TFT-LCD color filter is 360×360. However, for depth map embedding in equal proportions, cover frame of simulated Cone and TFT-LCD color filter was resized to 512×512.

Images for the real microscopic objects were obtained using Microscope Control System (MCS). This system comprises of a personal computer, a frame grabber board Matrox Meteor-II, a CCD Camera (SAMSUNG CAMERA SCC-341), and a microscope (NIKON OPTIPHOT-100S). Software is used to acquire images by controlling the lens position through a step motor driver MAC 5000 having a 2.5 nm step length.

The original pixel intensities and their DCT coefficients for the sequences corresponding to the pixel position (140, 140) of TFT-LCD color filter object are shown in Fig. 3. The peaks of intensity variations appear at about frame no. 42. However, these peaks may not be obvious from DCT energy in fig. 4. For this purpose, the effect of absolute difference of AC and DC component is shown in Fig. 5, which smoothes-out the curves and clearly shows the peaks. Now which peak to consider? For this purpose, Fig. 6 shows the modified AC components being transformed into the eigenspace. The curves for the first components are smoother and have greater discriminating power with respect to focus values. Therefore, we maximize the first principal component for depth map generation. Fig. 7(b) and Fig. 8(b) show the resultant depth maps for the test objects; simulated Cone and TFT-LCD color filter.

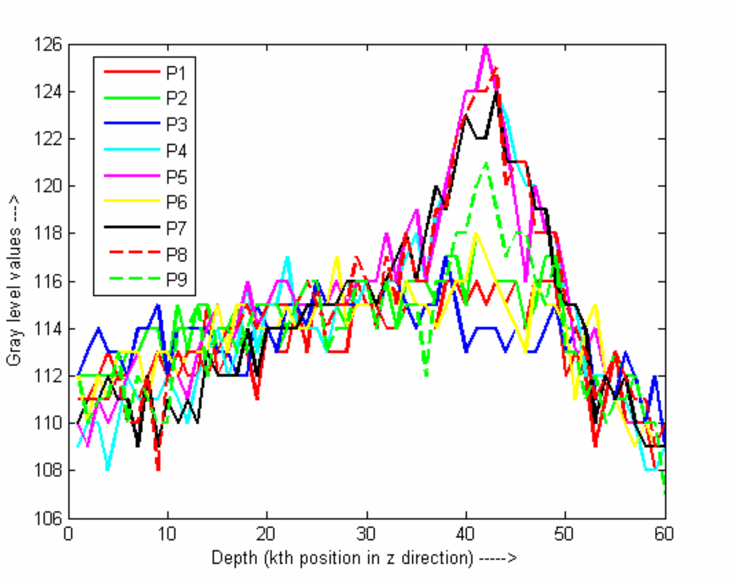


Fig. 3 Original pixel intensities for the sequences corresponding to the pixel position (140, 140) of TFT-LCD color filter object. The peaks of intensity variations appear at about frame no. 42 [25].

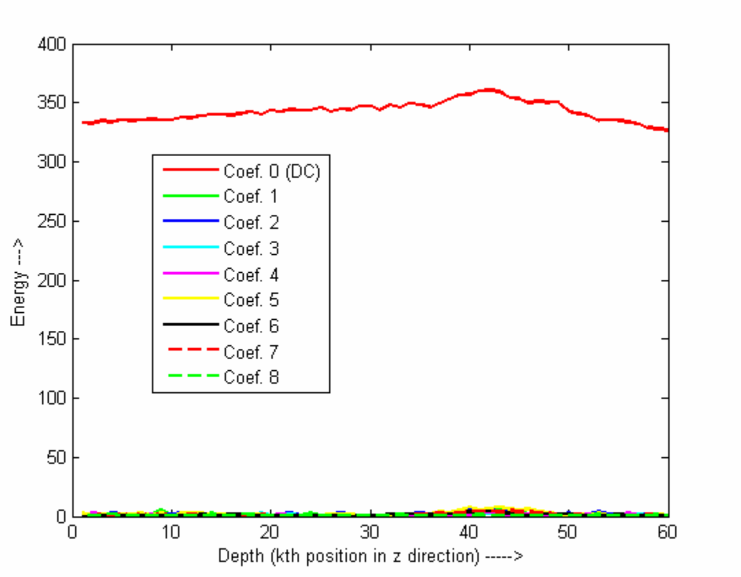


Fig. 4 DCT coefficients of original pixel intensities for the sequences corresponding to the pixel position (140, 140) of TFT-LCD color filter object [25]

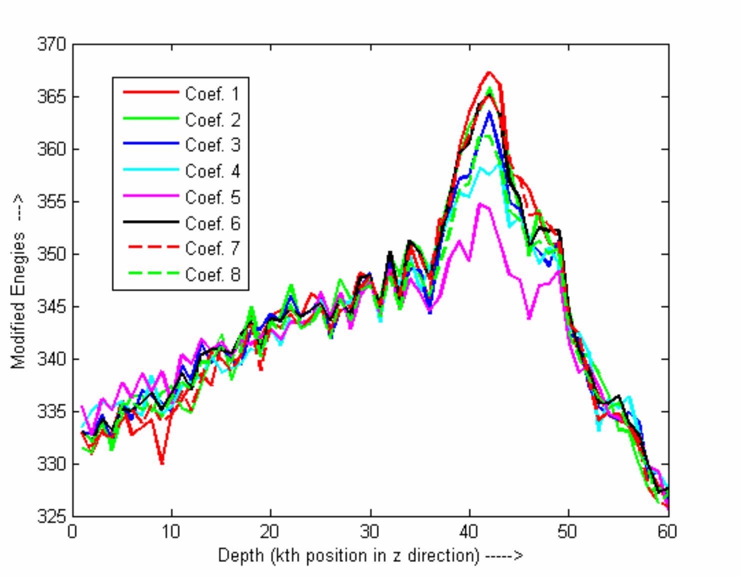


Fig. 5 Modified AC energy: The effect of absolute difference of AC and DC component [25]

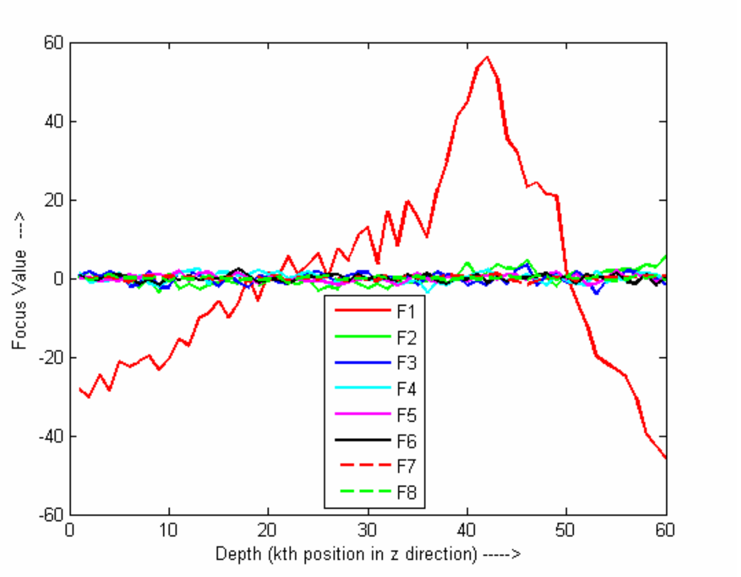
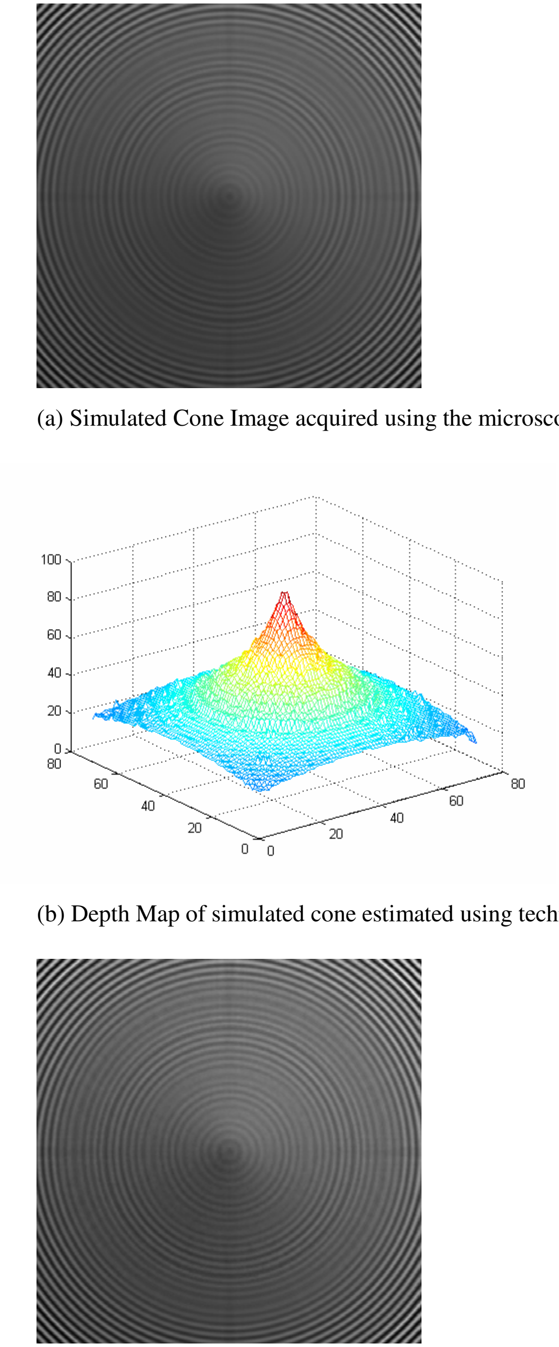


Fig. 6 Modified AC energy: Transformation into eigenspace for the point (140,140) of TFT-LCD color filter [25]

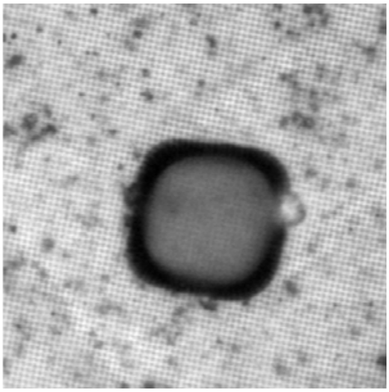


(a) Simulated Cone Image acquired using the microscopic control system

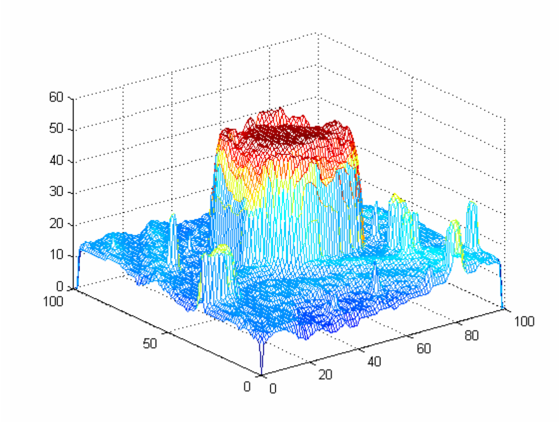
(b) Depth Map of simulated cone estimated using technique in section 4

(c) Watermarked Image with T_{MAX}=16, CIE=20, bpp=0.648499 & PSNR=40.109043

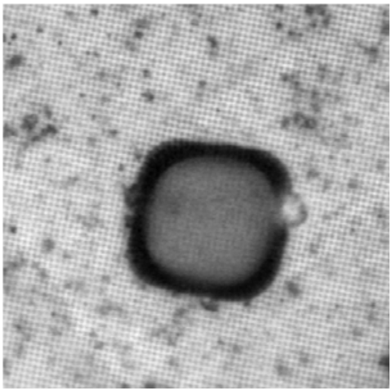
Fig. 7



(a) TFT–LCD color filter image acquired using the microscopic control system



(b) Depth map of TFT – LCD color filter estimated using the technique in section 4



(c) Watermarked Image with T_{MAX}=16, CIE=10, bpp= 0.648499 & PSNR=40.838859

Fig. 8

7 Experimental Results and Discussion

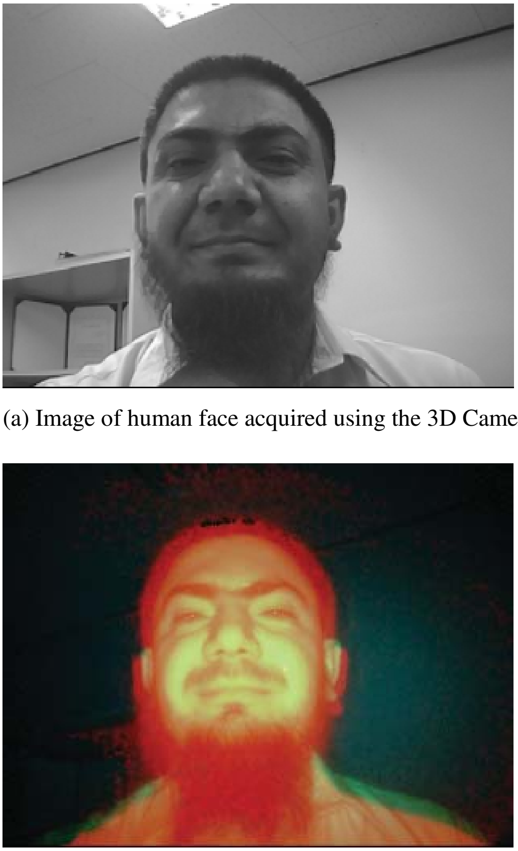
In addition to generating depth maps of objects using SFF, in our experiments we have also included depth map of human objects generated through a 3D camera. The 3D camera used to obtain depth map has two separate sensors for capturing image of the object and computing its depth map. The depth estimation technique is based on the Time-Of-Flight (TOF) principle. The depth information is captured by emitting pulses of infra-red light to the object in the scene and sensing the reflected light from the object surface. Fig. [9[a, b], 10[a, b], 11[a, b]] shows the images and their corresponding depth maps.

We applied reversible watermarking method described in section 5 to secretly hide the depth map into its corresponding image. The secret key used for random permutation is assumed to be provided to the watermark extractor through a private channel. Integer wavelet transform exploiting CDF (2,2) scheme is employed for transformation to frequency domain. All cover images were resized to 512 x 512. The value of T_{MAX} can be varied between 2 and 20 where as the value of maximum allowed change in error MAX_{CIE} can be varied between 1 and 500. Value of T_{MAX} and MAX_{CIE} directly control the quality of watermarked image. Higher the value, more the distortion and vice versa. Watermarked versions of all the images are shown in Fig. [7(c), 8(c), 9(c), 10(c), 11(c)] respectively. It is to be noted that the depth map of the image is resized to fit the available capacity.

In order to further elaborate the working of the algorithm we show T_{MAP} of all the images for different values of MAX_{CIE} in Fig. 12. Black values indicate that the block has not been selected for embedding where as white values indicate that the threshold value for that block approaches the T_{MAX} of 16. It can be observed from the figure that lower the value of MAX_{CIE} smaller the value of threshold for different blocks in the image. As we increase the MAX_{CIE} to 30, threshold for different blocks in the image tends to approach T_{MAX} allowing for higher embedding capacity.

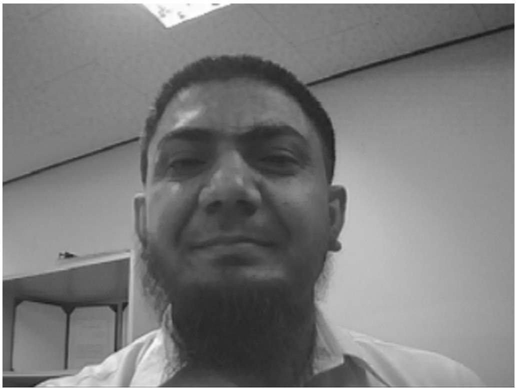
Tables 1, 2, 3, 4 and 5 compare the performance of proposed technique with fixed threshold method [8] for the five images in Fig. [7(a), 8(a), 9(a), 10(a), 11(a)]. It can be observed that the proposed method provides better imperceptibility for the same bit per pixel (bpp) compared to fixed threshold approach besides providing higher embedding capacity. This margin of improvement is clearly visible at low and high embedding rates providing high PSNR values. Even small improvement obtained in PSNR (in terms of dB) with adaptive threshold technique becomes significant when we consider medical and military applications.

It should also be noted that with the fixed threshold method in [8] the maximum embedding capacity that can be achieved is 0.75 bits per pixel (bpp) where as in our case the maximum payload that can be achieved is 0.9375 bpp for a 512 x 512 image in a single pass.



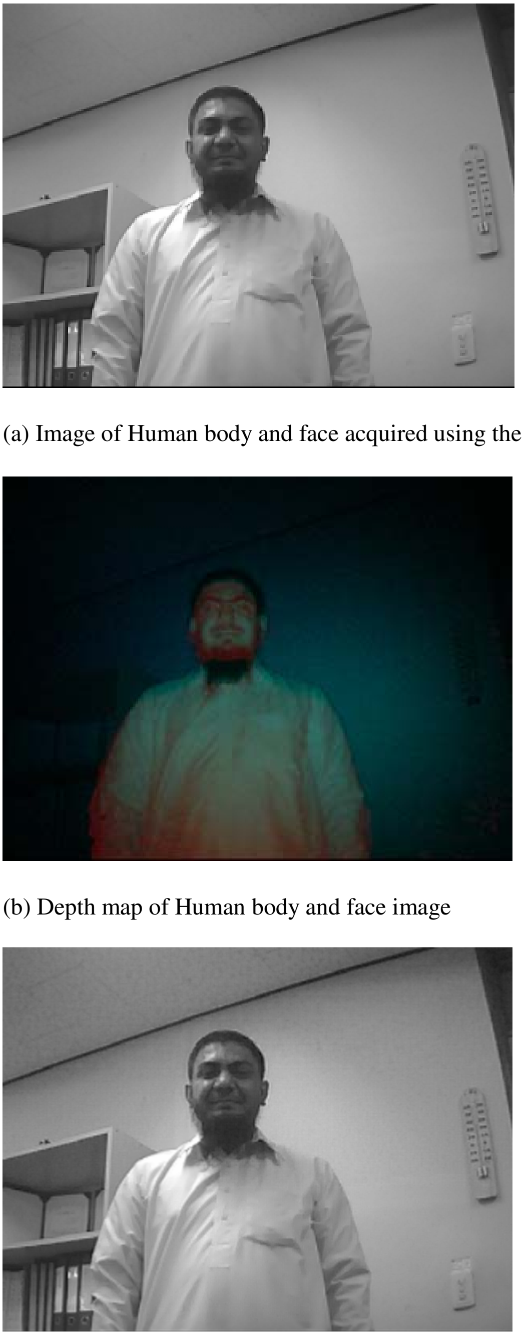
(a) Image of human face acquired using the 3D Camera

(b) Depth Map of the human face image



(c) Watermarked Image with T_{MAX}=10, CIE=7, bpp=0.686646 & PSNR=45.345813

Fig. 9



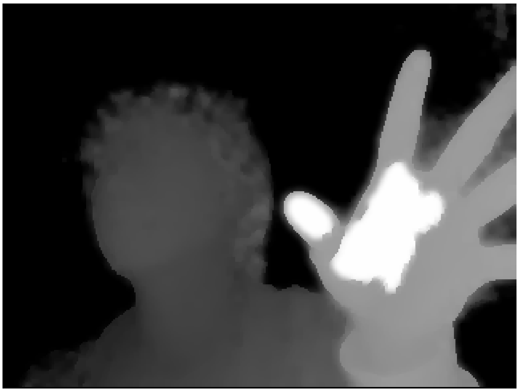
(a) Image of Human body and face acquired using the 3D Camera

(b) Depth map of Human body and face image

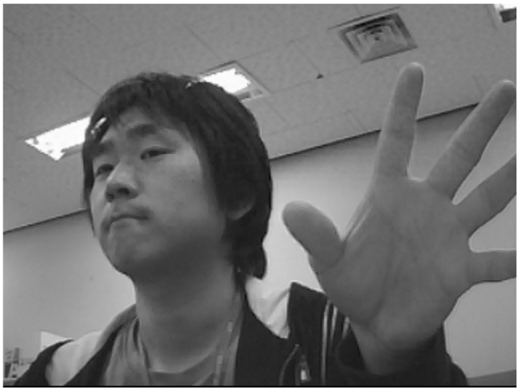
(c) Watermarked Image with T_{MAX}=10, CIE=7, bpp=0.782013 & PSNR=43.435003

Fig. 10

(a) Image of Human hand and face acquired using a 3D camera

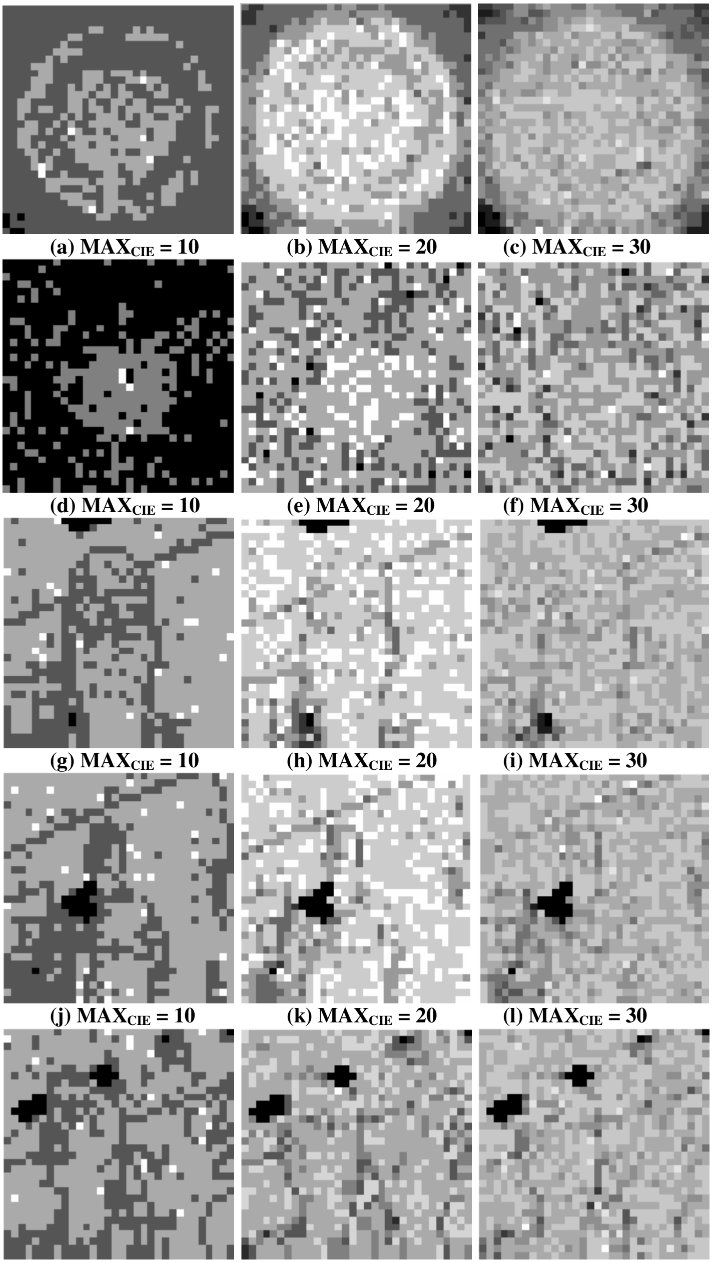


(b) Depth map of human hand and face image



(c) Watermarked Image with T_{MAX}=16, CIE=20, bpp= 0.686646 & PSNR=43.473308

Fig. 11



(a) MAX$_{CIE}$ = 10 (b) MAX$_{CIE}$ = 20 (c) MAX$_{CIE}$ = 30

(d) MAX$_{CIE}$ = 10 (e) MAX$_{CIE}$ = 20 (f) MAX$_{CIE}$ = 30

(g) MAX$_{CIE}$ = 10 (h) MAX$_{CIE}$ = 20 (i) MAX$_{CIE}$ = 30

(j) MAX$_{CIE}$ = 10 (k) MAX$_{CIE}$ = 20 (l) MAX$_{CIE}$ = 30

(m) MAX$_{CIE}$ = 10 (n) MAX$_{CIE}$ = 20 (o) MAX$_{CIE}$ = 30

Fig. 12 Shows the T$_{MAP}$ matrix of all the images used in our experiments for different values of MAX$_{CIE}$. 1st row contains T$_{MAP}$ for cone, 2nd row contains T$_{MAP}$ for TFT-LCD filter, 3rd row contains T$_{MAP}$ for Human face, 4th row contains T$_{MAP}$ of human body and 5th row contains T$_{MAP}$ of human hand and face image.

Table 1 Performance Comparisons for Fig. 8(a) TFT-LCD Color Filter

Fixed Threshold (*T=14)*		**Adaptive Threshold**	
PNSR	**BPP**	**PNSR**	**BPP**
39.37714	0.400356	41.47493	0.457764
37.22321	0.546167	41.05698	0.572205
36.04077	0.634967	40.89073	0.63324
35.37958	0.687978	40.68816	0.686646
34.84122	0.732722	39.67472	0.732422
34.56637	0.744344	38.40762	0.782013
34.38894	0.746622	38.20868	0.801086
34.31683	0.746656	34.78006	0.929363

Table 2 Performance Comparisons for Fig. 7(a) Cone

Fixed Threshold (*T=10)*		**Adaptive Threshold**	
PNSR	**BPP**	**PNSR**	**BPP**
42.72379	0.356262	43.41042	0.354767
42.31313	0.462006	42.78803	0.465393
41.83109	0.57309	42.17381	0.579834
40.93145	0.686951	41.29000	0.747681
40.66355	0.714691	39.59058	0.872150

Table 3 Performance Comparisons for Fig. 9(a) Face

Fixed Threshold(*T=12)*		**Adaptive Threshold**	
PNSR	**BPP**	**PNSR**	**BPP**
48.736197	0.369752	47.384043	0.366211
46.124895	0.649236	45.589576	0.648499
45.039516	0.692708	45.295915	0.694275
43.567226	0.717173	44.662219	0.732422
41.569356	0.732391	43.395306	0.816345
41.229346	0.734325	42.869815	0.852131
40.539727	0.7375	39.167237	0.889778

Next we compare PSNR vs bits per pixel (bpp) in Fig. 13 for the cone image of Fig. 7(a) when embedding is performed at different values of MAX_{CIE}. It can be seen that as we increase MAX_{CIE} from 10 to 30 the PSNR drops for the same bits per pixel value where as the increase in MAX_{CIE} allows for more data bits to be embedded as the embedding capacity increases with increase in MAX_{CIE}.

Effect of changing the block size from 16 × 16 to 32 × 32 can be seen in Fig. 14 for the face image in Fig. 9(a). For lower values of bits per pixel, the higher block size seems to perform better but as more and more data bits are embedded the

Table 4 Performance Comparisons for Fig. 10(a) Body

Fixed Threshold(*T=12)*		**Adaptive Threshold**	
PNSR	**BPP**	**PNSR**	**BPP**
48.781856	0.357738	47.4339	0.354767
46.154726	0.633046	45.32325	0.63324
44.925667	0.678264	45.07095	0.671387
43.92903	0.696389	44.80434	0.694275
42.485972	0.715139	44.20836	0.732422
41.04788	0.727708	43.44995	0.782013
40.608714	0.730317	42.55512	0.834026
39.920128	0.733968	38.63993	0.880859

Table 5 Performance Comparisons for Fig. 11(a) Hand

Fixed Threshold(*T=12)*		**Adaptive Threshold**	
PNSR	**BPP**	**PNSR**	**BPP**
46.05943	0.34338	46.66805	0.354767
45.28275	0.46545	45.97268	0.457764
44.9318	0.535641	45.37735	0.572205
44.15483	0.668972	44.30776	0.701904
43.69346	0.73782	42.7991	0.801086
43.5211	0.746712	38.50983	0.876736

smaller block size outperforms the other. With larger block size, threshold value adaptively selected for that block may be higher than that of the smaller block size thus allowing for more and more coefficients to be used for embedding data as bits per pixel crosses 0.65, thus decreasing the PSNR rapidly as compared to smaller block size.

Finally we compare the proposed adaptive threshold based embedding against Xuan's [26] distortion less embedding, Tian's [16] Difference Expansion (DE) and Xuan Lossless data hiding using fixed threshold [8] when used in context of depth map embedding. Fig. 15 shows the comparison in terms of PSNR vs bits per pixel (bpp). Embedding in case of Xuan's [8] technique was performed in the LSB's of the integer wavelet coefficients where as in case of Tian's pairing of the pixels was done horizontally and embedding was performed only once.

Our watermark embedding and extraction methods are simple and computationally efficient. The embedding module can be employed in the form of a chip in a 3D camera. Similarly, a decoder module can be employed at the receiver side to extract the depth map after transmission. The authorized person knowing the secret keys can then easily extract and use the depth map, and recover the image.

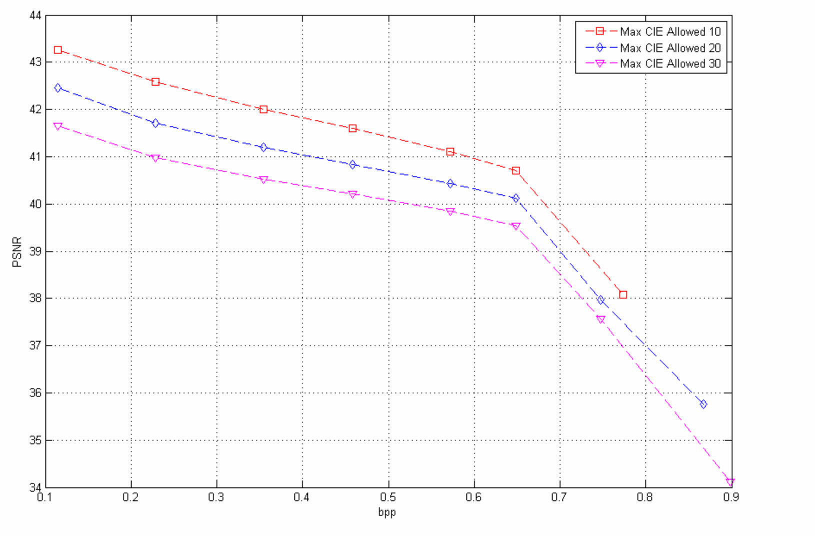


Fig. 13 PSNR vs bits per pixel (bpp) plot demonstrating the effect of using different values of $\mathrm{MAX}_{\mathrm{CIE}}$ while performing embedding in the Cone image of Fig. 7(a)

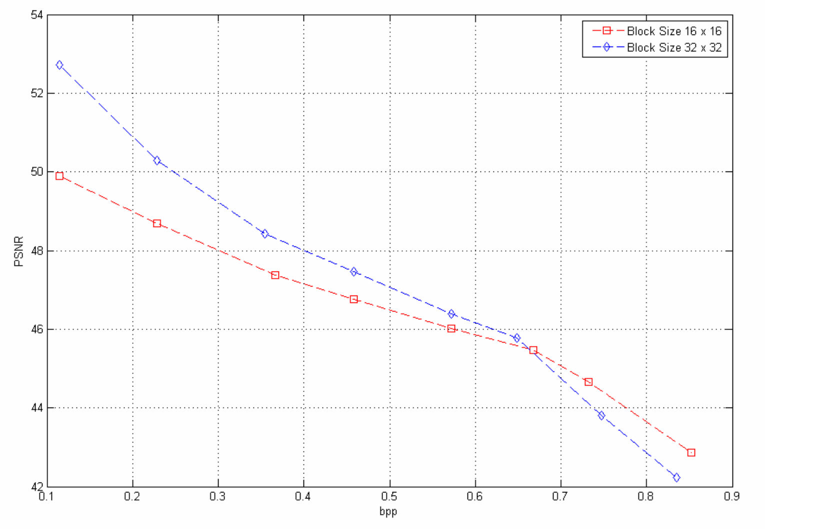


Fig. 14 Demonstrates the effect of varying the block size between 16 x 16 and 32 x 32 on face image of Fig. 9(a)

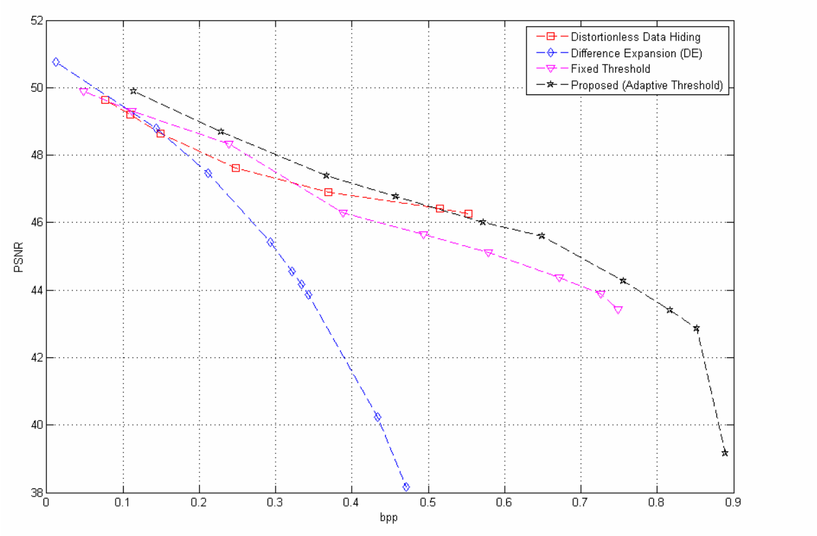


Fig. 15 Compares the performance of the proposed watermark embedding technique with Distortion-less data hiding of Xuan [26], Difference Expansion (DE) of Tian [16] and Fixed Threshold technique of Xuan [8] for the image in Fig. 9(a). Proposed technique visibly outperforms the state of the art in terms of quality and embedding capacity.

8 Potential Applications and Future Prospects

We envision several applications of our proposed idea, which are discussed below in analogy to the existing technologies for the same specific application.

(a) Depth maps could be embedded as a watermark in face data bases both to protect the data base and enhance the performance of the recognition system. Very recently, Wang et al. [21] showed that fusion of appearance image and passive stereo depth map is helpful in improving the performance of a face recognition system.

(b) Hologram stickers are used for verification, security, and even as a covert entity [22]. The complex optical patterns that they contain encode information about the depth and photographic appearance of the image. However, creating the master security hologram (originator) requires precision optical instruments, lasers and special photosensitive materials, which may be costly and time consuming. Our proposed approach could be used for the same purpose with an advantage of high security and the fact that it does not need some precision materials or precious machinery for extracting the embedded information.

(c) If the depth map is generated through a standard approach and encrypted through a hash function, the proposed approach could be used for authentication related applications [23]. This is because the depth maps have an inherent association with the pixel intensity distributions.

(d) Similarly, passive stereo depth map of a face can be embedded as a watermark in identity cards provided to employees [24]. This may help in thwarting any illicit manipulation of the image on the identity card. It may also help in providing depth map related face information for any subsequent processing.

(e) In applications, such as adaptive robotics, continuously updated depth maps are highly valuable for perceiving the local environment and taking safety measures. Such valuable information may need to be communicated safely between robots as well as between high-security sensing fields for fast and safe cooperation. Secret and safe embedding of depth maps could also be employed in security cameras to assist in separating intruders out from complex backgrounds.

(f) In case of mechanical or materials engineering related applications, if we examine a rough surface of a material, we can focus on the peaks and see these clearly. However, the lower parts of the object will be out of focus and blurred [7]. Depth map information obtained for such applications might be confidential in certain applications and therefore, should be secretly embedded in the out of focus image. Similarly, the proposed idea could also work for hiding important and confidential information in their corresponding 2D images, for example in case of electron, ultrasound, field ion emission, scanning tunneling, and atomic force microscopy.

(g) Content based image retrieval systems can use the embedded depth map information for extracting features in 3-dimensions which can then be indexed for querying later. This will significantly improve the retrieval performance of such systems and would equip them with a new level of interpretation and analysis.

9 Conclusions

In this chapter we have described a reversible watermarking technique capable of embedding depth maps of the acquired scene in its corresponding 2D image. 3D cameras equipped with self embedding capability can have potential applications in medical, military, and law enforcement related image processing, etc. To improve the imperceptibility of the existing reversible watermarking scheme, we use an adaptive threshold based algorithm operating on integer-to-integer wavelet transform coefficients. We also show that by employing PCA in the DCT domain, we can better exploit the variations in energy and thus generate improved depth maps. The technique has been tested by embedding depth maps, generated using the 3D camera and from a sequence of 2D images using shape from focus. In addition, the same idea can be used for video watermarking and authentication related applications of the cover data. Further, self embedded images obtained in this manner can significantly improve the performance of content based image retrieval applications as they add another dimension for analysis, comparison and retrieval.

References

[1] Khan, A., Tahir, S.F., Majid, A., Choi, T.S.: Machine Learning based Adaptive Watermark Decoding in View of an Anticipated Attack. Pattern Recognition 41(8), 2594–2610 (2008)

[2] Morikowa, G., Tokura, G.: Picture taking apparatus and method of controlling same. US Patent and Trademark Office, 2008/0025574 A1 (January 31, 2008)

[3] `http://www.electronicstalk.com/news/ado/ado108.html` (June 2, 2008)

[4] `http://www.ces.philips.com/press_release_watermarking.html` (June 2, 2008)

[5] Fridrich, I., Goljan, M., Du, R.: Invertible authentication. In: Proc. SPIE, Security and Watermarking of Multimedia Contents, San Jose, CA, USA, January 2001, pp. 197–208 (2001)

[6] Bilal, A.M., Choi, T.S.: Application of Three Dimensional Shape from Image Focus in LCD/TFT Displays Manufacturing. IEEE Trans. Consumer Electronics 53(1), 1–4 (2007)

[7] Jones, T.: Optical microscopy, software to the rescue: Visual inspection tools, technology represent the first line of defense when it comes to process and quality controls. Metal Finishing 105(2), 50–53 (2007)

[8] Xuan, G., Shi, Q.Y., Yang, C., Zhen, Y., Zou, D.: Lossless Data Hiding Using Integer Wavelet Transform and Threshold Embedding Technique. In: IEEE International Conference on Multimedia and Expo., July 2005, pp. 1520–1523 (2005)

[9] Luen, P., Rau, P., Chen, D.: Effects of Watermark and Music on Mobile Message Advertisements. Int. J. Human-Computer Studies 64(9), 905–914 (2006)

[10] Takahashi, T.: The present and future of telemedicine in Japan. International Journal of Medical Informatics 61(2), 131–137 (2001)

[11] Schweitzer, W., Häusler, M., Bär, W., Schaepman, M.: Evaluation of 3D surface scanners for skin documentation in forensic medicine: comparison of benchmark surfaces. BMC Medical Imaging 7(1) (2007)

[12] Ohba, K., Pedraza, J.C., Tanie, O.K., tsuji, M., Yamada, S.: Microscopic vision system with all-in-focus and depth images. Machine Vision and Applications 15(2), 56–62 (2003)

[13] Lee, S., Yoo, C.D., Kalker, T.: Reversible Image Watermarking Based on Integer-to-Integer Wavelet Transform. IEEE Transactions on Information Forensics and Security 2(3), 321–330 (2007)

[14] Vleeschouwer, C.D., Delaigle, J.F., Macq, B.: Circular interpretation of bijective transformations in lossless watermarking for media asset management. IEEE Trans. Multimedia 5(1), 97–105 (2003)

[15] Celik, M.U., Sharma, G., Tekalp, A.M., Saber, E.: Reversible data hiding. In: Proc. IEEE ICIP, Rochester, USA, September 2002, pp. 157–160 (2002)

[16] Tian, J.: Reversible data embedding using a difference expansion. IEEE Transactions on Circuits and Systems for Video Technology 13(8), 890–896 (2003)

[17] Malik, A.S., Choi, T.S.: Consideration of illumination effects and optimization of window size for accurate calculation of depth map for 3D shape recovery. Pattern Recognition 40(1), 154–170 (2007)

[18] `http://ieeexplore.ieee.org/iel5/2/4519918/04519930.pdf?tp=&isnumber=4519918&arnumber=4519930` (June 2, 2008)

[19] Mahmood, M.T., Choi, W.J., Choi, T.S.: DCT and PCA based method for shape from focus. In: Gervasi, O., Murgante, B., Laganà, A., Taniar, D., Mun, Y., Gavrilova, M.L. (eds.) ICCSA 2008, Part II. LNCS, vol. 5073, pp. 1025–1034. Springer, Heidelberg (2008)

[20] Calderbank, R., Daubechies, I., Sweldens, W., Yeo, B.L.: Wavelet transforms that map integers to integers. Appl. Comput. Harmonic Anal. 5(3), 332–369 (1998)

[21] Wang, J.G., Kong, H., Sung, E., Yau, W.Y., Teoh, E.K.: Fusion of Appearance Image and Passive Stereo Depth Map for Face Recognition Based on the Bilateral 2DLDA. EURASIP Journal on Image and Video Processing Article ID 38205, 11 (2007)

[22] `http://www.securityhologram.com/about.php` (June 2, 2008)

[23] Chamlawi, R., Khan, A., Idris, A.: Wavelet Based Image Authentication and Recovery. Journal of Computer Science and Technology 22(6), 795–804 (2007)

[24] Digimarc Corporation, Enhancing Personal Identity Verification with Digital Watermarks,
`http://csrc.nist.gov/piv-program/`
`FIPS201-Public-Comments/digimarc.pdf` (June 2, 2008)

[25] Khan, A., Ali, A., Mahmood, M.T., Usman, I., Choi, T.S.: Variable Threshold based reversible watermarking: Hiding depth maps. In: 4th IEEE/ASME International Conference on Mechatronics, Embedded Systems and Applications (MESA 2008), China, October 2008, pp. 59–64 (2008)

[26] Xuan, G., Zhu, J., Chen, J., Shi, Y.Q., Ni, Z., Su, W.: Distortionless data hiding based on integer wavelet transform. IEEE Electronics Letters 38(25), 1646–1648 (2002)

Audio Watermarking: More Than Meets the Ear

Nedeljko Cvejic, Dejan Drajic, and Tapio Seppänen

Abstract. This chapter gives an overview of the digital audio watermarking systems, including description of recently developed watermarking algorithms and an overview on the existing applications that use different audio watermarking methods. Audio watermarking algorithms are characterised by six essential properties, namely: perceptual transparency, watermark bit rate, robustness, blind/informed watermark detection, security and computational complexity. Psychoacoustic models of the HAS that are exploited in order to preserve the subjective quality of the watermarked audio during the watermarking process are shortly reviewed. The most common watermark embedding techniques ranging from the simple LSB scheme to the various spread spectrum methods are presented. Application areas that have recently been developed and possible future applications areas are listed as well.

1 Introduction

The focus of this chapter is the watermarking of digital audio (i.e. audio watermarking), including description of developed watermarking algorithms and insights of effective strategies against attacks on audio watermarking. Even though the number of published papers on watermarking and information hiding increased sharply from

Nedeljko Cvejic
Department of Engineering, University of Cambridge Trumpington Street,
CB2 1PZ Cambridge, United Kingdom
e-mail: nc332@cam.ac.uk

Dejan Drajic
Ericsson doo, Vladimira Popovica 6, 11070, Belgrade, Serbia
e-mail: dejan.drajic@ericsson.com

Tapio Seppänen
Computer Engineering Laboratory, University of Oulu, P.O. Box 4500, 4STOTKT,
90014 Oulu, Finland
e-mail: tapio.seppanen@ee.oulu.fi

M. Grgic et al. (Eds.): Rec. Advan. in Mult. Sig. Process. and Commun., SCI 231, pp. 523–550.
springerlink.com

1992, algorithms were primarily developed for digital images and video sequences (Bender, 1996; Cox, 2001); interest and research in audio watermarking started slightly later (Hartung, 1999; Swanson, 1999). In the past few years, several algorithms for embedding and extraction of watermarks in audio sequences have been presented. It is clear that audio watermarking initially started as a sub-discipline of digital signal processing, focusing mainly on convenient signal processing techniques to embed additional information to audio sequences. This included the investigation of suitable transform domain for watermark embedding and schemes for imperceptible modification of the host audio. Only recently watermarking has been placed to stronger theoretical foundation, becoming a more mature discipline with proper base in both communication modelling and information theory.

Watermarking algorithms can be characterised by a number of defining properties (Cox, 2001). In this chapter, six requirements are highlighted, that are important for audio watermarking algorithms (Arnold, 2003). For example, amount of data that can be embedded transparently into an audio sequence is considerably lower than the amount that can be hidden in images as audio signal has a dimension less than two-dimensional image files. All of the developed algorithms take advantage of perceptual properties of the human auditory system (HAS) in order to add watermark into a host signal in a perceptually transparent manner. Embedding additional information into audio sequences is more tedious task than in the case of images, due to dynamic supremacy of the HAS over human visual system (Bender, 1996). Psychoacoustic models of the HAS that are exploited in order to preserve the subjective quality of the watermarked audio during the watermarking process are shortly reviewed.

A literature survey of audio watermarking algorithms that form the mainstream research is presented as well. The algorithms are categorized by the method used for watermark detection and extraction, with references to specific algorithms using different signal domains for watermark embedding.

When the perceptual transparency requirement has been fulfilled, design objective is to increase robustness and achieve a practical watermark bit rate. Section 3 presents application areas for the audio watermarking algorithms, while Section 7 summarizes gives an overview of the chapter.

2 Basic Definitions and Terms in Digital Watermarking

Digital watermarking has been proposed as a new, alternative method to enforce the intellectual property rights and protect digital media from tampering. It involves a process of embedding into a host signal a perceptually transparent digital signature, carrying a message about the **host signal** in order to "mark" its ownership. The digital signature is called the **digital watermark**. The digital watermark contains data that can be used in various applications, including digital rights management, broadcast monitoring and tamper proofing. Although perceptually transparent, the existence of the watermark is indicated when watermarked media is passed through an appropriate watermark detector. A watermark, which usually consists of a binary data sequence, is inserted into the host signal in the **watermark embedder**. Thus, a watermark embedder has two inputs; one is the watermark message (usually accompanied by a secret key) and the other is the host signal (e.g. image, video clip, audio sequence etc.). The output of the watermark embedder is the **watermarked signal**, which cannot be perceptually discriminated from the host signal. The watermarked signal is then usually recorded or broadcasted and later presented to the **watermark detector**. The detector determines whether the watermark is present in the tested multimedia signal, and if so, what message is encoded in it. The research area of watermarking is closely related to the fields of information hiding (Johnson, 2001; Anderson, 1998) and steganography (Johnson, 1998; Katzenbeisser, 1999). Therefore, we can define **watermarking systems** as systems in which the hidden message is related to the host signal and **non-watermarking systems** in which the message is unrelated to the host signal. On the other hand, systems for embedding messages into host signals can be divided into **steganographic systems**, in which the existence of the message is kept secret, and **non-steganographic systems**, in which the presence of the embedded message does not have to be secret.

Audio watermarking initially started as a sub-discipline of digital signal processing, focusing mainly on convenient signal processing techniques to embed additional information to audio sequences. This included the investigation of a suitable transform domain for watermark embedding and schemes for the imperceptible modification of the host audio. Only recently watermarking has been placed to a stronger theoretical foundation, becoming a more mature discipline with a proper base in both communication modelling and information theory.

3 Applications of Digital Audio Watermarking

The basic goal is that embedded watermark information follows the watermarked multimedia and endures unintentional modifications and intentional removal attempts. The relative importance of the described properties significantly depends on the application for which the algorithm is designed. For copy protection applications, the watermark must be recoverable even when the watermarked signal undergoes a considerable level of distortion, while for tamper assessment applications, the watermark must effectively characterise the modification that took place. In this section, several application areas for digital watermarking will be presented and advantages of digital watermarking over standard technologies examined.

