Detecting Significant Changes in Image Sequences

Sergii Mashtalir and Olena Mikhnova

Abstract In this chapter the authors propose an overview on contemporary artificial intelligence techniques designed for change detection in image and video sequences. A variety of image features have been analyzed for content presentation at a low level. In attempt towards high-level interpretation by a machine, a novel approach to image comparison has been proposed and described in detail. It utilizes techniques of salient point detection, video scene identification, spatial image segmentation, feature extraction and analysis. Metrics implemented for image partition matching enhance performance and quality of the results, which has been proved by several estimations. The review on estimation measures is also given along with references to publicly available test datasets. Conclusion is provided in relation to trends of future development in image and video processing.

Keywords Artificial intelligence • Machine vision • Image recognition • Video processing • Spatio-Temporal segmentation • Salient points • Regions of interest • Voronoi diagrams

1 Introduction

Image processing is traditionally related to artificial intelligence issues that try imitating mental activity of visual information perception. Despite of variety of existing methods and content presentation models, the main challenge they face is the gap between information retrieved at a low level and semantic interpretation at a high level required for efficient understanding. In context of bridging the 'semantic gap' paradigm, inspired by tending to richness of human visual perception, image similarity evaluation should be sufficiently well-defined in terms of feature spaces

S. Mashtalir
Kharkiv National University of Radio Electronics, Kharkiv, Ukraine

O. Mikhnova (✉)
Kharkiv Petro Vasylenko National Technical University of Agriculture, Kharkiv, Ukraine
e-mail: elena_mikhnova@ukr.net

© Springer International Publishing AG 2017

A.E. Hassanien et al. (eds.), *Multimedia Forensics and Security*,
Intelligent Systems Reference Library 115, DOI 10.1007/978-3-319-44270-9_8

and, at the same time, it should provide enough meaningful information for semantic associations. Low-level features include without limitation color and texture analysis, motion, area and shape analysis. High-level features include but not limited to eigenimages, deformable intensity surfaces, intensity surface curvature histogram. Feature extraction is usually implemented through artificial intelligence methods among which are neural networks, genetic algorithms, clustering, statistical analysis, etc. Many of these popular approaches are briefly touched in this chapter.

A set of more complicated questions arises when comparing a bunch of images. These assume not only a single image recognition, but also some comparison procedure that requires differentiation of lighting conditions and camera characteristics with which an image was shot. Image size and resolution also have dramatic impact on the observed changes in image sequences. By an 'image sequence' we mean video frames or any consequentially shot pictures with small time lapse between the shots. (Actually, any video is just a huge number of static images that change each other in dynamics with the speed up to 25-30 times per second.) All the mentioned above is the reason why image and video processing attracts more and more research and development efforts. Still it turns out hard to distinguish objects from a background under noisy conditions, shades and overlapping between them. A short overview on recent achievements in searching for significant content or redundancy elimination (e.g. finding similar content, least common content, best representatives, duplicate removal) is provided.

To overcome the aforementioned problems of overlapping and distinguishing, spatio-temporal segmentation is studied. Spatial image segmentation corresponds to partitioning an image field of view into tessellations of arbitrary or strictly defined shape, regions of interest, or real objects. Particular attention is given to Voronoi diagrams as a geometrical apparatus for image segmentation with further comparison and change detection using specialized metrics. This approach was deeply studied in authors' scientific research while executing governmental and commercial research projects. It turned out to be much more efficient than traditional object-based segmentation and salient point analysis for the purpose of video frame summarization. One of higher order Voronoi diagram properties lies in ability to limit initial number of salient points with simultaneous increase in a number of Voronoi regions. All in all, it is a reasonable compromise between segmentation into real objects and analyzing separate points, being in fact an approximation of segmented regions found by points. Point-based approaches are also examined.

Temporal segmentation is referred to scene boundary detection when speaking about image sequence from video. A brief overview of these techniques is also provided. Changes between video frames emerge quite quickly, and the greatest challenge of real-time qualitative processing still remains. Nearly half a thousand state-of-the-art articles and books were analyzed to make this short review on image processing. By analyzing those cutting-edge approaches and methods, the main common drawbacks were revealed. Conclusions concerning benefits and disadvantages of the examined algorithms and basic tendencies of their development are marked in this chapter.

The main objective of the chapter is to provide a comprehensive overview on the recent trends in image and video processing, and content change detection in particular, provide a complete list of traditional test collections and a review on image processing evaluation techniques. Along with brief observation of legacy image and video processing techniques, the authors' approach to change detection is given in more detail. Numerous schemata, comparative tables, figures and formulas provided in this chapter aid in understanding the given material.

2 Background Feature Analysis

Understanding of image content can be presented as an attempt of finding relation between initial images and real-world models. Transition from initial images to models decreases the amount of information contained in an image to a limited sufficient amount of data concerning the object of interest. As a rule, the whole process is divided into several phases. Under this proviso, several levels of image presentation should be considered. The lowest level contains initial data which interpretation is performed at higher levels. The boundary between those levels is not severe. Some authors provide more detailed separation into sub-levels. With this, information flow does not usually have a single direction. Sometimes this process is iterative and has several cycles, which enables changing the way of algorithm running by taking intermediary results into account.

Though, such a hierarchy of image processing is quite often simplified to only two levels. The lower level provides direct processing of initial information, and the higher level establishes understanding and interpretation of image content. Low-level techniques usually do not use any knowledge about image content. In any case, initial data will be presented as a set of matrices in the lowest level of processing as well as final results that will be also in a matrix form [1]. Different weights can be assigned to low-level features, and sometimes even dynamically changing conditional weights can be implemented.