3.1 Ownership Protection

In the ownership protection applications, a watermark containing ownership information is embedded to the multimedia host signal. The watermark, known only to the copyright holder, is expected to be very robust and secure (i.e. to survive common signal processing modifications and intentional attacks) so the owner can demonstrate the presence of this watermark in case of dispute to demonstrate his ownership. Watermark detection must have a very small false alarm probability. On the other hand, ownership protection applications require small embedding capacity of the system, because the number of bits that can be embedded and extracted with small probability of error does not have to be large.

3.2 Proof of Ownership

It is even more demanding to use watermarks not only to in the identification of the copyright ownership, but as an actual proof of ownership. The problem arises when adversary uses editing software to replace the original copyright notice with his own one and then claims to own the copyright himself. Can the owner protect his rights and avoid the cost of copyright registration by applying a watermark to his multimedia file? In the case of early watermark systems, the problem was that the watermark detector was readily available to adversaries. As elaborated in (Cox, 2001), anybody that can detect a watermark can probably remove it as well. Therefore, because adversary can easily obtain a detector, he can remove owners watermark and replace it with his own. To achieve the level of the security necessary for proof of ownership, it is indispensable to restrict the availability of detector. When an adversary does not have the detector, the removal of a watermark can be made extremely difficult.

However, even if owners watermark cannot be removed, an adversary might try to undermine the owner. As described in (Cox, 2001), an adversary, using his own watermarking system, might be able to make it appear as his watermark data was present in the owners original host signal. This problem can be solved using a slight alteration of the problem statement. Instead of a direct proof of ownership by embedding e.g. a Dave owns this image watermark signature in the host image; algorithm will instead try to prove that the adversary's image is derived from the original watermarked image. Such an algorithm provides indirect evidence that it is more probable the real owner owns the disputed image, because he is the one who has the version from which the other two were created.

3.3 Authentication and Tampering Detection

In the content authentication applications, a set of secondary data is embedded in the host multimedia signal and is later used to determine whether the host signal was tampered. The robustness against removing the watermark or making it undetectable is not a concern as there is no such motivation from attacker's point of view. However, forging a valid authentication watermark in an unauthorized or tampered host signal must be prevented. In practical applications it is also desirable to locate (in time or spatial dimension) and to discriminate the unintentional modifications (e.g. distortions incurred due to moderate MPEG compression) from content tampering itself. In general, the watermark embedding capacity has to be high to satisfy the need for more additional data than in ownership protection applications. The detection must be performed without the original host signal because either the original is unavailable or its integrity has yet to be established. This kind of watermark detection is usually called a non-coherent detection.

3.4 Fingerprinting

Additional data embedded by watermark in this application is used to trace the originator or recipients of a particular copy of multimedia file. For example, watermarks carrying different serial or ID numbers are embedded in different copies of music CDs or DVDs before distributing them to a large number of recipients. The algorithms implemented in fingerprinting applications must show high robustness against intentional attacks and signal processing modifications such as lossy compression or filtering. Fingerprinting also requires good anti-collusion properties of the algorithms, i.e. it is not possible to embed more than one ID number to the host multimedia file, otherwise detector is not able to distinguish which copy is present. The embedding capacity required by fingerprinting applications is in the range of the capacity needed in copyright protection applications, with a few bits per second.

3.5 Broadcast Monitoring

A variety of applications for audio watermarking are in the field of broadcasting. Watermarking is an obvious alternative method of coding identification information for an active broadcast monitoring. It has the advantage of being embedded within the multimedia host signal itself, rather than exploiting a particular segment of the broadcast signal. Thus, it is compatible with the already installed base of broadcast equipment, including digital and analogue communication channels. The primary drawback is that embedding process is more complex than a simple placing data into file headers. There is also a concern, especially on the part of content creators, that the watermark would introduce distortions and degrade visual or audio quality of multimedia. A number of broadcast monitoring watermark-based applications are already available on commercial basis. These include program type identification; advertising research, broadcast coverage research etc. Users are able to receive a detailed proof of the performance information that allows them to:

1. Verify that the correct program and its associated promos aired as contracted;
2. Track advertising within programming;
3. Automatically track multimedia within programs using automated software online.

4 Requirements for Audio Watermarking Algorithms

Watermarking algorithms can be characterised by a number of defining properties (Cox, 2001). In this section, six of them will be highlighted, that are important for audio watermarking algorithms. The relative importance of a particular property is application-dependent and in many cases even the interpretation of a watermark property varies with the application.

4.1 Perceptual Transparency

In most of the applications, the watermark-embedding algorithm has to insert additional data without affecting the perceptual quality of the audio host signal (Zwicker, 1999). Fidelity of the watermarking algorithm is usually defined as perceptual similarity between the original and watermarked audio sequence. However, quality of the watermarked audio is usually degraded, either intentionally by an adversary or unintentionally in the transmission process, before a person perceives it. In that case, it is more adequate to define fidelity of a watermarking algorithm as perceptual similarity between watermarked audio and host audio at the point at which they are presented to a consumer (Cox, 2001). In perceptual audio coding, the quality of codecs often is evaluated by comparing an original signal, called reference, with its coded version. This general principle is applied to the quality evaluation of the watermarking systems as well. Instead of evaluating the coded version (as is the case in codec quality assessment) the watermarked version is analyzed. There are three objective measurement methods usually utilized for quality evaluation of the watermarked audio tracks. Those quality measurement systems are “Perceptual Audio Quality Measure” (PAQM) (Beerends, 1992), the system “Perceptual Evaluation of Audio Quality” (PEAQ) (ITU-R, 1998) and selected parameters of the “Noise to Mask Ratio” (NMR) (Brandenburg, 1992) measurement system.

1. PAQM derives an estimate of the signals on the cochlea and compares the representation of the reference signal with that of the signal under test. The weighted difference of these representations is mapped to the five-grade impairment scale as used in the testing of speech and audio coders. Table 1 shows this Subjective Grades (SG) scale (Sporer, 1996).

Table 1 Subjective Grades (SG) scale

SG	Description
5.0	Imperceptible
4.0	Perceptible, but not annoying
3.0	Slightly annoying
2.0	Annoying
1.0	Very annoying

2. The PEAQ system has been developed in order to get a perceptual measurement scheme that estimates the results of real world listening tests as faithfully as possible. In listening tests for very high quality signals, the test subjects sometimes confuse coded and original signal and grade the original signal below a SG of 5.0.

Therefore the difference between the grades for the original signal and the signal under test is used as a normalized output value for the result of the listening test. Table 3 also lists the corresponding Subjective Diff-Grades (SDG), which are the output values of the PEAQ system.

Table 2 Subjective Diff-Grades (SDG)

SG	Description
0.0	Imperceptible
-1.0	Perceptible, but not annoying
-2.0	Slightly annoying
-3.0	Annoying
-4.0	Very annoying

3. Overall NMR - total value expressed in dB indicates the averaged energy ratio of the difference signal with respect to a just masked signal (masking threshold). Usually, at NMR values below -10 dB there is no audible difference between the processed and the original signal.

In addition to objective measurements listening tests are usually performed as well. A number of audio sequences, that represent a broad range of music genres, are used as tests signals; usual duration of test clips is 10-20 s. In the first part of the test, participants listen to the original and the watermarked audio sequences and are asked to report dissimilarities between the two signals, using a 5-point impairment scale: (5: imperceptible, 4: perceptible but not annoying, 3: slightly annoying, 2:annoying 1: very annoying).

Results of the test should show the lowest and the highest value from the impairment scale and average MOS for given audio excerpt. In the second part, test participants are randomly presented with unwatermarked and watermarked audio clips and were asked to determine which one the watermarked one. Values near to 50% show that the two audio clips (original audio sequence and watermarked audio signal) cannot be discriminated.

4.2 Watermark Bit Rate

One of the most important properties of an audio watermarking system is watermarked bit rate, usually determined by specific demands of the application the system is designed for. Bit rate of the embedded watermark is number of embedded bits within a unit of time and is usually given in bits per second (bps).

In some applications, e.g. hiding speech in audio or compressed audio stream in audio, algorithms have to be able to embed watermarks with bit rate that is a significant fraction of the host audio bit rate, up to 150 kbps. It is a well-known fact in the audio compression community that only a few bits per sample are needed to represent music with quality near to compact disc quality music (Johnston, 1988). This implies that for uncompressed music, a significant level of noise can be injected into the signal without it being perceptible to the end user.

Contrary to the compression methods, where this fact is utilized to decrease the file size of the audio clip, in information hiding it is used to maximize the bit rate of the inserted watermark inside the perceptual requirements of the HAS. High capacity hiding algorithms are usually not robust to signal processing modifications of the watermarked audio. However, authors in (Chou, 2001) described system with watermark bit rate of 100 kbps, which does not cause distortion of the host audio sequence and is able to perfectly extract the hidden bits at a signal-to noise ratio of 15 dB.

Some audio watermarking applications, as copy control, require insertion of serial number or author ID, with average bit rate of up to 0.5 bps (Cox, 2001). On other hand, such applications demand a very high level of robustness and usually have to survive all the common signal processing modifications. For broadcast monitoring watermark bit rate is higher, caused by necessity of embedding of ID signature of a commercial within the first second at the start of the broadcast clip, with average bit rate up to 15 bps (Cox, 2001).

4.3 Robustness

Robustness of the algorithm is defined as ability of the watermark detector to extract the embedded watermark after common signal processing manipulations. Detailed overview of robustness tests is given in Section 5. Applications usually require robustness in the presence of a predefined set of signal processing modifications, so that watermark can be reliably extracted at the detection side. For example, in radio broadcast monitoring, embedded watermark need only to survive distortions caused by the transmission process including dynamic compression and low pass filtering, as watermark detection is done directly from the broadcast signal. On the other hand, in some algorithms robustness is completely undesirable and those algorithms are labelled fragile audio watermarking algorithms.

The ultimate goal of any watermarking system is reliable watermark extraction. In general, extraction reliability for a specific watermarking scheme relies on features of the original data, on the embedding distortion and on the attack distortion. Watermark extraction reliability is usually analysed for different levels of attack distortion and fixed data features and embedding distortion. Different reliability measures are used for watermark decoding and watermark detection.

In the performance evaluation of the watermark decoding, digital watermarking is considered as a communication problem. A watermark message is embedded into host signal and must be reliably decodable from the received signal. The decoding reliability is usually described by the word error probability (WER) or by the bit-error probability (BER).

Watermark detection is defined as the decision whether the received data is watermarked (hypothesis H_1) or not watermarked (hypothesis H_0). In general, both hypotheses cannot be separated perfectly. Thus, we define the probability p_{fp} (false positive) as the case of accepting H_1 when H_0 is true and the probability p_{fn} of accepting H_0 when H_1 is true (false negative). In many applications, the hypothesis test must be designed to ensure a limited false positive probability, e.g. $p_{fp} < 10^{-12}$ was proposed for watermark detection in the context of DVD copy protection (Cox et al., 2001). Another option for evaluation of watermark detection is the investigation of the total detection error probability p_e, which measures both possible error types.

4.4 Blind and Informed Watermark Extraction

The complete process of embedding and extraction of the watermark is modelled as communications channel where watermark is distorted due to presence of strong interference and channel effects. Strong interference is caused by presence of the host audio and channel effects correspond to signal processing operations.

In some applications, detection algorithm may use the original host audio to extract watermark from the watermarked audio sequence (informed extraction). It often significantly improves detector performance, in that the original audio can be subtracted from the watermarked copy, resulting in the watermark sequence alone. However, if detection algorithm does not have access to the original audio (blind extraction) and this inability substantially decreases the amount of data that can be hidden in the host signal.

In most blind watermarking schemes, e.g. blind spread spectrum watermarking, the host signal is considered as interfering noise during the watermark extraction. Nevertheless, recently it has been realized that blind watermarking can be modelled as communication with side information at the encoder. This has been published in (Chen, 1999) and (Cox, 1999) independently. The main idea is that, although the blind receiver does not have access to the host signal, the encoder can exploit his knowledge of host signal to reduce the influence of the host signal on the watermark detection and decoding.

4.5 Security

Security measures the impact on the detection capability of intentional processing dedicated to a certain class of watermarking techniques. They are sometimes called malicious attacks in the sense that the pirates know perfectly well the watermark embedding and detection algorithms, and they look for flaws in this targeted technique.

Watermark algorithm must be secure in the sense that an adversary must not be able to detect the presence of embedded data, let alone remove the embedded data. Security of watermark process is interpreted in the same way as security of encryption techniques, using the Kerckhoffs' principle. Hence, the security of the crypto-system must only stem from storing the secret key in a safe place, the rest of the system being public. The system cannot be broken unless the authorized user has access to a secret key that controls watermark embedding. An unauthorized user should be unable to extract the data in a reasonable amount of time even if he knows that the host signal contains a watermark and is familiar with the exact watermark embedding algorithm. Security requirements vary with application and the most stringent are in cover communications applications and in some cases data is encrypted prior to embedding into host audio.

4.6 Computational Complexity and Cost

Implementation of an audio watermarking system is tedious task and it depends on the business application involved. The principal issue from technical point of view is computational complexity of embedding and detection algorithms and number of embedders and detectors used in the system. E.g. in broadcast monitoring embedding and detection must be done in real time, while in copyright protection applications time is not a crucial factor for practical implementation. One of the economic issues is design of embedders and detectors, which can be implemented as hardware or software plug-ins is difference in processing power of different devices (laptop, PDA, mobile phone).

5 HAS-Based Perceptual Transparency

Watermarking of audio signals is more challenging compared to watermarking of images or video sequences, due to wider dynamic range of the HAS in comparison with human visual system (HVS). The HAS perceives sounds over a range of power greater than one billion to one and a range of frequencies greater than one thousand to one. Sensitivity of the HAS to the additive white Gaussian noise (AWGN) is high as well; this noise in a sound file can be detected as low as 70 dB below ambient level. On the other hand, opposite to its large dynamic range, HAS contains a fairly small differential range. As a result, loud sounds generally tend to mask out weaker sounds. Additionally, it is insensitive to a constant relative phase change in a static waveform, and some specific spectral distortions are interpreted as natural. It is important to take into account interaction of different frequencies and the subsequent processing of HAS to have a profound understanding of the relation between acoustic stimuli and hearing sensation. Auditory perception is based on critical band analysis in the inner ear where a frequency-to-place transformation takes place along the basilar membrane. The power spectra of the received sounds are not represented on a linear frequency scale but on limited frequency bands called critical bands (Zwicker, 1999). The auditory system is usually modelled as a band pass filterbank, consisting of band pass filters with bandwidths around 100 Hz for bands with central frequency below 500 Hz and up to 5000 Hz for bands placed at high frequencies. If we limit the highest frequency to 24000 Hz, 26 critical bands have to be taken into account. Table 3 gives an overview of the first 24 critical bands and corresponding frequencies inside the HAS frequency range.

Table 3 Critical bands and corresponding frequencies

z/Bark	f_{low} [Hz]	f_{up} [Hz]	$\triangle f$[Hz]	z/Bark	f_{low} [Hz]	f_{up} [Hz]	△f [Hz]
0	0	100	100	13	2000	2320	320
1	100	200	100	14	2320	2700	380
2	200	300	100	15	2700	3150	450
3	300	400	100	16	3150	3700	550
4	400	510	110	17	3700	4400	700
5	510	630	120	18	4400	5300	900
6	630	770	140	19	5300	6400	1100
7	770	920	150	20	6400	7700	1300
8	920	1080	160	21	7700	9500	1800
9	1080	1270	190	22	9500	12000	2500
10	1270	1480	210	23	12000	15500	3500
11	1480	1720	240	24	15500		
12	1720	2000	280				

Critical bands are an essential model for description of the auditory sensation as they show the nonlinear behaviour of the HAS. Two analytical expressions are used to describe the relation of critical band rate and critical bandwidth over the HAS frequency range:

$$z = 13 \cdot \arctan\left(0.76\frac{f}{kHz}\right) + 3.5 \cdot \arctan\left(\frac{f}{7.5kHz}\right)^2 \ [Bark] \tag{1}$$

$$\triangle f_G = 25 + 75\left[1 + 1.4\left(\frac{f}{kHz}\right)^2\right]^{0.69} \ [Hz] \tag{2}$$

Two properties of the HAS dominantly used in watermarking algorithms are frequency (simultaneous) masking and temporal masking. The concept of using perceptual holes of the HAS is taken from wideband audio coding (MPEG 1 compression, layer 3, usually called mp3). In the compression algorithms, the holes are used decrease the amount of bits needed to encode audio signal, without causing perceptual distortion to the audio. On the other hand, in information hiding scenario, masking properties are used to embed additional bits into existing bit stream, again without generating audible noise in the host audio sequence.

5.1 Frequency Masking

Frequency masking is a frequency domain phenomenon where a low level signal (the maskee) can be made inaudible (masked) by a simultaneously appearing stronger signal (the masker), if masker and maskee are close enough to each other in frequency (Zwicker, 1999). A masking threshold can be derived below which any signal will not be audible. Without a masker, a signal is inaudible if its sound pressure level (SPL) is below the threshold in quiet, which depends on frequency and covers a dynamic range of more than 70 dB as depicted in the lower curve of Figure 1. The masking threshold depends on the masker and on the characteristics of masker and maskee (narrowband noise or pure tone).

For example, with the masking threshold for the SPL equal to 60 dB, masker in Figure 1 at around 1 kHz, the SPL of the maskee can be surprisingly high it will be masked as long as its SPL is below the masking threshold. The slope of the masking threshold is steeper toward lower frequencies; in other words, higher frequencies tend to be more easily masked than lower frequencies. It should be pointed out that the distance between maskee SPL and masking threshold is smaller in noise-masks-tone case than in tone-masks-noise case, due to sensitivity of the HAS toward additive noise. Noise and low-level signal components are masked inside and outside the particular critical band if their SPL is below the masking threshold (Zwicker, 1999). Noise can arise from coding, inserted watermark sequence, aliasing distortions, etc.

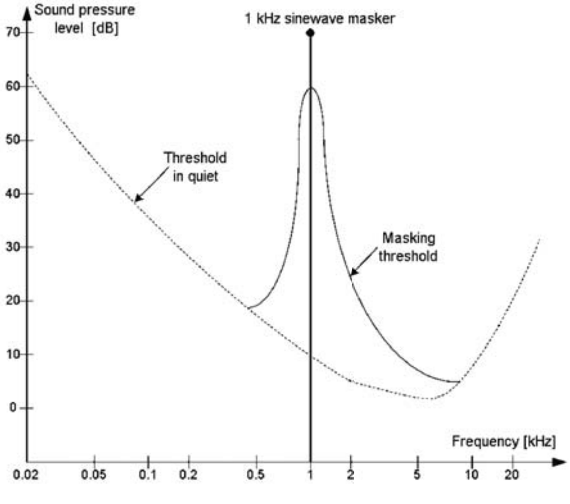


Fig. 1 Frequency masking in the human auditory system (HAS)

The qualitative sketch of Figure 2 gives more details about the masking threshold. The distance between the SPL of the masker (masking tone SPL in Figure 2) and the minimum masking threshold is called signal-to-mask ratio (SMR). Its maximum value is at the left end of the critical band. Let $SNR(m)$ be the signal-to-noise ratio resulting from watermark insertion in the subband m; the perceivable distortion in a given subband is then measured by the noise-to-mask ratio (NMR):

$$NMR(m) = SMR - SNR(m) \tag{3}$$

The noise-to-mask ratio $NMR(m)$ expresses the difference between the watermark noise in a given subband and the level where a distortion may just become audible. If a watermarking system needs to embed inaudible watermarks, $NMR(m)$ value in dB must be kept negative during watermark embedding. Thus, within a critical band, noise caused by watermark embedding (given as quantization noise in Figure 2) will be inaudible as long as bands SNR is higher than its SMR. It is clear that embedding of a watermark with higher amplitude will cause a decrease in the SNR value and increase the SPL of the noise above the minimum threshold level. This description is the case of masking by only one masker. If the source signal consists of many simultaneous maskers, a global masking threshold can be computed. It describes the threshold of just noticeable distortion (JND) as a function of frequency. The calculation of the global masking threshold is based on the high-resolution short-term amplitude spectrum of the audio signal, sufficient for critical band-based analysis. In a first step, all the individual masking thresholds are determined, depending on signal level, type of masker (tone or noise) and frequency range. After that, the global masking threshold is determined by adding all individual masking thresholds and the threshold in quiet. The effects of frequency masking reaching over critical band bounds must be included in the calculation as well. Finally, the global signal-to-noise ratio is determined as the ratio of the maximum of the signal power and the global masking threshold.

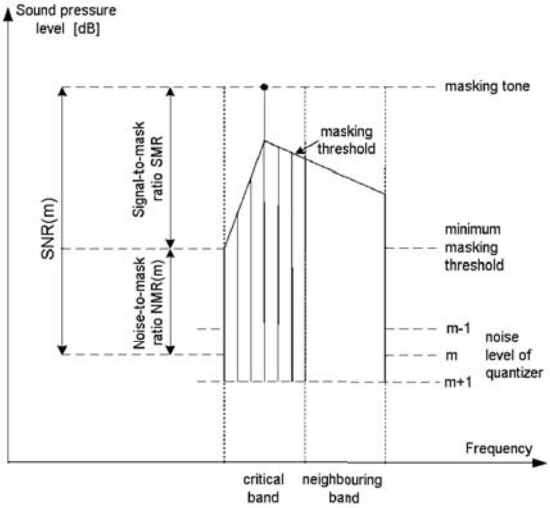


Fig. 2 Definition of signal-to-noise (SNR) ratio and signal-to-mask ratio (SMR)

5.2 Temporal Masking

In addition to frequency masking, two phenomena of the HAS in time domain have an important role in human auditory perception. Those are pre-masking and post-masking in time. The temporal masking effects appear before and after a masking signal has been switched on and off, respectively (Figure 3). Temporal masking is the characteristic of the HAS where a maskee is inaudible due to a masker, which has just disappeared, or even after masker which is about to appear. Therefore, if the SPL of the maskee is below the curve drawn in Figure 3 it will not be perceived by the HAS, because of the temporal masking. The temporal masking threshold increases as the appearance of the masker is approaching and, decreases as the appearance of the masker has passed. The duration of the pre-masking phenomenon is significantly less than one-tenth that of the post-masking, which is in the interval of 50 to 200 milliseconds. Both pre- and post-masking have been exploited in MPEG audio compression algorithm and in the most significant audio watermarking methods.

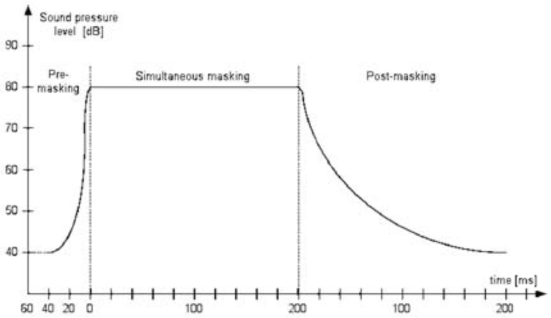


Fig. 3 Temporal masking phenomena

6 Audio Watermarking Algorithms

Watermarking algorithms were primarily developed for digital images and video sequences; interest and research in audio watermarking started slightly later. In the past few years, several concepts for embedding and extraction of watermarks in audio sequences have been presented. A large majority of the developed algorithms take advantage of perceptual properties of the human auditory system (HAS) in order to add watermark into a host signal in a perceptually transparent manner. A broad range of embedding techniques goes from simple least significant bit (LSB) scheme to the various spread spectrum methods.

Watermark embedder design consists of adjusting the watermark signal to satisfy the perceptual transparency and simultaneously maximize the power of the watermark signal to provide high robustness. It usually contains a psychoacoustic analysis block that provides the embedding algorithm with frequency masking threshold, maximum allowable phase difference, temporal masking threshold or similar parameters necessary for optimal watermark embedding. Selection of the particular psychoacoustic analysis block depends on the domain used for watermark embedding in a specific algorithm.

After the watermarked signal is generated it is subjected to common audio signal distortions, including dynamic compression, filtering, and perceptual coding. The effect of those distortions on the embedded watermark is usually considered to be in the form of stationary additive Gaussian noise, although many watermark attacks are more appropriately modeled as fading-like (Kundur, 2001). A well-defined model for the distortion introduced by certain attack is a necessary precondition for design of an optimal watermark detector.

The ultimate goal of any watermarking system is reliable watermark extraction. It is important to make term separation between watermark decoding and watermark detection during the watermark extraction. Communicating a watermark message is the essence of embedding and decoding of a digital watermark while verifying whether the received audio sequence is watermarked or not is watermark detection.

6.1 Least Significant Bit Coding

One of the earliest techniques studied in the information hiding and watermarking area of digital audio (as well as other media types (Fridrich, 2001; Lee, 2000; Fridrich, 2002) is LSB coding (Yeh, 1999). A simple approach in watermarking of the audio sequences is to embed watermark data by alternation of the certain bits of the digital audio stream, having the amplitude resolution of 16 bits per sample. It usually does not use any psychoacoustics model to perceptually weight the noise introduced by LSB replacement. However, there are some advanced methods of LSB coding (Lee, 2000; Cvejic, 2002) that introduce a certain level of perceptual shaping.

The watermark encoder usually selects a subset of all available host audio samples chosen by a secret key. The substitution operation on the LSBs is performed on this subset. Extraction process simply retrieves the watermark by reading the value of these bits. Therefore, the decoder needs all the samples of the watermarked audio that were used during the embedding process. The random selection of the samples used for embedding introduces low power additive white Gaussian noise. As noted in Section 5, HAS is very sensitive to the AWGN and that fact limits the number of LSBs that can be imperceptibly modified. The main advantage of the method is a very high watermark channel capacity; use of only one LSB of the host audio sample gives capacity of 44.1 kbps (all samples used). The obvious disadvantage is extremely low robustness, due to fact that random changes of the LSBs destroy the coded watermark (Mobasseri, 1998). In addition, it is very unlikely that embedded watermark would survive digital to analogue and subsequent analogue to digital conversion. As no calculation-demanding transformation of the host signal in the basic version of this method needs to be done, this algorithm has a very small computational complexity. This permits the use on this LSB in real-time applications. This algorithm is a good basis for steganographic applications for audio signals and a base for steganalysis of digital media (Chandramouli, 2001; Dumitrescu, 2003).

6.2 Watermarking of the Phase of Audio

Algorithms that embed watermark into the phase of the host audio do not use masking properties of the HAS, but the fact that the human auditory system has a low sensitivity to relative phase change (Bender, 1996). There are two main approaches used in watermarking of the host signals phase, phase coding (Bender, 1996; Ruiz, 2000) and phase modulation (Ciloglu, 2000; Tilki, 1997; Lancini, 2002).

6.2.1 Phase Coding

The basic phase coding method was presented in (Bender, 1996). The basic idea is to split the original audio stream into blocks and embed the whole watermark data sequence into the phase spectrum of the first block. One drawback of the phase coding method is considerably low payload as only the first block is used for watermark embedding. In addition, the watermark is not dispersed over the entire data set available, but is implicitly localized and can thus be removed easily by the cropping attack. It is a non-blind watermarking method, which limits the number of applications it is suitable for.

6.2.2 Phase Modulation

Watermark insertion in this method is performed using independent multiband phase modulation (Kuo, 2002; Gang, 2001]. The original signal is segmented into M blocks containing N samples using overlapping windows:

$$win(n) = sin\left(\frac{\pi(n+0.5)}{N}\right) 0 \leq n \leq N-1 \tag{4}$$

To ensure perceptual transparency by introducing only small changes in the envelope, the performed phase modulation has to satisfy the following constraint:

$$\left|\frac{\triangle\phi(z)}{\triangle z}\right| < 30^{o} \tag{5}$$

where $\phi(z)$ denotes the signal phase and z is the Bark scale. Each Bark constitutes one critical bandwidth; conversion of frequency between Bark and Hz is given in Table 3. Using a long block size N (e.g. $N = 2^{14}$) algorithm attains a slow phase change over time. The watermark is converted into phase modulation by having one integer Bark scale carry one message bit of the watermark. Each message bit is first represented by a phase window function, which centres at the end of the corresponding Bark band and spans two adjacent Barks. The phase window function is defined as follows:

$$\phi(z) = \sin^2\left(\frac{\pi(z+1)}{2}\right), -1 \leq z \leq 1 \tag{6}$$

Denote $a_1, a_2, , a_I$ the sequence of weights used for watermark embedding the kth block of host audio. The sign $a_k \in \{-1,1\}$ of the phase window function is determined by the kth watermark bit $m_k \in \{-1,1\}$. The total phase modulation is obtained as linear combination of the overlapped phase window functions:

$$\Phi_k(z) = \sum_{j=1}^{J} a_k[j]\phi(z-j), 0 \le z \le J \tag{7}$$

Using the $\Phi_k(z)$ the bits are embedded into the phases in the kth audio block by multiplying the Fourier coefficients with the phase modulation function.

$$\mathbf{A}_{wk}[f] = \mathbf{A}_{ok}[f] \times e^{j\Phi[f]} \tag{8}$$

with the frequency in Hz. The watermarked signal is computed by inverse Fourier transformation of the modified Fourier coefficients $\mathbf{A}_{wk}$. All the blocks are windowed and overlap-added to create watermarked signal. The robustness of the modulated phase can be increased by using n_z Bark values carrying one watermark bit.

6.3 Echo Hiding

A number of developed audio watermarking algorithms (Huang, 2002; Ko, 2002; Foo, 2001) are based on echo hiding method, described for the first time in (Bender, 1996). Echo hiding schemes embed watermarks into a host signal by adding echoes to produce watermarked signal. Echo hiding audio watermarking algorithm is a blind watermarking algorithm, designed especially for audio signals (it is not used in image or video watermarking). It is highly robust against standard watermarking attacks and the watermark bit rate of several tens of bps.

The nature of the echo is to add resonance to the host audio, therefore the acute problem of sensitivity of the HAS towards the additive noise is circumvented in this method. After the echo has been added, watermarked signal retains the same statistical and perceptual characteristics. The offset (or delay) between the original and watermarked signal is small enough that the echo is perceived by the HAS as an added resonance. The four major parameters, initial amplitude, decay rate, one offset and zero offset are given in Figure 4.

Watermark embedding process can be represented as a system that has one of two possible system functions. In the time domain, the system functions are discrete time exponential differing only in the delay between impulses. Processing host signal through any kernel in Figure 4 will result in an encoded signal. The delay between the original signal and the echo is dependent on the kernel being used, δ_1 if the one kernel is used and δ_0 if the zero kernel is used.

The host signal is divided into smaller portions for encoding more than one bit. Each individual portion can then be considered each as an independent signal and echoed with the desired bit. The final watermarked signal (containing several bits) is composite of all independently encoded signal portions. A smooth transition

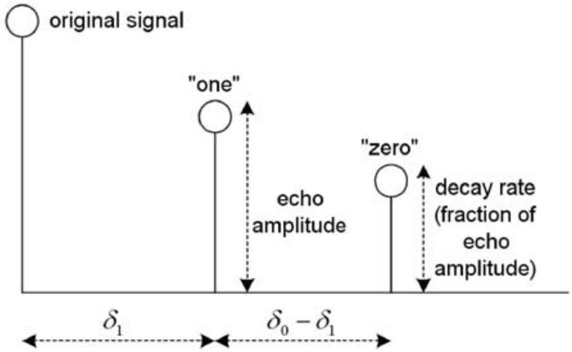


Fig. 4 Parameters of echo embedding watermarking method

between portions encoded with different bits should is adjusted using different methods to prevent abrupt changes in the resonance in the watermarked signal. Information is embedded into a signal by echoing the original signal with one of two delay kernels. Therefore, extraction of the embedded information is to detect the spacing between the echoes. The magnitude of the autocorrelation of the encoded signals cepstrum:

$$F^{-1}\{\log\left(|F(x)|^2\right)\} \tag{9}$$

where F represents the Fourier Transform and F^{-1} the inverse Fourier Transform can be examined at two locations, corresponding to the delays of the one and zero kernel, respectively. If the autocepstrum is greater at δ_1 than it is at δ_0, embedded bit is decoded as one. For multiple echo hiding, all peaks present in the autocepstrum are detected. The number of peaks corresponding to the delay locations of the one and zero kernels are then counted and compared. If there are more peaks at the delay locations for the one echo kernel, the watermark bit is decoded as one. Increased robustness of watermark algorithm requires high-energy echoes to be embedded which increases audible distortion. There are several modifications to the basic echo-hiding algorithm. (Xu, 1999) proposed a multi-echo embedding technique to reduce the possibility of echo detection by third parties. The technique has clear constraints regarding the increase of the robustness, as the audio timbre is noticeably changed with the sum of pulse amplitude (Oh, 2001). (Oh, 2001) proposed echo kernel comprising multiple echoes by both positive and negative pulses with different offsets (closely located) in the kernel, of which the frequency response is plain in lower bands and large ripples in high frequency.

6.4 Spread Spectrum

In a number of the developed algorithms (Bassia, 2001; Neubauer, 1998; Cox, 1997; Kirovski, 2003; Swanson, 1998), watermark embedding and extraction are carried out using spread-spectrum (SS) technique. SS sequence can be added to the host audio samples in time domain (Bassia, 2001; Cvejic, 2001), to FFT coefficients (Swanson, 1998; Ikeda, 1999; Seok, 2001), in subband domain (Kirovski, 2001; Li, 2000; Tachibana, 2002), to cepstral coefficients (Lee, 2000; Li, 2000) and in compressed domain (Neubauer, 2000; Cheng, 2002). If embedding takes place in a transform domain, it should be located in the coefficients invariant to common watermark attacks as amplitude compression, resampling, low pass filtering, and other common signal processing techniques. The idea is that after the transform, any significant change in the signal would significantly decrease the subjective quality of the watermarked audio. Thus, spread spectrum watermarking is a extremely robust, blind watermarking algorithm, with the watermark bit rate from a few bps to a several tens of bps.

Watermark is spread over a large number of coefficients and distortion is kept below the just noticeable difference level by using occurrence of masking effects of the human auditory system. Change in each coefficient can be small enough to be imperceptible, because correlator detector output still has a high signal to noise ratio, as it despreads the energy present in a large number of coefficients.

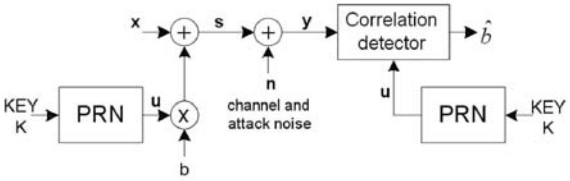


Fig. 5 General model for SS-based watermarking

A general system for SS-based watermarking is shown in Figure 5. Vector **x** is considered to be the original host signal already in an appropriate transform domain. The vector **y** is the received vector, in the transform domain, after channel distortions. A secret key **K** is used by a pseudo random number generator (Furon, 2003; Tefas, 2003) to produce a spreading sequence **u** with zero mean and whose elements are equal to $+\sigma_u$ or $-\sigma_u$. The sequence **u** is then added to or subtracted from the signal **x** according to the variable *b*, where *b* assumes the values of +1 or -1 according to the bit (or bits) to be transmitted by the watermarking process (in multiplicative algorithms multiplication operation is performed instead addition (Barni, 2003), The signal **s** is the watermarked audio signal. A simple analysis of SS-based watermarking leads to a simple formula for the probability of error. Thus, if we consider the definitions of inner product and norm:

$$\langle \mathbf{x}, \mathbf{u} \rangle = \sum_{i=0}^{N-1} x_i u_i \qquad \mathit{and} \qquad \|x\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} \tag{10}$$

where *N* is the length of the vectors **x, s, u, n**, and **y** in Figure 5.

Without loss of generality, we can assume that we are embed one bit of information in a vector **s** of N transform coefficients. That bit is represented by the variable b, whose value is either +1 or -1. Embedding is performed by

$$\mathbf{s} = \mathbf{x} + b\mathbf{u} \tag{11}$$

The distortion in the embedded signal is defined by $\|$**s-x**. It is easy to see that for the embedding equation (11), we have

$$D = \|b\mathbf{u}\| = \|\mathbf{u}\| = \sigma_u \tag{12}$$

The watermark communication channel is modelled as additive noise $\mathbf{y} = \mathbf{s} + \mathbf{n}$, and watermark extraction is usually performed by calculation of the normalized sufficient statistic (Box, 1978) r:

$$r = \frac{\langle \mathbf{y}, \mathbf{u} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle} = \frac{\langle b\mathbf{u} + \mathbf{x} + \mathbf{n}, \mathbf{u} \rangle}{\sigma_u^2} = b + x + n \tag{13}$$

and estimating the embedded bit as $\hat{b} = sign(r)$, where $x = \langle \mathbf{x}, \mathbf{u} \rangle / \|u\|$ and $n = \langle \mathbf{n}, \mathbf{u} \rangle / \|u\|$. Simple statistical models for the host audio **x** and the attack noise **n** are assumed. Both vectors are modelled as uncorrelated white Gaussian random processes (Box, 1978):

$$x_i \approx N(0, \sigma_x^2) \qquad and \qquad n_i \approx N(0, \sigma_n^2) \tag{14}$$

Then, it is easy to show (Box, 1978) that the sufficient statistic r is also Gaussian variable, i.e.:

$$r \approx N(m_r, \sigma_r^2), \qquad m_r = E[r] = b\sigma_r^2 = \frac{\sigma_x^2 + \sigma_n^2}{N\sigma_u^2} \tag{15}$$

Specifically, let us elaborate the case when b is equal to 1. In that case, an error occurs when $r < 0$, and therefore, the error probability p is given by

$$p = Pr\{\hat{b} < 0 | b = 1\} = \frac{1}{2} erfc\left(\frac{m_r}{\sigma_r \sqrt{2}}\right) = \frac{1}{2} erfc\left(\sqrt{\frac{\sigma_u^2 N}{2\left(\sigma_x^2 + \sigma_n^2\right)}}\right) \tag{16}$$

where $erfc()$ is complementary error function. The equal error probability is obtained under the assumption that $b = -1$. A plot of that probability as a function of the SNR m_r/σ_r is given in Figure 6. For example, from Figure 6, it is clear that if an error probability lower than 10^{-3} is needed, than we get:

$$\frac{m_r}{\sigma_r} > 3 \equiv N\sigma_u^2 > 9\left(\sigma_x^2 + \sigma_n^2\right) \tag{17}$$

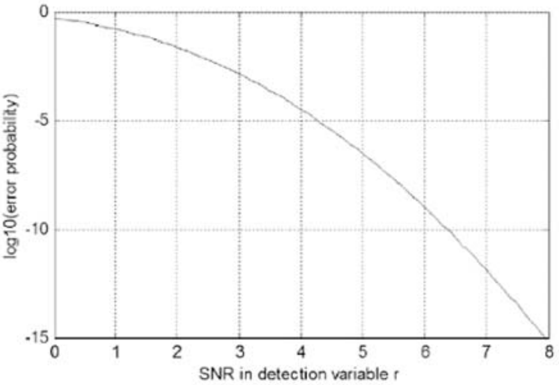


Fig. 6 Error probability as a function of the SNR

or more generally, to achieve an error probability p we need:

$$N\sigma_u^2 > 2\left(erfc^{-1}(p)\right)^2\left(\sigma_x^2+\sigma_n^2\right) \tag{18}$$

The equation above shows that we can make a trade-off between the length of the spreading sequence N and the energy of the spreading sequence σ_u^2. It lets us to simply compute either N or σ_u^2, given the other variables involved.

6.5 Patchwork Method

The patchwork technique was first presented in (Bender, 1996), for embedding watermarks in images. It is a statistical method based on hypothesis testing and relying on large data sets. As a second of CD quality stereo audio contains 88200 samples, patchwork approach is applicable for watermarking of audio sequences as well. The watermark embedding process uses a pseudorandom process to insert a certain statistic into host audio data set, which is extracted with the help of numerical indices (like the mean value) describing the specific distribution. The method is usually applied in a transform domain (Fourier, DCT, wavelet) in order to spread the watermark in time domain and to increase robustness against signal processing modifications (Sugihara, 2001; Arnold, 2000; Yeo, 2003). Patchwork algorithm does not require the original host signal in the process of watermark detection (blind watermarking detection). Watermark bit rate is 1-10 bps, if a high robustness in the presence of attacks is required. Watermark embedding steps are summarized as follows:

1. Map the secret key and the watermark to the seed of a random number generator. After that, generate an index set whose elements are pseudo-randomly selected integer values from $[K_1, K_2]$, where $1 \leq K_1 \leq K_2 \leq N$. Note that two index sets, I_0 and I_1, are needed to denote watermark bits 0 and 1, respectively. The choice of K_1

and K_2 is a crucial step in embedding the watermark because these values control the trade-off between the robustness and the inaudibility of the watermark.

2. Let $F = F_1,,F_N$ be the coefficients whose subscript denote frequency range from the lowest to the highest frequencies. Define $A = a_1,,a_n$ as the subset of F whose subscript corresponds to the first n elements of the index set I_0 or I_1 according to the embedded code with similar definition for $B = b_1,,b_n$ with the last n elements, that is $a_i = F_I$ and $b_i = F_{I_n+I}$, for $i = 1,,n$. 3. Calculate the sample means $\bar{a}$ and $\bar{b}$, respectively and the pooled sample standard error:

$$S = \sqrt{\frac{\sum_{i=1}^{n}(a_i - \bar{a})^2 + \sum_{i=1}^{n}(b_i - \bar{b})^2}{n(n-1)}} \tag{19}$$

4. The embedding function presented below introduces a location-shift change

$$a_i^* = a_i + sign(\bar{a} - \bar{b})\sqrt{C}\frac{S}{2} \qquad and \qquad b_i^* = b_i - sign(\bar{a} - \bar{b})\sqrt{C}\frac{S}{2} \tag{20}$$

where C is a constant and *sign* is the signum function. This function makes the large value set larger and the small value set smaller so that the distance between two sample means is always larger than $d = \sqrt{C}S$.

5. Finally, replace the selected elements a_i and b_i by a_i^* and b_i^*, respectively, and then apply the inverse transformation.

Since the proposed embedding method introduces relative changes of two sets in location, a natural test statistic which is used to decide whether or not the watermark is embedded should concern the distance between the means of A and B.

6.6 Methods Using Various Characteristics of the Host Audio

Several audio watermarking algorithms developed in the recent years use different statistical properties of the host audio and modify them in order to embed watermark data. Those properties are pitch values, number of salient points, difference in energy of two adjacent blocks etc. However, modifications of the host signal statistical properties do influence the subjective quality of the audio signal and have to be performed in a way that does not produce distortions above the audible threshold. Usually, these methods are robust to signal processing modifications, but offer low watermark capacity. Paper (Xu, 1999) introduced content-adaptive segmentation of the host audio according to its characteristics in time domain. Since the embedding parameters are dependent of the host audio, it is along the right direction to increase tamper resistance. The basic idea is to classify the host audio into a predetermined number of segments according to its properties in time domain, and encode each segment with an embedding scheme, which is designed to best suit this segment of audio signal, according to its features in frequency domain. In paper (Lemma, 2003), the temporal envelope of the audio signal is modified according to the watermark. A number of signal processing operations are needed for embedding a multibit payload watermark. First, the filter extracts the part of the audio signal that is suitable to carry the watermark information. The watermarked audio signal is then obtained by adding an appropriately scaled version of the product of watermark and filtered host audio to the host signal. Watermark detector consists of two stages: the symbol extraction stage and the correlation and decision stage.

Algorithm presented in (Kaabneh, 2001) embeds the watermark by deciding for each mute period in the host audio whether to extend it by a predefined value. In order to detect the watermark, the detector must have access to the original length of all mute periods in the host audio. Method described in (Hiujuan, 2002) uses pitch scaling of the host audio, realized using short time Fourier transform, to embed the watermark. The correlation ratio, computed during embedding procedure is quantized with different quantization steps in order to embed bit 0 and 1 of the watermark stream. In papers (Hsieh, 2002; Mansour, 2001) salient points are used as basis for watermark embedding resistant to desynchronization attacks. Salient point is defined as the energy fast climbing part of the host audio signal; it defines the synchronization point for the watermarking process without embedding additional synchronization tags. Embedding of the watermark bits in (Hsieh, 2002) is performed using statistical mean manipulation of the cepstral coefficients and in (Mansour, 2001) by changing the distance between two salient points. Algorithms presented in (Hiujuan, 2002; Xu, 2002) use feature extraction of the host audio signal in order to tailor specific embedding algorithm for the given segment of the host audio. In (Hiujuan, 2002) authors use neural networks for feature extraction and classification, while in (Xu, 2002) feature extraction is done using a nonlinear frequency scale technique. The algorithm proposed in (Lie, 2001) embeds watermarks using relative energy relations between consecutive sample sections of the host audio in time domain. Discontinuities between boundaries of adjacent sections that would cause significant audible distortions are blurred using progressive weighting near section boundaries.

Section 6 gives an overview of the state of the art audio watermarking algorithms. Generally, developed algorithms use the masking properties of the HAS in order to inaudibly embed watermark into a host signal. The broad range of algorithms goes from the simple LSB scheme to the various spread spectrum methods.

7 Summary

This chapter gives an overview of digital audio watermarking systems, including description of developed watermarking algorithms and a number of examples of application of the audio watermarking methods. Audio watermarking algorithms are characterised by a number of defining properties, ranging from robustness requirements to computational complexity and cost of implementation. The relative importance of a particular property is application-dependent and in many cases even the interpretation of a watermark property varies with the application. Psychoacoustic models of the HAS that are exploited in order to preserve the subjective quality of the watermarked audio during the watermarking process are shortly reviewed. In the past few years, several concepts for embedding and extraction of watermarks in audio sequences have been presented. A large majority of the developed algorithms uses the properties of the HAS described in the chapter in order to inaudibly embed watermark into a host signal. The broad area of embedding techniques that ranges from simple LSB scheme to the various spread spectrum methods is presented as well.

8 Additional Reading

Readers interested in the state-of-the-art audio watermarking algorithms are encouraged to read the following articles that cover the latest developments in the area:

Wu S., Huang J., Huang D., Shi Y.Q. (2005) Efficiently self-synchronized audio watermarking for assured audio data transmission. IEEE Transactions on Broadcasting, 51(1), 69-76.

Ko B.-S., Nishimura R., Suzuki Y (2005) Time-spread echo method for digital audio watermarking. IEEE Transactions on Multimedia, 7(2), 212-221.

Lee H.S., Lee W.S. (2005) Audio watermarking through modification of tonal maskers, ETRI Journal, 27(5), 608-615.

Zaidi A., Boyer R., Duhamel P. (2006) Audio watermarking under desynchronization and additive noise attacks. IEEE Transactions on Signal Processing, 54(2), 570-584.

Lie W.-N., Chang L.-C. (2006) Robust and high-quality time-domain audio watermarking based on low-frequency amplitude modification, IEEE Transactions on Multimedia, 8(1), 46-59.

Li W., Xue X., Lu P. (2006) Localized audio watermarking technique robust against time-scale modification. IEEE Transactions on Multimedia, 8(1), 60-69.

Wang X.-Y., Zhao H. (2006) A novel synchronization invariant audio watermarking scheme based on DWT and DCT, IEEE Transactions on Signal Processing, 54(12), 4835-4840.

Liu Y.-W., Smith J.O. (2007) Audio watermarking through deterministic plus stochastic signal decomposition. EURASIP Journal on Information Security, no. 75961.

Malik H., Ansari R., Khokhar A. (2008) Robust audio watermarking using frequency-selective spread spectrum. IET Information Security, 2(4), 129-150.

Wang X.-Y., Niu P.-P., Qi W. (2008) A new adaptive digital audio watermarking based on support vector machine, Journal of Network and Computer Applications, 31(4), 735-749.

Wang, H., Nishimura, R., Suzuki, Y., Mao, L. (2008) Fuzzy self-adaptive digital audio watermarking based on time-spread echo hiding. Applied Acoustics, 69(10), 868-874.

Xiang S., Kim H.J., Huang J. (2008) Audio watermarking robust against time-scale modification and MP3 compression, Signal Processing, 88(10), 2372-2387.

Erelebi E., Bataki L. (2009) Audio watermarking scheme based on embedding strategy in low frequency components with a binary image. Digital Signal Processing, 19(2), 265-277.

Deshpande, A., Prabhu, K.M.M. (2009) A substitution-by-interpolation algorithm for watermarking audio, Signal Processing, 89(2), 218-225.

References

[1] Arnold, M.: Audio watermarking: features, applications and algorithms. In: Proc. IEEE International Conference on Multimedia and Expo., New York, pp. 1013–1016 (2000)

[2] Arnold, M., Huang, Z.: Blind detection of multiple audio watermarks. In: Proc. International Conference on Web Delivering of Music, Florence, Italy, pp. 12–19 (2001)

[3] Arnold, M., Wolthusen, S., Schmucker, M.: Techniques and Applications of Digital Watermarking and Content Protection. Artech House (2003)

[4] Barni, M., Bartolini, F., De Rosa, A., Piva, A.: Optimum Decoding and Detection of Multiplicative Watermarks. IEEE Transactions on Signal Processing 51(4), 1118–1123 (2003)

[5] Bassia, P., Pitas, I., Nikolaidis, N.: Robust Audio Watermarking in the Time Domain. IEEE Transactions on Multimedia 3(2), 232–241 (2001)

[6] Beerends, J., Stemerdink, J.: A Perceptual Audio Quality Measurement Based on a Psychoacoustic Sound Representation. Journal of Audio Engineering Society 40(12), 963–972 (1992)

[7] Bender, W., Gruhl, D., Morimoto, N., Lu, A.: Techniques for data hiding. IBM Systems Journal 35(3), 313–336 (1996)

[8] Box, G.E.P.: Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building. John Wiley Sons, Chichester (1978)

[9] Brandenburg, K., Sporer, T.: NMR and Masking Flag: Evaluation of Quality using Perceptual Criteria. In: Proc. International Audio Engineering Society Conference on Audio Test and Measurement, Portland, OR, pp. 169–179 (1992)

[10] Chandramouli, R., Memon, N.: Analysis of LSB based image steganography techniques. In: Proc. IEEE International Conference on Image Processing, Thessalonica, Greece, pp. 1019–1022 (2001)

[11] Chen, B., Wornell, B.: Dither modulation: a new approach to digital watermarking and information embedding. In: Proc. of SPIE: Security and Watermarking of Multimedia Contents, San Hose, CA, pp. 342–353 (1999)

[12] Cheng, S., Yu, H., Xiong, Z.: Enhanced spread spectrum watermarking of MPEG-2 AAC. In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, Orlando, FL, pp. 3728–3731 (2002)

[13] Chou, J., Ramchandran, K., Ortega, A.: High capacity audio data hiding for noisy channels. In: Proc. International Conference on Information Technology: Coding and Computing, Las Vegas, NV, pp. 108–111 (2001)

[14] Ciloglu, T., Karaaslan, S.: An improved all-pass watermarking scheme for speech and audio. In: Proc. IEEE International Conference on Multimedia and Expo., New York, pp. 1017–1020 (2000)

[15] Cox, I., Kilian, J., Leighton, F., Shamoon, T.: Secure spread spectrum watermarking for multimedia. IEEE Transactions on Image Processing 6(12), 1673–1687 (1997)

[16] Cox, I., Miller, M., McKellips, A.: Watermarking as communications with side information. Proceedings of the IEEE 87(7), 1127–1141 (1999)

[17] Cox, I., Miller, M.: Electronic watermarking: the first 50 years. In: Proc. IEEE Workshop on Multimedia Signal Processing, Cannes, France, pp. 225–230 (2000)

[18] Cox, I., Miller, M., Bloom, J.: Digital Watermarking. Morgan Kaufmann Publishers, San Francisco (2001)

[19] Cvejic, N., Keskinarkaus, A., Seppänen, T.: Audio watermarking using m-sequences and temporal masking. In: Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New York, pp. 227–230 (2001)

[20] Cvejic, N., Seppänen, T.: Increasing the capacity of LSB-based audio steganography. In: Proc. IEEE International Workshop on Multimedia Signal Processing, St. Thomas, VI, pp. 336–338 (2002)

[21] Dumitrescu, S., Wu, W., Wang, Z.: Detection of LSB steganography via sample pair analysis. IEEE Transactions on Signal Processing 51(7), 1995–2007 (2003)

[22] Fridrich, J., Goljan, M., Du, R.: Distortion-Free Data Embedding for Images. In: Moskowitz, I.S. (ed.) IH 2001. LNCS, vol. 2137, pp. 27–41. Springer, Heidelberg (2001)

[23] Fridrich, J., Goljan, M., Du, R.: Lossless Data Embedding - New Paradigm in Digital Watermarking. Applied Signal Processing 2002(2), 185–196 (2002)

[24] Foo, S.W., Yeo, T.H., Huang, D.Y.: An adaptive audio watermarking system. In: Proc. IEEE Region 10 International Conference on Electrical and Electronic Technology, Phuket Island-Langkawi Island, Singapore, pp. 509–513 (2001)

[25] Furon, T., Duhamel, P.: An Asymmetric Watermarking Method. IEEE Transactions on Signal Processing 51(4), 981–995 (2003)

[26] Gang, L., Akansu, A., Ramkumar, M.: MP3 resistant oblivious steganography. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Salt Lake City, UT, pp. 1365–1368 (2001)

[27] Hartung, F., Kutter, M.: Multimedia Watermarking Techniques. Proceedings of the IEEE 87(7), 1079–1107 (1999)

[28] Hiujuan, Y., Patra, J.C., Chan, C.W.: An artificial neural network-based scheme for robust watermarking of audio signals. In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, Orlando, FL, pp. 1029–1032 (2002)

[29] Hsieh, C.T., Sou, P.Y.: Blind cepstrum domain audio watermarking based on time energy features. In: Proc. International Conference on Digital Signal Processing, Santorini, Greece, pp. 705–708 (2002)

[30] Huang, D.Y., Yeo, Y.H.: Robust and Inaudible Multi-echo Audio Watermarking. In: Proc. IEEE Pacific-Rim Conference on Multimedia, Taiwan, China, pp. 615–622 (2002)

[31] Ikeda, M., Takeda, K., Itakura, F.: Audio data hiding use of band-limited random sequences. In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, Phoenix, AZ, pp. 2315–2318 (1999)

[32] ITU-R Draft new Recommendation ITU-R BS. Method for objective measurements of perceived audio quality (1998)

[33] Johnston, J.: Estimation of perceptual entropy using noise masking criteria. In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, New York, pp. 2524–2527 (1988)

[34] Kaabneh, K.A., Youssef, A.: Muteness-based audio watermarking technique. In: Proc. International Conference on Distributed Computing Systems, Phoenix, AZ, pp. 379–383 (2001)

[35] Kirovski, D., Malvar, H.: Robust Covert Communication over a Public Audio Channel Using Spread Spectrum. In: Moskowitz, I.S. (ed.) IH 2001. LNCS, vol. 2137, p. 354. Springer, Heidelberg (2001)

[36] Kirovski, D., Malvar, H.: Spread-Spectrum Watermarking of Audio Signals. IEEE Transactions on Signal Processing 51(4), 1020–1033 (2003)

[37] Ko, B.S., Nishimura, R., Suzuki, Y.: Time-spread echo method for digital audio watermarking using PN sequences. In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, Orlando, FL, pp. 2001–2004 (2002)

[38] Kundur, D., Hatzinakos, D.: Diversity and Attack Characterization for Improved Robust Watermarking. IEEE Transactions on Signal Processing 29(10), 2383–2396 (2001)

[39] Kuo, S.S., Johnston, J., Turin, W., Quackenbush, S.: Covert audio watermarking using perceptually tuned signal independent multiband phase modulation. In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, Orlando, FL, pp. 1753–1756 (2002)

[40] Lee, Y.K., Chen, L.H.: High capacity image steganographic model. IEE Proceedings on Vision, Image and Signal Processing 147(3), 288–294 (2000)

[41] Lancini, R., Mapelli, F., Tubaro, S.: Embedding indexing information in audio signal using watermarking technique. In: Proc. EURASIP-IEEE Region 8 International Symposium on Video/Image Processing and Multimedia Communications, Zadar, Croatia, pp. 257–261 (2002)

[42] Lee, S.K., Ho, Y.S.: Digital audio watermarking in the cepstrum domain. IEEE Transactions on Consumer Electronics 46(3), 744–750 (2000)

[43] Lemma, A.N., Aprea, J., Oomen, W., Van de Kerkhof, L.: A Temporal Domain Audio Watermarking Technique. IEEE Transactions on Signal Processing 51(4), 1088–1097 (2003)

[44] Li, X., Yu, H.: Transparent and robust audio data hiding in subband domain. In: Proc. International Conference on Information Technology: Coding and Computing, Las Vegas, NV, pp. 74–79 (2000)

[45] Li, X., Yu, H.: Transparent and robust audio data hiding in cepstrum domain. In: Proc. IEEE International Conference on Multimedia and Expo., New York, pp. 397–400 (2000)

[46] Lie, W.N., Chang, L.C.: Robust and high-quality time-domain audio watermarking subject to psychoacoustic masking. In: Proc. IEEE International Symposium on Circuits and Systems, Sydney, Australia, pp. 45–48 (2001)

[47] Malvar, H., Florencio, D.: Improved Spread Spectrum: A New Modulation Technique for Robust Watermarking. IEEE Transactions on Signal Processing 51(4), 898–905 (2003)

[48] Mansour, M., Tewfik, A.: Audio watermarking by time-scale modification. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Salt Lake City, UT, pp. 1353–1356 (2001)

[49] Mobasseri, B.: Direct sequence watermarking of digital video using m-frames. In: Proc. International Conference on Image Processing, Chicago, IL, pp. 399–403 (1998)

[50] Neubauer, C., Herre, J., Brandenburg, K.: Continuous Steganographic Data Transmission Using Uncompressed Audio. In: Aucsmith, D. (ed.) IH 1998. LNCS, vol. 1525, pp. 208–217. Springer, Heidelberg (1998)

[51] Neubauer, C., Herre, J.: Audio Watermarking of MPEG-2 AAC Bit streams. In: Proc. Audio Engineering Society Convention, Paris, France (2000)

[52] Oh, H.O., Seok, J.W., Hong, J.W., Youn, D.H.: New Echo Embedding Technique for Robust and Imperceptible Audio Watermarking. In: Proc. IEEE International Conference on Acoustic, Speech and Signal Processing, Salt Lake City, UT, pp. 1341–1344 (2001)

[53] Ruiz, F., Deller, J.: Digital watermarking of speech signals for the national gallery of the spoken word. In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, Istanbul, Turkey, pp. 1499–1502 (2000)

[54] Seok, J.W., Hong, J.W.: Audio watermarking for copyright protection of digital audio data. Electronics Letters 37(1), 60–61 (2001)

[55] Sporer, T.: Evaluating Small Impairments with mean Opinion Scale-Reliable or just a guess. In: Proc. Audio Engineering Society Convention, Los Angeles, CA (1996)

[56] Stein, J.: Digital Signal Processing: A Computer Science Perspective. Wiley-Interscience, Hoboken (2000)

[57] Steinebach, M., Petitcolas, F., Raynal, F., Dittmann, J., Fontaine, C., Seibel, S., Fates, N., Ferri, L.: StirMark benchmark: audio watermarking attacks. In: Proc. International Conference on Information Technology: Coding and Computing, Las Vegas, NV, pp. 49–54 (2001)

[58] Sugihara, R.: Practical capacity of digital watermark as constrained by reliability. In: Proc. International Conference on Information Technology: Coding and Computing, Las Vegas, NV, pp. 85–89 (2001)

[59] Swanson, M., Zhu, B., Tewfik, A., Boney, L.: Robust audio watermarking using perceptual masking. Signal Processing 66(3), 337–355 (1998)

[60] Swanson, M., Zhu, B., Tewfik, A.: Current state-of-the-art, challenges and future directions for audio watermarking. In: Proc. IEEE International Conference on Multimedia Computing and Systems, Florence, Italy, pp. 19–24 (1999)

[61] Tachibana, R., Shimizu, S., Kobayashi, S., Nakamura, T.: An audio watermarking method using a two-dimensional pseudo-random array. Signal Processing 82(10), 1455–1469 (2002)

[62] Tefas, A., Nikolaidis, A., Nikolaidis, N., Solachidis, V., Tsekeridou, S., Pitas, I.: Performance Analysis of Correlation-Based Watermarking Schemes Employing Markov Chaotic Sequences. IEEE Transactions on Signal Processing 51(7), 1979–1994 (2003)

[63] Tilki, J., Beex, A.: Encoding a hidden auxiliary channel onto a digital audio signal using psychoacoustic masking. In: Proc. IEEE South East Conference, Blacksburg, VA, pp. 331–333 (1997)

[64] Xu, C., Wu, J., Sun, Q., Xin, K.: Applications of Watermarking Technology in Audio Signals. Journal Audio Engineering Society 47(10) (1999)

[65] Xu, C., Wu, J., Sun, Q.: Robust Digital Audio Watermarking Technique. In: Proc. International Symposium on Signal Processing and its Applications, Brisbane, Australia, pp. 95–98 (1999)

[66] Xu, C., Feng, D.: Robust and efficient content-based digital audio watermarking. Multimedia Systems 8(5), 353–368 (2002)

[67] Yeh, C.H., Kuo, C.J.: Digital Watermarking through Quasi m-Arrays. In: Proc. IEEE Workshop on Signal Processing Systems, Taipei, Taiwan, pp. 456–461 (1999)

[68] Yeo, I.K., Kim, H.J.: Modified Patchwork Algorithm: A Novel Audio Watermarking Scheme. IEEE Transactions on Speech and Audio Processing 11(4), 381–386 (2003)

[69] Yu, H., Kundur, D., Lin, C.Y.: Spies, Thieves, and Lies: The Battle for Multimedia in the Digital Era. IEEE Multimedia 8(3), 8–12 (2001)

[70] Zwicker, E., Fastl, H.: Psychoacoustics: Facts and models. Springer, Heidelberg (1999)

A Survey of Music Structure Analysis Techniques for Music Applications

Namunu C. Maddage, Haizhou Li, and Mohan S. Kankanhalli

Abstract. Music carries multilayer information which forms different structures. The information embedded in the music can be categorized into time information, harmony/melody, music regions, music similarities, song structures and music semantics. In this chapter, we first survey existing techniques for the music structure information extraction and analysis. We then discuss how the music structure information extraction helps develop music applications. Experimental studies indicate that the success of long term music research is based on how well we integrate domain knowledge of relevant disciplines such as musicology, psychology and signal processing.

1 Introduction

Music which carries multilayer information is a universal language for sharing thoughts and sensations across different communities and cultures. Understanding not only the formation of music structures but also how they stimulate our minds has been a highly motivated research focus over centuries. Past publications in computer music research [57][56][51][52] highlight the importance of understanding the ingredients on which the music structure is formed, for developing music applications such as music streaming, music protection and right management (watermarking), music therapy, multimedia documentation, and music representation techniques (music summarization, compression, genre and language classification and artist identification).

Namunu C. Maddage
Electrical and Computer Engineering,
Royal Melbourne Institute of Technology (RMIT) University, Melbourne, 3001
e-mail: namunu.maddage@rmit.edu.au

Haizhou Li
Institute for Infocomm Research (I^2R), 1 Fusionopolis Way, #08-05 South Tower,
Connexis, Singapore 138632
e-mail: hli@i2r.a-star.edu.sg

Mohan S. Kankanhalli
School of Computing, National University of Singapore, Singapore 117543
e-mail: mohan@comp.nus.edu.sg

M. Grgic et al. (Eds.): Rec. Advan. in Mult. Sig. Process. and Commun., SCI 231, pp. 551–577.
springerlink.com

Early music research mainly focused on symbolic music like MIDI (Music Instruments for Digital Interface), which requires a small storage space and has the access to music score information (beat structure, melody, harmony, music sources- tracks, tempo, etc). Recently an effective XML based music content descriptor (text based) was proposed to represent multilayer structural information in the music (Pinto, A. and Haus, G. 2007 [87], Baratè, A. and Ludovico, L. A. 2004 [7]). Early music information retrieval (MIR) systems were mainly implemented using those symbolic databases (Zhu 2004. [134], Typke et al 2004. [116]). Dannenberg [23] in 1984 proposed solo instrumental line (audio) and text score alignment algorithm in which the notes in the solo instrumental line are detected using pitch detection algorithm and match/align then with the score using dynamic programming algorithm (Sakeo and Chiba 1978 [94]). Takeda et. al (2004) [111] proposed a probabilistic approach to detect rhythm and tempo from MIDI score sheets. Since score information is available, the key challenges in music structure analysis and retrieval are to correctly match or find similar melodic or other structural patterns. However due to the recent advancements in high bandwidth data transmission, large data storage and high performance computing, present text based computing algorithms are able to adequately overcome such challenges. Thus the research focus is shifted onto raw data, such as real[1] sound recordings/ real[1] music in which the score information is not available.

In this chapter we first present a survey of computing techniques that have been developed for extracting music structure information in real sound recordings. Then we also discuss some of these music applications where there is a need to extract information in the music structures. From the composition point of view, music is a combination of multilayered information. Jourdain (1997) [54] discussed the formation of different levels of music structures and how they lead our imaginations. He described how the combinations of physical[2] music units such as sound and tone, combine together to formulate the physical[2] music structures such as melody harmony and composition. Then combinations of those physical[2] music structures construct higher order structures such as music performance, listening, understanding and ecstasy. These higher order music structures also known as music semantics are difficult to quantify. Similarly we also conceptually visualize music as different information layers in a pyramidal model as depicted in Figure 1. Since the foundation of the music is the time, the bottom layer of the pyramid represents the time information (beats, tempo, and meter). Second layer represents the harmony / melody formed by playing musical notes simultaneously; third layer describes the music regions (i.e. pure vocal–PV, pure instrumental-PI, instrumental mixed vocal IMV and silence -S); forth layer and above represent the music semantics such as song structure, moods, music imagine etc. Each information layer forms some kind of music structure. Mining information in music structures is an inter-disciplinary research which mostly relay on musicology, psychology, perception and signal processing. The music research community has been exploring different methodologies to

[1] Score information is not embedded with the real sound recordings or real music.

[2] Physical music units and physical music structure are measurable and relatively easy to define.

extract the information in the music structure. These methodologies incorporate perceptual, psychological and statistical characteristics of the music signals.

Survey in this book chapter intends to give the overall idea of the advances in computer music research and the challenges in solving real world music problems. We discuss information analysis and extraction methodologies according to the information layers as shown in Figure 1. Thus sections 2, 3, 4, and 5 explain existing techniques for time information extraction, melody and harmony analysis, music region detection, and music similarity and semantic detection respectively. We then discuss music applications in section 6 and conclude the chapter in section 7.

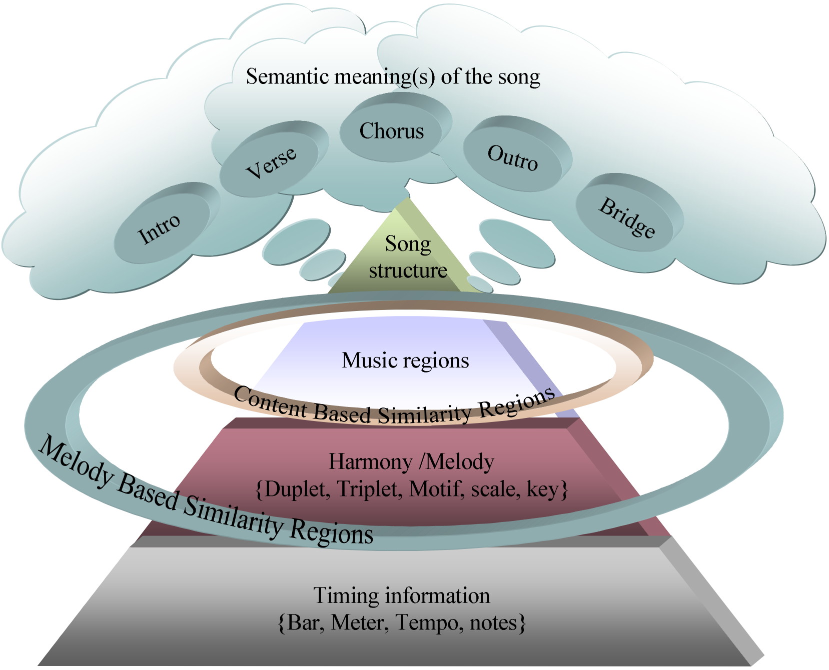

Semantic meaning(s) of the song
Chorus
Verse
Outro
Intro
Bridge
Song structure
Music regions
Content Based Similarity Regions
Harmony /Melody
{Duplet, Triplet, Motif, scale, key}
Melody Based Similarity Regions
Timing information
{Bar, Meter, Tempo, notes}

Fig. 1 Information grouping in the music structure model

2 Time Information Extraction (Beats, Accents, Onsets, Meter, Tempo)

Music time information include position details of the beats, accents and onsets, and information flow details i.e. meter and tempo. The duration of the song is measured by number of *Bars*. The steady throb to which one could clap while listening to a song is called the *Beat*. *Accents* are the beats which are stronger than the others and number of beats from one accent to the adjacent accents is equal and it divides the music into equal measure. Thus, the equal measure of number of beats from one accent to another is called the *Bar*. In a music phrase, the words or

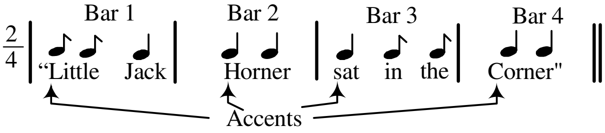


Fig. 2 Rhythmic flow of words

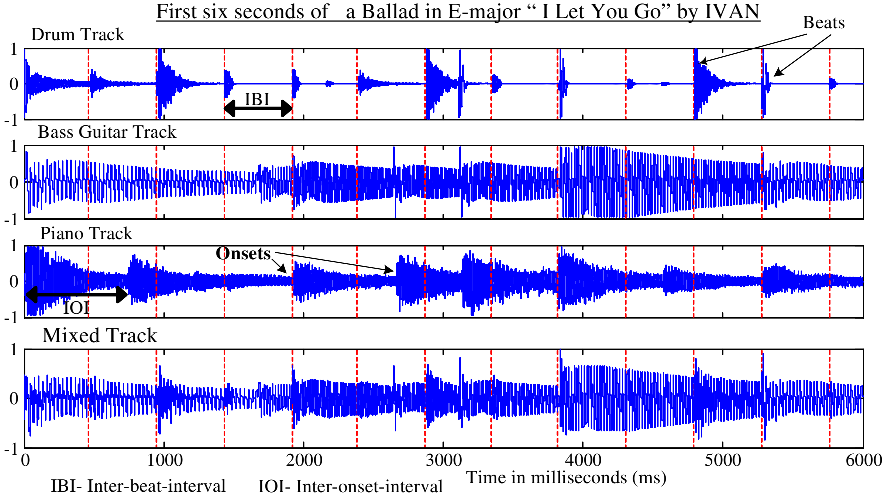


Fig. 3 First 6 seconds of instrumental tracks (Drum, Bass guitar, Piano) and edited final tract (mix of all the tracks) of a ballad (meter- 4/4 and tempo -125 BPM) "I Let You Go" sung by Ivan.

syllables are usually positioned on beats [92]. In Figure 2, we have shown the time alignment between musical notes and the words. Since accents are placed over the important syllables the *Meter* of this musical phrase is 2/4, i.e. two quarter beats per bar [92]. Tempo is the number of beats per minute (BPM).

Initial research on rhythm tracking focused on symbolic music data i.e. MIDI. Allen and Dannenberg (1990) [3] proposed a real time beat tracking / prediction technique for music in MIDI format; beam search tool was used for matching and grouping equally placed music notes and their patterns. Since the score information is not available in real sound recordings, first we need to extract information such as position, duration etc., about the notes and beats, in order to compute time information. Figure 3 shows 6 seconds of drum, bass guitar, piano and mixed (drum + bass guitar + piano) tracks of a song. Dotted vertical lines show the beat positions (see the drum track). Drum track is considered as the time stamp in a song because most of the percussion instruments create high energy impulses, which are significant in the signal. These dotted lines are aligned with many of the onset[3] positions as shown in the bass guitar and piano tracks. The distance between the adjacent dotted lines is known as the inter-beat-intervals (IBIs). Detection of the strong and weak beat patterns and their repetitions, further help to calculate the meter and tempo of the music (Scheirer 1998 [94]).

In the Figure, beats and many of the onsets are clearly visible in the solo instrumental tracks. However, when instrumental tracks are mixed, then it's difficult to clearly observe the beats and onsets (see the mixed track), which further make time information extraction in the mixed track more challenging. Since mixed music track is the one mostly available (commercially), we commonly refer to this edited /mixed final track as the real sound recording. In this chapter, we are particularly interested in music information extraction techniques that take real sound recording as the input. Accurate detection of beats and onsets is the first step for the computation of meter and tempo. Separation of instrumental tracks from the real sound recordings is an ideal and the prime step which can enhance the accuracy of beat and onset detection. However present techniques have not yet been mature enough for track / source separation. Thus sub-band decomposition technique has commonly been proposed assuming that beats and onsets in the different instrumental tracks are noticeable at different frequency bands. Existing algorithms commonly impose the following assumptions in order to detect the beat, onset meter or tempo, to simplify the problem. Such assumptions restrict the music to a constant rhythm structure.

- Meter doesn't change in the music
- Tempo doesn't change in the music
- Tempo is within a fixed range

The basic steps for time information extraction are depicted in Figure 4. Energy transient analysis, for the sub-band signals in both time and frequency domains, is considered the main step for onsets and beat detection. By applying the above assumptions, we further refine the detected onset positions and compute meter or tempo or the both. Next we explain how these steps are implemented in practice.

Dixon (2001) [27], Gouyon et al (2002) [47], Scaringella and Zoia (2004) [96] measured time-domain local peaks to detect the onsets. Bello and Sandler (2003) [10], Davies and Plumbley (2004) [25] analyzed frequency fluctuations to detect

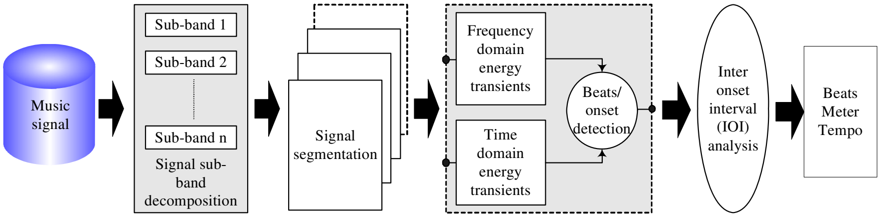

Music signal
Sub-band 1
Sub-band 2
Sub-band n
Signal sub-band decomposition
Signal segmentation
Frequency domain energy transients
Time domain energy transients
Beats/ onset detection
Inter onset interval (IOI) analysis
Beats Meter Tempo

Fig. 4 The basic steps in extracting time information

[3] Ideally onset is the starting position of a music note. Beat is an impulse generated by a percussion instrument.

onsets in the music. In order to increase the energy fluctuation resolution, signal is first decomposed into sub-bands (Scheirer 1998 [94], Duxburg et al. 2002 [30], Uhle and Herre 2003 [119], Alonso et al 2003 [2]).

Scheirer (1998) [94] passed the sub-band signals through combo filter bank to analyze tempo and beats in the signal. The algorithm, which is claimed as "perceptual model", was tested on 60 different classes of genres. Dixon (2001) [27] clustered the inter-onset intervals (IOI) to detect tempo and meter in classical, Jazz and popular music. Cemgil et al. (2001) [19] proposed stochastic dynamic system for tempo tracking. Tempo was modeled as a hidden state variable of the system. Gouyon et al (2002) [47] analyzed energy transients to construct an IOI histogram and compute the position of the smallest rhythmic pulse units name "Tick" from peak tracker. Jensen and Andersen (2003) [51] introduced a beat probability vector to keep the track of previous beat intervals and enhance the beat prediction capabilities. For beat tracking they constrained the tempo in the (50~200) BPM range. Alonso et al (2003) [2] used noise/harmonic decomposition technique to estimate tempo of 54 excerpts of different music genres where they claimed 96% of averaged accuracy. Signal was first decomposed into sub-bands using FIR filter bank. Then sub-band signals were projected into noise subspace to detect periodicities. These detected periodicities indicate the tempo. To detect tempo, time signature, Uhle and Herre (2003) [119] first detected the sub-band onsets from half wave rectified amplitude envelops of each sub-band. Based on inter-onset interval (IOI) analysis, rhythm characteristics of the signal were detected. Proposed time information extraction algorithms by Goto in 1994 [43] and in 2001 [45], assumed music to have 4/4 meter and tempo to be in the 61~185 BPM range. His initial work in 1994 was focused on beat tracking for signals with drum tracks where onsets were detected by analyzing frequency transients. Based on inter-onset interval (IOI) histogram, beats were predicted in the song. Later Goto (2001) [45] proposed real time beat tracking system which deals with music without drum sounds. In this system, information about onset times, chord change and drum patterns were taken into consideration for identifying the inter-beat-intervals (quarter-note level, half-note level and bar level). However, the frequency resolution (21.53Hz) used for detecting chord changes may not be sufficient enough because at the lower octaves, the differences of fundamental frequencies (F0s) of the notes can be as low as 1Hz (see Table 1). Maddage et al (2004) [69], combined both Duxburg et al. 2002 [30] and Goto (2001) [45]'s method to detect shortest note in popular music. Instead of linear signal decomposition, they used octave based signal sub-band decomposition for onset detection.

To detect both onsets and beats, Gao and Lee (2004) [39] used 12 Mel frequency cepstral coefficients (MFCCs) extracted from music segments (23.3 ms, 50% overlapped) and fed them into maximum a posterior algorithm. This prediction method inadequately accounted the relationship between beats and the meter, since without knowing the meter of a song, it is difficult to identify the beats from the detected onsets. To track the tempo, Davies and Plumbley (2004) [25] first detected onsets. Then beats were predicted by running auto-correlation function (ACF) over detected onsets. Sethares and Staley (2001) [97] measured periodicities and meter of the music by projecting octave based decomposed music signals into set of non-orthogonal periodic sub spaces. Tzanetakis et al. (2002) [117] computed a beat histogram to characterize the different music genres. Beat histogram was computed from wavelet based decomposed sub-band signals. Sethares et al. (2005) [99] discussed two beat tracking methods based on both Bayesian decision framework and gradient strategy. Pikrakis et al. (2004) [84] analysed acoustic level self-similarities using Mel frequency cepstral coefficient (MFCC) feature to extract meter and tempo in 300 music of Greek dance folklore music and neighbouring Eastern music traditions. They assumed that the meter remains constant throughout the music and the tempo varies between 40~330 BPM. Wang and Vilermo (2001) [120] proposed compressed domain beat tracking system based on Modified Discrete Cosine Transform (MDCT) coefficient feature extracted from both full band and sub-bands. Inter interval (IOI) histogram was used to select the correct beats.

Discussion

Onset and beat position are the prime information in the music timing/ time structure. Detection of the short time energy transitions in both time and frequency domains, is the commonly used technique for beat and onset detection. However due to polyphonic and heterogeneous source nature, it's difficult to accurately detect short time energy transition positions in the full band signal, resulting higher inaccuracies in beat and onset position detection. Thus proposed alternative approaches, which assume constant tempo and follow the main steps such as signal decomposition, beat and onset detection in sub-band signals, and fusion of the information of the detected sub-band, have indicated higher accuracy in beat and onset detection.

Another challenge is to identify the beat positions from the onsets. Beats are closely related to tempo and meter. Therefore, researches commonly assume meter to be 4/4 i.e. 4 quarter beats per bar and the tempo to be limited to a certain range, in the process of identifying beats.

In the music composition melody/ harmony contours, duration of music regions such as vocal regions and the duration of semantic regions such as chorus and verse in the song structure have certain relationships in terms of durations as well as their repetitions (Music Tech [111]). Thus by applying the knowledge of these relationships, we can effectively detect tempo and meter detection. As shown in Figure 1, based on music composition, the information layers melody/harmony, music region and semantic regions in song structure are above the time information layer. Thus incorporation of the knowledge in these information layers for time information extraction is considered as top down approach to the task. Recent efforts along this path (Goto 2001 [45]) suggest efforts that the combination of bottom up and top down approaches are more effective in the time information extraction.

3 Melody and Harmony Analysis

Playing music notes according to timing information (meter and tempo), forms complex tonal structures, *Chords* and *Key* [92] . Figure 5 illustrates the correlations between music notes, chords and key. Set of notes on which the piece is built is known as the *Key*. Playing a music note at a time (single instrumental track) creates a melody line. Playing more than 2 music notes simultaneously creates a harmonic line.

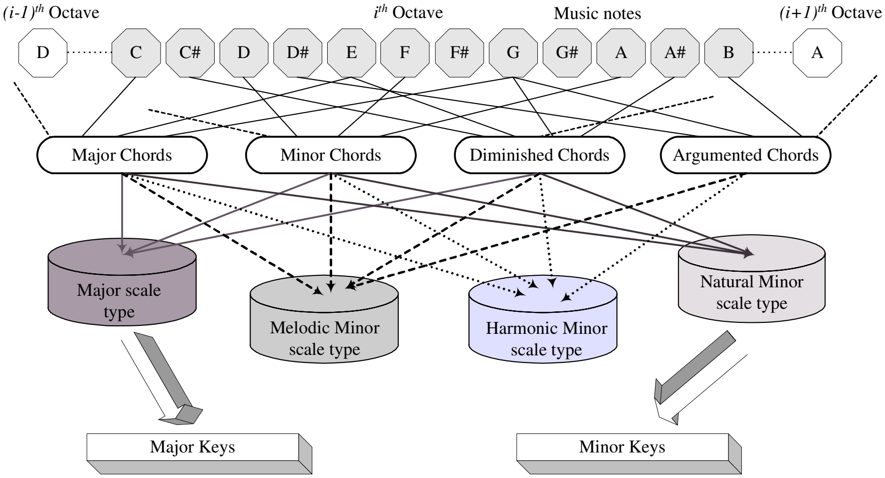


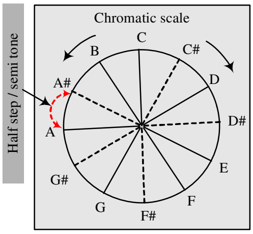
Fig. 5 Overview of top down relationship of notes, chords and key

Different approaches have been discussed in the literature to detect notes, chords and key based on the detection of fundamental frequency (F0) and harmonic components of the notes. The foundations of these approaches are based on the early 20th century research on psychological representation of tones.

Stevens et al. (1937) [103], Stevens and Volkmann (1940) [106] described pitch perception as continuous psychological effect, which is proportional to the magnitude of the frequency (i.e. *pitch height*). The octave, which is the basis of the music tonal system, has been studied by many researchers (Dowling and Harwood 1986 [28]). It was proved that tones differed by an octave interval are psychologically closely related (Allen 1967 [1], Bachem 1954 [6], Deutsch 1973 [26]). Based on this evidence, circular representation of the music pitch across octaves was proposed (Bachem 1950 [5], Shepard 1964 [102], Krumhansl 1979 [64], Bharucha and Stoeckig 1986 [12] & 1987 [13]). This circular representation is known as *chroma cycle* (Rossing et al 2001 [91]). Table 1(left) shows the music note arrangement in chromatic scale. Table 1(right) show the F0s of the music notes in each octave and chroma cycle repeats in each octave. The frequency ratio of a physical octave is 2:1 (Rossing et al 2001[91]). However, cognitive experiments have highlighted that this subjective octave ratio is closed to 2:1 at lower frequencies but increases with the frequency and exceeds physical octave ratio by 3% at about 2 kHz (Ward 1954 [124], Terhardt 1974 [111], Sundberg and Lindqvist 1973 [108], Ohgushi 1978 [82] & 1983 [83]). Musically trained/untrained listeners (Ward 1954 [124]) and number of music cultures (Dowling and Harwood 1986 [28]) presented this octave enlargement effect. In order to study this octave enlargement, Ohgushi (1978 [82]

& 1983 [83]), Hartmann (1993) [50], McKinney and Delgutte (1999) [76] suggested an octave matching scheme based on a temporal model which predicts the octave enlargement effect. However, constant 2:1 physical octave ratio is commonly practiced to simplify the complexity in musicology.

Table 1 Chroma cycle (Left); Fundamental frequencies (F0s) of the music notes in different octaves (Right). F0s are based on the concert pitch A4=440Hz



Octave		~ B1	C2B2	C3B3	C4B4	C5B5	C6B6	C7B7	C8B8	All the higher octaves in the (8192 ~ 22050) Hz frequency range
Freq-range (Hz)			64~128	128~256	256~512	512~1024	1024~2048	2048~4096	4096~8192	
12 Pitch class notes	C	0 ~ 64	65.406	130.813	261.626	523.251	1046.502	2093.004	4186.008	
	C#		69.296	138.591	277.183	554.365	1108.730	2217.460	4434.920	
	D		73.416	146.832	293.665	587.330	1174.659	2349.318	4698.636	
	D#		77.782	155.563	311.127	622.254	1244.508	2489.016	4978.032	
	E		82.407	164.814	329.628	659.255	1318.510	2637.02	5274.04	
	F		87.307	174.614	349.228	698.456	1396.913	2793.826	5587.652	
	F#		92.499	184.997	369.994	739.989	1479.978	2959.956	5919.912	
	G		97.999	195.998	391.995	783.991	1567.982	3135.964	6271.928	
	G#		103.826	207.652	415.305	830.609	1661.219	3322.438	6644.876	
	A		110.000	220.000	440.00	880.000	1760.000	3520.000	7040.000	
	A#		116.541	233.082	466.164	932.328	1864.655	3729.310	7458.62	
	B		123.471	246.942	493.883	987.767	1975.533	3951.066	7902.132	

ISO 16 standard specifies A4 = 440Hz and it is called as concert pitch

The listening tests also revealed that octave judgments for music tones over 2 kHz is difficult. Pitch perception experiments conducted by Ritsma (1967) [87] concluded that fundamental frequencies in the 100-400Hz range and their 3rd,4th, and 5th harmonics, which cover up to 2kHz frequency range produce well-defined pitch perception in human ears. Biasutti (1997) [14] conducted hearing test using 12 subjects to find the frequency limits (lower and upper) of musicians to identify major and minor triads. These lower and upper limits were found in (120~3000) Hz frequency range. Ward (1954) [124], Attneave and Olson (1971) [4] have also acknowledged that the upper limit in music pitch is in the range of 4-5 kHz. Thus, the useful upper limit of the F0 of the tones produced by music instruments is set below 5 kHz. The highest tone (C7) of the piano has a frequency of 4186 Hz.

There are two approaches discussed in the literature for representing music pitch. Goldstein (1973) [42] and Terhardt (1974 [111] & 1982 [114]), proposed two psycho-acoustical approaches: harmonic representation and sub-harmonic representation of complex tones respectively. In Goldstein's pitch representation, music tone is characterized by fundamental frequency (F0) with harmonic partials. Terhardt proposed that each separable component of a complex tone generates eight sub-harmonics and the frequency of the most commonly generated sub-harmonics determines the perceived pitch. Houtgast (1976) [48] also claimed that listeners can discriminate sub-harmonics in the higher frequencies. Laden and Keefe (1989) [65] examined the issue of the representation of music pitch by training a neural net to classify music chord types (Major, Minor and Diminished). They claimed that psycho-acoustical representation of music pitches has advantages in encoding information concerning chord inversions and spectral content than Pitch Class representation. Maddage et al. (2006) [72], also examined the pitch class profile and psycho acoustical representations of music pitch for the chord detection. The experiments highlighted that F0, sub-harmonics and harmonics information are important for the chord detection.

Moorer (1975) [79], analyzed harmonic content of music signals using optimum-comb periodicity detector to identify music notes. Though his work was limited to a mixture of duets, his fundamental research work for music signal analysis and transcription is useful for research in the music community. Recently Poliner et al. (2007) [88] discussed melody transcription systems which were used for benchmarking competitions in 2004 and 2005 at the MIREX [80].

It is assumed that the set of music tones consists of a finite set of pitches (Deutsch 1999 [26] and Rossing et al. 2001 [91]) in most of the music communities. An octave is divided into 12 tones (dodecaphonic notes[4] or 12 pitch class notes) which are approximately equally spaced in terms of log frequency and the interval between adjacent pitches is called a half step or semitone. Two tones separated by 12 half-step form an octave interval, with a frequency ratio of approximately 2:1. Krishnaswamy (2003) [63] investigated the claim, "There are some musicologists who maintain more than 12 intervals per octave" using Indian classical (Carnatic) music. In his examination, he found only 12 distinctive pitches per octaves.

Twelve-pitch class profile arrangement has been commonly used to characterize the music notes, chords in the literature: Krumhansl (1979) [64], Bharucha (1986 [12] & 1987 [13]), Fujishima (1999) [38], Goto (2001) [44], Sheh and Ellis (2003) [100], Shenoy et al. (2004) [101], Maddage et al (2004) [69] and Yoshioka et al. (2004) [132]. Fujishima (1999) [38] compared the nearest neighbour classifier and weighted sum matching methods to identify the chord from music frames, which are characterized with 12 dimension chroma vectors. Sheh and Ellis (2003) [100] modeled 24 dimensions of both pitch class profile (PCP) feature and MFCCs with HMM for chord detection. Shenoy et al. (2004) [101] discussed a rule based technique to detect the key of a song (only 4/4 music) by identifying Major and Minor chords. Inter-beat-interval frames were characterized by chroma vectors (12 dim) which accounted 5 octaves (C2~B6). Then authors ran 16 bar length window to detect the key. Su and Jeng (2001) [107] represented the harmonic contents of chords types (i.e. Major, Minor, Argument, Diminished) in time-frequency map using wavelets and modeled them in self-organizing map.

Melody of the music is created by playing solo notes with time. Goto (2001) [44] proposed adaptive tonal model to detect both melody line and the bass line which is independent of the number of sources in the CD recording. Szczerba and Czyżewski (2002) [110] ran autocorrelation within the signal frames of single instrumental line to calculate the pitch of the music. Klapuri (2003) [62] used spectral characteristics in a recursive algorithm for multiple F0 frequency estimation. Eggink and Brown (2004) [31] proposed a method to extract melody line from the complex audio using the knowledge of the signal source and the fundamental frequency (F0) detection technique. Kameoka et al. (2004) [58] modeled the harmonic structure in polyphonic music using tied-GMM.

Previous methods for harmonic structure analysis commonly utilized linear frequency transformation techniques such as Discrete Fourier Transform (DFT). Brown (1991) [15] discussed the importance of the octave fashion temporal behaviours in

[4] Dodecaphonic notes are the twelve tones (C, C$^{\#}$, D, D#, E, F, F$^{\#}$, G, G$^{\#}$, A, A$^{\#}$, B) in an octave.

music signals. She highlighted that the non-linear frequency analysis is more sensitive to the frequency components of music tones. In her method, she used wider window to calculate lower octave frequencies and smaller window for higher octave frequency. Results revealed that constant Q transformation performs better identifying F0s and harmonics in music tones than DFT. Zhu et al. (2005) [135] proposed music scale, root and key determination method based on pitch profile features and tone clustering algorithm. Instead of Fourier transformation, Constant Q transformation (Brown 1991 [15]) was used to extract 12 dimension pitch class profile feature from 11.6ms frames covering 7 octaves (27.5 ~3520) Hz.

Discussion

Melody and harmonic contours constructed by identifying all the notes played progressively and simultaneously. The steps, extraction of fundamental pitches, their associated both harmonic sub-harmonic structures and their relationships such as chroma cycle are the essential steps for note identification. Recent efforts on octave based acoustic information processing indicate high effectiveness in the melody and harmony extraction (Brown 1991 [15], Maddage et al 2006 [72], Zhu et al 2005 [135]). Then applying music knowledge such as Key, which explains the not combinations / chords that are played, can effectively correct the errors in the detected melody or harmony.

4 Music Region Detection

As shown in Figure 6, based on the sources, popular music can be classified into pure instrumental (PI) regions, pure vocal (PV) region, instrumental mixed vocal (IMV) region and silence (S) region. Music region content analysis is important for both semantic information extraction (as shown in the 4th layer and above in Figure 1) and developing music applications such as singer identification, music synthesis, transcription, etc. Generally, PV regions are rare in the popular music. Therefore, both PV and IMV regions can be considered as Vocal (V) regions. Many of the past work focused on the analysis of the content in the instrumental region. Those methods considered that whole music content is solely instrumental music. Therefore, the focus was to identify instrumental type (string, bowing, blowing, brass, percussion etc.) and the names of the instruments played in the instrumental music. In this section, we survey the past researches about both the content analysis in the instrumental region and the techniques used for detecting the music regions (PI, PV, IMV and S) in the music.

For timbre identification, Coei et al. (1994) [22] trained a Self-Organizing-Map (SOM) with MFCC feature extracted from isolated music tones which were generated from 40 different music instruments. Brown and Cooke (1994) [16] built a system to recognize instruments in which note similarity “brightness” and onset asynchrony were used to group duets, played by synthesized brass and clarinet. Kaminskyj and Materka (1995) [54] extracted amplitude envelop features from an octave isolated tones to identify guitar, piano, marimba and accordion. The instrument classification abilities of a feed-forward neural network with a K-nearest

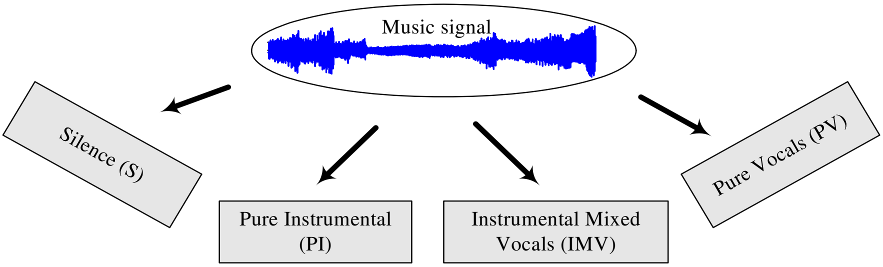


Fig. 6 Regions in the music

neighbour classifier were compared. They found both classifiers performed with nearly 98% accuracy.

Kashino and Murase (1998) [60] designed a note transcription (pitch and instrument name) system, which initially matched the input note with notes in the database. Then probabilistic network was employed as music context integration for instrument identification. To identify music instruments, Fujinaga (1998) [36] trained a K-nearest neighbour classifier with features extracted from 1338 spectral slices representing 23 instruments playing a range of pitches.

Martin (1999) [73] trained a Bayesian network with different type of features such as spectral feature, pitch, vibrato, tremolo features, and note characteristic features to recognize the non-percussive music instruments. Eronen et al. (2000) [32] used similar hierarchical classification approach earlier proposed by Martin (1999) [73] for instrument recognition.

Few instrument recognition systems have been proposed to operate on real music recordings. Brown (1999) [18] trained two Gaussian Mixture Models (GMMs) with constant-Q Cepstral coefficients extracted from oboe and saxophone using approximately one minute music data each. Dubnov and Rodet (1998) [29] used quantized MFCC feature vectors to characterize the music played on instruments. Then they used clustering algorithm to group the similar vectors and measured the similarity of the instruments. Marques (1999) [75] built a classifier to recognize flute, clarinet, harpsichord, organ, piano, trombone, violin and bagpipes. MFCCs were extracted from the recordings of solo instruments to train the classifier.

The vocal cord is the oldest music instrument. Both human auditory physiology and perceptual apparatus have evolved to develop a high level of sensitivity to the human voice. It is noticed in our survey (Maddage, 2005 [71]) that over 65% of music content in the popular music consists of singing voice. Thus singing voice carry strong messages emotional or intellectual (Xu et al. 2005 [130]). Speech processing techniques have limitations when they are applied to singing voice analysis because speech and singing voice differ significantly in terms of their production as well as their perception by human ears (Miller 1986[78], Sundberg 1987 [109]). Unlike in speech, most of the content in singing is voiced (Kim and Brian 2002 [61]). Singing voice has wider range of dynamic characteristics such as pitch, harmonics and sub-harmonics than speech (Saitou et al 2002 [93]). F0 fluctuation in singing voice is larger and more rapid than those in the speech (Miller 1986[78], Sundberg 1987 [109]).

For singing voice detection, Berenzweig and Ellis (2001) [11] used probabilistic features generated from Cepstral coefficient based acoustic model. Acoustic model is constructed as multilayer perceptual neural network with 2000 hidden units. Kim and Brian (2002) [61] first filtered the music signal using IIR bandpass filter (200~2000 Hz) to highlight the energy of the vocal region. Then vocal regions were detected analysing high amount of harmonicity of the filtered signal using an inverse comb filter bank. Zhang and Kuo (2001) [136] used a simple threshold, which was calculated using energy, average zero crossing, harmonic coefficients and spectral flux features, to find the starting point of the vocal part of the music. The similar technique was applied to detect the semantic boundaries of the online audio data; i.e. speech, music and environmental sounds for the classification. Tsai et at. (2003) [115] trained 64-mixture vocal GMM and 80-mixture non-vocal GMM with MFCCs extracted from 32ms time length with 10ms overlapped training data (216 song tracks). Bartsch and Wakefield (2004) [9] extracted spectral envelopes to characterize 10ms frames of vocal notes played by 12 singers. Then singers were identified using quadratic classifier. It is noted that in these experiments only data from vocal regions are used.

Tzanetakis (2004) [118] applied bootstrapping technique for vocal detection problem. He used spectral shape features (Centroid, Roll off, Relative Sub-band Energy) to characterize vocal/instrumental frames of 10 Jazz songs and tested with different classifiers; Logistic Regression, Neural Net classifiers, SVM, Nearest Neighbours, Bayes, and J48. Leung and Ngo (2004) [66] characterized music segments using 39 dimensional perceptual linear predictive coding (PCP) features and classified them into vocal class or instrumental class using SVM classifier.

Many current research efforts have considered the rhythm-based segmentation for music content analysis (Maddage et al 2004 [69] 2006 [73], Nwe et al 2004 [80], Ellis and Poliner 2006 [32]). Nwe et al (2004) [80], claimed Log Frequency Power Coefficients (LFPCs), which tap the spectral strengths from 12 logarithmically spaced band pass filters in 130Hz to 16 kHz frequency range, perform better than MFCCs. Mel scale is a subjective scale constructed from perceived subjective pitches where as octave scale /music scale arranges the audible frequencies in octaves (Stevens *et al.,* 1937 [104]). Fletcher (1931) [34] used octave scale for music and speech signal analysis. Brown and Puckette, (1992) [16], Maddage *et al.,* (2004) [71] and (2006) [73] reported that octave scale information extraction is effective in characterizing music information. It was empirically highlighted that cepstral feature extracted from octave scale (known as octave scale cepstral coefficients OSCC) is more capable of characterising music contents than cepstral coefficients extracted from mel scale, which is more popular in speech processing. Thus these evidences suggest the analysis of the temporal properties in the music signals such as short time energy, zero crossing, F0s, harmonics and sub-harmonics can be enhanced when the octave scale information extraction is considered.

Discussion

Due to heterogeneous source nature, detection of vocal and instrument regions in music is a very challenging task. Many of the previous techniques followed speech processing techniques which have fixed length signal segmentation (20-50ms frame size), acoustic features such as MFCC, spectral characteristics (centroid, roll off, relative sub-band energy etc.) for region content characterization, and statistical models such as GMM, Neural network and SVM for region classification steps. The performances of these techniques are limited because music knowledge has not been effectively exploited. Improvement in music region detection relays on effective ways of music segmentation, acoustic features, and statistical modeling techniques. Recent experiments highlight that rhythm based signal segmentation and octave scale spectral features such as OSCC feature can improve the music region detection accuracy.

5 Music Similarity and Semantic Detection

In the previous sections, we have discussed existing methods for analyzing music structural information such as time, melody, harmony, music regions and they are conceptually visualized as information layers in the pyramidal model shown in Figure 1. As mentioned below, music contents can be grouped into different similarity regions.

- Beat cycle : repeated beat patterns
- Melody/harmony based similarity : repetition of the melody/chord patterns in music (see Figure 1)
- Vocal similarity : similar phrases but different melody/harmony
- Content based similarity – both similar harmonic/melody line and vocal line
- Semantic level similarity : music pieces or excerpts that creates similar auditory sensation

Intra music similarity describes the content repetitions within the music piece. Whereas, inter music similarity describes structural similarities among several music pieces. Intra and inter music similarity analysis has been a highly focused research in the recent years, mainly due to strong commercial interested in developing applications such as music search engines for music content distributors.

In the initial research, feature-based similarity marching algorithms were proposed for detecting repeated patterns in music. Dennenberg and Hu (2002) [23] proposed chroma based and autocorrelation based techniques to detect the melody line in the music. Repeated segments in the music were identified using Euclidean distance similarity matching and clustering of the music segments. Goto (2003) [46] and Bartsch and Wakefield (2001) [7] used pitch sensitive chroma-based features to detect repeated sections (i.e. - chorus) in the music. Foote et al. (2002) [34] constructed a similarity matrix and Cooper and Foote (2002) [21] defined a global similarity function based on extracted MFCCs to find the most salient sections in the music. Logan and Chu (2000) [67] used clustering and hidden Markov model (HMM) to detect the key phrases in the choruses. Lu and Zhang (2003) [68] estimated the most repetitive segment of the music clip based on high-level features (occurrence frequency, energy and positional weighting) calculated from MFCC and octave-based on the spectral contrast.

Pikrakis et al. (2003) [84] used context dependent dynamic time warping algorithm to find the music patterns in monophonic environment. Xu et al. (2005) [130] used an adaptive clustering method based on LPC and MFCC features to find the repeated contents in the music. Xi et al. (2005) [128] extracted choruses which have both similar melody and phrase, to create music summary. Chai and Vercoe (2003) [20] characterized the music with pitch, spectral and chroma based features and then analyzed recurrent structure to generate music thumbnail. Gao et al. (2004) [40] proposed a rhythm based clustering technique to detect some kind of music structures. Then music frames were characterized with 12 dimension MFCCs and clustered them into similar music group. However, those works haven't discussed the characteristics of the similarity groups detected in their algorithm.

Maddage et al (2004) [69] and (2005) [71] discussed a method to detect content based similarity and melody based similarity regions in the music. Melody-based similarity regions are defined as the regions, which have similar pitch contours constructed from the chord patterns. Content-based similarity regions are the regions, which have both similar vocal content and melody. Corresponding to the popular song structure [111], the Chorus sections and Verse sections in a song are considered as the content-based similarity regions and melody-based similarity regions respectively. Then rule based method used to detect the semantic regions in the popular song structure (Maddage et al 2004 [69]). Paulus, and Klapuri (2008) [84] used MFCC, chroma, and rhythmgram features characterise the music content and the measure the content similarities in the probabilistic space to detect chorus and verses in the music. They highlighted incorporation of music knowledge for content modelling is more effective in similarity analysis.

Music mood is another semantic descriptor, which would be highly useful in music therapy, and music recommendation. There are few research work in this direction and they are all focused on predefined mood classification. Lu and Zhang (2006) [69] modeled predefined mood classes using spectral shape and spectral contrast features and GMMs.

Discussion

Music similarity analysis at acoustic, content and semantic levels is very useful for designing search engines.. Many of the earlier systems used chroma features such as pitch class profile feature (PCP) and timbre features such as MFCC, for similarity analysis. These features cover acoustic similarities in the music. When music knowledge is incorporated with these acoustic similarities, we are able to analyse content based similarities such as similar music phrases, choruses, similar chord patterns, etc. Capturing different similarities at content level is essential for designing semantic descriptors such as genre, mood etc. Statistical modeling techniques, such as GMM, SVM, NN and clustering have been commonly used in the earlier systems for modeling these semantic descriptors.

6 Music Applications

In this section, we outline how information in the music structure is useful for applications.

6.1 Lyrics Identification and Music Transcription

Transcription of music details the composition of the music piece. Thus tools which transcribe music information are useful in the music education for understanding music content. Music transcription also serves as a semantic indexing scheme in music search engines. Lyrics identification has applications such as production of Karaoke music and music documentary, music summarization. Due to high signal complexity, lyrics transcription remains very challenging task. Thus researches alternatively focus on the lyrics (text) and music (audio) alignment task (Wang et. al 2004. [122] and Fujihara et. al 2006. [36])

6.2 Music Genre Recognition

Genre which describes the style of the music can be used as a semantic indexing scheme for clustering /organizing music databases. Thus genre is useful for music content characterizazion and summarization applications, as well as digital right management applications. Previous genre recognition techniques mainly used low level acoustic features such as MFCC, spectral characteristics, and statistical modelling techniques such as HMM, GMM, SVM and neural networks (NN) to classify music into predefined music genre classes such as POP, ROCK, JAZZ, CLASSIC (Scaringella e al. 2006. [97], Soltau et al. 1998 [104], Han et al. 1998 [47], Pye 2000. [89], Jiang et al. 2002. [54], Tzanetakis and Cook 2002. [117]). Since music genre directly related to music style, its essential to incorporate time, melody/harmony and music region contents and different content similarity relationships in the genre recognition modelling frameworks. Towards music structure incorporation, Xu et al (2003) [130], proposed support vector machine (SVM) based hierarchical genre modelling technique and Shaoxi et al (2004) [126] incorporated beat structure information for genre classification.

6.3 Music Summarization

Advertising a music summary instead of the full record can be a way to prevent the illegal music downloading and many music record publishes manually create music summaries which is time consuming and labour intensive. Therefore, people look into automatic music summarization solutions. May be due to the legal issues, a music summary is an excerpt of the original song. Thus the key challenge is to find an excerpt to reflect the essence of the song. Summary making is subjective. A number of techniques have been proposed for music summary generation (Logan and Chu (2000) [67], Xu et al (2002) [128], Lu and Zhang (2003) [68], Chai and Vercoe (2003) [20]). All these approaches have difficulties detecting boundaries of content-based similarity regions and avoiding the repetitions in the summary. Thus accurate extraction of high level song structure information (Maddage et al 2004 [69]) which is formulated using semantic events such as Intro, Verse, Chorus, Bridge, Middle eighth and Outro, is useful for summary making.

6.4 Music Search

Like any other search engine, music information search engines are also very useful in the music education, digital right management and entertainment industry. Queries for the music search engines are in the form of text, humming, or a music excerpt. In the query by humming systems, the melody based match techniques have been commonly employed. In these systems, melodies of the query and the music in the archive are extracted using F0 tracking algorithms (Ghias et al 1995 [41]). Then either string matching techniques (McNab et al 2000 [77]) or statistical models such as Hidden Markov Models (Shifrin et al 2002 [103]) are employed to find the similarities between the query and the achieved music. The effectiveness of these music search algorithms is lower because other structure similarities in the music signals such as beat, rhythm, vocal similarities etc. have not been taken accounted. Maddage et al. (2006) [73] proposed music structure based search algorithm, in which search space includes time, melody/harmony and vocal information. Search space can be expanded by incorporating more semantic information.

6.5 Music Streaming

There has been a greater concern regarding how to stream media contents in real-time, over different networks which keep acceptable quality of service (QoS) at the receivers end. The objective of packet loss recovery schemes in the audio streaming is to reconstruct the data packets so that the received audio is perceptually indistinguishable or sufficiently similar to the original audio (Perkins et al 1998 [84] and Wah et al 2000 [124]). In the networks, sender-receiver based schemes, especially media specific FEC schemes, are more common in the music streaming. Earlier proposed error concealment techniques for music streaming mainly focused on reconstructing packets with percussion signals (Wang et al 2004 [123], Wang et al 2003 [121], Wyse et al 2003 [126]). However, those methods are inefficient reconstructing packets, which contain signals other than percussion sounds such as vocals, instrumental. Music structure analysis is able to reveal different levels of similarities such as content or melody based. Taking music similarities into consideration in error concealment schemes, we are able to reduce the bandwidth by avoiding the retransmission of the packets.

6.6 Music Compression

The greatest challenge in music compression is how to deal with the trade-off between the size of the music file (storage space) and the loss in the signal information (perceptual quality). MPEG-1, MPEG-2, ATRAC-2, ATRAC -3 and DOLBY AC-3 are some existing compression techniques. Earlier research on music perception reveals a strong relationship between the intervallic structure (harmonics) of the music tones and our cognitive mechanism. Thus applying octave scale music content characterization techniques together with rhythm based signal segmentation technique we can design higher perceptual quality compression scheme.

Compared with conventional audio compression techniques such as MP3, which produces a 5:1 compression ratio, incorporation of music structure analysis (especially with semantic similarities in the music) produces much higher compression ratios, which can reach up to10:1 or even higher.

6.7 Watermarking Scheme for Music

Existing watermarking techniques are evenly applied to the entire music content. But they may not detect the watermark on the randomly clipped song section. We can use music structure information to design a robust watermarking scheme so that there is a high probability that the watermark can be detected in any possible music extract of the song. For example, listeners can remember and recall chorus sections better than verse sections. So, there is a high probability that the song clips contain choruses rather than verses. With this knowledge, we can design a watermark scheme which gives higher priority to watermarking chorus rather than verses. A song is measured in terms of bars, and all music content fluctuations are synchronized according to the beat structure. Thus, beat space level watermarking is better than evenly distributed watermarking from the point of view of testing a section of the music (Xu et al 2007 [132]).

6.8 Computer Aid Tools for Music Composers and Analyzers

It is beneficial to have computer music tools which can assist musicians to analyze not only others' music but also their own music. Music transcription and summarization can aid understand the music. Students who are trained to become musicians mostly analyze the music composed by others. They incorporate pieces of those music to bridge the gap between creating new music and making mixes out of bits and pieces. The ideas that are presented in this chapter allow us to produce a complete set of tools which help musicians analyze music from a completely objective point of view and a logical prospective. Today, DJs (Disc Jockeys) are making mixers by manually cutting vocal phrases and music pieces from various songs. However, with the help of computer-aid music structure analysis tools, they can effectively analyze much bigger music archives in a short time and find interesting music clips. By combining these extracted clips we can create remix version of songs which sound much more organic and naturally miraculous that the original one.

6.9 Music for Video Applications

Automatic insertion of music into video content, explores efficient ways of making different audio-visual documentaries. For example, TV channels like National Geographic would like to have an audio-visual platform which can assist in the selection of different background music for visual content. Entertainment channels like MTV may be interested in making automatic music videos, and sport channels like ESPN may consider generating music sport videos. To formulate these audio-visual applications, it is required to understand both audio and video contents. Music structure analysis can help in selecting suitable music for the visual content.

7 Conclusion

Overall music structure consists of different layers of information such as time information, melody/harmony, music regions, song structures and higher order semantics. Previous methods for music structure analysis commonly followed fixed length signal segmentation, feature extraction to characterize the signal frames and statistical modelling techniques. These general steps are similar to those adopted in the speech signal processing. However, the important argument is that how well these speech processing techniques are suitable for music signal processing, knowing that speech and music significantly differ each other from both production and perception. Unlike speech signals, music signals are produced by heterogeneous sources with much more dynamic.

Recent music research has indicated that the incorporation of music composition knowledge is beneficial for music information modeling. Along this direction, rhythm, based signal segmentation and signal separation, music production based such as octave scale information characterization, layer wise information modeling, have been proposed for music structure analysis.

One of the important issues is how to evaluate different algorithms that are developed for music information extraction. MIREX [80] has been actively conducting benchmarking competition for music algorithm evaluations. Implementing robust and accurate music structure analysis algorithms will have great impact not only on the music entertainment industry but also on the other important sectors such as education and personal health care. For example, music therapy can be useful for healing patients who are suffering from deceases in every part of the body. There for not only designing robust and accurate content extraction/ modeling algorithms, but reshaping the scope to incorporate the needs in other disciplines such as psychology can drive the computer music research with great momentum.

References

[1] Allen, D.: Octave Discriminability of Musical and Non-musical Subjects. Journal of the Psychonomic Science 7, 421–422 (1967)

[2] Alonso, M., Badeau, R., David, B., Richard, G.: Musical Tempo Estimation using Noise Subspace Projections. In: Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, New York, October 19-22 (2003)

[3] Allen, P.E., Dannenberg, R.B.: Tracking Musical Beats in Real Time. In: Proc. of the International Computer Music Conference (ICMA), Glasgow, pp. 140–143 (1990)

[4] Attneave, F., Olson, R.: Pitch as a Medium: A New Approach to Psychophysical Scaling. American Journal of Psychology 84, 147–166 (1971)

[5] Bachem, A.: A Tone Height and Tone Chroma as Two Different Pitch Qualities. International Journal of Psychonomics (Acta Psychological) 7, 80–88 (1950)

[6] Bachem, A.: Time Factors in Relative and Absolute Pitch Determination. Journal of the Acoustical Society of America (JASA) 26, 751–753 (1954)

[7] Baratè, A., Ludovico, L.A.: An XML-based Synchronization of Audio and Graphical Representations of Music Scores. In: Proc. 8th International Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS 2007 (2007)

[8] Bartsch, M.A., Wakefield, G.H.: To Catch a Chorus: Using Chroma-based Representations for Audio Thumbnailing. In: Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, New York, October 21-24 (2001)

[9] Bartsch, M.A., Wakefield, G.H.: Singing Voice Identification Using Spectral Envelope Estimation. IEEE Transaction on Speech and Audio Processing 12(2), 100–109 (2004)

[10] Bello, J.P., Sandler, M.B.: Phase-Based Note Onset Detection for Music Signals. In: Proc. International conference on Acoustics, Speech, and Signal processing (ICASSP), Hong Kong, April 6-10 (2003)

[11] Berenzweig, A.L., Ellis, D.P.W.: Location singing voice segments within music signals. In: Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, New York, October 21-24, 2001, pp. 119–122 (2001)

[12] Bharucha, J.J., Stoeckig, K.: Reaction Time and Musical Expectancy: Priming of Chords. Journal of Experimental Psychology: Human Perception and Performance 12, 403–410 (1986)

[13] Bharucha, J.J., Stoeckig, K.: Priming of Chords: Spreading Activation or Overlapping Frequency Spectra? Journal of Perception and Psychophysics 41(6), 519–524 (1987)

[14] Biasutti, M.: Sharp Low-and High-Frequency Limits on Musical Chord Recognition. Journal of Hearing Research 105, 77–84 (1997)

[15] Brown, J.C.: Calculation of a Constant Q Spectral Transform. Journal of the Acoustical Society of America (JASA) 89, 425–434 (1991)

[16] Brown, J.C., Puckette, M.S.: An efficient algorithm for the calculation of a constant Q transform. Journal of Acoustic Society America 92(5), 1933–1941 (1992)

[17] Brown, J.C., Cooke, M.: Perceptual Grouping of Musical Sounds: A Computational Model. Journal of New Music Research 23, 107–132 (1994)

[18] Brown, J.C.: Computer identification of musical instruments using pattern recognition with Capstral coefficients as features. Journal of Acoustic Society America 105(3), 1933–1941 (1999)

[19] Cemgil, A.T., Kappen, H.J., Desain, P.W.M., Honing, H.J.: On tempo tracking: Tempogram representation and Kalman filtering. Journal of New Music Research 29(4), 259–273 (2001)

[20] Chai, W., Vercoe, B.: Music Thumbnailing via Structural Analysis. In: Proc. ACM International conference on Multimedia (ACM MM), Berkeley, CA, USA, November 2-8, 2003, pp. 223–226 (2003)

[21] Cooper, M., Foote, J.: Automatic Music Summarization via Similarity Analysis. In: Proc. 3rd International Symposium of Music Information Retrieval (ISMIR), Paris, France, October 13-17 (2002)

[22] Cosi, P., DePoli, G., Prandoni, P.: Timbre characterization with Mel- Cepstrum and neural nets. In: Proc. of International Computer Music Conference (ICMC), Aarhus, Denmark, September 12 - 17, pp. 42–45 (1994)

[23] Dannenberg, R.B.: An On-Line Algorithm for Real-Time Accompaniment. In: Proc. International Computer Music Conference, pp. 193–198 (1984)

[24] Dannenberg, R.B., Hu, N.: Discovering Musical structure in Audio Recordings. In: Proc. 2nd International Conference of Music and Intelligence (ICMAI), Edinburgh, Scotland, UK, September 12-14, 2002, pp. 43–57 (2002)

[25] Davies, M.E.P., Plumbley, M.D.: Causal Tempo Tracking of Audio. In: Proc. of 5th International Symposium/Conference of Music Information Retrieval (ISMIR), Barcelona, Spain, October 10-15 (2004)

[26] Deutsch, D.: The Psychology of Music, 2nd edn. Series in Cognition and Perception. Academic Press, San Diego (1999)

[27] Dixon, S.: Automatic Extraction of Tempo and Beat from Expressive Performances. Journal of New Music Research 30(1), 39–58 (2001)

[28] Dowling, W.J., Harwood, D.L.: Music Cognition. Series in Cognition and Perception. Academic Press, San Diego (1986)

[29] Dubnov, S., Rodet, X.: Timbre Recognition with Combined Stationary and Temporal Features. In: Proc. International Computer Music Conference (ICMC), Michigan, USA, October 1-6 (1998)

[30] Duxburg, C., Sandler, M., Davies, M.: A Hybrid Approach to Musical Note Onset Detection. In: Proc. of 5th International Conference on Digital Audio Effects (DAFx 2002), Hamburg, Germany, September 26-28 (2002)

[31] Eggink, J., Brown, G.J.: Extracting Melody Lines from Complex Audio. In: Proc. of 5th International Symposium/Conference of Music Information Retrieval (ISMIR), Barcelona, Spain, October 10-15 (2004)

[32] Ellis, D.P.W., Poliner, G.E.: Identifying 'Cover Songs' with Chroma Features and Dynamic Programming Beat Tracking. In: Proc. International Conference on Acoustics, Speech, and Signal Processing, ICASSP (2006)

[33] Eronen, A., Klapuri, A.: Musical Instrument Recognition Using Cepstral Coefficients and Temporal Features. In: Proc. of International Conference on Acoustic, Speech and Signal Processing (ICASSP), Istanbul, Turkey, June 05-09 (2000)

[34] Fletcher, H.: Some Physical Characteristics of Speech and Music. Journal of Acoustical Society of America 3(2), 1–26 (1931)

[35] Foote, J., Cooper, M., Girgensohn, A.: Creating Music Video using Automatic Media Analysis. In: Proc. International ACM Conference on Multimedia (ACM MM), Juan-les-Pins, France, December 1-6 (2002)

[36] Fujihara, H., Goto, M., Ogata, J., Komatani, K., Ogata, T., Okuno, H.G.: Automatic Synchronization between Lyrics and Music CD Recordings based on Viterbi Alignment of Segregated Vocal Signals. In: Proc. IEEE International Symposium on Multimedia (ISMIR),

[37] Fujinaga, I.: Machine Recognition of Timbre Using Steady-state Tone of Acoustic Musical Instruments. In: Proc. International Computer Music Conference (ICMC), Michigan, USA, October 1-6, pp. 207–210 (1998)

[38] Fujishima, T.: Real-time Chord Recognition of Musical Sounds: A System using Common Lisp Music. In: Proc. of International Computer Music Conference (ICMC), 1999, Beijing, pp. 464–467 (1999)

[39] Gao, S., Lee, C.H.: An Adaptive Learning Approach to Music Tempo and Beat Analysis. In: Proc. of International Conference on Acoustic, Speech and Signal Processing (ICASSP), Montreal, Canada, May 17-21 (2004)

[40] Gao, S., Lee, C.H., Zhu, Y.: An Unsupervised Learning Approach to Music Event Detection. In: Proc. of IEEE International Conference on Multimedia and Expo. (ICME), Taipei, Taiwan, June 27-30 (2004)

[41] Ghias, A., Logan, J., Chamberlin, D., Smith, B.C.: Query By Humming: Musical Information Retrieval in an Audio Database. In: 3rd ACM International conference on Multimedia (ACM MM), San Francisco, California, USA, November 5-9, pp. 231–236 (1995)

[42] Goldstein, J.L.: An Optimum Processor Theory for the Central Formation of the Pitch of Complex Tones. Journal of the Acoustical Society of America (JASA) 54, 1496–1516 (1973)

[43] Goto, M., Muraoka, Y.: A Beat Tracking System for Acoustic Signals of Music. In: Proc. 2nd ACM International Conference on Multimedia, San Francisco, California, USA, October 15-20, pp. 365–372 (1994)

[44] Goto, M.: A Predominant F0 Estimation Method for CD Recordings: MAP Estimation using EM Algorithm for Adaptive Tone Models. In: Proc. of International conference on Acoustics, Speech, and Signal processing (ICASSP), Sault lake city, Utah, May 7-11, pp. 3365–3368 (2001)

[45] Goto, M.: An Audio-based Real-time Beat Tracking System for Music With or Without Drum-sounds. Journal of New Music Research 30(2), 159–171 (2001)

[46] Goto, M.: A Chorus-Section Detecting Method for Musical Audio Signals. In: Proc. International conference on Acoustics, Speech, and Signal processing (ICASSP), Hong Kong, April 6-10 (2003)

[47] Gouyon, F., Herrera, P., Cano, P.: Pulse-Dependent Analyses of Percussive Music. In: Proc. International Conference on Virtual, Synthetic and Entertainment Audio (AES 22), Espoo, Finland, June 15-17 (2002)

[48] Han, K.P., Pank, Y.S., Jeon, S.G., Lee, G.C., Ha, Y.H.: Genre Classification System on TV Sound Signals Based on a Spectrogram Analysis. IEEE Transaction on Consumer Electronics 55(1), 33–42 (1998)

[49] Houtgast, T.: Sub-Harmonic Pitches of a Pure Tone at Low S/N Ratio. Journal of the Acoustical Society of America (JASA) 60(2), 405–409 (1976)

[50] Hartmann, W.: On the Origin of the Enlarged Melodic Octaves. Journal of the Acoustical Society of America (JASA) 93, 3400–3409 (1993)

[51] International Conference on Computer Music Research

[52] International Society for Music Information Retrieval

[53] Jensen, K., Andersen, T.H.: Real-time beat estimation using feature extraction. In: Wiil, U.K. (ed.) CMMR 2003. LNCS, vol. 2771, pp. 13–22. Springer, Heidelberg (2004)

[54] Jiang, D.N., Lu, L., Zhang, H.J., Tao, J.H., Cai, L.H.: Music Type Classification by Spectral Contrast Feature. In: Proc. of IEEE International Conference on Multimedia and Expo. (ICME), Lausanne, Switzerland (2002)

[55] Jourdain, R.: Music, The Brain, and Ecstasy: How Music Capture Our Imagination. HarperCollins (1997)

[56] Journal of New Music Research

[57] Journal of the Acoustical Society of America Computer Music Journal (JASA)

[58] Kameoka, H., NIshimoto, T., Sagayama, S.: Separation of Harmonic Structures based on Tied Gaussian Mixture Model and Information Criterion for Concurrent Sounds. In: Proc. of International conference on Acoustics, Speech, and Signal processing (ICASSP), Montreal, Canada (May 2004)

[59] Kaminskyj, I., Materka, A.: Automatic Source Identification of Monophonic Musical Instrument Sounds. In: Proc. IEEE International Conference on Neural Networks, Perth, Australia, November 27-December 1, pp. 189–194 (1995)

[60] Kashino, K., Murase, H.: Music Recognition using Note Transition Context. In: Proc. of International conference on Acoustics, Speech, and Signal processing (ICASSP), Seattle, Washington, USA, May 12-15 (1998)

[61] Kim, Y.K., Brian, W.: Singer Identification in Popular Music Recordings Using Voice Coding Features. In: Proc. 3rd International Symposium of Music Information Retrieval (ISMIR), Paris, France, October 13-17 (2002)

[62] Klapuri, A.P.: Multiple Fundamental Frequency Estimation Based on Harmonicity and Spectral Smoothness. IEEE Transaction on Speech and Audio Processing 11(6), 804–816 (2003)

[63] Krishnaswamy, A.: Application of Pitch Tracking to South Indian Classical Music. In: Proc. of International conference on Acoustics, Speech, and Signal processing (ICASSP), Hong Kong, April 6-10 (2003)

[64] Krumhansl, C.L.: The Psychological Representation of Musical Pitch in a Tonal Context. Journal of Cognitive Psychology 11(3), 346–374 (1979)

[65] Laden, B., Keefe, D.H.: The Representation of Pitch in a Neural Net Model of Chord Classification. Computer Music Journal 13(4), 12–26 (Winter 1989)

[66] Leung, T.W., Ngo, C.W.: ICA-FX Features for Classification of Singing Voice and Instrumental Sound. In: Proc. International Conference on Pattern Recognition (ICPR), Cambridge, UK, August 23-26 (2004)

[67] Logan, B., Chu, S.: Music Summarization Using Key Phrases. In: Proc. International Conference on Acoustics, Speech, and Signal processing (ICASSP), Orlando, USA (2000)

[68] Lu, L., Zhang, H.J.: Automated Extraction of Music Snippets. In: Proc. ACM International Conference on Multimedia (ACM MM), Berkeley, CA, USA, pp. 140–147 (2003)

[69] Lu, L., Zhang, H.J.: Automatic Mood Detection and Tracking of Music Audio Signals. IEEE Transactions on Audio, Speech, and Language Processing 14(1) (January 2006)

[70] Maddage, N.C., Xu, C.S., Kankanhalli, M.S., Shao, X.: Content-based Music Structure Analysis with Applications to Music Semantic Understanding. In: Proc. International ACM Conference on Multimedia (ACM MM), New York, USA, October 10-16 (2004)

[71] Maddage, N.C.: Content-Based Music Structure Analysis. Ph.D. dissertation, School of Computing, National University of Singapore (2005)

[72] Maddage, N.C., Kankanhalli, M.S., Li, H.: A Hierarchical Approach for Music Chord Modelling based on the Analysis of Tonal Characteristics. In: IEEE International Conference on Multimedia & Expo. (ICME), Toronto, Canada, July 9-12 (2006)

[73] Maddage, N.C., Li, H., Kankanhalli, M.S.: Music Structure based Vector Space Retrieval. In: Proc. International Conference of ACM Special Interest Group on Information Retrieval (ACM SIGIR), pp. 67–74 (2006)

[74] Martin, K.D.: Sound-Source Recognition: A Theory and Computational Model. Ph.D. dissertation, Massachusetts Institute of Technology (MIT), Media Lab, Cambridge, USA (June 1999)

[75] Marques, J.: An Automatic Annotation System for Audio Data Containing Music. Master's Thesis, Massachusetts Institute of Technology (MIT), Media Lab, Cambridge, USA (1999)

[76] McKinney, M.F., Delgutte, B.: A Possible Neurophysiologic Basis of the Octave enlargement Effect. Journal of the Acoustical Society of America (JASA) 106(5), 2679–2692 (1999)

[77] McNab, R.J., Smith, L.A., Witten, I.H., Henderson, C.L.: Tune Retrieval in the Multimedia Library. Journal of Multimedia Tools and Applications 10(2-3), 113–132 (2000)

[78] Miller, R.: The Structure of Singing: System and Art in Vocal Technique. Wadsworth Group/Thomson Learning, Belmont California, USA (1986)

[79] Moorer, J.A.: On the Segmentation and Analysis of Continuous Musical Sound by Digital Computer. Ph.D. dissertation, Department of Computer Science, Stanford University (1975)

[80] Music Information Retrieval Evaluation eXchange (MIREX)

[81] Nwe, T.L., Wang, Y.: Automatic Detection of Vocal Segments in Popular Songs. In: Proc. of 5th International Symposium/Conference of Music Information Retrieval (ISMIR), Barcelona, Spain, October 10-15 (2004)

[82] Ohgushi, K.: On the Role of Spatial and Temporal Cues in the Perception of the Pitch of Complex Tones. Journal of the Acoustical Society of America (JASA) 64, 764–771 (1978)

[83] Ohgushi, K.: The Origin of Tonality and a Possible Explanation of the Octave Enlargement Phenomenon. Journal of the Acoustical Society of America (JASA) 73, 1694–1700 (1983)

[84] Paulus, J., Klapuri, A.: Music Structure Analysis using a Probabilistic Fitness Measure and an Integrated Musicological Model. In: Proc. International Symposium/Conference of Music Information Retrieval, ISMIR (2008)

[85] Perkins, C., Hodson, O., Hardman, V.: A Survey of Packet Loss Recovery Techniques for Streaming Audio. IEEE Network Magazine, 40–48 (September/October 1998)

[86] Pikrakis, A., Antonopoulos, I., Theodoridis, S.: Music Meter and Tempo Tracking from Raw Polyphonic Audio. In: Proc. of 5th International Symposium/Conference of Music Information Retrieval (ISMIR), Barcelona, Spain, October 10-15 (2004)

[87] Pinto, A., Haus, G.: A novel xml music information retrieval method using graph invariants. ACM Transactions on Information Systems (2007)

[88] Poliner, G., Ellis, D., Ehmann, A., Gómez, E., Streich, S., Ong, B.: Melody Transcription from Music Audio: Approaches and Evaluation. IEEE Transaction on Audio, Speech, and Language Processing 14(4), 1247–1256 (2007)

[89] Pye, D.: Content-Based Methods for the management of Digital Music. In: Proc. of International conference on Acoustics, Speech, and Signal processing (ICASSP), Istanbul, Turkey, June 05-09 (2000)

[90] Ritsma, R.J.: Frequency Dominant in the Perception of the Pitch of Complex Sounds. Journal of Acoustical Society of America 42(1), 191–198 (1967)

[91] Rossing, T.D., Moore, F.R., Wheeler, P.A.: Science of Sound, 3rd edn. Addison Wesley, Reading (2001)

[92] Rudiments and Theory of Music, The associated board of the royal schools of music, 14 Bedford Square, London, WC1B 3JG (1949)

[93] Saitou, T., Unoki, M., Akagi, M.: Extraction of F0 Dynamic Characteristics and Developments of F0 Control Model in Singing Voice. In: Proc. of the 8th International Conference on Auditory Display, Kyoto, Japan, July 02 – 05 (2002)

[94] Sakeo, H., Chiba, S.: Dynamic Programming Algorithm Optimization for Spoken Word Recognition. IEEE Transaction on Audio, Speech, and Language Processing 26(1), 43–49 (1978)

[95] Scheirer, E.D.: Tempo and Beat Analysis of Acoustic Music Signals. Journal of Acoustical Society of America 103(1), 588–601 (1998)

[96] Scaringella, N., Zoia, G.: A Real-Time Beat Tracker for Unrestricted Audio Signals. In: Proc. of the Conference of Sound and Music Computing (JIM/CIM), Paris, France, October 20-22 (2004)

[97] Scaringella, N., Zoia, G., Mlynek, D.: Automatic Genre Classification of Music Content. IEEE Signal Processing Magazine 23(2) (March 2006)

[98] Sethares, W.A., Staley, T.W.: Meter and Periodicity in Music Performance. Journal of New Music Research 30(2) (June 2001)

[99] Sethares, W.A., Morris, R.D., Sethares, J.C.: Beat Tracking of Musical Performances Using Low-Level Audio Features. IEEE Transactions on Speech and Audio Processing 13(2), 275–285 (2005)

[100] Sheh, A., Ellis, D.P.W.: Chord Segmentation and Recognition using EM-Trained Hidden Markov Models. In: Proc. 4th International Symposium of Music Information Retrieval (ISMIR), Baltimore, Maryland, USA, October 26-30 (2003)

[101] Shenoy, A., Mohapatra, R., Wang, Y.: Key Detection of Acoustic Musical Signals. In: Proc. of IEEE International Conference on Multimedia and Expo. (ICME), Taipei, Taiwan, June 27-30 (2004)

[102] Shepard, R.N.: Circularity in Judgments of Relative Pitch. Journal of the Acoustical Society of America (JASA) 36, 2346–2353 (1964)

[103] Shifrin, J., Pardo, B., Meek, C., Birmingham, W.P.: HMM-Based Musical Query Retrieval. In: Proc. of the 2nd Joint International Conference (ACM & IEEE-CS) on Digital Libraries (JCDL), Portland, Origone, USA, July 14-18, pp. 295–300 (2002)

[104] Soltau, H., Schultz, T., Westphal, M., Waibel, A.: Recognition of Music Types. In: Proc. of International conference on Acoustics, Speech, and Signal processing (ICASSP), Seattle, Washington, USA, May 12-15 (1998)

[105] Stevens, S.S., Volkmann, J., Newman, E.B.: A Scale for the Measurement of the Psychological Magnitude of Pitch. Journal of the Acoustical Society of America (JASA) 8(3), 185–190 (1937)

[106] Stevens, S.S., Volkmann, J.: The Relation of Pitch Frequency; a Relative Scale. Journal of the Acoustical Society of America (JASA) 53, 329–353 (1940)

[107] Su, B., Jeng, S.: Multi-Timbre Chord Classification using Wavelet Transform and Self-organized Map Neural Networks. In: Proc. of International conference on Acoustics, Speech, and Signal processing (ICASSP), Sault lake city, Utah, vol. V, pp. 3377–3380 (2001)

[108] Sundberg, J., Lindqvist, J.: Musical Octaves and Pitch. Journal of the Acoustical Society of America (JASA) 54, 922–929 (1973)

[109] Sundberg, J.: The Science of the Singing Voice. Northern Illinois University Press, Dekalb (1987)

[110] Szczerba, M., Czyżewski, A.: Pitch estimation Enhancement Employing Neural Network-Based Music Prediction. In: Proc. 6th IASTED International Conference on Artificial Intelligence and Soft Computing (ASC), Banff, Canada, July 17-19 (2002)

[111] MUSIC TECH, Ten Minute Master No 18: Song Structure, MUSIC TECH magazine, pp. 62–63 (October 2003), http://www.musictechmag.co.uk

[112] Takeda, H., NIshimoto, T., Sagayama, S.: Rhythm and Tempo Recognition of Music Performance from a Probabilistic Approach. In: Proc. 5th International Symposium of Music Information Retrieval (ISMIR), Barcelona, Spain, October 2004, pp. 357–364 (2004)

[113] Terhardt, E.: Pitch, Consonance and Harmony. Journal of the Acoustical Society of America (JASA) 55(5), 1061–1069 (1974)

[114] Terhardt, E.: Pitch of Complex Signals According to Virtual-Pitch Theory: Tests, Examples, and Predictions. Journal of the Acoustical Society of America (JASA) 71(3), 671–678 (1982)

[115] Tsai, W.H., Wang, H.M., Rodgers, D., Cheng, S.S., Yu, H.M.: Blind Clustering of Popular Music Recordings Based on Singer Voice Characteristics. In: Proc. 4th International Symposium of Music Information Retrieval (ISMIR), Baltimore, Maryland, USA, October 26-30 (2003)

[116] Typke, R., Veltkamp, R.C., Wiering, F.: Searching Notated Polyphonic Music Using Transportation Distances. In: Proc. International ACM Conference on Multimedia (ACM MM), New York, USA, October 10-16 (2004)

[117] Tzanetakis, G., Cook, P.: Music Genre Classification of Audio Signals. IEEE Transactions on Speech and Audio Processing 10(5), 293–302 (2002)

[118] Tzanetakis, G.: Song-Specific Bootstrapping of Singing Voice Structure. In: Proc. of IEEE International Conference on Multimedia and Expo. (ICME), Taipei, Taiwan, June 27-30 (2004)

[119] Uhle, C., Herre, J.: Estimation of Tempo, MicroTime and Time Signature from Percussive Music. In: Proc. of the 6th International Conference on Digital Audio Effects (DAFX 2003), London, UK, September 8-11 (2003)

[120] Wang, Y., Vilermo, M.: A Compressed Domain Beat Detection Using MP3 Audio Bitstreams. In: Proc. 9th ACM International Conference on Multimedia (ACM MM), Ottawa, Ontario, Canada, September 30 - October 5 (2001)

[121] Wang, Y., Ahmaniemi, A., Isherwood, D., Huang, W.: Content –Based UEP: A New Scheme for Packet Loss Recovery in Music Streaming. In: Proc. ACM International conference on Multimedia (ACM MM), Berkeley, CA, USA, November 2-8 (2003)

[122] Wang, Y., Kan, M.Y., Nwe, T.L., Shenoy, A., Yin, J.: LyricAlly: Automatic Synchronization of Acoustic Music Signals and Textual Lyrics. In: Proc. International ACM Conference on Multimedia (ACM MM), New York, USA, October 10-16 (2004)

[123] Wang, Y., Huang, W., Korhonen, J.: A Framework for Robust and Scalable Audio Streaming. In: Proc. International ACM Conference on Multimedia (ACM MM), New York, USA, October 10-16 (2004)

[124] Wah, B.W., Su, X., Lin, D.: A Survey of Error-Concealment Schemes for Real-Time Audio and Video Transmission over the Internet. In: IEEE International Symposium on Multimedia Software Engineering, Taipei, Taiwan, December 2000, pp. 17–24 (2000)

[125] Ward, W.: Subjective Musical Pitch. Journal of the Acoustical Society of America (JASA) 26, 369–380 (1954)

[126] Wyse, L., Wang, Y., Zhu, X.: Application of a Content-Based Percussive Sound Synthesizer to Packet Loss Recovery in Music Streaming. In: Proc. ACM International conference on Multimedia (ACM MM), Berkeley, CA, USA, November 2-8 (2003)

[127] Xi, S., Xu, C.S., Kankanhalli, M.S.: Unsupervised Classification of Music Genre Using Hidden Markov Model. In: Proc. of IEEE International Conference on Multimedia and Expo. (ICME), Taipei, Taiwan, June 27-30 (2004)

[128] Xi, S., Maddage, N.C., Xu, C.S., Kankanhalli, M.S.: Automatic music summarization based on music structure analysis. In: Proc. Acoustics, Speech, and Signal Processing (2005)

[129] Xu, C., Zhu, Y., Tian, Q.: Automatic Music Summarization Based on Temporal, Spectral and Cepstral Features. In: Proc. IEEE International Conference on Multimedia and Expo., Lausanne, Switzerland, August 26-29, pp. 117–120 (2002)

[130] Xu, C.S., Maddage, N.C., Shao, X., Cao, F., Tian, Q.: Musical Genre Classification Using Support Vector Machines. In: Proc. International Conference on Acoustics, Speech, and Signal processing (ICASSP), pp. V429–V432 (2003)

[131] Xu, C.S., Maddage, N.C., Shao, X.: Automatic Music Classification and Summarization. IEEE Transaction on Speech and Audio Processing 13, 441–450 (2005)

[132] Xu, C.S., Maddage, N.C., Shao, X., Qi, T.: Content-Adaptive Digital Music Watermarking based on Music Structure Analysis. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP) 3(1) (2007)

[133] Yoshioka, T., Kitahara, T., Komatani, K., Ogata, T., Okuna, H.G.: Automatic Chord Transcription with Concurrent Recognition of Chord Symbols and Boundaries. In: Proc. of 5th International Symposium/Conference of Music Information Retrieval (ISMIR), Barcelona, Spain, October 10-15 (2004)

[134] Zhu, Y.: Content-Based Music Retrieval by Acoustic Query. Ph.D. dissertation, Department of Computer Science, National University of Singapore (October 2004)

[135] Zhu, Y., Kankanhalli, M.S., Gao, S.: Music Key Detection for Musical Audio. In: Proc. 11th International Multimedia Modelling Conference (MMM), Melbourne, Australia, January 12-14 (2005)

[136] Zhang, T., Kuo, C.C.J.: Audio Content Analysis for Online Audiovisual Data Segmentation and Classification. IEEE Transaction on Speech and Audio Processing 9(4), 441–457 (2001)

Examining the Theoretical Schema of *Shimpai Muyou!* Narrative Learning Environment

Samiullah Paracha, Sania Jehanzeb, and Osamu Yoshie

Abstract. The chapter highlights the problem between narrativity and interactivity in multimedia or virtual environments, motivates the need of Human-Computer Interface design advice and then introduces combine knowledge of theatre, education and technology that will be used in the design process. The survey of Classical and Structuralist or Formalist narrative approaches for potential computer implementation reveals the mere focus on one aspect of narrative whilst simply neglecting the other. The question of balancing fabula and interactivity, within a virtual environment, is seemingly unresolved, despite the existing overhauled narrative models based on those theories. The chapter discussed the broad conceptual framework in which *Shimpai Muyou!* virtual learning environment shall be developed and explores how such a narrative learning environment facilitates Information Extraction and Emergent Narrative mechanisms. It hosts concepts from a vast variety of sources, as for instance, the learning theories of constructivism and constructionism from the realm of psychology, are at the core of our work. Likewise, Boal's Forum Theatre, theories of Narrative Role Playing and Drama in Education serve as raison d'être of this study. All of these approaches complete the embroidered theoretical framework of *Shimpai Muyou!* the potential of which lies in delivering a micro world to the learners where they can safely explore socially sensitive issues like *Ijime*-bullying, drug addiction and various delinquency matters.

Keywords: Multimedia, Narrative learning Environment, Empathy, Fabula.

1 Introduction

Virtual Reality (VR) is a combination of technologies whose objective is to integrate real and synthetic sensory signals and present them to the user in such a way as to evoke the sensation of being immersed in a virtual environment. Multimedia

Samiullah Paracha
Board of Education, Yufu City Hall 2F, Oita 879-5192, Japan
e-mail: samiullah.paracha@gmail.com

Sania Jehanzeb
Ritsumeikan Asia Pacific University, 1-1 Jumonjibaru, Beppu City, Oita 874-8577, Japan

Osamu Yoshie
IPS, Waseda University, Kitakyushu City, Fukuoka 808-0135, Japan

M. Grgic et al. (Eds.): Rec. Advan. in Mult. Sig. Process. and Commun., SCI 231, pp. 579–609.
springerlink.com

and VR are treated as separate fields, but this is a false distinction (Sutcliffe 2003). Within the applications of virtual reality and multimedia lies a wealth of confusing or even contradictory terminology that is inherent to any young technology. VR systems would be better described as virtual environment systems, since reality is often difficult to simulate, and it's usually better to simulate something other than "reality." Multimedia systems on the other hand would be better described specifically as what kinds of media are involved in the particular application. VR expands the visual aspects of multimedia and both fields are similar in that, because they are at root *multiperceptual, real-time interactive communication media* (Begault 1994).

The defining characteristic of multimedia or VR systems is not their visual realism, but their interactivity. The process of balancing authorial input and user freedom within a Virtual Environment (VE) can be a pursuit that is simultaneously mind-boggling and frustrating due to the complexity of the problem and the fact that humans, who inevitably mess up one's best-laid plans, are involved. However, solutions to these problems are crucial to the success of VR or multimedia systems. Any effort to create VE has always been motivated by some theories that provide guideline about the design of user presence, social agents, multimedia presentation and user support. These principles provide the basis around which entire edifice of algorithmic choices for computer simulation can be erected. We present a brief overview of various narrative approaches including classical, structuralist and formalist, in relation to the recent works produced by the AI Community and its flimsiness. In our view, the source of the difficulty is that they often seek after types of cognitive architectures, kinds of representations, and methods of inference that are based on some single simple process, theory, or principle. Despite their elegance, the authorial nature restricts the amount of freedom allowed by virtual environments to users. In addition, no single one of such approaches can capture the diversity of mechanisms needed to balance fabula and interactivity. So rather, seeking a unified theory, we seek instead to develop an architecture based on a multidisciplinary study that derives concepts mainly from drama, psychology and computer science.

In this chapter, we briefly describe the conceptual background and relevant research to support the construction of *Shimpai Muyou!* narrative learning environment which is under construction. It is a 3D state of the art educational intervention to counteract persistently proliferating violent behaviour (*Ijime*) in the Japanese schools. It is hoped that the Virtual Learning Environment (VLE), through empathic interaction with Autonomous Virtual Actors (AVAs), will allow children to explore *Ijime* problem through coping mechanisms in a safe and sound environment.

2 Context

The digital revolution is transforming every aspect of children's lives and it would not be unfair to assume that the new generation of children will be doing almost everything with computers (Tapscott 1998), including robotic toys (Druin and Hendler 2000). Persuasive computing is a kind of education technology which is deliberately designed to influence one's attitude or behaviour in a premeditated way of which VLEs are important part albeit, of limited scope as yet. Although there has been a widespread use of VLEs in practice to support exploration of history, heritage, field trips and creative writing yet they are deficient of providing any active engagement to their users and merely limited to passive exploration of the subject matter. It would be unfair to say that persuasive or experiential mode of learning has yet to be explored by this new medium.

Since the emphasis of persuasive education is basically to change attitudes, issues of emotional involvement and empathy are very prominent. And that, of course, is the reason of narrative role-playing has been widely adopted in today's learning activities e.g. Theatre-in-Education (TIE), puppet shows, cartoons etc (Paracha and Yoshie 2008-*a*). TIE targets persuasive and emotional learning, but schools do not have open access to TIE companies. Furthermore, the collective environment of theatre may constrain the individual child from exploring personally sensitive issues such as drug addiction, bullying, street-violence etc.

In order to contextualize our approach, we shall use cases from *Shimpai Muyou!* project that aims to apply synthetic characters and Emergent Narrative (EN) for children aged eight to twelve. The project domain is *Ijime*-bullying and empathy building with the user which will be a novel learning experience and a major educational requirement. The final product will comprise of a 3D real-time virtual environment where children are faced with virtual *Ijime* situations inspired by what really goes on in the Japanese schools. Each session with the child user will be composed of a series episodes and each segment will portray certain dramatic situation in *Ijime*-bullying context (e.g., a character being hit and then teased by a bully and a set of other children/characters). The episodes will be followed by intervals in which the user will intervene, evaluate the situation and suggest a possible course of action for the victim, thus, influencing what will happen next in the drama.

Our hypothesis is that by intervening through *Shimpai Muyou!* VLE, we shall be able to establish connection with children's own spontaneous ability as narrators and learners. Resultantly, it shall help them to develop and practice their conflict resolution and argumentation skills. It is urgently needed for the Japanese children who do not have these abilities well developed therefore, opt for suicide as a way out in situations of severe *Ijime*-bullying (Paracha et al. 2008-*b*). The Japanese public education is, unfortunately, devoid of wit and inventiveness, required for the holistic child development. Therefore, we want to target social and emotional learning of child education through our Narrative Learning Environment (NLE) on *Ijime*-domain to give them voice and energy to counteract the oppression that surrounds them.

3 Annotated Review of Literature

This section consists of works that have contributed to the development of my vision for *Shimpai Muyou*!. 'Creative Drama in the Classroom and Beyond' of Nellie McCaslin 1996, presents strong rationale for the use of creative drama in curriculum and many lessons, game and activity ideas for creative drama with children. It has helped me to understand the nature of a creative drama and its profoundness in the classroom setting. However, *Shimpai Muyou!* takes its inspiration mainly from Augusto Boal's Theatre of the Oppressed (TO), 1983 and Games for actors and non-actors, 1992. Undoubtedly, this had a tremendous impact on our ongoing work in its various manifestations. The reading of these two books has given me the idea that if Japanese children shall be given opportunity to create their own plays and present them as Forum Theater (FT), the learning would be more complete and deep because that would be really learning by practice.

Likewise, the masters thesis of David Shaw 2004, titled 'Aspects of Interactive Storytelling Systems', is indeed an inspiring and useful resource, for those who are interested to work towards solving the problem of integrating story with interactivity. He has surveyed a range of storytelling models and argued for a balance approach in Interactive Storytelling (IS) where the audience is an active participant taking the role of the protagonist in the story. David Shaw, 2004 pleaded for truly interactive storytelling environments must allow their audience the opportunity to shape the path of the story through their actions. A linear pre-written story cannot provide this level of freedom, as the story acts as a constraint to the audience's ability to choose their own path. In order to resolve the long standing conflict between the script writers or designers and the programmers of virtual environment, he used the metaphor of doors and keys to represent plot challenges. He has presented a prototype model to produce abstract story environments that would suit an interactive environment. David Shaw, 2004 asserts by quoting 'door and keys' that how his model can be used to generate plots for interactive stories.

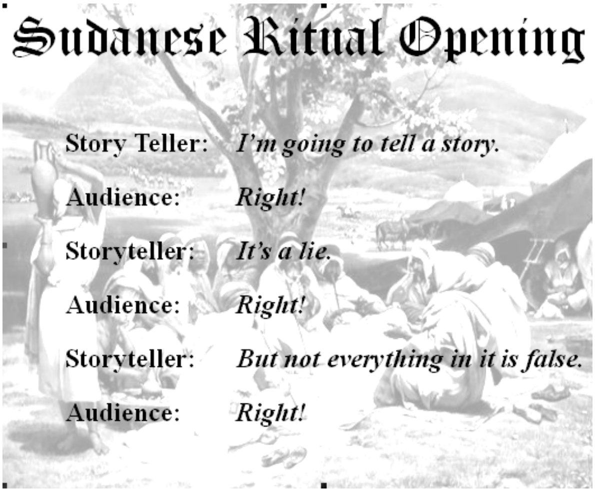
Another remarkable contribution that can be cited as a catalyst and an inspiring guide for my vision is the interdisciplinary doctoral dissertation of Alice C. Mello Cavallo 2008, titled 'Virtual Forum Theater: Creating and sharing drama to resolve conflicts'. Based upon Bertold Brecht's Epic Theater (1964) and Augusto Boal's TO (1983), she has developed a Virtual Forum Theatre (VFT) which is intended to provide a safe environment in which children can explore and react to injustice, oppression and conflict. The program consists of a face and storyboard editors, chat environment, and a media player combining development in Java and existing free-software. It is the first educational software that facilitates resolution of conflicts between youth through free on-line theater interactions.

In addition to all the foregoing, I went through VICTEC's complete listing (http://www.macs.hw.ac.uk/~ruth/pubs.html#narrative) on EN concept aiming to review narrative approaches and theories in an effort to assess their potential as suitable models for computational implementation. The website is a rich source of learning for those who want to investigate the major schools of narrative. Furthermore, it provides in-depth discussion on classical narrative theories and alternative interactive models according to the narrative requirements presented by VICTEC. The EN concept (Aylett 2006) is also defined and referred as an essential element of the VICTEC research project. The EN research carried out by VICTEC and particularly regarding the conditions and sources for the emergence of character based interactive narratives, highlighted the need for *Shimpai Muyou!* to consider similar use of AVAs and narrative approach.

4 Classical, Structuralist and Formalist Approaches

4.1 Plato

The traditional storytelling had been an effective medium for transmitting literature and culture from people to people before the invention of writing (Shaw, 2004). Lord AB 1960 noticed that the *Homeric Poems* (9th Century B.C) which include the classic *Iliad* and *Odyssey* have their origin in oral source. Plato (428 B.C) developed Diegesi*s* Narrative Model (DNM) which relates to the oral storytelling. In this form of narrative, the author directly addresses the audience (as obvious in Fig.1) and it involves pure narrative. Diegetic narrative occurs when the poet narrates a narrative as himself, and do not assign a character to it. The author may include elements which are not intended for the primary narrative, such as stories within stories; characters and events that may be referred to elsewhere or in historical contexts and that are outside the main story and are thus presented in an *extra-diegetic* situation. Despite its importance, mostly in academics and films, the DNM has been confuted by AI community on the grounds that it is purely authorial and interactivity does not fit properly (Paracha et al. 2008-*c*).



Sudanese Ritual Opening

Story Teller: *I'm going to tell a story.*

Audience: *Right!*

Storyteller: *It's a lie.*

Audience: *Right!*

Storyteller: *But not everything in it is false.*

Audience: *Right!*

Fig. 1 A simple representation of Plato's *Diegesis* Narrative Concept (Livo and Rietz 1986)

4.2 Aristotle

The first high-level narrative concept is said to be developed by Aristotle (Aristotle, 1998) that became the basis for understanding the interactive narrative. His main focus is to explain the nature of tragic drama, but he also refers to other art forms e.g., epic poetry, comedy, dithyrambic poetry, music, dancing, and painting. He claims that all these forms of "imitation" differ from each with respect to their objects, medium, and manner (Tomaszewski and Binsted, 2006). Aristotle classified drama action into six components (i.e. Plot, Character, Thought, Diction, Song and Spectacle) in respect to its significance to the tragedy order. It was he, who stated that action and behaviour can be portrayed through characters. Aristotle considers plot (Muthos) and characters (Mimesis) as the most crucial components of the narrative order.

One of Aristotle's fundamental ideas about drama (as well as other forms of literature) is that a finished play is an *organic whole.* He used the term *organic* to evoke an analogy with living things, insofar as a whole organism is more than the sum of its parts, all of the parts are necessary for life, and the parts have certain necessary relationships to one another. He identified six qualitative elements of drama and suggested the relationships among them in terms of formal and material causality. Aristotle's model is crucial because of its elegance and robustness of the categories and their causal relations. Following the causal relations through as one creates or analyzes a drama seems to automatically reveal the ways in which things should work or exactly how they have gone awry. Furthermore, it creates a disciplined way of thinking about the design of a play in both constructing and debugging activities. Because of its fundamental similarities to drama, human-computer activity can be described with a similar model, with equal utility in both design and analysis (Laurel, 1991).

4.3 Smiley

Sam Smiley 1971, studied the early spadework carried out in narratology by Aristotle and introduced the notion of formal and material causes between object(s), medium and manner. Aristotle's model provides the foundation for describing an art form in terms of its object(s), medium, and manner that has come a long way through different modifications. Aristotle listed the manner of Spectacle as the least essential to examine tragedy, whereas, Sam Smiley 1971 considered it as the integral component in narrativity.

4.4 Laurel

Laurel's work 1991 is actually an extension of Smiley's contributions to narratology that describes the framework of interactive narrative. She expounds by renaming some of the items in Smiley's model i.e. Action, Character, Thought, Language, Melody, and Spectacle and gives broader meaning to these elements in interactive narrative context (Fig. 2). The most powerful idea involved in her approach is that the computer can be studied from a rigorous humanistic perspective, using well-defined models established for other forms of art. This is part of the reason that Laurel recommends a thorough understanding of the principles being appropriated and applied, and names the *Poetics* an essential text for students of human-computer interaction. Brenda Laurel 1991 asserts that technologies offer new opportunities for creative, interactive experiences, and in particular, for new forms of drama. But these new opportunities will come to pass only if control of the technology is taken away from the technologist and given to those who understand human beings, human interaction, communication, pleasure, and pain.

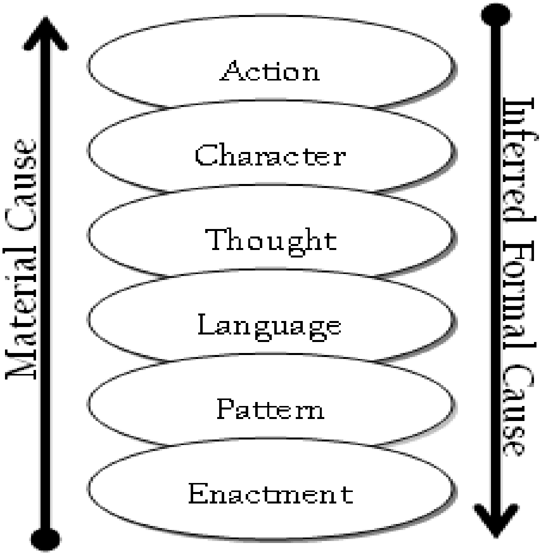


Fig. 2 Causal relations among elements of quantitative structure

4.5 Mateas

Michael Mateas 2001 endorses the position taken by Brenda Laurel 1991 however he adds the concept of Agency given by Murray which he defines as ‘the feeling of empowerment that comes from being able to take actions in the world whose effects relate to the player's intention (Fig. 3). In a role-playing drama, the story

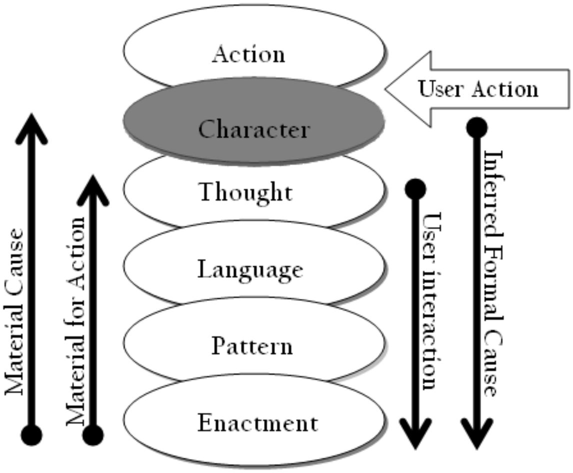


Fig. 3 Mateas's Model

emerges from the player's interaction which he represents with two new causal chains:1) Material for Action and 2) User Intention. According to Mateas 2001, during interaction process, the immersive environment supports certain action taken by the user from below successively the narrative provides certain authorial restrictions from the top. The user's interaction with other characters in the narrative is represented by a formal cause and agency can be achieved only when there is balance between the material and formal constraints. In short, what he wants to achieve is a common platform for interaction and the experience of agency to co-exist and grow together.

4.6 Reidl

Another significant contribution in this direction is the acyclic branching story structure developed by Riedl and Young 2006 that establish narrative consistency within an interactive narrative environment. Since, user intervention can influence the way in which a narrative unfolds, NLEs more often use a branching story structure where authorial input is interleaved with user judgment. Narrative mediation can be adopted as an alternative which unveils 'story as a linear progression of events with anticipated user actions and system-controlled agent actions together in a partially-ordered plan'. Nevertheless, the possibility of violating the story plan through user intervention and thus generating the need of an alternative story plan is always present. Riedl and Young 2006 pleaded that 'if it is powerful enough to express the same interactive stories as systems that use branching story structures, then linear narrative generation techniques can be applied to interactive narrative generation with the use of narrative mediation'. Riedl sketches out a proof that narrative mediation is at least as powerful as acyclic branching story structures (Fig. 4).

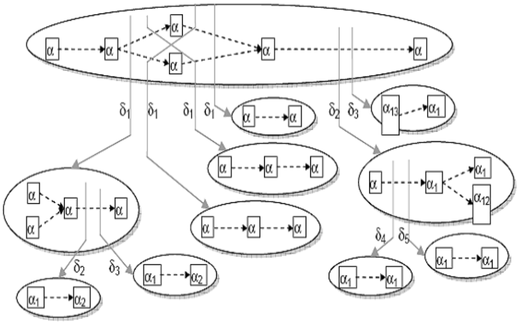


Fig. 4 A Narrative Mediation Tree

4.7 Propp

The first rule-based story generation approach that attracted the interest of AI Community was the classification system for Russian folktales developed by the

Russian folklorist Vladimir (Propp 1968). According to Propp, the structure of the plot is the most vital component of any folktale. The presence of characters is not a mandatory condition as the plot elements and the event-sequence. He defined these elements as “functions” and put forward thirty one functions which help classifying the structure of Russian folktales. Propp’s work greatly contributed to the understanding of plot structure thus, provided preliminary spadework to TEATRIX virtual story creation environment developed by Prada, Machado and Paiva, 2000. Based on DIE techniques (Heathcote 1984), constructionist approach (Papert 1990) and Propp’s list of functions, TEATRIX explores drama as a form of child’s education and development. The designers emphasized that in order to achieve sense of immersion and control of the dramatic plot, the participants should have the freedom to examine and modify the character’s mind. The NLE thus, allows children to freeze the action of a character, evaluate the role, goals, and previous actions of their avatars, reflect upon them and introduce necessary modifications.

4.8 Todorov

The French structuralist Todorov 1969, coined the term "narratology" for the structuralist analysis of any given narrative into its constituent parts to determine their function(s) and relationships. For these purposes, the story is *what* is narrated as usually a chronological sequence of themes, motives and plot lines; hence, the plot represents the logical and causal structure of a story, explaining why its events occur. The term *discourse* is used to describe the stylistic choices that determine *how* the narrative text or performance finally appears to the audience. One of the stylistic decisions may be to present events in non-chronological order, using flashbacks, for example, to reveal motivations at a dramatic moment.

Todorov developed plot relapses in algebric form, identifying and distinguishing narrative noun-subject (characters), narrative adjectives (situations) and narrative predicates (actions) (Louchart and Aylett, 2004). Todorov considers plot the actual focus of structural examination and describes it a movement from one state of equilibrium through a state of disequilibrium to a final state of equilibrium that is similar to, but not the same as, the first state of equilibrium. Prior to his work, narrative was examined in relation to theme and rhetoric i.e. script or text and the diction deployed to achieve that end. In contrast, Todorov proposed an examination that gives high priority to narrative syntax. The motivation was to understand the functions of plot in general i.e. *langue* underlying all plots and to differentiate between various types of plot i.e. *paroles* of this substratum langue.

4.9 Campbell

Joseph Campbell’s work in 1949 on the generalized pattern of narrative structure is another worth mentioning contribution. In his work, *The Hero with a Thousand Faces*, Campbell discovered the similarities amongst the hero tales across the world. He argued that the ‘*Journey of the Hero*’ stories follow identical patterns and structure. Fig. 6 summarizes the mythic hero’s adventures and identifies elements which can be found in cultures from all around the world and through all recorded time. The overall structure (Fig. 5) can be verbalized as a troika: “a separation from the world, a penetration to some source of power and a life-enhancing return” (Campbell 1949). The empirical basis of Campbell’s study is never stated explicitly nevertheless, his *Cyclical Diagram of Plot* has been highly appreciated by the script writers and designers of computer game and film industries.

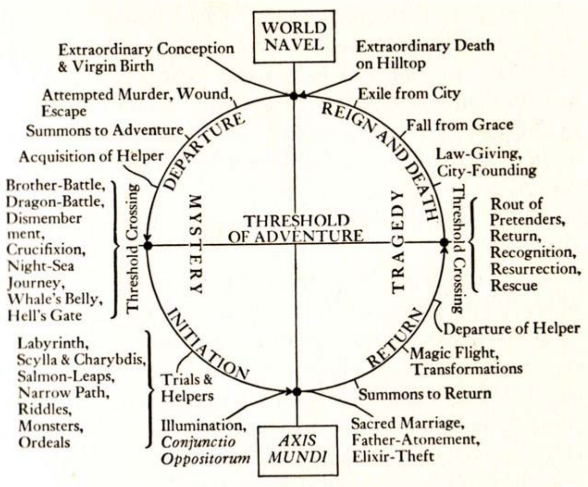

WORLD NAVEL
Extraordinary Conception & Virgin Birth
Attempted Murder, Wound, Escape
Summons to Adventure
Acquisition of Helper
Brother-Battle, Dragon-Battle, Dismemberment, Crucifixion, Night-Sea Journey, Whale's Belly, Hell's Gate
Threshold Crossing
DEPARTURE
MYSTERY
THRESHOLD OF ADVENTURE
TRAGEDY
REIGN AND DEATH
Extraordinary Death on Hilltop
Exile from City
Fall from Grace
Law-Giving, City-Founding
Threshold Crossing
Rout of Pretenders, Return, Recognition, Resurrection, Rescue
Departure of Helper
RETURN
Magic Flight, Transformations
Summons to Return
Sacred Marriage, Father-Atonement, Elixir-Theft
INITIATION
Labyrinth, Scylla & Charybdis, Salmon-Leaps, Narrow Path, Riddles, Monsters, Ordeals
Trials & Helpers
Illumination, Conjunctio Oppositorum
AXIS MUNDI

Fig. 5 Campbell’s Monomyth

5 Omissions and Tensions

Aristotle's *Poetics* has served as something rudimentary which has gone through many adjustments and re-adjustments, simply to acquire interactive narrative status. Propp’s list of functions on the other hand, is merely a focus on Russian folktale which is a limited story domain. Furthermore, TEATRIX that heavily draws on Propp’s approach although gives due consideration to user participation albeit, scenarios have to be pre-defined by the user before starting the drama. This, not only limits the chances of narrative to emerge through interaction, but also reduce the possibilities of meeting unexpected events in the story which can increase the immersion, dramatic and emotional effects as well as it can also help in achieving the resulted educational outcome (Hodhod and Kudenko 2007).

The analysis of classical, structuralist or formalist narrative approaches from Plato’s story conception and Aristotle’s plot emphasis, all the way down to Propp’s meta-structural narrative articulation, and Campbell’s cyclical diagram reveals their multi-authorial nature. Notwithstanding, the plentiful overhauled models that have been introduced, the problem of balancing fabula and interactivity in a virtual environment still remains subject to other thought. These approaches, no single one of these, can capture the diversity of mechanisms needed to balance fabula and interactivity, within a virtual learning environment. The authorial position does not allow narrative to emerge through interaction (Table 1).

Table 1 Laurel 1991 listed the six qualitative elements of structure in drama and in human-computer activity

Element	In Drama	In Human-Computer Activity
Action	The whole action being represented. The action is theoretically the same in every performance.	The whole action, as it is collaboratively shaped by system and user. The action may vary in each interactive session.
Character	Bundles of predispositions and traits, inferred from agents' patterns of choice.	The same as in drama, but including agents of both human and computer origin.
Thought	Inferred internal processes leading to choice:cognition, emotion, and reason.	The same as in drama, but including processes of both human and computer origin.
Language	The selection and arrangement of words; the use of language.	The selection and arrangement of signs, including verbal, visual, auditory, and other nonverbal phenomena when used semiotically.
Melody (Pattern)	Everything that is heard, but especially the melody of speech.	The pleasurable perception of pattern in sensory phenomena.
Spectacle (Enactment)	Everything that is seen.	The sensory dimensions of the action being represented: visual, auditory, kinesthetic and tactile, and potentially all others.

6 *Shimpai Muyou!* Conceptual Framework

In the sections that follow, we introduce the theoretical framework in which *Shimpai Muyou!* research shall be designed. We investigate the existing theories of constructivism (Piaget, 1977), constructionism (Papert, 1990), Boal's TO (1983) which are at the core of our study. Constructivism expounds the learning expectations from various age-groups and how these groups raise their reasoning caliber. It also describes the importance of team work and collaborative activities in resolving different problems through peer's feedback (Vygotsky, 1978). Constructionism regards information technology a learning tool and recommends ways for its positive use and growth.

Forum Theater (FT) presents a more solid constructivist learning process as compared to ordinary theatrical performance. Boal's Theatre of the Oppressed (TO) is an extension of Freire's Pedagogy of the Oppressed (Freire, 1972). He has used participatory theatre to develop awareness through modeling real-world situations and role-playing potential solutions (Cavallo and Couch 2004) in precisely the same way, as Freire, 1972 recognized the need of assisting participants to develop literacy and to become watchful critics of their surroundings. A few other perspectives have also been visited describing other academic work related to this research and finally we discuss how this work will influence the design of our NLE. Fig. 6 indicates the intersection of worlds and views that will shape *Shimpai Muyou!* as an educational tool.

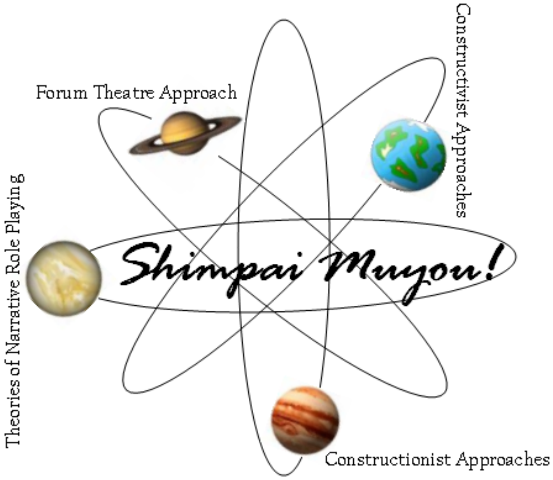

Forum Theatre Approach
Constructivist Approaches
Theories of Narrative Role Playing
Shimpai Muyou!
Constructionist Approaches

Fig. 6 *Shimpai Muyou!* Universe

6.1 The Constructivist and Constructionist Approaches

The learning aspect of *Shimpai Muyou!* has been derived from the constructivist and constructionist approaches. Learning theorists, developmental psychologists and pedagogues such as Piaget (1977), Vygotsky (1971), Papert (1990), Dewey (1938), Freire (1972), Duckworth (1987), Gardner (1973) and others from the open school movement give us insight into how to reshape education, create new environments, and put new tools, media, and technologies at the service of the growing child. Accordingly, learning should not be taken as an inert process of merely receiving information or accepting others' views and believes rather it is about holding one's own opinion and having the confidence to present and share it with others (Cavallo, 2008).

Constructivists believe that people are not just passive receptacle of information, but on the contrary, they build their own viewpoint based on their past experiences. According to Piaget 1977, knowledge is “to understand is to invent” i.e. experience acquired through interaction with the environment and not just a segment of information which is to be retrieved, encoded, memorized and later applied. The Constructionists build their arguments about learning and knowledge on constructivism. However, they argue that proper learning conditions are those when a person is deliberately engaged in constructing a public entity, whether it is a sand castle on the beach or a theory of the universe (Papert, 1990). They sketch out the entire learning process as idea formation, its transformation and transmission through media in a particular context. According to Papert, digital media and computer-based technology are objects with which to think, and extensions of one's construction artifacts; due to their inherent malleability to adapt both to the cognitive style of the learner as well as to the domain being investigated (Cavallo and Couch 2004).

Based on the above conceptual paradigms, expounded by the constructivist and constructionist learning theorists, we shall investigate how to sensitize children of the problems and dangers of *Ijime*-bullying; in particular how to augment their talents as mediators, actors and sympathizers through the use of technological theatre. VR may offer a profound learning environment in both constructionist and constructivist settings, because we want to investigate how effectively participatory theatre play its role in children's social and emotional learning. The results will greatly contribute to the development of progressive learning thus, empowering educational applications with positive bearing about cognitive learning. The *Shimpai Muyou!* virtual actors shall thus, be designed as an object to act with or upon. We shall study the impact of virtual environment on *Ijime* domain on children's learning and cognition processes and on their interaction with AVAs.

The prototype will be tested on primary schools children under Yufu City Board of Education to gather feedback from children and teachers about the character interactions; the design of the graphics; the design of AI responses; the animation of objects and characters. As children are the intended users of the application, it is important to incorporate their views, expectations and perspectives within the design process. Evaluation of the prototype shall be carried out intensively by applying both the Classroom Discussion Forum (CDF) technique and longitudinal and large-scale studies. Questionnaires shall be designed to assess general cognitive and affective empathy reactions (Paracha and Yoshie 2008-*e*). Two types of questionnaires will be distributed among the participants:

- Pre-test: Prior to interaction with the software.
- Post-test: Post interaction with the software.

The evaluation will employ three research methods i.e. traditional, observational and interview. These approaches will focus on the emotional and empathic aspects of learning process through the use of interactive 3D environment. Direct observation conducted by video taping (digital camera) will be used to collect the data. In order to record all of the user's interactions with the agents and the VE log files will be made use of, which will be synchronized for examination by a time code. The examination of child's interaction with the VE aims to determine the ease of use through direct observation during software use, questionnaires and facial expression examination. User satisfaction will be determined through focus groups' direct observation on user friendliness, difficulties with the software and the kind of choices user selects etc.

Likewise, users' emotional state and empathy towards synthetic characters will be determined through videotaping and by asking simple questions from children during the application use: whether or not the participants experienced fear, anger, joy, sadness etc. at particular moments whilst interacting with the Intelligent

Virtual Agents (IVAs). The feedback obtained through CDF will be utilized for the development of final version. The most significant aspect of this evaluation phase will be in determining the impact of this interaction on child's social behavior real life. This would be possible by putting into play the investigative pre-post test design. The user will be asked, a week before interaction with the prototype, about how they deal when confronted with street violence, *Ijime*, drugs etc. The children will be re-evaluated a week later, the engagement with our interactive narrative environment, to determine how far they have implemented those strategies in their lives i.e deterrent to victimization or when helping others. Comparison of "Pre- vs- Post Interview" would give us an insight into how far these options explored in the virtual space would have been utilized in the real situations (Paracha et al. 2009-*d*).

6.2 Forum Theatre Approach

Shimpai Muyou! virtual learning environment is based upon the work of Augusto Boal, a Brazilian theatre director, author, activist, teacher, and politician. Boal considers theatre a coin whose two sides are 'learning' and 'entertainment'. Theatre for that reason should be recognized as a learning environment for the audience and actors to seek knowledge as well as to entertain themselves. It is a learning platform which the modern day audience requires for its knowledge building. Boal's modification to traditional static theatre is actually a reaction to Aristotle's narrative approach and a movement to overhaul it. The Aristotelian play essentially portrays the world as quiescent. It reflects various events of life as something inevitable or necessary fate to which a particular person or thing is destined and as such beyond human power. Boal has pointed out various flaws in the classical narrative framework and regards it flaccid for social transformation. Aristotelian catharsis, in his opinion, takes away motivation from the stage and the audience necessary to induce change or to succeed in causing events to occur in certain way. Theater is therefore, not to paralyze people from acting, but instruct them to change what is not working for them. It is a safe space for rehearsing possible alternatives of action to be implemented in real life (Taussig and Schechner, 1994).

The Forum Theatre (FT) is a dynamic approach in the sense that it depicts the world as something changeable and describes ways of how it can be reversed. He has achieved more than this, in fact, he has generated practicality and vigor in his audience, so that the participants should not conceive a crisis or conflict as something unchangeable or beyond their control. Boal's concept of theatre inspires audience with motivation to intervene actively against conflicts and successfully alter their course. Theatre of the Oppressed (TO) is platform through which we can create drama based haunting issues which really matters to us e.g. oppression and inequalities. Boal's technique allows the spectators to stop the play when conflict arises or when they disagree with the scenarios performed on the stage. The spectator might go on stage and re-enact the piece or explain to the actor what should be done. It is a kind of learning environment where actors and spectators participate together to resolve real life issues or conflicts in an entertaining way. Spectators are encouraged to become ***spect-actors*** i.e. active participants who rehearse strategies for change (Cavallo and Couch 2004).

Forum Theatre is a collaborative effort of actors and spectators to create drama on hot political or community related issues. The beauty of this technique is that no textual details are required to be adhered throughout the length and breadth of the drama. But, it does not mean that actors do not need to rehearse the script that contains broad lines about possible dynamics of the play. The dramatic enactment lasts for five to ten minutes in which dramatists and actors together with the participants attempt to bring change in the social status quo by allowing disequilibrium that paves the way for action. Boal created the figure of the "joker" (Boal 1983) to lead the audience and actors who can be a Drama Manager (DM), a community leader, or group member having some experience in TO. The joker or DM keeps workable relationship with both the spectators and actors to lead rehearsals and works towards attainment of the dramatic goals. The demonstration takes place in front of "spect-actors", who are only interested in the subject of the play that is to bring change. They hope to open an avenue for any possible solution to the conflict being performed on the stage.

The DM invites the audience to interrupt and give advice at times of conflict. The facilitator sometimes indicates this moment of crisis and facilitates the "spect-actors" interventions, opinions and interactions. According to Boal (1992), the joker should decide together with the audience if the intervention works i.e resolves the conflict, or inadequate. The drama is performed once and the DM calls the participants to split into groups of four to give direction to the play. They are expected of presenting solutions to matters of conflict as described in the text of the play. The drama is re-played with the "spect-actor" who can stop the play, goes to the stage, and takes the role of protagonist or any other character. Furthermore, he can direct actors on how to portray a role and advise to defuse the conflict. Boal's intervention encourages audience to think, reflect and try out different options. The spectators are active participants who have the right to experience a dramatic situation and exercise their share to alter the course.

Another worth-mentioning aspect is the simplicity of stage, because Boal does not want to keep his audience in a state of illusion or fantasy. For instance, we observe in real life in times of crisis when it falls on the victim, the sky, the moon the birds etc become meaningless and the intensity of situation and its action towards settlement gets more important. In a similar manner, the play as expounded by Boal, is an 'object to think about' as compared to the sets, backgrounds, lightings, decorations and shades. He desires the "spect-actors" to be immersed totally in issue being presented and desires their attentions to be focused only on the subject matter of the drama should be their focus of attention (Boal, 1992). The participants' input, feedback, and intervention are the main features of FT. Therefore, the FT stage gives due importance to participation, roles and resolution at times of conflict rather stage decorations and outlooks.

As depicted in Fig.7, the author and the participants seem to write the play together in FT. The original drama-script can be modified through "on-the-spot feedback" obtained from the "spect-actors" interventions. Boal's idea behind intervention is to keep the room open for improvements in the script by incorporating participants' input through a continuous interruption system. The spectators should no longer delegate power to the actors to decide matters in their place instead, they

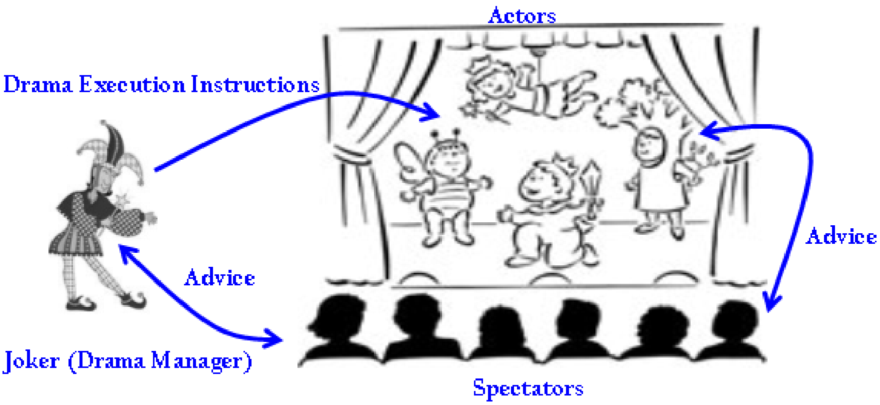


Fig. 7 A simple representation of Forum Theatre Approach

must think and act for themselves. He considers it an organizing tool which helps in bringing the participants together to look for resolution; a way out of oppression; to express their opinions and to liberate their voices (Cohen-Cruz, 1994).

Boal (1983) coined several terms for theater e.g. a "tool" or an "action" or a "rehearsal" to bring forth changes in the status quo. *Shimpai Muyou!* is an attempt in this direction to improve the quality of child's life in Japan by providing them with opportunity to try out options against *Ijime* other than committing suicides in the face of it. We believe that any particular study on narrativity has always tended to be limited in scope primarily, because it focused on one aspect of narrative and neglected the other. However, Boal's phenomenon combines several characteristics of narrativity and seemingly adhering to the principles of narrative i.e. characters, events and their association in the narrative context. These features will be practically manifested in *Shimpai Muyou!* NLE populated by IVAs.

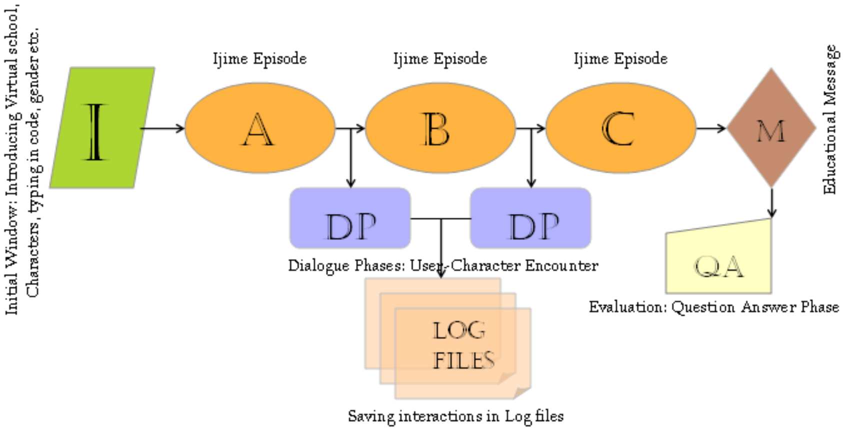

Initial Window: Introducing Virtual school, Characters, typing in code, gender etc.
I
Ijime Episode
A
Ijime Episode
B
Ijime Episode
C
M
Educational Message
DP
DP
Dialogue Phases: User-Character Encounter
QA
Evaluation: Question Answer Phase
LOG FILES
Saving interactions in Log files

Fig. 8 Interaction Sequence in *Shimpai Muyou!*

In the introductory phase of the virtual drama, the user meets his fictitious friend i.e. IVA. The user choice will play a central role in the selection of the character and should result in creating some sort of initial interface, where the character will tell its situation and ask for help. This will ensure an encouraging way of constructing empathic relations with the virtual actors. The user will then pursue the character as friend or supporter into a scenario. As a spectator the user has to watch various episodes on *Ijime* domain and during this time his encounters with the virtual actors will be restricted mainly to prevent IVAs from external manipulations. The system will allow the user to halt action a couple of times and take his virtual friend out of the scenario for advice (Fig. 8). It shall influence the behaviour of his favorite character in the coming scenarios. At the end of the episode, there will be a dialogue phase in which the virtual actor will communicate with the child user and seek his advice.

Within the initiated dialogue the user selects an advice from a list of coping options through a drop down menu. The user also explains why this specific option is a good choice in the given situation, by typing it in. The underlying idea is akin to the real life where advice is given to influence friend's ideas or actions. However, the IVAs are not bound to the advice given by user; rather operate autonomously to carry out particular action. This approach is in conformity with the theatrical genre of FT in which spectators take responsibility for characters in a drama and are able to interact with them in character between episodes of the drama. The rule here is that the character asks for help and the child-user then suggests one of a set of coping responses.

In *Shimpai Muyou!* the user will interact with the virtual actors or IVAs however, the episodes will be preceded by a non-interactive introduction about the characters and scenario (Fig.9). The system will allow interaction from third person perspective by a symbolic avatar as well as from first person perspective by a partly represented avatar (Paracha et al. 2009-*d*). The motives behind adopting such interaction sequence are: Firstly, this type of autonomy materializes our aim of providing realistic and credible immersive environment to the user without which the projection of empathic engagement would be unachievable. Secondly, the juvenile delinquency issues like *Ijime*, drug addiction, crimes etc are emotionally sensitive matters and role-plays on such domains in a classroom setting may physically hurt the user. But, in our case, if the user's avatar hits a virtual actor or *vice versā*, it would be in the virtual context and not in the actual sense.

Thus, we shall achieve our objective of providing safe and sound environment to the child user. The last but not least consideration is that absolute freedom in virtual context corrupts absolutely therefore, it would be more appropriate if we link the child user empathically with avatar which goes through delinquency situations rather than allowing the user to manipulate characters directly. It may hinder user's ability to contemplate motives and being carried away with the action may treat the environment merely as bullying game. The evidence of such type of user behaviour has been found in some systems which have been created with noble intention, but absolute freedom led to their devious manipulation at the hands of their users. TEATRIX can be quoted as one of these examples (Paracha et al. 2009-*d*).

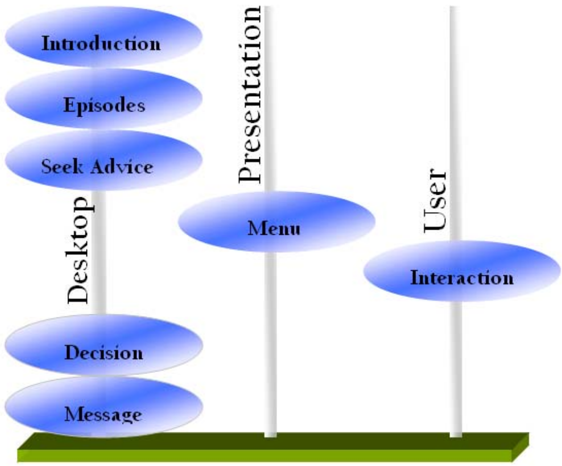


Fig. 9 Interaction Process in *Shimpai Muyo*u! virtual learning environment

Generally, the computer-based narratives are pre-scripted, with few alternative branches to explore. The same thing happens each time the user runs the system, so that it is hard to suspend his disbelief that these characters are somehow real. EN is build upon the model of improvisational drama rather than authored stories: an initial situation and characters with well-defined personalities and roles produces an unscripted interaction driven by real-time choices. Thus, similar, but not identical stories, are produced as different characters with different backgrounds are involved (Aylett et al. 2006). *Ijime*-bullying is naturally episodic and while each time differs in some aspects each time is also identical, making it a good candidate for an EN approach as shown in Fig. 10.

The *Shimpai Muyou!* stories shall be designed on the emergent narrative concept. The character's decisions and user's advice will determine as what is next i.e. whether, for instance, the IVA stands up to the bully or runs away. The requirement that the child influences characters also cry out for an emergent approach, since branching on every possible suggestion over a number of episodes would otherwise produce a combinatorial explosion, while the child will soon notice if a scripted agent would not respond to his advice or action. The agent framework of the IVAs will produce emotions in a most natural and believable way. The agent architecture would be developed in a way to be applied in a wider range of emergent dramas. Short and unscripted episodes divided by intervals will be designed during which the child user will advise the virtual *Ijime* victim. The advice will then influence the character's choices in the next episode. This position of the child user is in accordance with the concept of Boal's "*Spect-Actor*", who is an active participant of the play (Boal 1983).

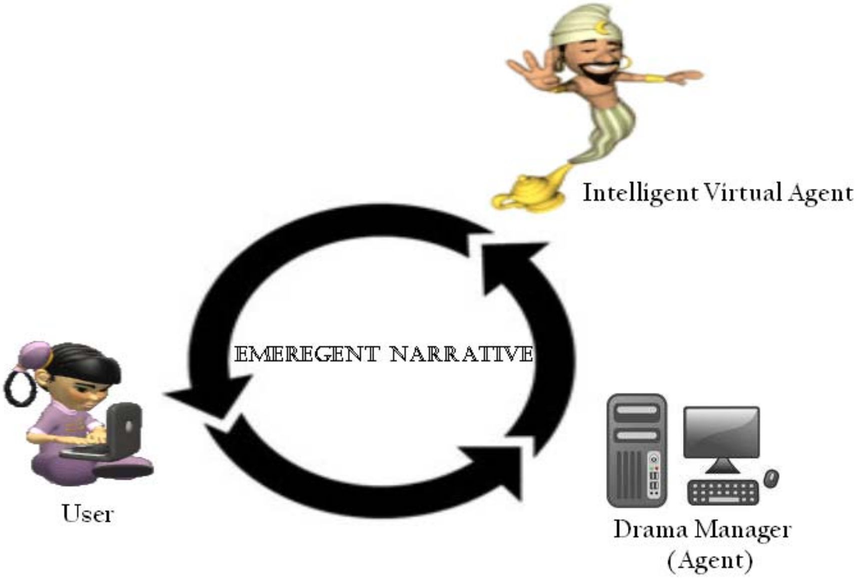


Fig. 10 A simple representation of narrative that emerges from interaction process

The architecture comprises of several agents, ensuing an Agent-based model that is systematized in Fig. 11. The flat list of major components of the system's architecture includes five modules:

1. Global Model
2. Display Model
3. Drama Manager (DM)
4. Narrative Action

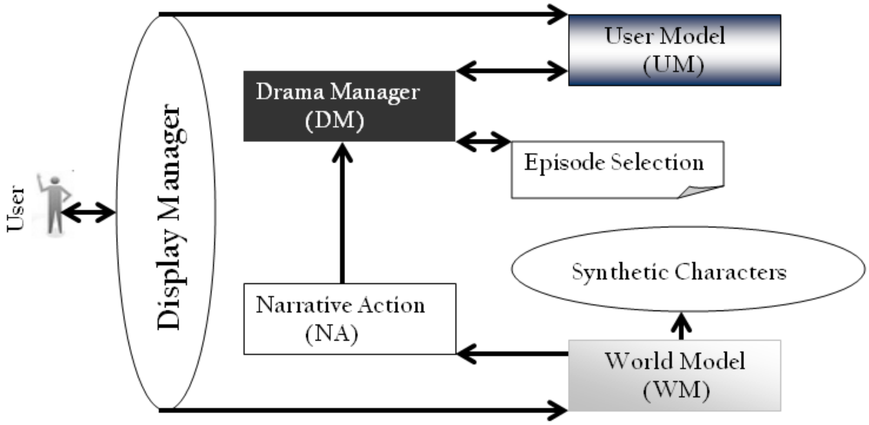


Fig. 11 General architecture of *Shimpai Muyou!* Agent-based Model

Global Model is the key source of information for the VLE. It carries the narrative structure and facilitates the communication between various other agents of the system. The two major agents, as such, are the *Drama Manager* (DM) and the Characters; whilst characters perform their role in the story, the drama manager ensures the attainment of educational goals. It contains fundamental elements of the narrative structure for instance, characters, goals, tasks, sub-tasks, segments, the states of the characters defined with predicates and information related to the material of the world of story. The *Display Model* will be responsible for displaying the actions and it will manage the interaction between the computer and the user.

The *Narrative Action* implies a dialog or performed in the virtual world, which has some narrative significance. These actions constitute the core of the narrative sequence. In other words, it is an action available to the author to describe episodes in the Introduction (set up the scene) of the episode, where they can be used to insert the characters of the story, set the camera to a particular position or narrate some text in the interface (Paracha et al. 2008-*c*).

The *Drama Manager* (DM) will be created, in conformity, with Boal's Forum Theatre approach, and it shall be assigned to perform some critical functions in *Shimpai Muyou!*. The prime function will be to control and facilitate the unfolding of the narrative as well as retrieve and save all interactions with the user in log files. The DM will have the prerogative to check the status of the NLE and intervene. It will invigilate the back and forth transactions between the agents and their narrative framework. It shall also keep track of all actions taken by the characters and the users. The most crucial function would be choosing appropriate episodes and for this, it performs a sequencing algorithm during the filtration process as diagrammatically represented in Fig. 12.

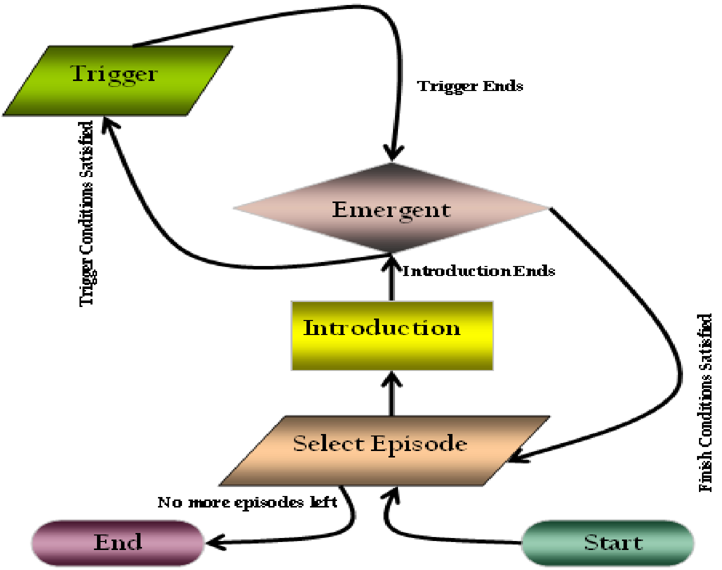


Fig. 12 Sequencing algorithm performed by DM during the filtration of episodes

In that manner, as mentioned earlier, the DM just after the Introductory Phase (typing in code, name, gender, characters introduction and schools) has to select Contentious Episodes first and then pick the Resolution Episodes. As for instance, we want the child user to advice the victim 'ask your friend for help', the episode 'Make Friends' must be played before this action. Therefore, the Resolution Episode shall be made available by the DM only, when Make Friend episode has already been watched by the child user. Furthermore, the DM will filter the logical conditions determining the context of an episode (type of episode, its location, the characters involved etc) and it fires triggers after each condition will be satisfied. Finally, it shall determine the ending conditions of an episode and it shall display the Educational Message, if there would be no further episode left.

6.3 Narrative Role Playing in Education

Narrative role playing is the most natural way of communication and a powerful technique to sense the world. It is developed at a young age through pretense play in which children set the stages and build the props that enable them to revisit, recast, and play out their fears and fancies (Ackermann, 2004). They use dolls and puppets to create role plays, using their own voices and facial expressions. The emotions are relatively uncharged and expressed in an almost calm way. In that manner, the children are trying to sense their new environment and through these kinds of one-to-one or solitary games they develop social skills necessary to communicate with this new world.

Creative drama (McCaslin, 1996), TIE and DIE (Heathcote, 1984) are effective ways of learning and child education (Jorgensen, 2000). According to Heathcote (1984) 'to use drama in education is to look for the precise dramatic pressure that will lead to a breakthrough, to a point where the students have to think about a problem in a new way, to fight for language adequate to the tension they feel. It is to literally bring out what children already know, but do not yet know they know'. She recommends drama a tool to expand child's awareness; to enable them to look at reality through fantasy; to see below the surface of actions to their meaning, and not just to produce plays. The objective of DIE is to comprehend -- what is being presented and not only playmaking. It is a way of learning therefore, it should give due role to the participants to share their experiences instead of focusing on communication between actors and audience.

Jorgensen (2000) describes, both playmaking and FT need an experienced facilitator to guide and deepen the learning of the participants. The participants are at the heart of the process and merely polished performance is not the goal. The main objective is improvising and acting out myriad solutions in order to learn experientially inside the drama parameters instead of outside passive mode. He also noticed the strong similarities between FT and TIE that imply demonstrating dramatic skills for audience. He also reflects on the differences and similarities between the drama teacher and Boal's joker. In his dissertation, Jorgensen (2000) gives examples of how he practices drama with children through side-coaching and teaching role playing narrative. He also considers FT's Joker system nonstructured and flexible as compared to the teacher taking the role of character with the students tends to be more strictly structured and controlled.

Boal (1992) regarded theater as a participants itself that implies democratization of communication as a medium. In other words, the roles can be performed by all classes of society and not a monopoly of professional artists. This democratization process also includes children, who being the first recipients of transformation, are not only active learners, but also very flexible for changes. It is natural therefore, to provide them the opportunity to make the world where they are growing up much better through role playing in a safe and sound environment. The world changes at a rather fast pace, especially in this technological era therefore, giving children the tool to anticipate, plan and rehearse these changes means empowering them to create a better living environment or world for themselves. They should rehearse how to solve the conflicts that arise from this fast change and help to shape a better world for the generations to come. Putting *Shimpai Muyou!* NLE in the hands of children will be an effort 'to enable them to explore and democratize drama through technological tools' (Cavallo A, 2008).

Role-playing games (RPGs) do not allow free expression, but they revolve around an orbit of ostentation. RPGs give illusion to their participants that they are living in a medieval time or a heaven and provide them with some moments of relief where they can escape the stress of daily life. In that situation one can dress up and give a set of behaviors to an avatar and pretend to live that life through the created character. Since this is a make-believe situation there is not much space to work on daily issues or oppression. Quite the contrary, one wants to forget the real life problems and the game might became addictive (Burrill, 2005).

Shimpai Muyou! on the other hand, has the objective of allowing children to perform real life problem in a safe environment in order to rehearse and find solutions for them. Conflicts are part of our lives and all children have conflicts with other children sometimes. It provides opportunities to learn how to negotiate, resolve and deal with other people. However, *Ijime*-bullying is different from occasional conflicts. Victimization refers to a child repeatedly subjected to acts of physical or verbal aggression with clear intention to humiliate him before others. The victim is usually weaker and *Ijime* is often carried out by a group of children. However, there are also subtle forms of bullying that hurt as much but are difficult to detect e.g. relational aggression or the systematic exclusion from peers.

In a virtual environment somewhere between a computer game and a real-life drama, children will be able to follow various characters as they go through *Ijime* scenarios and advise them on the best course of action. The idea is to provide pupils a better grasp of why bullies, victims, and bystanders act the way they do. We want them to use our system and feel empathy towards the victim characters and to take responsibility for the actions that take place. We will never be able to totally root out bullying but we hope we can make the lives of many pupils much better (Wolke *et al*, 2001). *Shimapi Muyou!* is therefore, a computer-based solution to the threat of *Ijime*-bullying. The motivation behind is to design a virtual reality school where in a safe environment children encounter the bullies and work out how to deal with them - without engaging themselves into the real troubles. If children can play out different alternatives of how to deal with *Ijime* safely in a virtual school this is likely to benefit them in real life.

7 Relevant Narrative Learning Environment Architectures

The section reviews NLEs of particular relevance to *Shimpai Muyou!*, covering STEVE, Carmen's Bright IDEAS, PUPPET, *FearNot!* and TEATRIX. In each case it examines the similarities to and differences from *Shimpai Muyou!,* covering overall architecture, agent architecture, use of narrative and user interaction. The objective does not require us to be comprehensive and to evaluate every VLE, assuming that this would be possible here. We have endeavored instead for the lessons *Shimpai Muyou!* project can learn from them, and in particular from the architectures used in these applications. The idea here is to avoid reinventing already existing good ideas which can be used in the project or expending effort solving problems which other groups have already solved.

7.1 STEVE

The STEVE system of Rickel and Johnson 2000, has been chosen for this investigation for following reasons. Firstly, it is among the oldest systems that have used 3D characters in a virtual environment for teaching purposes and confronted the basic problem of integrating a number of different types of components both from AI and graphics. Secondly, it employs both natural language and speech for interaction with the user, a substantial achievement given the less developed nature of speech recognition and text-to-speech systems at that time (Aylett 2002). However, it is significant to mention the differences also from *Shimpai Muyou!* perspective. In general, STEVE was developed for immersive training (using a head-mounted display) in the operation of gas turbines in the US Navy vessel. *Shimpai Muyou!* is not focusing on similar physical immersive experience, rather it utilizes standard desktop equipment that is normally available in schools. Furthermore, It is also, for reasons stated earlier, excluding the user from direct participation in episodes of narrative (Fig. 13).

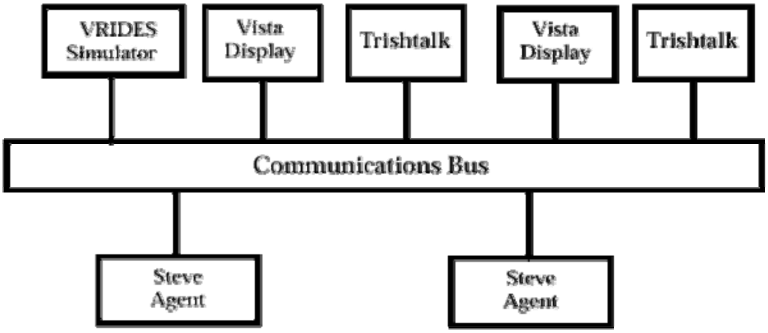


Fig. 13 STEVE overall architecture

In addition to that the STEVE domain is knowledge-based, and the teaching objective is for the student to learn how to operate machinery. In contrast, *Shimpai Muyou!* domain rests on feelings, behaviours and attitudes and the aim is to create empathy. Consequently, the STEVE system did not incorporate emotion, and interaction is driven by the structure of the task and the actions carried out in the procedure being taught. The internal architecture of a STEVE agent is of less relevance to our NLE, but the worth mentioning aspect is its implementation in a large publicly-available AI rule-based system called 'Soar'. The agent's activity is driven by an AI planning system implemented in 'Soar', which handles sequences of actions in operational procedures along with the logical pre-conditions and effects of each action (Aylett 2002).

To summarize, the system delivers though a related example of systems integration to our project, the nature and function of agents in STEVE do not match to those envisaged for *Shimpai Muyou!*. The orientation of its instruction is diametrically to ours, focusing on a task-based approach intended to convey procedural information. It is still not obvious if any detailed evaluation has been conducted to determine it as a live teaching system or merely intended as a demonstration of what can be accomplished.

7.2 Carmen's Bright IDEAS

Carmen's Bright IDEAS may be regarded as the most relevant virtual learning environment to *Shimpai Muyou!*, in many ways even though it is entirely scripted, totally language based and uses 2D graphics only. IDEAS is an acronym for the technique – **I**dentify problem, **D**evelop solutions, **E**valuate them, **A**pply one and **S**ee if it worked (Marsella 2000). The pedagogical objective here is to assist mothers of those patients, who frequently confront multiple problems at home and at work, apart from their fears for their sick child, to learn a cognitive problem-solving technique. In its concentration on the agents' emotional state, as well as its pedagogical objective of inculcating empathy between a real mother in this situation and virtual actor Carmen, the approach resembles to that being adopted in *Shimpai Muyou!*.

Furthermore, the stance of the user here exhibits similarity to *Shimpai Muyou!*. At intervals during the dialogue, Carmen gets three thought bubbles, each representing in summary her feelings about the discussion. The user selects one of the options and this then determines the direction the dialogue takes. The user is a spectator for much of the scenario but is able to participate in a limited way to influence one of the characters via its emotional response or the topic in its mind. One reason for this choice, equally valid for *Shimpai Muyou!* is to maintain a separation between the problems of agent and user. The idea is to evoke empathy at the same time not forcing the user to carry an extra burden (Aylett 2002). In a similar way, we want the child user to empathise without feeling that he or she is being bullied. The Gina agent as shown in Fig. 14, has morally instructive role in the system whereas, the agents to be used in *Shimpai Muyou!* scenarios are not intended to play such a didactic role.

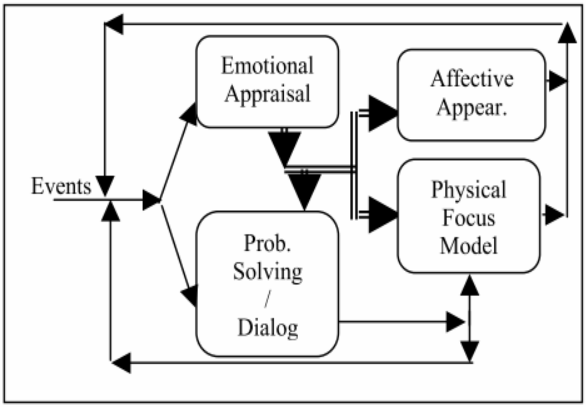


Fig. 14 Agent architecture in Carmen's Bright IDEAS

7.3 Puppet Theatre

In the *Puppet Theatre of the Virtual World* (Klesen et al. 2001) children can interact with believable virtual playmates in 3D environment and learn how they may respond to behaviors of others through experimenting (taking) actions of their own choice. The Theatre includes a range of life-like characters with personalities specified by the child, which can allow full improvisations with emergent behaviors in developing stories. The dramatic situation was simple compared to the one envisaged in *Shimpai Muyou!*. It involved two autonomous characters, a farmer and a cow. A third character, represented as a sheep, could be played as an avatar by the child user and used to change the status and attitude of the farmer or cow by interaction, depending on the sheep's mood, set by the child to positive, negative or neutral. Agents had no language capabilities, and there was no requirement for overall dramatic structure, merely a never-ending set of episodic interactions.

In that perspective, interaction history is much less important than in the proposed *Shimpai Muyou!* scenarios since there is no real beginning or end, and characters change but do not have to develop, as is usually the case in behaviorally-driven systems like this. The system was tested and deployed in some schools, but it is no longer in use. The lack of narrative structure was seen an issue, in particular the lack of any feeling of finishing, or of achieving any goals (Aylett 2002). It suggests that the idea of including different characters, in *Shimpai Muyou!* virtual learning environment, is a sensible one since it should allow a set of scenarios to have a natural end for a particular character.

7.4 FearNot!

The *FearNot!* application, developed by VICTEC (Aylett et al. 2005), offers a very relevant example of agent architecture and an agent oriented system to the *Shimpai Muyou!* project. The artistic and graphical orientations of *FearNot!*

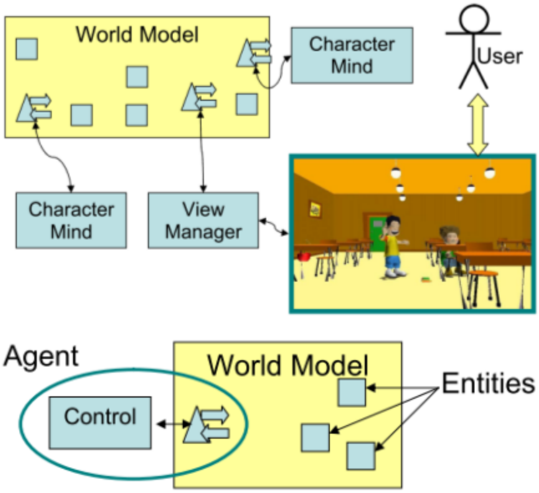


Fig. 15 *FearNot!* Agent within in the virtual space

though differ still yet the application exposes and highlights some relevant elements regarding believability and suspension of disbelief to *Shimpai Muyou!*.

In order to foster the essential illusion of reality, researchers at the *FearNot!* virtual learning environment developed agents of broad capabilities, as shown in Fig.15, such as the management and expression of internal goals, reactivity, emotion, natural language abilities, knowledge of agents (self and other) as well as of the simulated physical micro-world. The research undertaken within *FearNot!* regarding interactive EN, and particularly regarding the conditions and sources for the emergence of character based interactive narratives, highlighted the need for *Shimpai Muyou!* to consider the use of similar autonomous agents and narrative approach.

7.5 TEATRIX

TEATRIX also has several resemblances, in spite of their differences of goals, with *Shimpai Muyou!* system. The children create storyboards of existing fairy tales and posses some control over existing characters that are intelligent agents, but the environment, the creation and the problem solving capabilities, limits the learners to a very specific domain and does not allow enough interactions. TEATRIX is part of the Networked Interactive Media in Schools (NIMIS) program in Europe and is being used in schools in Germany, England and Portugal (Prada et al. 2000). The most revealing feature of TEATRIX from *Shimpai Muyou!* perspective is the reflection tool. During its evaluation, it was observed that without it, children could get so immersed in the folds of its narrative that they did not reflect very much on the emotional states of characters, resulting in

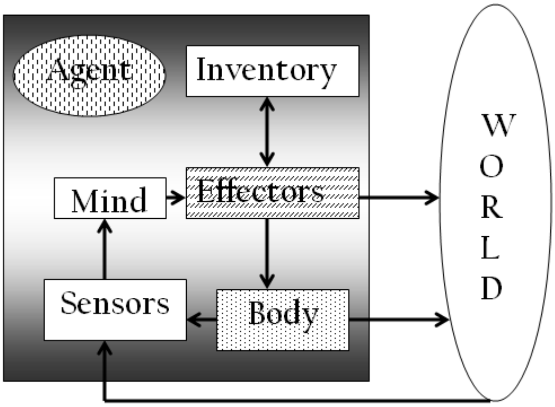


Fig. 16 Agent Components in TEATRIX

stories with surface actions only (Fig. 16). In other words, by halting action and asking specifically for such reflection, produces more detached user behaviour thus, user is forced to think objectively.

Shimpai Muyou! is perhaps unusual in requiring both good physical interaction between agents (pushing, stealing a trainer etc) and emotionally-driven language interaction (threats, apologies, abuse etc). Therefore, it seems likely that the *Shimpai Muyou!* architecture will have to 'combine some aspects of more behavioural architectures with those of more cognitively oriented architectures'(Aylett 2002).

8 Conclusion

Based on above theoretical framework the *Shimpai Muyou!* system will induce affective responses in children during *Ijime*-bullying situations, and thus, change their behaviour and cognitions. The application uses a 3D environment, intelligent agents with an *emotional model*, and *natural language processing* to a small extent. It is an attempt towards Positive Technology Development (PTD) that arose from the necessity to explore the use of technology on youth development programs and interventions to foster positive ways of children growth (Bers 2006). We hope that the system will encourage personal abilities in the pupils to deter *Ijime* at schools and engage them on meaningful real life issues. The NLE may also be considered a sort of Positive Youth Development (PYD) initiative as it will help in the growth of child's cognitive variability that develops as a consequence of interactions among the growing person and her biological, psychological, ecological (family, community, culture), and historical place (Lerner et al. 2005). By *Shimpai Muyou!* architecture in effect we shall have the opportunity to put in practice Friere's Constructivism, Papert's constructionism and Boal's participatory theatre to inculcate conflict resolution skills in children. It will also expose them to technology and augment their technical fluency, which is a very important skill in the actual cyber world.

Abbreviations Used in This Work

Abbreviations	**Terms**
AI	Artificial Intelligence
CDF	Classroom Discussion Forum
DNM	*Diegesis* Narrative Model
DM	Drama Manager
DIE	Drama-in-Education
EN	Emergent Narrative
FT	Forum Theatre
IE	Information Extraction
IVA	Intelligent Virtual Agents
IDEAS	**I**dentify problem, **D**evelop solutions, **E**valuate them, **A**pply one and **S**ee
Ijime	Japanese terminology “苛め” for “bullying”
NIMIS	Networked Interactive Media in Schools
NLE	Narrative Learning Environment
PTD	Positive Technology Development
PYD	Positive Youth Development
RPG	Role Playing Games
Shimpai Muyou!	Japanese terminology “心配無用” for “Don’t worry” or “Don’t be afraid”
TIE	Theater-in-Education
TO	Theatre of the Oppressed
VLE	Virtual Learning Environment
VE	Virtual Environment
VFT	Virtual Forum Theatre
VICTEC	Virtual ICT with Empathic Characters
VR	Virtual Reality
3D	Three Dimensional

References

Ackermann, E.: Constructing knowledge and transforming the world. In: Tokoro, M., Steels, L. (eds.) A learning zone of one’s own: Sharing representations and flow in collaborative learning environments. IOS Press, Amsterdam (2004)

Aristotle, Poetics. Trans. Kenneth McLeish. Nick Hern Books, London (1998)

Aylett, R., Louchart, S., Dias, J., Paiva, A., Vala, M., Woods, S., Hall, L.: Unscripted Narrative for Affectively Driven Characters. IEEE Computer Graphics and Applications 26(3), 42–52 (2006)

Aylett, R.: Deliverable 3.1.1/Report no.1/Version 1. IST-2001-33310 VICTEC (2002), http://www.macs.hw.ac.uk/victec/deliverables.htm (Accessed February 1, 2009)

Aylett, R., Louchart, S., Dias, J., Paiva, A., Vala, M.: FearNot! - an experiment in emergent narrative. In: Panayiotopoulos, T., Gratch, J., Aylett, R.S., Ballin, D., Olivier, P., Rist, T. (eds.) IVA 2005. LNCS (LNAI), vol. 3661, pp. 305–316. Springer, Heidelberg (2005)

Begault, D.: 3D Sound for Virtual Reality and Multimedia. Academic Press Professional, Inc., San Diego (1994)

Bers, M.: The role of new technologies to foster positive youth development. Applied Developmental Science 10(4), 200–219 (2006)

Boal, A.: Theatre of the oppressed. Civilizacao Brasileira, Rio de Janeiro (1983)

Boal, A.: Games for actors and non-actors. Routledge, London (1992)

Brecht, B.: Brecht on Theatre. The development of an aesthetic. Hill and Wang, New York (1964)

Burrill, D.A.: Out of the Box: Performance, Drama, and Interactive Software. Modern Drama 48(3), 492–512 (2005)

Campbell, J.: The Hero with a Thousand Faces. Pantheon Books, New York (1949)

Cavallo, A., Couch, A.: Virtual Forum Theater: a CSCL for underprivileged children. XXXI SEMISH - Integrated Seminar of Software and Hardware, Salvador, Brazil (2004), http://web.media.mit.edu/~mello/research.htm (Accessed December 24, 2008)

Cavallo, A.: Virtual Forum Theater: Creating and sharing drama to resolve conflicts. Unpublished doctoral dissertation in Drama, Tufts University (2008), http://web.media.mit.edu/~mello/research.htm (Accessed October 20, 2008)

Cohen-Cruz, J.: Mainstream or Margin? US activist performance and Theatre of the Oppressed. In: Schutzman, M., Cohen-Cruz, J. (eds.) Playing Boal, pp. 137–156. Routledge, London (1994)

Dewey, J.: Experience and Education. Collier Books, New York (1938)

Druin, A., Hendler, J.: Robots for kids: Exploring new technologies for learning. Academic Press, San Diego (2000)

Duckworth, E.: The Having of Wonderful Ideas and other essays on teaching and learning. Teachers College, Columbia University, New York (1987)

Freire, P.: Pedagogy of the oppressed. Herder and Herder, New York (1972)

Gardner, H.: The Arts and Human Development. John Wiley & Sons, Inc., New York (1973)

Heathcote, D.: In: Johnson, E. (ed.) Dorothy Heathcote: Collected writings on education in drama, Hutchinson, London (1984)

Hodhod, R., Kudenko, D.: Interactive Narrative for Adaptive Educational Games. In: Proceedings of The First York Doctoral Symposium on Computing, University of York, UK (2007)

Jorgensen, L.: Boal and youth theatre in the United States. Unpublished doctoral dissertation in Drama, Tufts University. In: Reference in Virtual Forum Theater: Creating and sharing drama to resolve conflicts, Alice Cavallo 2008. Unpublished doctoral dissertation in Drama, Tufts University, http://web.media.mit.edu/~mello/research.htm (Accessed October 20, 2008)

Klesen, M., Szatkowski, J., Lehmann, N.: A Dramatised Actant Model for Interactive Improvisational Plays. In: de Antonio, A., Aylett, R.S., Ballin, D. (eds.) IVA 2001. LNCS (LNAI), vol. 2190, pp. 181–194. Springer, Heidelberg (2001)

Laurel, B.: Computers as Theatre. Addison-Wesley Publishing, Reading (1991)

Lerner, R.M., Lerner, J.V., Almerigi, J.B., Theokas, C., Phelps, E., Gestsdottir, S., et al.: Positive Youth Development, Participation in Community Youth Development Programs, and Community Contributions of Fifth-Grade Adolescents: Findings From the First Wave of the 4-H Study of Positive Youth Development (2005)

Lerner, R.M., Lerner, J.V., Alermigi, J., Theokas, C., Phelps, E., Gestsdottir, S., Naudeau, S., Jelic, H., Alberts, A.E., Ma, L., Smith, L.M., Bobek, D.L., Richman, R.D., Simpson, L., Christiansen, E.D., Von Eye, A.: Positive youth development, participation in community youth development programs, and community contributions of fifth grade adolescents: Findings from the first wave of the 4-H Study of Positive Youth Development. Journal of Early Adolescence 25(1), 17–71 (2005)

Livo, N.J., Rietz, S.A.: Storytelling: Process and Practice. Littleton. Libraries Unlimited Inc., Colorado (1986)

Lord, A.B.: The Singer of Tales. Harvard Univ. Press, Cambridge (1960)

Louchart, S., Aylett, R.S.: Narrative Theory and Emergent Interactive Narrative. International Journal of Continuing Engineering Education and Life-long Learning, special issue on narrative in education 14(6), 506–518 (2004)

Marsella, S.: Pedagogical Soap, AAAI Fall Symposium Technical ReportFS-00-04, pp. 107–112. AAAI Press, Menlo Park (2000)

Mateas, M.: A preliminary poetics for interactive drama and games. Digital Creativity 12(3), 140–152 (2001)

McCaslin, N.: Creative Drama in the Classroom and Beyond, 6th edn. Longman Publishers, USA (1996)

Papert, S.: Introduction. In: Harel, I. (ed.) Constructionist Learning. MIT Media Laboratory, Cambridge (1990)

Paracha, S., Yoshie, O.: Being Spectator, Being Actor: Human Computer Interaction. In: Proceedings of IEE J., Hakodate, Japan, August 2008 (2008a)

Paracha, S., Khan, M.T.A., Yoshie, O.: Virtual Reality Problem. IEICE Technical Report ET2008-17~30, 57–62 (2008b)

Paracha, S., Mohammad, H.M., Khan, M.T.A., Mehmood, A., Yoshie, O.: Balancing Fabula & Interactivity: An Approach to VR Drama. In: Proceedings of the 4th IEEE ICET, Rawalpindi, Pakistan, October 2008 (2008c)

Paracha, S., Jehanzeb, S., Mehmood, A., Yoshie, O.: Virtual Reality Intervention: A Promising Deterrent to Juvenile Delinquency. In: Proceedings of IEEE-IC4 2009, Pakistan (2009d)

Paracha, S., Yoshie, O.: Combating Juvenile Delinquency with Empathic Agents. International Journal of Computer Science and Network Security 8(9), 196–205 (2008e)

Piaget, J.: The Essential Piaget. In: Gruber, H.E., Jacques Vonèche, J. (eds.) Basic Books, Inc., New York (1977)

Prada, R., Machado, I., Paiva, A.: TEATRIX: Virtual Environment for Story Creation. In: Gauthier, G., VanLehn, K., Frasson, C. (eds.) ITS 2000. LNCS, vol. 1839, p. 464. Springer, Heidelberg (2000), http://www.springerlink.com/content/jnff2ppe1tvdcwva/ (Accessed June 24, 2008)

Propp, V.: Morphology of the Folktale. University of Texas Press, Austin (1968)

Rickel, J., Johnson, W.L.: Task-Oriented Collaboration with Embodied Agents in Virtual Worlds. In: Cassell, J.S., Prevost, S. (eds.) Embodied Conversational Agents. MIT Press, Boston (2000)

Riedl, M.O., Young, R.M.: From Linear Story Generation to Branching Story Graphs. IEEE Computer Graphics and Applications 26(3), 23–31 (2006)

Shaw, D.: Aspects of Interactive Storytelling Systems. Masters thesis, The University of Melbourne (2004), `http://users.rsise.anu.edu.au/~davids/publications.html` (Accessed December 12, 2008)

Smiley, S.: Playwriting: The Structure of Action. Prentice-Hall, Inc., Englewood Cliffs (1971)

Sutcliffe, A.: Multimedia and Virtual Reality: Designing Multi-sensory User Interfaces. Lawrence Erlbaum, Hillsdale (2003)

Tapscott, D.: Growing Up Digital: The Rise of the Net Generation. McGraw Hill, New York (1998)

Taussig, M., Schechner, R.: Boal in Brazil, France, the USA. An interview with Augusto Boal. In: Schutzman, M., Cohen-Cruz, J. (eds.) Playing Boal, pp. 17–32. Routledge, London (1994)

Tzevtan, T.: Grammaire du Décameron. Mouton, The Hague (1969)

Tomaszewski, Z., Binsted, K.: A Reconstructed Neo-Aristotelian Theory of Interactive Drama. In: Workshop on Computational Aesthetics: Artificial Intelligence Approaches to Beauty and Happiness, National Conference on Artificial Intelligence (AAAI), Boston, Massachusetts (2006)

Vygotsky, L.S.: Mind in Society. Harvard University Press, Cambridge (1978)

Vygotsky, L.S.: Psychology of Art. MIT Press, Cambridge (1971)

Wolke, D., Woods, S., Schulz, H., Stanford, K.: Bullying and victimization of primary school children in South England and South Germany: Prevalence and school factors. British Journal of Psychology 92, 673–696 (2001)

Art – A Perfect Testbed for Computer Vision Related Research

Peter Peer and Borut Batagelj

Abstract. Contemporary art nowadays tries to exploit modern technology to better address and enlighten specific problems and ideas of our time. Our interest in wider impact of modern technology on society and the interest in contemporary art, brought our attention also to the applicative field of use of computer vision methods in art. This chapter walks us through a few projects, proving that art is definitely a perfect testbed for our research: 15 Seconds of Fame, Dynamic Anamorphosis, Virtual Skiing, Smart Wall, Virtual Dance and Virtual Painter, from face detection, motion following, depth recovery, touchless human-computer interaction to pop-art, constant eye gaze of a person on the portrait, regardless of where the spectator stands, immersion into different virtual worlds without the need for any special equipment.

1 Introduction

Human ability to function in multiple disciplines, communities is again becoming very important. Latest observations in computer vision community are also showing the need for collaboration between our community, computer graphics community and contemporary art community. We are talking about somekind of convergence, eventhough the idea has been around for some time now.

A wider discussion is always appreciated as similar problems are enlightened from different perspectives. In such a way a specific professional terminology

Peter Peer
Computer Vision Laboratory, Faculty of Computer and Information Science,
University of Ljubljana, Tržaška 25, 1000 Ljubljana, Slovenia
e-mail: `peter.peer@fri.uni-lj.si`

Borut Batagelj
Computer Vision Laboratory, Faculty of Computer and Information Science,
University of Ljubljana, Tržaška 25, 1000 Ljubljana, Slovenia
e-mail: `borut.batagelj@fri.uni-lj.si`

M. Grgic et al. (Eds.): Rec. Advan. in Mult. Sig. Process. and Commun., SCI 231, pp. 611–629.
springerlink.com

becomes community-independent (or at least less dependent), similar solutions are refactored, joint, improved, the ideas become clearer, new applicative areas are found.

By following this philosophy, we initiated collaboration between our faculty and the Academy of Fine Art and Design at the University of Ljubljana. To be more precise, the collaboration between the students of both faculties within the undergraduate subject Communication Methods at our faculty was initiated by prof. Franc Solina (Faculty of Computer and Information Science) and prof. Srečo Dragan (Academy of Fine Art and Design). The successfull collaboration resulted also in the establishment of the formal association called ArtNetLab [4], which core purpose is to support fusion of science and art [21, 22].

This is way the next section talks about the way the students collaborate and points to some latest projects. Because of our interest in contemporary art and also because of seeing art as the perfect testbed for our research, a new and interesting applicative area for us, we also developed some installations out of the scope of the mentioned undergraduate subject. Thus, in section 3 we present a project 15 Seconds of Fame, an interactive art installation which elevates the face of a randomly selected gallery visitor for 15 seconds into a work of art. In section 4 we describe another project called Dynamic Anamorphosis, an installation that enables constant eye gaze of a person on the portrait, regardless of where the spectator stands. The Virtual Skiing idea is discussed in section 5. The Smart Wall project, which provides a platform for a rapid prototyping of computer supported interactive presentations that sense human motion is presented in section 6. Section 7 reveals the Virtual Dance idea, which allows to define a set of visual markers and to associate them with visual icons in the interactive video. The conclusions are given in section 8, touching the discussed projects, future projects, ideas and applicability of mentioned solutions in other areas than art.

2 How Communities Collaborate?

At the Academy of Fine Art and Design they have a Department for Video and New Media. Postgraduate students studying at this department develop an idea that incorporates new technologies into it. Since the art students are less experienced with technical development, have less engineering skills, we form a project group around them. The project team normally consists of one art student and two to four undergraduate students from the Faculty of Computer and Information Science in their final year of study. The advisor is also assigned to them to help them successfully carry out the project in the given time-frame, to monitor the progress, to suggest, point out the right approach.

Each year we form around 10 teams and they have a bit less than a semester to finish the project, present it and write a final report. The projects are then also presented, exhibited at different festivals.

The Smart Wall and Virtual Dance ideas (sections 6 and 7) are actually two projects done in such teams last year. Other projects include titles like: Sinking;

Morphing; Demon; An Invitation for a 20th Century Dinner; Interactive Meeting – just to name a few dealing with computer vision methods in art. Some interesting projects are wider in the sense that they are integrating new technologies and not specifically computer vision, for instance: Touch Animation; Protocols, Communications and Expanded Media (A Place of Home); DataDune.

Everybody involved in the process gain a lot: an artist gets the possibility to put his/hers idea into life, experiment freely with new technology and try to invent better and new ways of interfacing with virtual worlds [12, 17], science/engineering students get hands on practice, go through the project phases, learn to listen to and understand the customer, they all learn a lot, and at the end they are all enthusiastic because of the working installation, product.

All descriptions of the mentioned projects are available on the Internet [4].

3 15 Seconds of Fame

15 Seconds of Fame is an interactive installation which every 15 seconds generates a new pop–art portrait of a randomly selected person from the audience [23, 25]. The installation was inspired by Andy Warhol's ironical statement that "In the future everybody will be famous for 15 minutes". The installation detects human faces in digital images of people who are standing in front of the installation. Pop-art portraits are then generated from randomly chosen faces in the audience by applying randomly selected filters. These portraits are shown in 15 second intervals on the flat-panel computer monitor, which is framed as a painting. Electronic copies of each displayed portrait can be ordered by e–mail.

3.1 Motivation

Warhol took faces from mass media, banal in their newspaper everydayness, and transformed them into portraits. Warhol portrayed in this fashion celebrities from arts and politics. The installation tries to make instant celebrities by reversing Warhol's process – making Warhol-like celebrity portraits of common people and putting them on the gallery walls to make the portraitees implicitly famous. Since 15 minutes would hardly make the installation interactive the fame interval was shortened to 15 seconds. The faces for the portraits made by the installation are selected by chance out of all people in front of the installation to allude that fame tends to be not only short-lived but also random. In his film and video projects Andy Warhol was in fact fascinated with celebrification of "nobodies" which marks the beginning of an era in which media attention became the new mirror of the individual's self-perception.

3.2 How the Installation Works?

The visible part of the installation consists of a computer monitor framed like a painting. A digital camera is hidden behind the frame so that only a round opening for the lens is visible. Pictures of gallery visitors which are standing in front of the installation are taken every 15 seconds by the digital camera using a wideangle lens setting (Fig. 1). The camera is connected to a computer, which detects all faces in each picture, randomly selects a single face, makes a pop-art portrait out of it and displays it for 15 seconds on the monitor.

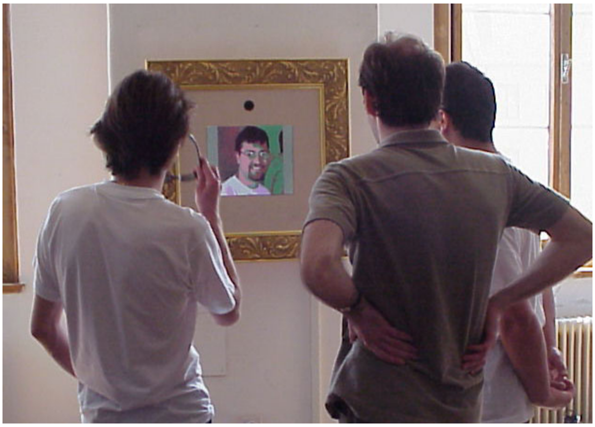


Fig. 1 People in front of the 15 Seconds of Fame installation

The color-based nature of our face detection makes it sensitive to illumination. Since it is not always possible to exhibit the installation under daylight or white-balanced studio illumination, we improved our face detection results by applying color compensation methods to make the whole system more flexible [25].

To make his celebrity portraits Warhol segmented the face from the background, delineated the contours, highlighted some facial features, started the process with the negative, overlaid the image with geometric color screens etc. We tried to achieve similar effects with a set of filters that achieve effects similar to segmentation. The filters drastically reduce the number of different colors by joining similar looking pixels into uniform regions. They combine three well known filters: posterize, color balance and hue-saturation with an additional process of random coloring. In this way, we achieve millions of different effects.

Our primary goal was not to mimic Andy Warhol's pop-art portraits per se, but to play upon the celebrification process and the discourse taking place in front of the installation. In comparison to other video camera based art installations, ours does not require exact positioning of observers due to automatic face detection with the additional benefit that a group of people can interact with the installation simultaneously. The interaction is technically very simple – no visible interface is actually involved – but unpredictable and socially revealing.

4 Dynamic Anamorphosis

In [24] we define the concept of dynamic anamorphosis. A classical or static anamorphic image requires a specific, usually a highly oblique view direction, from which the observer can see the anamorphosis in its correct form (Fig. 2). Dynamic anamorphosis adapts itself to the changing position of the observer so that wherever the observer moves, she/he sees the same undeformed image.

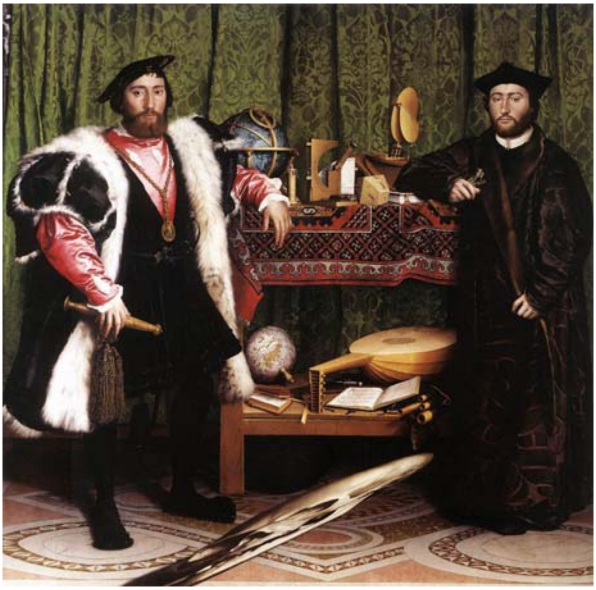


Fig. 2 On the bottom of the painting appears a diagonal blur, which appears as a human skull when viewed from the upper right (The Ambassadors by Hans Holbein)

4.1 Motivation

The dynamic changing of the anamorphic deformation, in concert with the movement of the observer, requires from the system to track the 3D position of the observer's head and the recomputation of the anamorphic deformation in real time. This is achieved using computer vision methods which consist of face detection/tracking of the selected observer and stereo reconstruction of its 3D position while the anamorphic deformation is modeled as a planar homography. Dynamic anamorphosis can be used in the context of art installation, in video conferencing to fix the problem of the missing eye contact and can enable an undistorted view in restricted situations. Anamorphosis serves as a model for the concept of the gaze, which suggests that visual appreciation rather than passive "looking" requires active "observing".

4.2 How the Installation Works?

We use a face detection method to determine the position of the user's face in the pictorial plane. Face detection is now a mature technology and can run in real-time. For head tracking we must detect the faces in every frame. To improve the tracking we use addition clues such as motion, skin color or near-infrared image. By using two or even more cameras and the principle of stereo reconstruction of distances we can further determine the position of the user's head in 3D space. The most difficult problem in stereo reconstruction is the correspondence problem – to find for a given point in the left image the corresponding point in the right image. Since the number of possible matches goes into thousands of points this is a computationally intensive task. The correspondence problem in this particular case is solved by finding faces in both images first. Next, only correspondences between faces need to be established.

We approach the stereo matching problem as a matching between homologous faces, instead of point matching. The main idea is to determinate a unique disparity value for the whole face region and no longer for individual pixels. After we detect the position of faces in both stereo images we construct a graph for each image where face blobs are represented as nodes in the graph. To find homologous faces in both stereo images we perform graph matching. The computational process is simple and fast since we consider only complete face regions.

At the end we deform the projected image of the face in such a way that it looks undeformed from the viewpoint of the observer (Fig. 3).

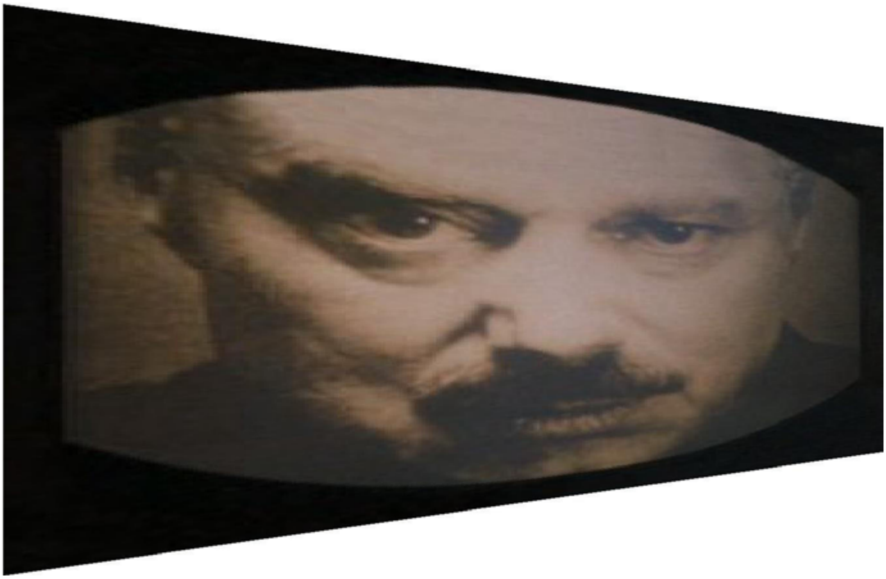


Fig. 3 Transformed frame of a video clip, when the user views it under 30° angle from the right (Big Brother from the film after George Orwell's novel 1984)

Dynamic anamorphosis disassociates the geometric space in which the user moves from the visual cues she/he sees, since wherever the observer moves, she/he sees the same image. The installation promotes a human face (Fig. 3) with the eye gaze directed straight ahead to meet the eyes of the installation user. It requires a dark room with the video projection over an entire wall so that the only visible cues seen by the user are given by the projection. The light reflected back into the room from the projected image must sufficiently illuminate the scene that face detection can be performed. Since the installation can truly be experienced only by a single user, the entrance to the room with the installation should be controlled.

5 Virtual Skiing

An interactive installation Virtual Skiing [26] enables a visual immersion into the feelings of gliding on snow through a winter landscape. The computer rendered winter landscape is displayed over the entire wall in front of the skier (Fig. 4). As on real skis you can regulate the speed of descent by changing the posture of your body so that the air resistance is decreased or increased. By shifting the weight of your body to the right or left ski you can make turns down the slope between the snow capped trees. The interface to the virtual world is implemented by computer vision techniques, which capture the posture of the skier's body using a video camera placed in front of him and processed on a PC in real time to drive the projected animation of the virtual slope.

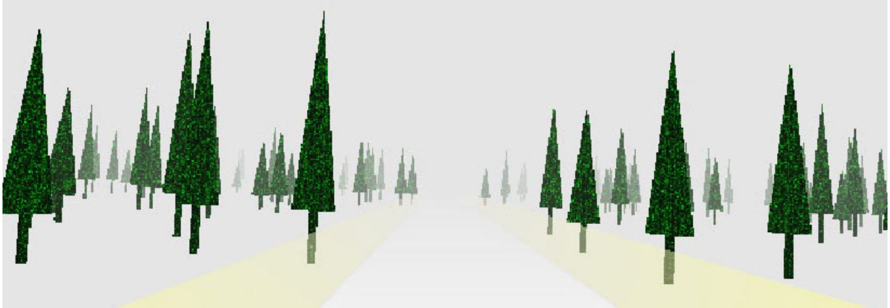


Fig. 4 The virtual slope is covered with sparsely populated trees among which the skier must find his way

5.1 Motivation

Real-time interaction of people with virtual environments is a well established concept but finding the right interface to do it is still a challenging task. Wearing different kinds of sensors attached to the body of the participants is often cumbersome. Computer vision offers the exiting possibility to get rid of such sensors and to record the body movements of participants using a camera [20]. People, their appearance (i.e. face), their emotions and the movements of their bodies are becoming on the other hand an important object of study in computer vision research [11].

The number of application areas for virtual environments is growing since the cost of technology for making virtual environments is in general going down. Sporting games in general are an attractive area for using virtual technology. Many training machines for cycling, running, rowing are enhanced with a virtual world to make the training more interesting. Instead of a static scene in a fitness room one can get the feeling of moving along a real scene or even to race against other real or virtual competitors. Virtual exercisers are sophisticated simulations that deliver the demands, stresses, and sensations of a sport or exercise with unprecedented verisimilitude and precision.

Skiing is a popular sport, which is unfortunately restricted to appropriate climactic and terrain conditions. Therefore various attempts have been made to bring this sport closer to anyone using modern technology. A very costly and direct method is to build an artificial slope with artificial snow or some other surface that enables sliding. Much more cost effective is now virtual technology although the whole ensemble of sensations experienced in the virtual world is not as realistic.

There have been quite a number of skiing games and skiing simulators played on a regular computer interface or on dedicated platforms such as the video game "Alpine Racer" by Namco. In the mid 1990's a special robotic platform was built by Jim Rodnunsky for the Vail ski center. The Ski and Snowboard Simulator is a surging, rolling, swaying, pitching, heaving, and yawing hydraulic recreation which took $4 million to develop. More recently, complete 3D models of existing ski centers have been build for promotional goals. They enable the user to freely move around and attain additional information or to follow a predefined path for the user to experience a virtual downhill race [1, 2].

5.2 How the Installation Works?

The virtual skiing installation consists of a video camera in front of the skier which records the skier's movements, a computer with a 3D graphics card, a sound system to convey the basic sound events (i.e. colliding with a tree) and a video projector to project the virtual slope.

From the grabbed video frame we first subtract the background. The background image is taken at the beginning and periodicity when there is no moving in front of the camera. In such a way we also get rid of the problems with changing illumination during the exhibition. Additionally we binarize the images to final separate the figure of the skier from the background (Fig. 5). The threshold for binarization is determined dynamically using a histogram method each time a new participant steps into the scene. The image subarea where binarization takes place can be adjusted interactively to suit different setups of the installation.

Each time a user enters the scene as seen from the camera, the height of his silhouette is recorded and the center of his silhouette is determined by finding a

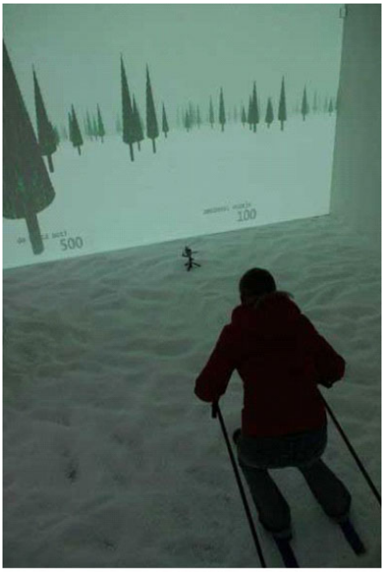


Fig. 5 The silhouette of the skier as result of the background subtraction and binarization of the input video image

pixel balance between his left and right side. As the skier shifts his body to the left or right side this initiates on modern carving skis a turn in the same direction. Turning can be therefore controlled just by shifting the center of the body/silhouette. When the user flexes his knees and lowers his body to achieve a lower air resistance the height of his silhouette is decreased and this is interpreted as an increase of the skier's speed (Fig. 6). The information of the body's position relative to the upright position is given to the skiing polygon rendering engine, which displays the virtual slope. In reality, the biomechanics of skiing is much more complicated but for the interaction with a virtual environment such actions are quite realistic.

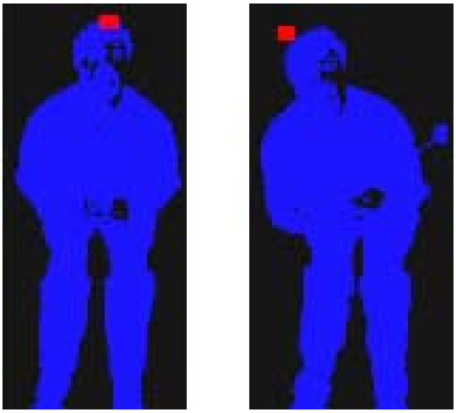


Fig. 6 The skier performs the same movements as on real skis to turn on the virtual ski slope. The position of the video camera can be seen at the bottom of the projection

When the user comes in the virtual world too close to a tree, a collision is triggered. The collision algorithm is quite simple and written to suit the dedicated terrain rendering engine. It first checks if any of the trees are close along the direction of the movement of the skier (z axis). Then it checks if any of these trees are also in the x axis range (x axis of the projected image). When the skier collides, three parameters are taken into consideration. The skiers speed, tree height and the side from which the skier hit the tree (left, right, or direct middle). The life points are decreased according to the skier's speed and tree height. The side of the tree is used to "bounce" the skier off the tree to the side that the skier came from. Hitting the tree on its left side, would bounce him back to the left.

The software is written in C++ and uses the DirectShow interface to capture video data. For displaying the terrain the cross-platform SDL library and a custom rendering engine based on OpenGL is used. The rendering of the virtual terrain is done in OpenGL using the cross-platform SDL library. The trees have been modeled in 3D Studio and imported into the rendering engine. At startup this engine computes random positions and heights for 400 trees, which make up all of the trees seen. Additionally, 7% of the trees are flagged as "hearts". This information is used when the user is skiing in the survival mode of the game play to draw hearts. Another mode is plain skiing where the user skis the virtual slopes with no obvious objective other than to enjoy the experience.

6 Smart Wall

The main goal of the Smart Wall project is to provide a platform for a rapid prototyping of computer supported interactive presentations that sense human motion. The system is composed by a front end application, where the developer defines a number of hot spots in a camera view, a Hotspot Processor, which senses the activity in each of the hot spots, and a Player, which displays interactive content triggered by the activity in hot spots. By associating actions or sequences of actions in the environment to actions in the interactive presentation, a variety of complex interactive scenarios can be developed and programmed with ease. Due to the modular architecture, the platform supports distributed interaction, connecting physical activity and content display at remote locations.

6.1 Motivation

Computer vision is widely used to sense the presence and actions of humans in the environment. Novel surveillance systems can reliably track and classify human activity, detect unusual events and learn and retrieve a number of biometric features. Due to the low cost and the ubiquity of personal video technology, the research has recently shifted towards developing novel user interfaces that use vision as the primary input. In the area of personal computing, the most prominent areas of research are desktop interfaces that track gestures [18], and novel multi touch interfaces which use camera based tracking of fingers in contact with a translucent display [13]. On a wider scale, human motion can be used to interact with smart environments [20], or to trigger smart public displays [3].

The major motivation of Smart Wall framework is to provide an open environment to develop and experiment with interactive presentations in public space. The interactive area can be any public area that is observed by a camera connected to a computer. The field of view of the camera can then be divided in a grid of cells which define the smallest areas where activity will be detected. The resolution and, consequently, the size and the number of the cells, is defined by the developer. Multiple neighboring cells on the grid can then be joined in a single "hot spot", representing thus an interactive entity that promotes its state to other modules.

Each hot spot can define an action that occurs when the activity in the hot spot area exceeds a certain threshold (i.e. a person enters the area). The presentation on the screen is controlled by such actions or by a sequence of actions.

6.2 How the Installation Works?

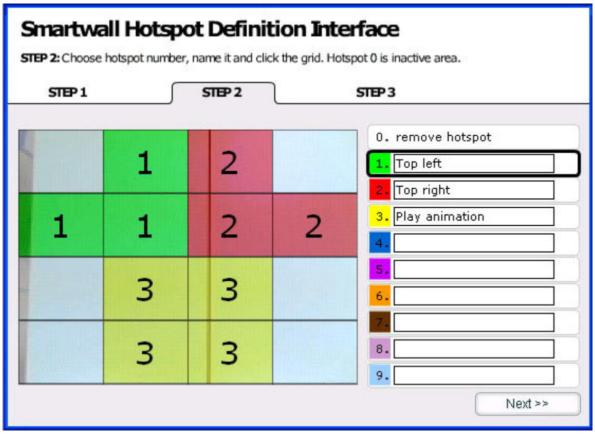
Smart Wall system consist of three parts: HotSpot Processor, HotSpot Definition Interface and Player. All three parts communicate through XML files and can be installed on different systems. Furthermore, a Player with any other content can be implemented and by reading generated XML a variety of other actions could be achieved.

HotSpot Definition Interface (HDI) is a Flash application for defining hotspots (Fig. 7). HDI imports the picture of the floor and overlays the floor with grid of selected resolution. In first step user can choose density of the grid and in second step user can define hotspots (a unique name can be applied for easier XML reading by users). In the last step HDI exports the XML via web service, so it can be saved on any computer.

HotSpot Processor is a C# application which uses AForge [15] library for computer vision. HotSpot processor input is a XML file from HDI in which hotspots are defined with the name, threshold and squares that are attached to it. HotSpot processor compares current image from the camera with the snapshot of the grid at the beginning or any other time if the user defined so. The output of the HotSpot Processor is a XML file in which every hotspot is presented with 0/1 value. If a change is detected on a hotspot then the hotspot value is 1 otherwise the value is 0.

The third part of the system is the Player. We used Adobe Flash in our project, as it can produce eye-catching, dynamic content and support many multimedia formats, however it can be implemented in any other technology. The player runs the presentation in loop and reads the HotSpot Processor output XML. When it detects a change on a hotspot it runs the action defined on the hotspot. HotSpot actions must be defined in the player.

Each part of the system can be used separately and for many applications (Fig. 8).



Smartwall Hotspot Definition Interface

STEP 2: Choose hotspot number, name it and click the grid. Hotspot 0 is inactive area.

STEP 1 | STEP 2 | STEP 3

	1	2	
1	1	2	2
	3	3	
	3	3	

0. remove hotspot
1. Top left
2. Top right
3. Play animation
4.
5.
6.
7.
8.
9.

Fig. 7 HotSpot Definition Interface (HDI)

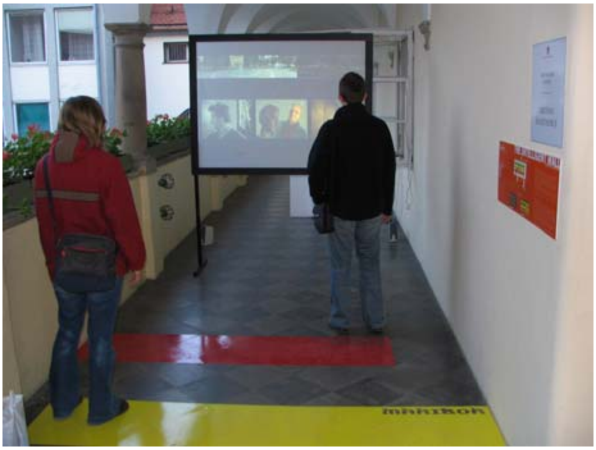


Fig. 8 Smart Wall in action

7 Virtual Dance

The Virtual Dance project [10] presents a system that is able to track dance movements and integrates the motion in an interactive video. The idea is to provide a flexible framework which allows the artist to set up an interactive virtual dance performance by defining markers, videos and interactive visual icons associated with markers. The system is then able to interact between the real movement and the virtual movement. We used standard tracking methods and modified them to support fast moving markers, small markers and discontinuous tracking of markers.

7.1 Motivation

Consequently, by dancing, the dancer creates two patterns of movement in two different spaces [19]. The first pattern is produced in the three dimensional real world. The second pattern is reflected in the two dimensional virtual space. These two representations are inseparable. The virtual dance is shown on the screen and in this way the real world dance extends to a new complementary dimension of expression [27].

7.2 The Real and the Virtual Dance

The real dance is recorded by the camera in real time (Fig. 9). While the dance is limited by space, no spatial constraints should be imposed by the camera and other technology that processes the video. Therefore, the system must handle tracking of markers that leave the camera field of view and reappear, or markers that are temporarily occluded by the dancer, by objects in the scene, or by other dancers. In that way, the dancer can also leave the scene and reappear multiple times during the performance, and the system must be able to robustly track the markers thorough the performance.

The virtual dance is produced as a consequence of real dance. The user defines the virtual space while the dancers dance, so she manipulates the virtual dance and its presentation [7]. The application produces a movement of the virtual representations

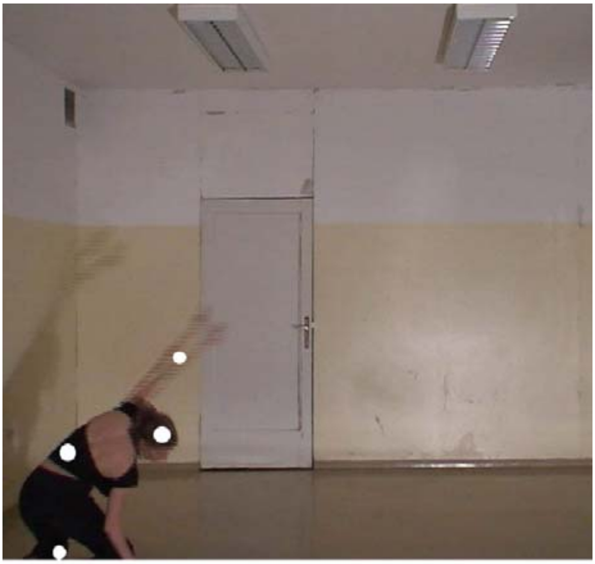


Fig. 9 The dancer is marked with four hypothetical markers. The positions of the markers were chosen by the dancer.

that is in correlation with markers on the dancer's body. The user interactively chooses the representation of a marker in the virtual video dance. The representation can be a picture or an animated two dimensional sequence, e.g. a movie clip of the dancer representing a pre-recorded movement. The user can also choose the background of the virtual dance and can change the size and the colour of virtual markers (Fig. 10).

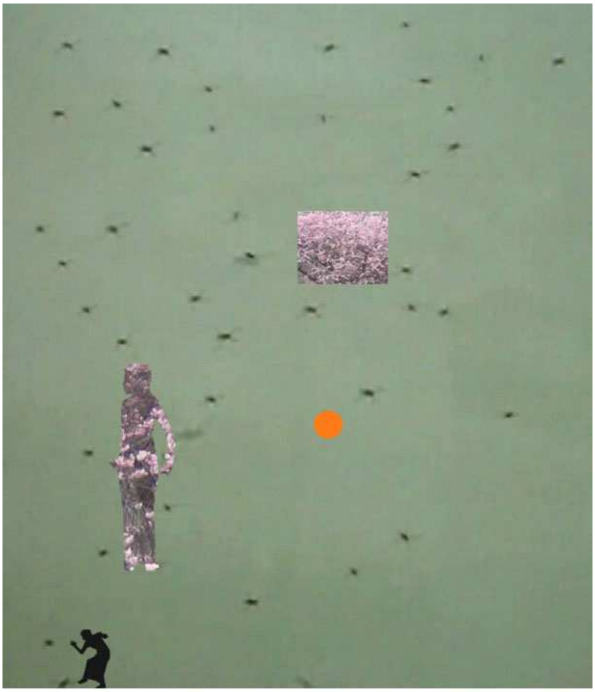


Fig. 10 This figure shows the virtual dance. Four markers were chosen in the real world, thus, the virtual world has also four markers.

The real dance and its virtual presentation are inseparably connected because of real time video processing [9]. Every movement in real world immediately produces a movement in virtual world. Dancers can observe the virtual dance that is produced by their movement [6]. So the dancers can also interact with the virtual space through their dance. A dancer can observe the virtual dance and produce a brand new story and movement in the virtual world. If she chooses some particular presentation of her markers in the virtual dance and if she moves in some special way, she can produce a new story in the virtual world which is not focused on tracking the real world movement, but it becomes a virtual story that has no connection with dance but is still directed by the dancer [5].

7.3 How the Installation Works?

The application can read the video either directly from the camera or from a prerecorded video file. The application then reads images from the input video, and, if markers are defined, each of the frames is processed by the marker tracker. In the pause mode, the video can be also displayed on the screen to support the definition of the marker areas (Fig. 11).

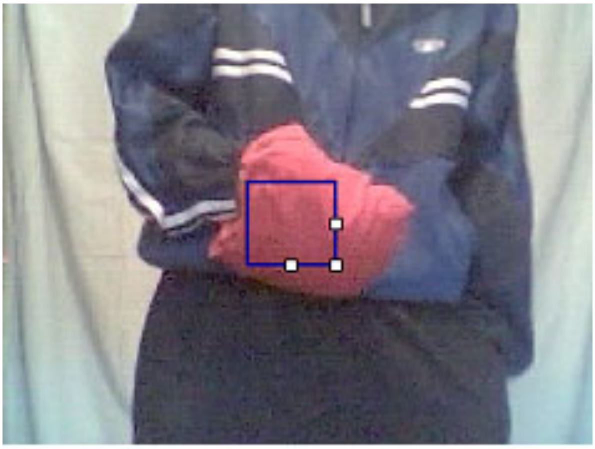


Fig. 11 This figure shows an example of input video that is shown on the screen. The user has chosen that the marker will be the red textile and therefore positioned a selection rectangle on that textile.

To represent a marker's appearance, we calculate the RGB histogram of chosen area on the picture where the marker is positioned. We apply a weighing schema; for example, if marker is mostly red, we give a high weight value to the red component, and a low weight value to the green and the blue component. The histogram is then a representation of the weighted distribution of red, green and blue values on the marker area (Fig. 12).

For every frame, we calculate the histogram back projection with respect to the histogram of the tracked marker. The function returns a probability of a pixel belonging to a marker area for each pixel in the image.

After we calculate the back projection, we calculate the marker position by using the Mean Shift algorithm [8]. We also consider the estimated position and size of the marker in the previous frame and modified the original Mean Shift algorithm to meet our demands. The original mean shift uses the back projection and the previous bounding box parameters of the marker. Then it finds the position inside the rectangle which has the highest probability of being part of the marker. This probability is

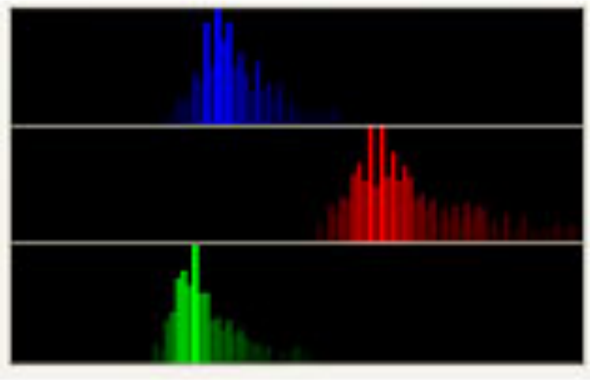


Fig. 12 Histogram of the marker area shown in 11

calculated considering all probabilities of pixels that are inside the rectangle. Then, the center of the bounding box rectangle is moved over that pixel with the highest probability, and the same calculation is performed again. The calculation repeats until the rectangle is moved less than a user defined parameter at the last iteration, or for a maximum of *N* iterations.

The original Mean Shift does not track well fast moving objects, small objects, and objects that have left video and have appeared again. These are very frequent events in dancing, where the camera is able to record only a part of the dance floor. So we made a change to the original Mean Shift algorithm. Since we cannot rely on a search only in the immediate neighbourhood, the search has to be initiated on a number of areas covering the whole image.

We divide the picture to 23 parts vertically and 23 parts horizontally. Then, the centre of rectangle is moved in every of 23×23=529 parts of the picture and the probability of marker being inside the moved rectangle is calculated. That way we find the position on the picture where is the highest probability that there is the marker if we move the original rectangle there.

The new position of rectangle represents the basis to calculate the new position using the Mean shift procedure. Once the application has found the position of the marker on the image, the position is written to a XML file.

The module that displays the virtual dance reads the position of markers from XML files and updates the position of virtual markers on the screen. The artist can select whether a marker is represented by a picture or by an animated sequence. She can also choose the size and the colour of markers. The background is also selected by the user and can display a static image or a video. New pictures and videos for markers and background can be simply added by updating XML files. The resulting virtual world and the virtual dance can be then integrated into the performance shown to the public.

8 Conclusions

The main idea behind this work is to prove or at least give a very good indicator that art is a perfect testbed for computer vision based research in many cases, giving you a very multidisciplinary feeling. This indicator is provided through the summarization of the projects in the Computer Vision Laboratory dealing with the convergence of topics mostly in computer vision, graphics and contemporary art communities.

Let us look back a bit. The installation 15 Second of Fame detects human faces in digital images of people who are standing in front of the installation. Pop-art portraits are then generated from randomly chosen faces in the audience by applying randomly selected filters.

The dynamic changing of the anamorphic deformation, in concert with the movement of the observer, requires from the system to track the 3D position of the observer's head and the recomputation of the anamorphic deformation in real time.

The installation Virtual Skiing provides a user with the opportunity to interact with a virtual space. The virtual skier can navigate down the virtual slope just by changing the posture of his body. The possible actions are turn right/left and change of the speed of descent. In the process of navigating down the slope the skier must avoid the sparsely arranged trees. The interface is very intuitive since the skier just repeats the actions that he knows from real skiing and learns to control his movement in the virtual world in less than a minute. The system works in real time and is very robust to any influences from the environment, such as change of illumination. Other games for the same virtual skiing setup could be introduced. A race course with gates could be implemented so that the skiers could compete and try to finish the course in the shortest time. A virtual model of an existing slope could be used as a promotional means for that ski resort. Virtual skiing is just one possible application of such computer vision based interface.

We introduced the Smart Wall framework for controlling different presentations only with movements in the particular hotspots. The applicability of the Smart Wall idea is very diverse, for instance, we used it also in computer games [16].

With the Virtual Dance project we presented one of the possible connections between dance, video and technology. In ideal illumination conditions, the marker does not have to be a separate add-on that is worn by a dancer, but can be defined as a part of the costume or as a part of the body that is distinguished enough to be recognized by the system. The tracked person is not necessarily aware of the tracking. Obviously, we can use the system not only for tracking in dance, but also in many other applications.

Another example of immersion into a virtual world is given by the installation Virtual Painter, which is a work in progress. As suggested by the name itself, a user of it can paint without the actual paints, while the painting forms itself on the computer screen, it can be stored, printed. As in the previous cases the camera is observing the user in front of the installation and tracks hers/his palm. The movement of the palm is then transformed into streaks in the painting. With the second hand you could simply switch between the colors on the pallet by raising the hand. An example of the interaction with the installation is given in Fig. 13.

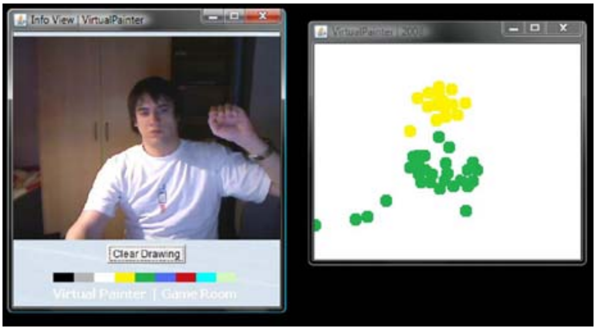


Fig. 13 Painting in the Virtual Painter installation

We could most probably go on and on with counting up application areas of introduced approaches, but for the end we would like to mention one more, which uses unconventional human-computer interaction through camera: helping disabled persons to use computers, helping them towards better life. The author of the initial version of the Virtual Painter was inspired by it and consequently initiated AIADA (Artificial Intelligence Aided Disable Assistance) project, which enables for instance writing in MS Word, drawing in MS Paint, manipulating components of the operating system, talking, and controlling home appliances by tracking head movements (Fig. 14) [14].

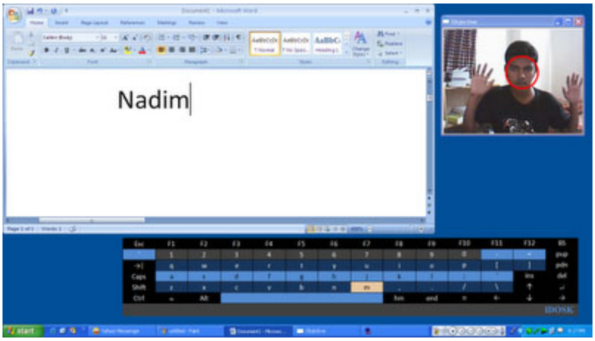

Nadim

Fig. 14 Writing in MS Word by tracking head movements

References

[1] Almer, A., Stelzl, H.: Multimedia Visualisation of Geoinformation for Tourism Regions based on Remote Sensing Data. ISPRS – Technical Commission IV/6, ISPRS Congress Ottawa (2002)

[2] Altland, G.: Virtual Skiing Environment (1997), `http://www-personal.umich.edu/~galtland/skiVR/skiVR.html` (Accessed February 2, 2009)

[3] Batagelj, B., Ravnik, R., Solina, F.: Computer vision and digital signage. In: ICMI 2008. Association for Computing Machinery, New York (2008)

[4] Bovcon, N., Vaupotič, A.: ArtNetLab (2000–2009), `http://black.fri.uni-lj.si/` (Accessed February 2, 2009)

[5] Burtnyk, N., Wein, M.: Interactive skeleton techniques for enhancing motion dynamics in key frame animation. Communications of the ACM 19, 564–569 (1976)

[6] Calvert, T.: Animating dance. In: Graphics Interface 2007, Canada (2007)

[7] Calvert, T., Wilke, L., Ryman, R., Fox, I.: Applications of Computers to Dance. IEEE Computer Graphics and Applications 25, 6–12 (2005)

[8] Comaniciu, D., Meer, P.: Mean Shift: A Robust Approach Toward Feature Space Analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 24, 603–619 (2002)

[9] Csuri, C.: Computer animation. ACM SIGGRAPH Computer Graphics 9, 92–101 (1975)

[10] Dovgan, E., Čigon, A., Šinkovec, M., Klopčič, U.: A system for interactive virtual dance performance. In: 50th International Symposium ELMAR 2008, Zadar, Croatia (2008)

[11] Essa, I.A.: Computers seeing people. Artificial Intelligence Magazine 20, 69–82 (1999)

[12] Grau, O.: Virtual art, from illusion to immersion. MIT Press, Cambridge MA (2003)

[13] Han, J.Y.: Low–cost multi-touch sensing through frustrated total internal reflection. In: UIST 2005: ACM symposium on User interface software and technology (2005), doi:10.1145/1095034.1095054

[14] Jahangir, N.: AIADA (Artificial Intelligence Aided Disable Assistance) project (2008), `http://nadim0112358.blogspot.com/2008/07/project-aiada.html` (Accessed February 2, 2009)

[15] Kirillov, A.: AForge .NET framework (2009), `http://www.aforgenet.com/` (Accessed February 2, 2009)

[16] Kreslin, R., Dežman, D., Emeršič, Ž., Peer, P.: Use of Unconventional User Interfaces Based on Computer Vision in Computer Games. In: potočnik, B. (ed.) Rosus 2009, Maribor, Slovenia (2009)

[17] Levin, T.Y., Frohne, U., Weibel, P.: CTRL [SPACE], Rhetorics of Surveillance from Bentham to Big Brother. MIT Press, Karlsruhe, ZKM and Cambridge (2002)

[18] Nielsen, M., Moeslund, T.B., Storring, M., Granum, E.: Gesture Interfaces. HCI Beyond the GUI. Morgan Kaufmann, San Francisco (2008)

[19] Norman, I.B., Stephen, W.S.: Digital Representations of Human Movement. ACM Computing Surveys (CSUR) 11, 19–38 (1979)

[20] Pentland, A.: Smart rooms. Scientific American 274, 68–76 (1996)

[21] Solina, F.: Internet based art installations. Informatica 24, 459–466 (2000)

[22] Solina, F.: ArtNetLab – the essential connection between art and science. In: Gržinić, M. (ed.) The future of computer arts & the history of The International Festival of Computer Arts, Maribor (2004)

[23] Solina, F.: 15 seconds of fame. Leonardo (Oxf.) 37, 105–110, 125 (2004)

[24] Solina, F., Batagelj, B.: Dynamic anamorphosis. In: Enactive 2007 enaction in arts: International Conference on Enactive Interfaces, Grenoble, France (2007)

[25] Solina, F., Peer, P., Batagelj, B., Juvan, S., Kovač, J.: Color-Based Face Detection in the "15 Seconds of Fame" Art Installation. In: International Conference on Computer Vision / Computer Graphics Collaboration for Model-based Imaging, Rendering, image Analysis and Graphical special Effects MIRAGE 2003, Paris, France (2003)

[26] Solina, F., Batagelj, B., Glamočanin, S.: Virtual skiing as an art installation. In: 50th International Symposium ELMAR 2008, Zadar, Croatia (2008)

[27] Wilke, L., Calvert, T., Ryman, R., Fox, I.: From dance notation to human animation: The LabanDancer project: Motion Capture and Retrieval. Computer Animation and Virtual Worlds 16, 201–211 (2005)

A Survey of Image Processing Algorithms in Digital Mammography

Jelena Bozek, Mario Mustra, Kresimir Delac, and Mislav Grgic

Abstract. Mammography is at present the best available technique for early detection of breast cancer. The most common breast abnormalities that may indicate breast cancer are masses and calcifications. In some cases, subtle signs that can also lead to a breast cancer diagnosis, such as architectural distortion and bilateral asymmetry, are present. Breast abnormalities are defined with wide range of features and may be easily missed or misinterpreted by radiologists while reading large amount of mammographic images provided in screening programs. To help radiologists provide an accurate diagnosis, a computer-aided detection (CADe) and computer-aided diagnosis (CADx) algorithms are being developed. CADe and CADx algorithms help reducing the number of false positives and they assist radiologists in deciding between follow up and biopsy. This chapter gives a survey of image processing algorithms that have been developed for detection of masses and calcifications. An overview of algorithms in each step (segmentation step, feature extraction step, feature selection step, classification step) of the mass detection algorithms is given. Wavelet detection methods and other recently proposed methods for calcification detection are presented. An overview of contrast enhancement and noise equalization methods is given as well as an overview of calcification classification algorithms.

1 Introduction

Detection and diagnosis of breast cancer in its early stage increases the chances for successful treatment and complete recovery of the patient. Screening mammography is currently the best available radiological technique for early detection of breast cancer [1]. It is an x-ray examination of the breasts in a woman who is asymptomatic. The diagnostic mammography examination is performed for symptomatic women who have an abnormality found during screening mammography. Nowadays, in most hospitals the screen film mammography is being replaced with digital mammography. With digital mammography the breast image is captured

Jelena Bozek, Mario Mustra, Kresimir Delac, and Mislav Grgic
University of Zagreb, Faculty of Electrical Engineering and Computing
Department of Wireless Communications
Unska 3/XII, HR-10000 Zagreb, Croatia
e-mail: jelena.bozek@fer.hr
http://www.mammoimage.org/

M. Grgic et al. (Eds.): Rec. Advan. in Mult. Sig. Process. and Commun., SCI 231, pp. 631–657.
springerlink.com

using a special electronic x-ray detector which converts the image into a digital mammogram for viewing on a computer monitor or storing. Each breast is imaged separately in craniocaudal (CC) view and mediolateral-oblique (MLO) view shown in Figure 1(a) and Figure 1(b), respectively. The American College of Radiology (ACR) Breast Imaging Reporting and Data System (BI-RADS) suggests a standardized method for breast imaging reporting [2]. Terms have been developed to describe breast density, lesion features and lesion classification. Screening mammography enables detection of early signs of breast cancer such as masses, calcifications, architectural distortion and bilateral asymmetry.

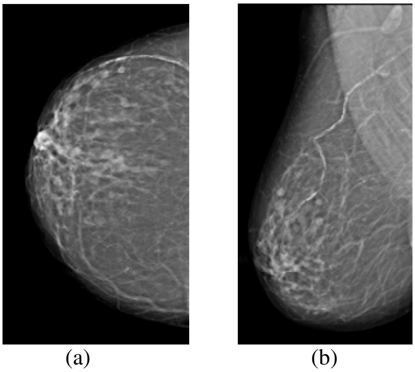


Fig. 1 Two basic views of mammographic image: (a) craniocaudal (CC) view, (b) mediolateral-oblique (MLO) view

A mass is defined as a space occupying lesion seen in at least two different projections [2]. If a potential mass is seen in only a single projection it should be called 'Asymmetry' or 'Asymmetric Density' until its three-dimensionality is confirmed. Masses have different density (fat containing masses, low density, isodense, high density), different margins (circumscribed, microlobular, obscured, indistinct, spiculated) and different shape (round, oval, lobular, irregular). Round and oval shaped masses with smooth and circumscribed margins usually indicate benign changes. On the other hand, a malignant mass usually has a spiculated, rough and blurry boundary. However, there exist atypical cases of macrolobulated or spiculated benign masses, as well as microlobulated or well-circumscribed malignant masses [3]. A round mass with circumscribed margins is shown in Figure 2(a).

Calcifications are deposits of calcium in breast tissue. Calcifications detected on a mammogram are an important indicator for malignant breast disease but are also present in many benign changes. Benign calcifications are usually larger and coarser with round and smooth contours [2]. Malignant calcifications tend to be numerous, clustered, small, varying in size and shape, angular, irregularly shaped and branching in orientation [1]. Calcifications are generally very small and they may be missed in the dense breast tissue. Another issue is that they sometimes have low contrast to the background and can be misinterpreted as noise in the inhomogeneous background [4]. Fine pleomorphic clustered calcifications with high probability of malignancy are shown in Figure 2(b).

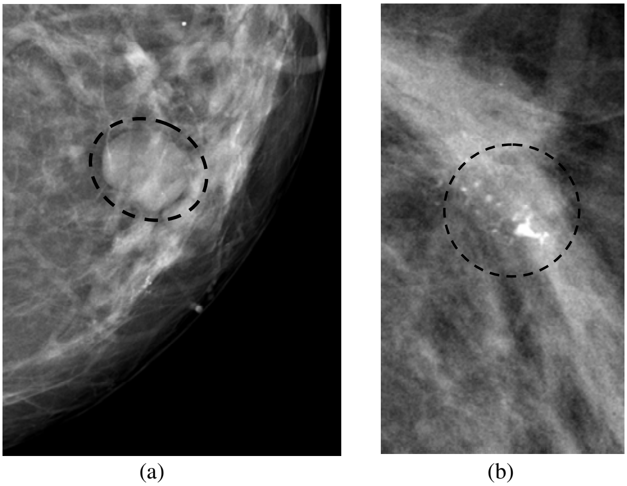


Fig. 2 Examples of abnormalities: (a) round mass with circumscribed margins, (b) fine pleomorphic clustered calcifications

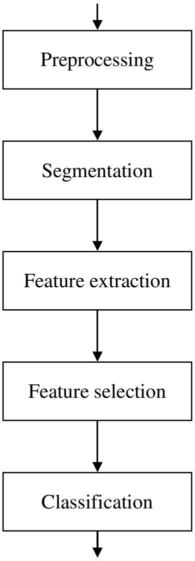
Architectural distortion is defined as distortion of the normal architecture with no definite mass visible, including spiculations radiating from a point and focal retraction or distortion at the edge of the parenchyma [2]. Architectural distortion of breast tissue can indicate malignant changes especially when integrated with visible lesions such as mass, asymmetry or calcifications. Architectural distortion can be classified as benign when there is a scar and soft-tissue damage due to trauma.

Asymmetry of breast parenchyma between the two sides is useful sign for detecting primary breast cancer. Bilateral asymmetries of concern are those that are changing or enlarging or new, those that are palpable and those that are associated with other findings, such as microcalcifications or architectural distortion [5]. If a palpable thickening or mass corresponds to an asymmetric density, the density is regarded with a greater degree of suspicion for malignancy.

As mentioned, breast lesions have a wide range of features that can indicate malignant changes, but can also be part of benign changes. They are sometimes indistinguishable from the surrounding tissue which makes the detection and diagnose of breast cancer more difficult. Radiologist's misinterpretation of the lesion can lead to a greater number of false positive cases. 65-90% of the biopsies of suspected cancers turn out to be benign [6]. Thus, it is important to develop a system that could aid in the decision between follow-up and biopsy. The use of computers in processing and analyzing biomedical images allows more accurate diagnosis by a radiologist. Humans are susceptible to committing errors and their analysis is usually subjective and qualitative. Objective and quantitative analysis facilitated by the application of computers to biomedical image analysis leads to a more accurate diagnostic decision by the physician [7]. Computer-aided detection (CADe) and computer-aided diagnosis (CADx) systems can improve the results of mammography screening programs and decrease number of false positive cases.

Most image processing algorithms consist of a few typical steps depicted in Figure 3. The screen film mammographic images need to be digitized prior the image processing. This is one of the advances of digital mammography where the image can be directly processed. The first step in image processing is the preprocessing step. It has to be done on digitized images to reduce the noise and improve the quality of the image. Most digital mammographic images are high quality images. Another part of the preprocessing step is removing the background area and removing the pectoral muscle from the breast area if the image is a MLO view. The segmentation step aims to find suspicious regions of interest (ROIs) containing abnormalities. In the feature extraction step the features are calculated from the characteristics of the region of interest. Critical issue in algorithm design is the feature selection step where the best set of features are selected for eliminating false positives and for classifying lesion types. Feature selection is defined as selecting a smaller feature subset that leads to the largest value of some classifier performance function [8]. Finally, on the basis of selected features the false positive reduction and lesion classification are performed in the classification step.

Fig. 3 Typical steps in image processing algorithms



In the case of mammographic image analysis, the results produced using a certain method can be presented in a few ways. The interpretation being mostly used is the confusion matrix (1) or just the number of true positives (TPs) and false positives (FPs). The confusion matrix consists of true negative (TN), false positive (FP), false negative (FN) and true positive (TP).

$$C = \begin{bmatrix} TN & FP \\ FN & TP \end{bmatrix} \tag{1}$$

There are some often mentioned terms such as accuracy (2), precision (3), sensitivity or true positive rate (TPR) (4) and false positive rate (FPR) (5).

$$accuracy = \frac{TP + TN}{TN + FP + FN + TP} \tag{2}$$

$$precision = \frac{TP}{FP + TP} \tag{3}$$

$$TPR = \frac{TP}{FN + TP} \tag{4}$$

$$FPR = \frac{FP}{TN + FP} \tag{5}$$

Area *A* under the ROC (Receiver Operating Characteristic) curve [9] gives the information of how successful the classification is. ROC curve is determined by true positive (TP) and false negative (FN) results of an experiment. The larger the area (total area is 1.00) the better the classification is. In the case of *A*=1.00, the detection performance is 100% with zero false positive detected objects at the same time. ROC curves are often used for classification tasks because they can give a good description of the overall system performance. It is worth mentioning that the area under ROC curve can be maximized without really improving the classification success. Random guessing will result in area *A*=0.5 which can be artificially boosted to some higher values close to 1.0 [10]. This, of course, will give false results and therefore results presented using only ROC curves should be taken with caution. Figure 4 shows the example of the ROC curve; *A* denotes the area under the curve that demonstrates the quality of classification.

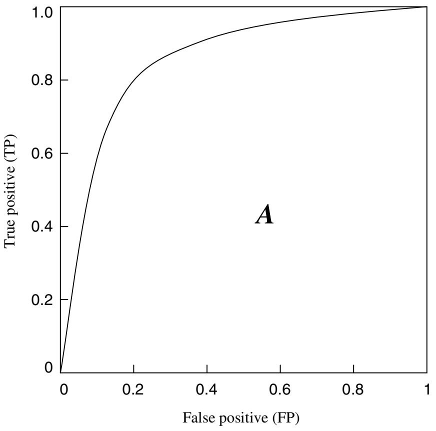


Fig. 4 A typical Receiver Operating Characteristic (ROC) curve

In this chapter the algorithms for detection of two most common signs of breast cancer, masses and calcifications are presented. Algorithms for architectural distortion detection and bilateral asymmetry detection are often a part of mass detection algorithms. Thus, a detailed description of those algorithms is not given. Also, there exist algorithms specially designed for architectural distortion detection and bilateral asymmetry detection, but due to the lack of space they will be described in detail in our future survey paper. The organization of the chapter is as follows. In Section 2 some of the recent algorithms for mass detection are presented. Subsections 2.1, 2.2 and 2.3 provide overview of algorithms in segmentation step, feature extraction and selection steps and classification step, respectively. Section 3 is devoted to microcalcification detection algorithms. Subsections 3.1 and 3.2 outline the Wavelet detection methods and other recently proposed methods. Subsection 3.3 gives an overview of contrast enhancement and noise equalization methods and subsection 3.4 gives an overview of calcification classification algorithms. Finally, Section 4 summarizes and concludes the chapter.

2 Mass Detection Algorithms

As already defined, a mass is space occupying lesion seen in at least two different projections defined with wide range of features that can indicate benign changes but can also be a part of malignant changes. Masses with round, smooth and circumscribed margins usually indicate benign changes while masses with spiculated, rough and blurry margins usually indicate a malignant mass. Some researchers have focused mainly on the detection of spiculated masses because of their high likelihood of malignancy. A benign round mass is shown in Figure 5(a) and malignant spiculated mass is shown in Figure 5(b).

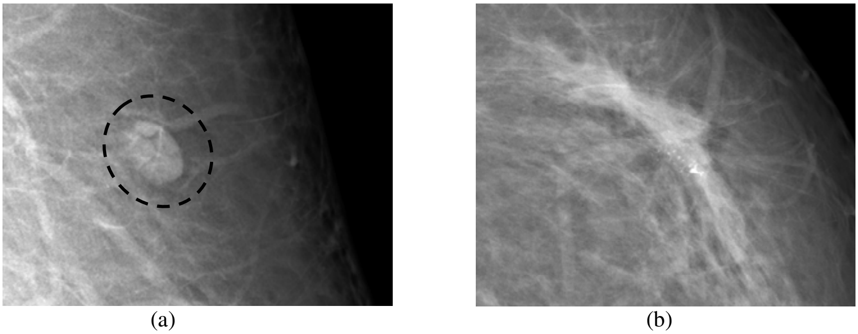


Fig. 5. An example of: (a) round mass, (b) spiculated mass

Algorithms for breast mass detection in digital mammography usually consist of several steps: segmentation, feature extraction, feature selection and classification. In the segmentation step regions of interest (ROIs) that contain abnormalities are segmented from the normal breast tissue. In the second stage of the algorithm each ROI is characterized with the set of features. In the feature selection step the best set of features are selected and in the classification step suspicious ROIs are classified as benign masses or malignant masses.

2.1 Segmentation

The aim of the segmentation is to extract ROIs containing all masses and locate the suspicious mass candidates from the ROI. Segmentation of the suspicious regions on a mammographic image is designed to have a very high sensitivity and a large number of false positives are acceptable since they are expected to be removed in later stage of the algorithm [4]. Researchers have used several segmentation techniques and their combinations.

2.1.1 Thresholding Techniques

Global thresholding [11] is one of the common techniques for image segmentation. It is based on the global information, such as histogram. The fact that masses usually have greater intensity than the surrounding tissue can be used for finding global threshold value. On the histogram, the regions with an abnormality impose extra peaks while a healthy region has only a single peak [6]. After finding a threshold value the regions with abnormalities can be segmented. Global thresholding is not a very good method to identify ROI because masses are often superimposed on the tissue of the same intensity level. Global thresholding has good results when used as a primary step of some other segmentation techniques.

Local thresholding is slightly better than global thresholding. The threshold value is defined locally for each pixel based on the intensity values of its neighbor pixels [6]. Multiple pixels belonging to the same class (pixels at the periphery of the mass and pixels at the center of the mass) are not always homogenous and may be represented by different feature values. Li et al. [12] used local adaptive thresholding to segment mammographic image into parts belonging to same classes and an adaptive clustering to refine the results.

Matsubara et al. [13] developed an adaptive thresholding technique that uses histogram analysis to divide mammographic image into three categories based on the density of the tissue ranging from fatty to dense. ROIs containing potential masses are detected using multiple threshold values based on the category of the mammographic image.

Dominguez and Nandi [14] performed segmentation of regions via conversion of images to binary images at multiple threshold levels. For images in the study, with grey values in the range [0, 1], 30 levels with step size of 0.025 were adequate to segment all mammographic images.

Varela et al. [15] segmented suspicious regions using an adaptive threshold level. The images were previously enhanced with an iris filter.

Li et al. [16] used adaptive gray-level thresholding to obtain an initial segmentation of suspicious regions followed by a multiresolution Markov random field model-based method.

2.1.2 Region-Based Techniques

Markov random field (MRF) or Gibbs random field (GRF) is one of the segmentation methods in iterative pixel classification category. MRFs/GRFs are statistical methods and powerful modeling tools [16]. Székely et al. [17] used MRF in "fine" segmentation to improve the preliminary results provided by the "coarse" segmentation. In "coarse" segmentation the feature vector is calculated and passed to a set of decision trees that classifies the image segment. After the "fine" segmentation they used a combination of three different segmentation methods: a modification of the radial gradient index method, the Bézier histogram method and dual binarization to segment a mass from the image.

Region growing and region clustering are also based on pixel classification. In region growing methods pixels are grouped into regions. A seed pixel is chosen as a starting point from which the region iteratively grows and aggregates with neighboring pixels that fulfill a certain homogeneity criterion. Zheng et al. [18] used an adaptive topographic region growth algorithm to define initial boundary contour of the mass region and then applied an active contour algorithm to modify the final mass boundary contour.

Region clustering searches the region directly without initial seed pixel [6]. Pappas [19] used a generalization of *K*-means clustering algorithm to separate the pixels into clusters based on their intensity and their relative location. Li et al. [12] used an adaptive clustering to refine the result attained from the localized adaptive thresholding. Sahiner et al. [20] used *K*-means clustering algorithm followed by object selection to detect initial mass shape within the ROI. The ROI is extracted based on the location of the biopsied mass identified by a qualified radiologist. Initial mass shape detection is followed by an active contour segmentation method to refine the boundaries of the segmented mass.

2.1.3 Edge Detection Techniques

Edge detection algorithms are based on the gray level discontinuities in the image. Basis for edge detection are gradients or derivatives that measure the rate of change in the gray level. Rangayyan [7] described standard operators for edge detection such as Prewitt operator, Sobel operator, Roberts operator and Laplacian of Gaussian (LoG) operator.

Fauci et al. [21] developed an edge-based segmentation algorithm that uses iterative procedure, a ROI Hunter algorithm for selecting ROIs. ROI Hunter algorithm is based on the search of relative intensity maximum inside the square windows that form the mammographic image.

Petrick [22] used Laplacian of Gaussian filter in conjunction with density-weighted contrast enhancement (DWCE). DWCE method enhances the structures within the mammographic image to make the edge detection algorithm able to detect the boundaries of the objects.

Zou et al. [23] proposed a method that uses gradient vector flow field (GVF) which is a parametric deformable contour model. After the enhancement of mammographic images with adaptive histogram equalization, the GVF field component with the larger entropy is used to generate the ROI. In the Figure 6 an example of GVF with and without enhancement is given.

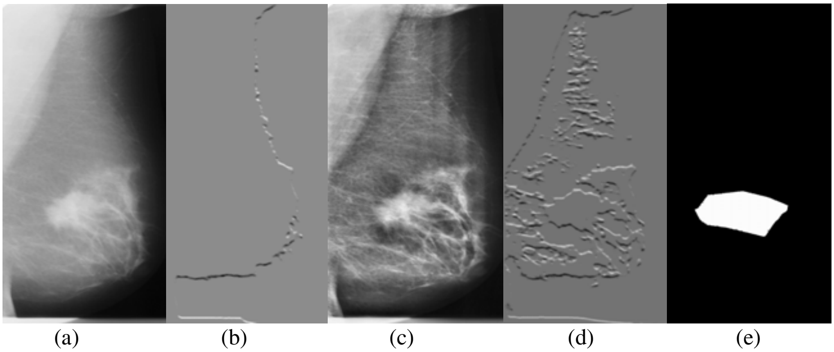


Fig. 6 An example of GVF: (a) the original mammographic image, (b) the horizontal GVF component generated from (a), (c) the enhanced image through adaptive histogram equalization, (d) the horizontal GVF component of (c), (e) generated mass mask [23] © IEEE

Ferreira et al. [24] used active contour model (ACM) based on self-organizing network (SON) to segment the ROI. This model explores the principle of isomorphism and self-organization to create flexible contours that characterizes the shapes in the image.

Yuan et al. [25] employed a dual-stage method to extract masses from the surrounding tissues. Radial gradient index (RGI) based segmentation is used to yield an initial contour close to the lesion boundary location and a region-based active contour model is utilized to evolve the contour further to the lesion boundary.

2.1.4 Hybrid Techniques

Stochastic model-based image segmentation is a technique for partitioning an image into distinctive meaningful regions based on the statistical properties of both gray level and context images. Li et al. [26] employed a finite generalized Gaussian mixture (FGGM) distribution which is a statistical method for enhanced segmentation and extraction of suspicious mass areas. They used FGGM distribution to model mammographic pixel images together with a model selection procedure based on the two information theoretic criteria to determine the optimal number of image regions. Finally, they applied a contextual Bayesian relaxation labeling (CBRL) technique to perform the selection of the suspected masses. The examples of the segmentation results are shown in Figure 7.

Ball and Bruce [27] segmented suspicious masses in polar domain. They used adaptive level set segmentation method (ALSSM) to adaptively adjust the border threshold at each angle in order to provide high-quality segmentation results. They extended their work in [28] where they used spiculation segmentation with level sets (SSLS) to detect and segment spiculated masses. In conjunction with level set segmentation they used Dixon and Taylor line operator (DTLO) and a generalized version of DTLO (GDTLO).

Hassanien and Ali [29] developed an algorithm for segmenting spiculated masses based on pulse coupled neural networks (PCNN) in conjunction with fuzzy set theory.

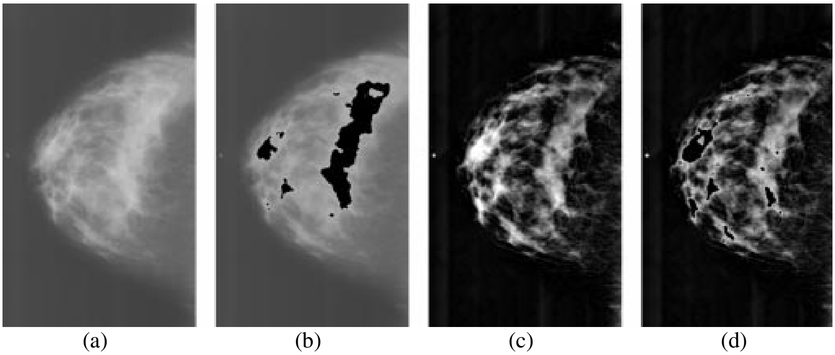


Fig. 7 Examples of normal mixed fatty and glandular mammogram: (a) Original mammogram, (b) Segmentation result based on the original mammogram, (c) Enhanced mammogram, (d) Result based on the enhanced mammogram [26] © IEEE

2.2 Feature Extraction and Selection

In the feature extraction and selection step the features that characterize specific region are calculated and the ones that are important are selected for the classification of the mass as benign or malignant. The feature space is very large and complex due to the wide diversity of the normal tissues and the variety of the abnormalities [6]. Some of the features are not significant when observed alone, but in combination with other features can be significant for classification. Li et al. [30] proposed general guidelines for feature extraction and selection of significant features: discrimination, reliability, independence and optimality. They divided features into three categories: intensity features, geometric features and texture features. Cheng et al. [6] gave a detailed list of features in each category.

Bellotti et al. [31] characterized ROI by means of textural features computed from the gray level co-occurrence matrix (GLCM), also known as spatial gray level dependence (SGLD) matrix.

Varela et al. [15] used features based on the iris filter output, together with gray level, texture, contour-related and morphological features. The best performance was provided with the combination of seven features. Namely, the maximum mean iris filter output, the mean value of the enhanced filter output, the average gray level value of the segmented region, isodense, size, eccentricity and compactness.

Yuan et al. [25] used three groups of features in their study. The first group included features characterizing spiculation, margin, shape and contrast of the lesion. The second group consisted of texture features and the third group included a distance feature calculated as a Euclidean distance from the nipple to the center of the lesion. They used a linear stepwise feature selection method with a Wilks lambda criterion to select a subset features for the classification task.

Sahiner et al. [20] developed an algorithm for extracting spiculation feature and circumsribed margin feature. Both features had high accuracy for characterizing mass margins according to BI-RADS descriptors.

Ball and Bruce [28] used features that include patient's age, morphological features, statistical features and features based on the segmentation boundary and the rubber band straightening transform.

Timp and Karssemeijer [32] proposed temporal feature set consisted of complete set of single view features together with temporal features. They extracted 38 single view features and grouped them into 12 main categories according to the type of characteristic they represent. The features from the current mammographic image are combined with the features calculated from the corresponding region in mammographic image taken in previous exam to provide temporal information. Temporal features are obtained by subtracting prior from current feature values which resulted in 29 temporal features. To select a best subset from the temporal feature set the sequential forward floating selection (SFFS) is used. In their later work, Timp et al. [33] designed two kinds of temporal features: difference features and similarity features. Difference features measured changes in feature values between corresponding regions in the prior and the current view. Similarity features measured whether two regions are comparable in appearance.

Fauci et al. [21] extracted 12 features from segmented masses. Some features gave the geometrical information, others provided shape parameters. The criterion for feature selection was based on morphological differences between pathological and healthy regions.

Rangayyan et al. [34] proposed methods to obtain shape features from the turning angle functions of contours. Features are useful in the analysis of contours of breast masses and tumors because of their ability to capture diagnostically important details of shape related to spicules and lobulations.

Li et al. [26] applied a contextual Bayesian relaxation labeling (CBRL) technique to perform the selection of suspected masses. The large improvement in classification was obtained after several iterations of CBRL algorithm.

Nandi et al. [35] used five stand-alone feature selection algorithms: Kullback-Leibler divergence (KLD), Kolmogorov-Smirnov test (K-S test), Students *t* test (*t* test), sequential forward selection (SFS) and sequential backward selection (SBS) to narrow the pool of features for classification step. They concluded that the shape measure of fractional concavity was the most important feature for the classifier.

Hupse and Karssemeijer [36] compared two feature selection criterions: Wilks lambda and the mean sensitivity of the FROC (free response operating characteristic) curve, both criterions with and without feature normalization. The feature selection method that performed best was the method in which Wilks lambda was used as selection criterion in combination with the use of normalized features.

Kim and Yoon [37] evaluated recursive feature elimination-based support vector machines (SVM-RFE) to improve classification accuracy. SVM-RFE incorporates feature selection in a recursive elimination manner to obtain a ranking of features that are particularly meaningful to SVMs and the top ranked features are chosen for classification. SVM-RFE has revealed that using only a subset of the 22 BI-RADS and gray level features facilitated increased CAD accuracy compared to using all 22 features.

2.3 Feature Classification

In feature classification step masses are classified as benign or malignant using the selected features. Various methods have been used for mass classifications. Some of the most popular techniques are artificial neural networks and linear discriminant analysis.

Varela et al. [15] merged the feature set into a backpropagation neural network (BNN) classifier to reduce the number of false positives. Their results yielded a sensitivity of 88% at an approximate false positive rate per image of 1 when considering lesion-based evaluation and sensitivity of 94% at 1.02 false positive findings per image when considering case-based evaluation.

Li et al. [38] merged the selected features using a Bayesian artificial neural network (BANN) classifier to generate an estimate of the probability of malignancy. The merged features showed a statistically significant improvement as compared to the individual features in the task of distinguishing between benign and malignant masses. The performance of the method yielded an A value under the ROC curve of 0.83 with a standard error of 0.02.

Fauci et al. [21] performed classification by means of an artificial neural network (ANN) with 12 input neurons, a number of hidden neurons which is tuned to obtain the best classification performance and one output neuron. The output neuron provides the probability that the ROI is pathological. Their adopted neural network was a feed-forward back-propagation supervised network trained with gradient descent learning rule with *momentum*. *Momentum* represents a sort of inertia which is added to quickly move along the direction of decreasing gradient, thus avoiding oscillations around secondary minima. Their results (A=0.85 with standard error of 0.08) were comparable with the performance obtained by commercial CAD [39].

Ball and Bruce [28] analyzed feature vector using generalized discriminant analysis (GDA) to provide a non-linear classification and to classify masses as spiculated or not. The features extracted from spiculated masses are classified as benign or malignant using k nearest neighbor (k-NN) and maximum likelihood (ML) classifiers. They showed that the k-NN classifier outperformed the ML classifier slightly in terms of higher overall accuracy and fewer numbers of false negatives. Using 1-NN or 2-NN classifier they achieved 93% overall accuracy with three FP and one FN. Using ML classifier they achieved 92% overall accuracy with three FP and two FN.

Nandi et al. [35] introduced genetic programming and adapted it for classification of masses. The genetic programming classifier performed well in discriminating between benign and malignant masses with accuracies above 99.5% for training and typically above 98% for testing.

Mu et al. [40] proposed a 2-plane learning method for binary classification, named as the strict 2-surface proximal (S2SP) classifier. They proposed the S2SP classifier for both linear and nonlinear pattern classification. The S2SP classifier improved the accuracy of discriminating between benign and malignant masses based on features that provided weak performance using classical pattern recognition methods. The linear classification yielded performance of A=0.97 and in the case of nonlinear classification the performance of A=1.0.

Li et al. [16] used fuzzy binary decision tree (FBDT) based on a series of radiographic, density-related features. They classified ROIs as normal or suspicious. Their results indicate that their approach might be particularly accurate and effective for small tumors (≤10 mm in size) which are not palpable or easily distinguishable in mammographic images. Their algorithm achieved 90% sensitivity with two false positives per image.

Krishnapuram et al. [41] proposed a multiple-instance learning (MIL) algorithm that automatically selects a small set of diagnostically useful features. The algorithm is more accurate than the support vector machine classifier.

For improving classification performance the classifier ensembles can be used. The classification decision is initially made by several separate classifiers and then combined into one final assessment. West et al. [42] investigated the effect of classifier diversity (the number of different classifiers in ensemble) on the generalization accuracy of the ensemble. Their results demonstrated that most of the improvement occurred with ensembles formed from 3-5 different classifiers. The most effective ensembles formed in their research resulted from a small and selective subset of the population of available classifiers, with potential candidates identified by jointly considering the properties of classifier generalization error, classifier instability and the independence of classifier decisions relative to other ensemble members.

3 Microcalcification Detection Algorithms

Calcifications are calcium deposits inside the breast. They can be roughly divided in two major groups: macrocalcification and microcalcifications. Macrocalcifications are, as expected, large calcium deposits, while microcalcifications are tiny calcium deposits. Macrocalcifications are usually not linked with the development of breast cancer and that is the reason why no special attention is being devoted to them. On the other hand, detection of microcalcifications is very important for the early breast cancer detection. Microcalcifications are usually associated with extra cell activity in the breast tissue. The extra cell activity does not have to be cancerous and it usually is not, but if microcalcifications are grouped in clusters, that can be a sign of developing malignant tumor. Scattered microcalcifications are usually a part of benign breast tissue. In mammograms calcifications are seen as bright dots of different sizes. The exact position of microcalcifications can not be predicted, as well as their number. Microcalcifications can be grouped in clusters, but also more often they are found to be stand-alone. Detection of microcalcifications is a very challenging task for radiologists as well as for computer-aided detection software. Development of the information technology and computers influenced mammography by giving the possibility of producing high quality digitized images with good resolution. Good spatial resolution is very important for microcalcification detection because of their actual size that can be as small as 100 μm. In digital mammography which is being mostly used today, displays with high resolution are necessary for delivering sharper images richer with details to radiologists.

CADe software tries to make diagnosis process easier and almost automatically. In mammography applications, one of the most important tasks for CADe is to detect the presence of microcalcifications, especially clustered ones, because they can be the early sign of possible cancer. Since microcalcifications are small and randomly scattered in breast tissue it is possible for a radiologist to overlook them. In that case CADe software should give good results by producing less false negative (FN) results. There also lays the biggest problem of CADe software, because radiologists could possibly overlook some microcalcifications trusting the software detection accuracy too much, which would again give the false negative results. The general microcalcification detection process is shown in Figure 8. After the image enhancement, region of interest (ROI) should be detected. Feature extraction and selection are the next two steps. Finally, the decision algorithm based on selected features provides detection.

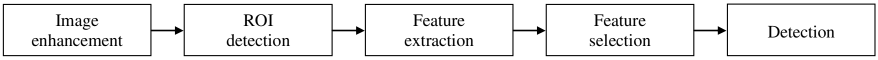


Fig. 8 The general microcalcification detection algorithm

During the past two decades, many methods for microcalcification detection have been presented. In this chapter only few commonly used methods published in recent papers will be described. Methods that will be presented combine the use of wavelet analysis, contrast enhancement, noise equalization, higher order statistics and classifications for benign and malignant differentiation.

3.1 Wavelet Detection Methods

Many of the recently presented methods for microcalcification detection use wavelet-based algorithms [43-50]. The beginning of wavelet method usage for microcalcification detection was in late 1990s. Wavelet-based subband image decomposition is used to detect and isolate clusters of microcalcifications from the surrounding tissue. Since microcalcifications are rather small objects, they correspond to the high-frequency components. Standard dyadic wavelet decomposition filters the original image to the desired level producing sub-images. Sub-images can contain combinations of lowpass and highpass filter components. Next step is to determine which of these new sub-images contains the best results. After suppressing low-frequency components using wavelet decomposition filters of the desired level, the image is reconstructed and the process of microcalcification detection can begin. This is the generalized approach used in most wavelet-based methods. Differences between them lay in different decomposition and detection process but also in the most important step and that is the reduction of many false positive results. FP results can occur very often in microcalcification detection because detection threshold should be set rather sensitive in order to detect as many high-frequency objects as possible since microcalcifications can be scattered in the breast tissue.

Wavelet transform is used to construct time-frequency representation of a certain dataset. Fourier transform gives only frequency content but can not localize objects of certain frequency in the image. Wavelet transform is therefore superior because it gives both frequency content and exact position of the object in the image. Discrete wavelet transform (DWT) is commonly used in image processing. The dyadic wavelet transform decomposes the original image into sub-images using the desired wavelet function called "mother wavelet" that is scaled to get so called "daughter wavelets" and translated through the image. Decomposition at each scale subdivides the frequency range as shown in Figure 9.

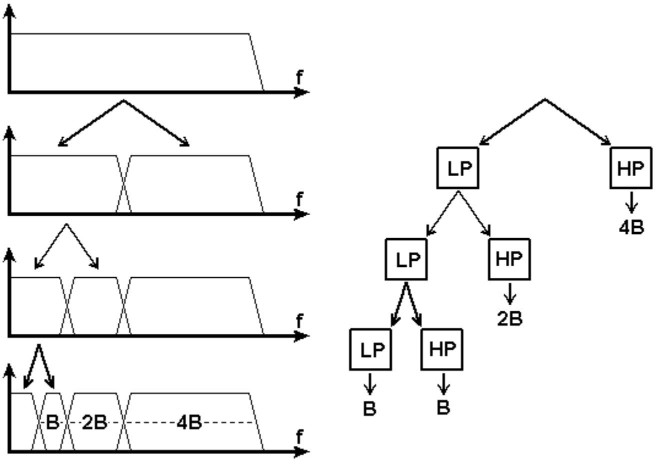


Fig. 9 Frequency range subdivision obtained with dyadic wavelet decomposition up to third level

The sub-images obtained using wavelet decomposition are often noted with "LL", "HL", "LH" and "HH". "LL" is the approximation image, "LH" and "HL" are the horizontal and vertical detail images and "HH" is the diagonal detail image. Figure 10 shows the analogy of notation and corresponding images (test image "Lena" in this example) after first level wavelet decomposition.

LL	HL
LH	HH

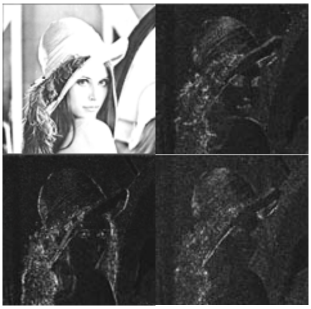


Fig. 10 The analogy of notation and corresponding images

In 1996 Strickland and Hahn [43] presented the two-stage method based on wavelet transform for detecting and segmenting calcifications. They have used HH and the combination of LH+HL sub-bands in the detection process. Detected pixel in HH and LH+HL are dilated and then weighted before proceeding with the inverse wavelet transform. By this, individual microcalcifications are greatly enhanced in the output image. After that the straightforward thresholding can be applied to segment microcalcifications.

In 1998 Wang and Karayiannis [44] proposed a very similar method that uses wavelet decomposition in order to suppress low-frequency components. These components are suppressed by reconstructing the image from HH, LH and HL sub-bands making high-frequency objects distinct from the background. The proposed method uses Daubechies wavelets; one is “DAUB 4” with only 4 coefficients and the second is “DAUB 20” with 20 coefficients. Shorter ("DAUB 4") wavelet filters produced more high frequency results as the output with more false positives than longer ones ("DAUB 20").

Salvado et al. proposed another method for the microcalcification detection that uses wavelet analysis and contrast enhancement [45]. Wavelets are used here again to avoid the tradeoff between time and frequency resolution in Fourier representation. The proposed method has the following steps: histogram analysis, 2D DWT analysis, noise removal and low-frequency band elimination, image enhancement and finally image reconstruction. The DWT uses Daubechies-6 orthogonal wavelet with 10 levels of decomposition. Both contrast enhancement operators, linear and multiscale adaptive non-linear, are integrated in the wavelet transform. For result verification the MIAS database [46] was used. Figure 11(a) and 11(b) shows results that this method produces on a mammogram with localized dense tissues. Microcalcifications after the enhancement procedure are shown inside the ellipse in Figure 11(b).

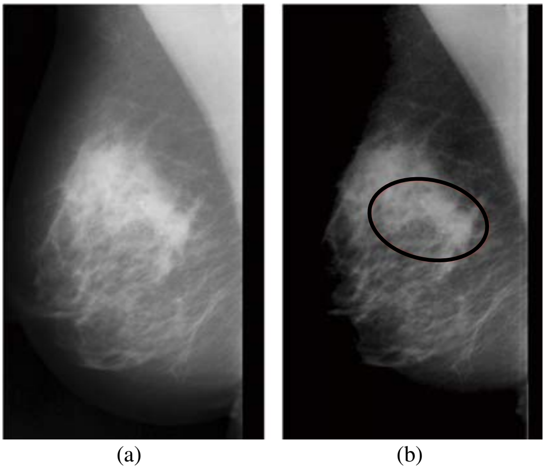


Fig. 11 (a) Original mammographic image, (b) results of the proposed method for microcalcification detection with contrast enhancement [45] © IEEE

The computerized scheme for detection of the microcalcifications clusters in mammographic images using wavelets is presented in [47]. The proposed detection algorithm consists of four steps: creation of negative image, decomposition of the negative image using wavelet transform, creating a binary image of approximation coefficients and pre-detecting microcalcifications and finally, identifying clusters of pixels after applying the threshold. Each image has been decomposed using the Daubechies Wavelets (db2, db4, db8 and db16). The detection accuracy is claimed to be up to 80%.

The method that combines the use of wavelet transform and morphology is presented in [48]. The results of both morphology and wavelet transform are combined using the logical AND operation. This approach gives lower true positive rate (TPR) but at the same time less false positives (FPs). Results obtained using MIAS database gave TPR of 80.2% with 2.5 false microcalcifications per region of interest.

Wavelets can also be used for enhancement of microcalcifications in mammograms [49]. For image enhancement, this method proposes 3 steps. First step is computing an adapted multiresolution decomposition of the image into wavelet coefficients using integrated wavelet transform. Second step is applying a local enhancement operator *E* on the calculated wavelet coefficients. The final step is image reconstruction. In this method, microcalcifications are approximated by a Gaussian form. Enhancement is done on the discrete decompositions, called integrated wavelets. Figure 12(a) shows original image and Figure 12(b) shows results of microcalcifications enhancement obtained using the proposed method. The method is tested on the image set provided by Department of Radiology at the University of Nijmegen, the Netherlands.

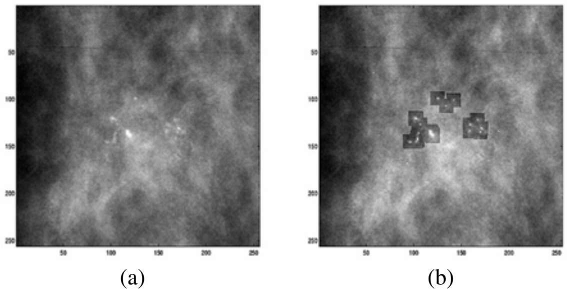


Fig. 12 (a) Original image, (b) results of microcalcification enhancement [49] © IEEE

Another method for image enhancement and denoising that uses wavelets has been presented in [50]. Enhancement of microcalcifications and suspicious masses is done using adaptive gain algorithm and fine noise estimation. The adaptive gain is calculated at each scale to adequately enhance coefficients at each level of decomposition. For image denoising, a wavelet shrinkage denoising algorithm with adaptive threshold setting is applied. The proposed method is tested on DDSM database.

3.2 Other Recently Proposed Methods

Besides the use of wavelets contrast enhancement methods with noise estimation, other approaches have also been used to detect microcalcifications. Sankar and Thomas proposed the method that uses fractal modeling of mammograms based on mean and variance to detect microcalcifications [51]. This method was tested on 28 mammograms from the MIAS database and produced the following results: TPR=82% with an average of 0.214 negative clusters per image.

The semiautomatic segmentation method [52] allows some interaction of the radiologist. For the detection of microcalcification this method also uses Daubechies 6 wavelets as a central component of the system. Detected components that have a higher or lower spatial frequency than the spatial frequency of calcifications are zeroed. The result is band pass filtered version of the input image. Segmentation of calcifications can be done by applying a threshold on the filtered image. Instead of that approach, this method proposes an interactive segmentation stage done by the radiologist. The results of the fully automatic and the semiautomatic segmentation on images from the DDSM base [53] are evaluated using areas under the ROC curve. The fully automatic segmentation gave results of A=0.80 while the semiautomatic gave significantly higher results of A=0.84.

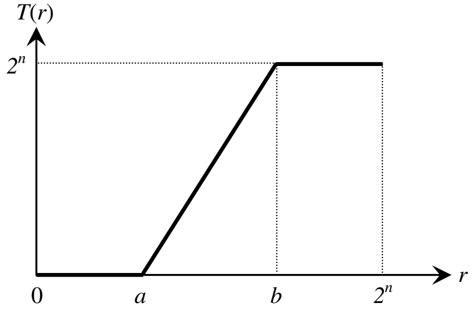
Another microcalcification detection method [54] presented in 2007 uses a different approach. The first stage of the process is extraction of zones that potentially correspond to microcalcifications by analyzing the distribution of brightness over the mammogram. The second stage is identification of clusters as ROIs. The final stage is retrieving the information that might have been lost in the previous stages. The method is tested using DDSM database.

Detection of microcalcification by meta-heuristic algorithms was proposed by Thangavel and Karnan [55]. This method uses the meta-heuristic methods such as Ant Colony Optimization (ACO) and Genetic Algorithm (GA) for identification of suspicious regions in mammograms. The method relies on the property of bilateral asymmetry. If the structural asymmetries between the left and the right breast are stronger, possibilities for microcalcifications are higher. Bilateral subtraction is used to determine the structural asymmetry. In the first step mammogram images are enhanced using median filter, pectoral muscle region is removed and breast border is detected. GA is then applied to enhance the detected border. The next step is image alignment using the border points and nipple position. After that images are subtracted to extract the suspicious region. The algorithms are tested on the entire MIAS database (322 mammograms which equals to 161 pairs). Results are presented using FROC (Free-Response Receiver Operating Characteristics) curve. The authors claim this method achieves the area under the curve A=0.94 for the proposed algorithm, which are very good results for a fully automatic method.

3.3 Contrast Enhancement and Noise Equalization

Contrast enhancement is of very high importance in x-ray imaging. It helps in making diagnosis more accurate. Contrast enhancement can be done globally and locally. Global contrast enhancement uses transforming function which can be shown as a look up table (LUT). One of the simplest examples is contrast stretching. The transforming function for linear rescaling shown in Figure 13 stretches the part of the image histogram where amplitudes that contain important information are placed across the whole amplitude range [56]. Figure 13 shows transforming function that takes values from *r* and stretches [*a*, *b*] to [0, 2^n], where *T(r)* is the transforming function, *a* the amplitude that will be displayed as black at the output and *b* the amplitude that will be displayed as 100% white at the output.

Fig. 13 Transforming ramp-function for contrast enhancement [56]



As expected, global contrast enhancement will do the change in image contrast regardless of image contents. Global contrast enhancement can generally be observed as some kind of histogram equalization method. Local contrast enhancement methods are more suitable in the field of digital mammography. The main reason for that is the size and uniformity of the image. Mammographic images, as well as other types of x-ray images have histograms of similar shape. There are two sets of components that dominate in the histogram. There are components that make the background (amplitudes around zero) and components of which the objects are consisted. For contrast enhancement, both global and local, background should be removed to make the set of amplitudes in image histogram narrower. In many cases pectoral muscle is also removed because it presents a large and rather uniform area so it can present an additional problem in histogram based contrast enhancement. Earlier presented methods like [57] used global gray level threshold as an initial processing stage and then a local adaptive thresholding technique. The method proposed by [58] uses local thresholding calculating the difference between the local maximum and mean gray levels. The aim of this approach is to highlight all bright image structures. Regions in the image should be properly enhanced, because under-enhancement can result in false negatives (FNs) and over-enhancement in false positives (FPs) [59]. Contrast enhancement is necessary for making microcalcifications stand out from the breast tissue in a dense breast.

Noise equalization is a very important step in the process of microcalcification detection. Microcalcification can very easily be mixed with image noise and therefore not detected or produce a false-positive result. Most methods described in literature use some kind of noise-dependent thresholding. In some cases threshold is determined locally and in some globally. Noise in images occurs mostly because of the fluctuations in photon fluence at the detector. In digital x-ray images in general, quantum noise is dominant. Pixel values in digital mammography are linearly proportional to the amount of detected photons. Due to the detector inhomogeneity, the anode heel effect and other sources of variation, noise properties vary across the image. Because of that, noise across the image can not be modeled with the fixed constant. Modeling with the fixed constant would lead to over- or underestimated image noise and could produce more FP or FN results. Nonuniform noise model proposed by G. van Schie and N. Karssemeijer takes the properties of noise variation into account. This method is based on subdividing a mammogram into square tiles of adaptive width and finding a square root noise model per tile [60].

3.4 Classification of Microcalcifications

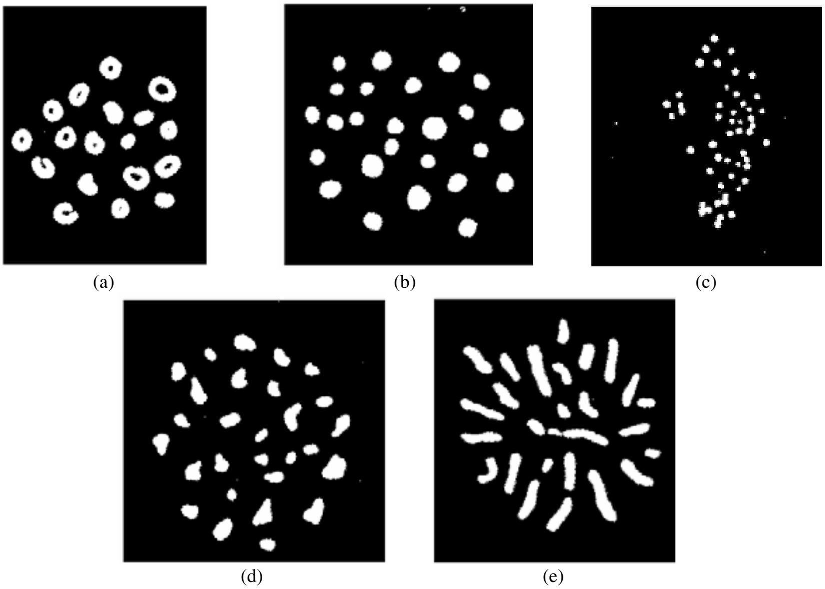
Besides detecting microcalcifications, another challenging task is automatic classification of microcalcifications. Classification should give the answer whether microcalcification is benign or malignant. For classification purposes, many classifiers have been used. Some commonly used classification methods are: neural networks, Bayesian classification, K-nearest neighbor classifiers, support vector machine and different decision trees. In this chapter only a few classifiers presented in the recent time will be described. De Santo et al. [61] used multiple classifier system. One classifier is devised for the classification of the single microcalcifications while the second one classifies the entire cluster. The classifier for single microcalcification evaluates the following features: compactness, roughness, border gradient strength and local contrast. The classifier for clusters of microcalcifications evaluates the following features: mass density of the cluster, average mass of the microcalcifications and the centre mass of the cluster, standard deviation of the masses of the microcalcifications and standard deviation of distance between microcalcifications and center of mass. Some typical microcalcification shapes of different form and possibility to be malignant are shown in Figure 14.

Combining these two proposed classifiers into the “Multiple Expert System” gave better results than each classifier by itself and the total recognition ratio of about 75% for benign and malignant clusters.

Support vector machine (SVM) is a form of machine learning algorithm. It is directly derived from the statistical learning theory [62]. SVM is based on the principle of risk minimization that is conducted by minimizing the generalization error. Generalization error is made by the learning machine on the test data set that is different from the training data set and has no overlapping [63]. To make SVM function properly, i.e. to avoid overfitting, the decision boundary should not correspond too good to the training data set. Some special user defined parameters are presented to avoid the possible overfitting.

Another approach in microcalcification classification uses content-based image retrieval technique [64]. The proposed method consists of two steps:

1. retrieving similar mammogram images from a database by using learning based similarity measure;
2. classifying the query mammogram image based on retrieved results (retrieval-driven classification).



(a) (b) (c)

(d) (e)

Fig. 14 Some typical microcalcification shapes (adapted from [61]): (a) rings-always benign, (b) circular, with various sizes, smooth border-frequently benign, (c) pulverulent-sometimes malignant, (d) granule-like, with various sizes and shapes, irregular border-malignant almost always, (e) vermicular-typically malignant © Elsevier B.V.

Similar cases are being used to improve a numerical classifier's performance and adaptive support vector machine is being used to improve classification performance. Experimental setup to test this method used 600 image pairs that have been scored in a human observer study for training the similarity function. 200 mammograms were then used as test samples. Results representation has been done using ROC curves. Content based image-retrieval technique gave the following results: A=0.7752 using SVM and A=0.8223 using adaptive SVM.

A different approach in microcalcification classification was presented by Hadjiiski et al. [65]. Their work is based on development of CAD systems for assisting radiologists in classification of breast lesions. After detection and segmentation of microcalcification, the classification has been done using 5 morphological features that describe size, density and shape. The presented method gave results of A=0.77 for 117 two-view pairs what seems to be slightly better than the expert radiologist classification.

A new particle swarm optimization algorithm for feature selection has also been recently presented [66]. For the segmentation of microcalcification from the enhanced mammographic image the New Particle Swarm Optimization (NPSO) algorithm hybrid with Markov Random Field (MRF) is being used. Classification is being done using a three-layer Backpropagation Network hybrid with NPSO (BPN-NPSO). Performance of the algorithm was evaluated also using the ROC analysis like the two methods previously described. The results show that the NPSO algorithm selects features better than Genetic Algorithm.

In the process of microcalcification detection, wavelet based methods are being mostly used. Proposed methods work more or less in the same way and give acceptable results. Microcalcifications are very small objects and to detect them it is necessary to extract high frequency components. Wavelet transform, as mentioned before, also gives the spatial information of the detected object and that is the main reason why it is so successful in this area. Also there have been proposed some other methods for microcalcification detection but their number is significantly smaller. Those other methods presented in this chapter are contrast based, one that uses particle swarm optimization, noise estimation and some that use combination of two or more approaches. Automatic classification is another issue that needs to be solved. There have been some different approaches using different classifiers. The future is to show if it is possible to obtain areas under ROC curve greater than 0.8.

4 Conclusion

Breast cancer is one of the major causes of death among women. Digital mammography screening programs can enable early detection and diagnose of the breast cancer which reduces the mortality and increases the chances of complete recovery. Screening programs produce a great amount of mammographic images which have to be interpreted by radiologists. Due to the wide range of breast abnormalities' features some abnormalities may be missed or misinterpreted. There is also a number of false positive findings and therefore a lot of unnecessary biopsies. Computer-aided detection and diagnosis algorithms have been developed to help radiologists give an accurate diagnosis and to reduce the number of false positives. There are a lot of algorithms developed for detection of masses and calcifications. In this chapter, algorithms that are commonly used and the ones recently developed were presented. Over the years there has been an improvement in the detection algorithms but their performance is still not perfect. The area under the ROC curve is rarely above 90% which means that there are still many false positive outputs. Possible reason for such a performance may be the characteristics of breast abnormalities. Masses and calcifications are sometimes superimposed and hidden in the dense tissue which makes the segmentation of correct regions of interest difficult. Another issue is extracting and selecting appropriate features that will give the best classification results. Furthermore, the choice of a classifier has a great influence on the final result and classifying abnormalities as benign or malignant is a difficult task even for expert radiologists. Further developments in each algorithm step are required to improve the overall performance of computer-aided detection and diagnosis algorithms.

Acknowledgments

The work in this chapter was conducted under the research project "Intelligent Image Features Extraction in Knowledge Discovery Systems" (036-0982560-1643), supported by the Ministry of Science, Education and Sports of the Republic of Croatia.

The authors would like to thank Prof. Boris Brljacic, Dr. Renata Huzjan Korunic and Mr. Milan Grzan from the Department of Diagnostic and Interventional Radiology, University Hospital "Dubrava", Zagreb, Croatia, for providing digital mammographic images.

References

[1] Acha, B., Rangayyan, R.M., Desautels, J.E.L.: Detection of Microcalcifications in Mammograms. In: Suri, J.S., Rangayyan, R.M. (eds.) Recent Advances in Breast Imaging, Mammography, and Computer-Aided Diagnosis of Breast Cancer. SPIE, Bellingham (2006)

[2] American College of Radiology (ACR): ACR Breast Imaging Reporting and Data System, Breast Imaging Atlas, 4th edn., Reston, VA, USA (2003)

[3] Rangayyan, R.M., Ayres, F.J., Desautels, J.E.L.: A Review of Computer-Aided Diagnosis of Breast Cancer: Toward the Detection of Subtle Signs. Journal of the Franklin Institute 344(3-4), 312–348 (2007)

[4] Sampat, M.P., Markey, M.K., Bovik, A.C.: Computer-Aided Detection and Diagnosis in Mammography. In: Bovik, A.C. (ed.) Handbook of Image and Video Processing. Elsevier Academic Press, Amsterdam (2005)

[5] de Paredes, E.S.: Atlas of Mammography, 3rd edn. Lippincott Williams & Wilkins, Philadelphia (2007)

[6] Cheng, H.D., Shi, X.J., Min, R., Hu, L.M., Cai, X.P., Du, H.N.: Approaches for Automated Detection and Classification of Masses in Mammograms. Pattern Recognition 39(4), 646–668 (2006)

[7] Rangayyan, R.M.: Biomedical Image Analysis. CRC Press LLC, Boca Raton (2005)

[8] Jain, A.K., Duin, R.P.W., Mao, J.: Statistical Pattern Recognition: A Review. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(1), 4–37 (2000)

[9] Metz, C.E.: Basic principles of ROC analysis. Seminars in Nuclear Medicine, 283–298 (October 1978)

[10] Long, P.M., Servedio, R.A.: Boosting the Area Under the ROC Curve. In: Advances in Neural Information Processing Systems 20, Conference Proceedings (December 2007)

[11] Brzakovic, D., Luo, X.M., Brzakovic, P.: An approach to automated detection of tumors in mammograms. IEEE Transactions on Medical Imaging 9(3), 233–241 (1990)

[12] Li, L.H., Qian, W., Clarke, L.P., Clark, R.A., Thomas, J.: Improving Mass Detection by Adaptive and Multi-Scale Processing in Digitized Mammograms. Proceedings of SPIE—The International Society for Optical Engineering 3661 1, 490–498 (1999)

[13] Matsubara, T., Fujita, H., Endo, T., et al.: Development of Mass Detection Algorithm Based on Adaptive Thresholding Technique in Digital Mammograms. In: Doi, K., Giger, M.L., et al. (eds.) pp. 391–396. Elsevier, Amsterdam (1996)

[14] Dominguez, A.R., Nandi, A.F.: Enhanced Multi-Level Thresholding Segmentation and Rank Based Region Selection for Detection of Masses in Mammograms. In: IEEE International Conference on Acoustics, Speech and Signal Processing 2007, ICASSP 2007, Honolulu, HI, April 15-20, pp. 449–452 (2007)

[15] Varela, C., Tahoces, P.G., Méndez, A.J., Souto, M., Vidal, J.J.: Computerized Detection of Breast Masses in Digitized Mammograms. Computers in Biology and Medicine 37, 214–226 (2007)

[16] Li, H.D., Kallergi, M., Clarke, L.P., Jain, V.K., Clark, R.A.: Markov Random Field for Tumor Detection in Digital Mammography. IEEE Transactions on Medical Imaging 14(3), 565–576 (1995)

[17] Székely, N., Tóth, N., Pataki, B.: A Hybrid System for Detecting Masses in Mammographic Images. IEEE Transactions on Instrumentation and Measurement 55(3), 944–951 (2006)

[18] Zheng, B., Mello-Thoms, C., Wang, X.H., Gur, D.: Improvement of Visual Similarity of Similar Breast Masses Selected by Computer-Aided Diagnosis Schemes. In: 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI 2007, April 12-15, pp. 516–519 (2007)

[19] Pappas, T.N.: An Adaptive Clustering Algorithm for Image Segmentation. IEEE Transactions on Signal Processing 40(4), 901–914 (1992)

[20] Sahiner, B., Hadjiiski, L.M., Chan, H.P., Paramagul, C., Nees, A., Helvie, M., Shi, J.: Concordance of Computer-Extracted Image Features with BI-RADS Descriptors for Mammographic Mass Margin. In: Giger, M.L., Karssemeijer, N. (eds.) Proc. of SPIE Medical Imaging 2008: Computer-Aided Diagnosis, vol. 6915 (2008)

[21] Fauci, F., Bagnasco, S., Bellotti, R., Cascio, D., Cheran, S.C., De Carlo, F., De Nunzio, G., Fantacci, M.E., Forni, G., Lauria, A., Torres, E.L., Magro, R., Masala, G.L., Oliva, P., Quarta, M., Raso, G., Retico, A., Tangaro, S.: Mammogram Segmentation by Contour Searching and Massive Lesion Classification with Neural Network. In: 2004 IEEE Nuclear Science Symposium Conference Record, Rome, Italy, October 16–22, vol. 5, pp. 2695–2699 (2004)

[22] Petrick, N., Chan, H.P., Sahiner, B., Wei, D.: An Adaptive Density Weighted Contrast Enhancement Filter for Mammographic Breast Mass Detection. IEEE Transactions on Medical Imaging 15(1), 59–67 (1996)

[23] Zou, F., Zheng, Y., Zhou, Z., Agyepong, K.: Gradient Vector Flow Field and Mass Region Extraction in Digital Mammograms. In: 21st IEEE International Symposium on Computer-Based Medical Systems, CMBS 2008, Jyvaskyla, June 17-19, pp. 41–43 (2008)

[24] Ferreira, A.A., Nascimento Jr., F., Tsang, I.R., Cavalcanti, G.D.C., Ludermir, T.B., de Aquino, R.R.B.: Analysis of Mammogram Using Self-Organizing Neural Networks Based on Spatial Isomorphism. In: Proceedings of International Joint Conference on Neural Networks, IJCNN 2007, Orlando, Florida, USA, August 12-17, pp. 1796–1801 (2007)

[25] Yuan, Y., Giger, M.L., Li, H., Sennett, C.: Correlative Feature Analysis of FFDM Images. In: Giger, M.L., Karssemeijer, N. (eds.) Proc. of SPIE Medical Imaging 2008: Computer-Aided Diagnosis, vol. 6915 (2008)

[26] Li, H., Wang, Y., Liu, K.J.R., Lo, S.B., Freedman, M.T.: Computerized Radiographic Mass Detection C Part I: Lesion Site Selection by Morphological Enhancement and Contextual Segmentation. IEEE Transactions on Medical Imaging 20(4), 289–301 (2001)

[27] Ball, J.E., Bruce, L.M.: Digital Mammographic Computer Aided Diagnosis (CAD) using Adaptive Level Set Segmentation. In: Proceedings of the 29th Annual International Conference of the IEEE EMBS, Cité Internationale, Lyon, France, August 23-26, pp. 4973–4978 (2007)

[28] Ball, J.E., Bruce, L.M.: Digital Mammogram Spiculated Mass Detection and Spicule Segmentation using Level Sets. In: Proceedings of the 29th Annual International Conference of the IEEE EMBS, Cité Internationale, Lyon, France, August 23-26, pp. 4979–4984 (2007)

[29] Hassanien, A.E., Ali, J.M.: Digital Mammogram Segmentation Algorithm Using Pulse Coupled Neural Networks. In: Proceedings of the Third International Conference on Image and Graphics, ICIG 2004 (2004)

[30] Li, H., Wang, Y., Ray Liu, K.J., Lo, S.C.B., Freedman, M.T.: Computerized Radiographic Mass Detection—Part II: Decision Support by Featured Database Visualization and Modular Neural Networks. IEEE Transactions on Medical Imaging 20(4) (April 2001)

[31] Bellotti, R., De Carlo, F., Tangaro, S., Gargano, G., Maggipinto, G., Castellano, M., Massafra, R., Cascio, D., Fauci, F., Magro, R., Raso, G., Lauria, A., Forni, G., Bagnasco, S., Cerello, P., Zanon, E., Cheran, S.C., Lopez Torres, E., Bottigli, U., Masala, G.L., Oliva, P., Retico, A., Fantacci, M.E., Cataldo, R., De Mitri, I., De Nunzio, G.: A Completely Automated CAD System for Mass Detection in a Large Mammographic Database. Medical Physics 33(8), 3066–3075 (2006)

[32] Timp, S., Karssemeijer, N.: Interval Change Analysis to Improve Computer Aided Detection in Mammography. Medical Image Analysis 10, 82–95 (2006)

[33] Timp, S., Varela, C., Karssemeijer, N.: Temporal Change Analysis for Characterization of Mass Lesions in Mammography. IEEE Transactions on Medical Imaging 26(7), 945–953 (2007)

[34] Rangayyan, R.M., Guliato, D., de Carvalho, J.D., Santiago, S.A.: Feature Extraction from the Turning Angle Function for the Classification of Contours of Breast Tumors. In: IEEE Special Topic Symposium on Information Technology in Biomedicine, Iaonnina, Greece, October 2006, 4 pages (2006) CDROM

[35] Nandi, R.J., Nandi, A.K., Rangayyan, R.M., Scutt, D.: Genetic Programming and Feature Selection for Classification of Breast Masses in Mammograms. In: Proceedings of the 28th IEEE EMBS Annual International Conference, New York City, USA, August 30-September 3 (2006)

[36] Hupse, R., Karssemeijer, N.: Feature Selection for Computer-Aided Detection: Comparing Different Selection Criteria. In: Giger, M.L., Karssemeijer, N. (eds.) Proc. of SPIE Medical Imaging 2008: Computer-Aided Diagnosis, vol. 6915, 6915 691503-1 (2008)

[37] Kim, S., Yoon, S.: Mass Lesions Classification in Digital Mammography using Optimal Subset of BI-RADS and Gray Level Features. In: 6th International Special Topic Conference on ITAB, 2007, Tokyo, pp. 99–102. IEEE, Los Alamitos (2008)

[38] Li, H., Giger, M.L., Yuan, Y., Lan, L., Sennett, C.A.: Performance of CADx on a Large Clinical Database of FFDM Images. In: Krupinski, E.A. (ed.) IWDM 2008. LNCS, vol. 5116, pp. 510–514. Springer, Heidelberg (2008)

[39] Malich, A., Marx, C., Facius, M.: Tumour Detection Rate of a New Commercialy Available Computer- Aided Detection System. Eur. Radiology 11(12), 2454–2459 (2001)

[40] Mu, T., Nandi, A.K., Rangayyan, R.M.: Strict 2-Surface Proximal Classifier with Application to Breast Cancer Detection in Mammograms. In: IEEE ICASSP 2007, pp. II 477–480 (2007)

[41] Krishnapuram, B., Stoeckel, J., Raykar, V., Rao, B., Bamberger, P., Ratner, E., Merlet, N., Stainvas, I., Abramov, M., Manevitch, A.: Multiple-Instance Learning Improves CAD Detection of Masses in Digital Mammography. In: Krupinski, E.A. (ed.) IWDM 2008. LNCS, vol. 5116, pp. 350–357. Springer, Heidelberg (2008)

[42] West, D., Mangiameli, P., Rampal, R., West, V.: Ensemble Strategies for a Medical Diagnostic Decision Support System: a Breast Cancer Diagnosis Application. European Journal of Operational Research 162, 532–551 (2005)

[43] Strickland, R.N., Hahn, H.I.: Wavelet transforms for detecting microcalcifications in mammograms. IEEE Transactions on Medical Imaging 15(2), 218–229 (1996)

[44] Wang, T.C., Karayiannis, N.B.: Detection of Microcalcifications in Digital Mammograms Using Wavelets. IEEE Transactions on Medical Imaging 17(4), 498–509 (1998)

[45] Salvado, J., Roque, B.: Detection of Calcifications in Digital Mammograms using Wavelet Analysis and Contrast Enhancement. In: IEEE International Workshop on Intelligent Signal Processing 2005, Faro, Portugal, September 2005, pp. 200–205 (2005)

[46] Suckling, J., Parker, J., Dance, D.R., Astley, S., Hutt, I., Boggis, C.R.M., Ricketts, I., Stamatakis, E., Cernaez, N., Kok, S.L., Taylor, P., Betal, D., Savage, J.: The Mammographic Image Analysis Society Digital Mammogram Database. In: Proceedings of the 2nd International Workshop on Digital Mammography, York, England, July 10-12, pp. 375–378. Elsevier Science, Amsterdam (1994)

[47] Juarez, L.C., Ponomaryov, V., Sanchez, R.J.L.: Detection of Microcalcifications in Digital Mammograms Images Using Wavelet Transform. In: Electronics, Robotics and Automotive Mechanics Conference, September 2006, vol. 2, pp. 58–61 (2006)

[48] Song, L., Wang, Q., Gao, J.: Microcalcification detection using combination of wavelet transform and morphology. In: Proceedings of the 8th International Conference on Signal Processing, ICSP 2006, vol. 4, pp. 16–20 (2006)

[49] Heinlein, P., Drexl, J., Schneider, W.: Integrated Wavelets for Enhancement of Microcalcifications in Digital Mammograph. IEEE Transactions on Medical Imaging 22(3), 402–413 (2003)

[50] Mencattini, A., Salmeri, M., Lojacono, R., Frigerio, M., Caselli, F.: Mammographic Images Enhancement and Denoising for Breast Cancer Detection Using Dyadic Wavelet Processing. IEEE Transactions on Instrumentation and Measurement 57(7), 1422–1430 (2008)

[51] Sankar, D., Thomas, T.: Fractal Modeling of Mammograms based on Mean and Variance for the Detection of Microcalcifications. In: Proceedings of the 2007 International Conference on Computational Intelligence and Multimedia Applications, Sivakasi, India, December 2007, pp. 334–338 (2007)

[52] Elter, M., Held, C.: Semiautomatic segmentation for the computer aided diagnosis of clustered microcalcifications. In: Proc. SPIE, San Diego, CA, USA, February 2008, vol. 6915, 691524-691524-8 (2008)

[53] Heath, M., Bowyer, K., Kopans, D., Moore, R., Kegelmeyer Jr., P.: The Digital Database for Screening Mammography. In: Proceedings of the 5th International Workshop on Digital Mammography, Canada, June 11-14, pp. 212–218. Medical Physics Publishing (2001)

[54] Lopez-Aligue, F.J., Poveda-Pierola, A., Acevedo-Sotoca, I., Garcia-Urra, F.: Detection of Microcalcifications in Digital Mammograms. In: 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS 2007, Lyon, France, August 22-26, pp. 3906–3909 (2007)

[55] Thangavel, K., Karnan, M.: Computer Aided Diagnosis in Digital Mammograms: Detection of Microcalcifications by Meta Heuristic Algorithms. GVIP Journal 5(7), 41–55 (2005)

[56] Mustra, M., Grgic, M., Delac, K.: Efficient Presentation of DICOM Mammography Images using Matlab. In: Proceedings of the 15th International Conference on Systems, Signals and Image Processing (IWSSIP), Bratislava, Slovakia, June 25-28, pp. 13–16 (2008)

[57] Nishikawa, R.M., Giger, K.L., Doi, K., Vyborny, C.J., Schmidt, R.A.: Computer-aided detection of clustered microcalcifications on digital mammograms. Medical & Biological Engineering & Computing 33(2), 174–178 (1995)

[58] Neiber, H., Müller, T., Stotzka, R.: Local Contrast Enhancement for the Detection of Microcalcifications. In: IWDM 2000, Canada, pp. 598–604 (2000)

[59] Cheng, H.D., Cai, X., Chen, X., Hu, L., Lou, X.: Computer-aided detection and classification of microcalcifications in mammograms: a survey. Pattern Recognition 36(12), 2967–2991 (2003)

[60] van Schie, G., Karssemeijer, N.: Detection of Microcalcifications Using a Nonuniform Noise Model. In: Krupinski, E.A. (ed.) IWDM 2008. LNCS, vol. 5116, pp. 378–384. Springer, Heidelberg (2008)

[61] De Santo, M., Molinara, M., Tortorella, F., Vento, M.: Automatic classification of clustered microcalcifications by a multiple expert system. Pattern Recognition 36(7), 1467–1477 (2003)

[62] Vapnik, V.: Statistical Learning Theory. Wiley, New York (1998)

[63] Wei, L., Yang, Y., Nishikawa, R.M., Jiang, Y.: A study on several Machine-learning methods for classification of Malignant and benign clustered microcalcifications. IEEE Transactions on Medical Imaging 24(3), 371–380 (2005)

[64] Yang, Y., Wei, L., Nishikawa, R.M.: Microcalcification Classification Assisted by Content-Based Image Retrieval for Breast Cancer Diagnosis. In: IEEE International Conference on Image Processing 2007, ICIP 2007, September 16-19, vol. 5, pp. 1–4 (2007)

[65] Hadjiiski, L., Filev, P., Chan, H.-P., Ge, J., Sahiner, B., Helvie, M.A., Roubidoux, M.A.: Computerized Detection and Classification of Malignant and Benign Microcalcifications on Full Field Digital Mammograms. In: Krupinski, E.A. (ed.) IWDM 2008. LNCS, vol. 5116, pp. 336–342. Springer, Heidelberg (2008)

[66] Geetha, K., Thanushkodi, K., Kishore Kumar, A.: New Particle Swarm Optimization for Feature Selection and Classification of Microcalcifications in Mammograms. In: International Conference on Signal Processing Communications and Networking, ICSCN 2008, January 4-6, pp. 458–463 (2008)

Author Index