Very often, color information is assumed (because it does not depend on an angle of a shot image and resolution). Though, to use only color is usually not enough for efficient data presentation. It is evident that images with similar color distribution may have completely different content, so treating them the same way will be a huge mistake. Color information only may be sufficient for a priori limited application domain. Color features may be analyzed by histograms that depict frequency of one or another color tone, means, maximum or minimum values of a particular color channel or a local range of a histogram which decreases processing time. Ohta features can be used for color analysis [2]. They were designed to show the level of intensity, the difference between red and blue components, and green excess as follows:

$$\begin{aligned}&\text{intensity } = \frac{r+g+b}{3}\ ;\\ &\text{red } - \text{blue difference } = r-b\ ;\\ &\text{green\_excess} = (2g-r-b)\end{aligned} \tag{1}$$

where r is a red component present in an image with RGB color schema; g is a green component present in an image with RGB color schema; b is a blue component respectively.

Aside from additive color presentation, subtractive color model CMYK can also be used, such models as HSB, HSV, YUV, Luv and Lab are used more seldom.

Texture features contain information about spatial distribution of color tone changes in a local image area. To put this another way, texture describes structure or a pattern in an image area. To characterize texture, any of 28 texture features described in detail by Haralick et al. [3] and Deselaers et al. [4] can be used. The only thing to consider, while working with these features, is the fact of their high correlation between each other. Using an entropy example, spatial connectivity between frame pixel intensity can be shown:

$$E = -\sum_{c=1}^{h} u_c \log_2 u_c \tag{2}$$

where h is a number of color tones available; u_c is the frequency of pixels with the tone c in the whole image or a local area being analyzed [3].

Aside from the aforementioned, there are methods for texture calculation based on auto regression, Markov chains, mathematical morphology, fractals, wavelets, etc. Estimation of relevant location in an image is also an important low-level feature that is considered more and more often despite of its computational complexity. Along with relevant location, density of motion flow, speed of this flow and trajectory can be analyzed. There are several groups of methods for motion analysis: methods of block comparison, methods of phase correlation, optical flow methods. In the first group of methods, blocks are obtained after image division into non-intersecting areas with further comparison between consecutive images. Phase correlation methods assume motion estimation using DCT (Discrete Cosine Transform). Though, lately the latter group of methods based on optical flow calculation is used more often. The equation for optical flow between pixels of two images shot consequentially in the moment of time t and $t + \Delta t$ respectively can be written as follows:

$$I(x+\Delta x,\ \ y+\Delta y,\ \ t+\Delta t) \approx I(x,\ y,\ t) + \frac{\partial I}{\partial x}\Delta x + \frac{\partial I}{\partial y}\Delta y + \frac{\partial I}{\partial t}\Delta t \tag{3}$$

where Δ specifies time incremental step and motion between corresponding coordinates of two images; $I(x,\ y,\ t)$ specifies pixel intensity with coordinates $(x,\ y)$.

Currently, there are many algorithms for optical flow calculation. Most of them are enlisted in Middlebury database [5] which aims at performance estimation of optical flow algorithms [6]. The main advantage of this database is that it provides information on rating for optical flow algorithms by comparing their quality on different (and the same for each algorithm) test samples. Despite this relative performance is calculated without normalization under processor powers and other hardware accelerators, it gives an excellent overview for the mentioned above algorithms. The database is constantly updated by novel algorithms, which makes it so popular among image processing community of researchers. By December 2009 there were only 24 algorithms here, in December 2012 it was enhanced up to 77 algorithms, by the end of 2013–90 algorithms were presented there. By the time of writing this chapter, in May 2015, there were 114 algorithms in the database.

Needless to say that these algorithms provide outstanding results from analytical point of view, but all of them are too bulky for real-time processing due to the necessity of spatial and temporal derivative computation between image pixels. The procedure is very time consuming compared with any other feature set implementation. Moreover, extra procedures should be incorporated to discriminate camera motion from object motion. One more shortage of optical flow implementation lies in a fact that any derivative computations are sensitive to noises of different nature.

Object motion trajectories can be calculated with differential images, they even may not include motion direction. To present information about motion direction at a machine level, cumulative and differential images are used, that are a sequence of images with the first one being a sample image. Such type of images enables acquisition of some other temporal properties of motion, motion of small objects and slow motion. Cumulative differential image values show frequency and qualitative difference from the sample grayscale image [1].

Another interesting approach analyses structural features. An object border is ‘flooded’ using water-filling algorithm, flooding time is considered along with border length or perimeter. In general, areas with the greatest intensity change are meant to be object borders. Methods based on metrics are considered the simplest ways for computing shape, border or object boundaries. Metrics provide opportunities for finding area and perimeter (object shape and its border) of selected objects for which primary segmentation is needed. Spatial segmentation ensures selecting uniform image areas which, as a rule, are objects or their parts. The easiest algorithms imply segmentation under certain thresholds or mean values of intensity, enlarging areas with close values of intensity, applying filters [7].

There are also a bit more complicated techniques for edge detection like Fourier descriptor, Zernike moment, Freeman chain code, wavelet analysis, Roberts cross operator, Sobel operator, Kirsch operator, Prewitt operator, Canny detector, etc. Many of them are based on intensity gradients. In addition, object form estimation can be made by different geometrical characteristics. Despite of variety of existing techniques, under condition of object overlapping with non-stationary background in an image sequence, lots of them provide pure results with too detailed segmentation into a huge number of small areas that actually are not that significant form image content point of view.

In case when it comes to image understanding by a machine, high-level algorithms are implied. They are grounded on knowledge, objectives and plans for reaching the desired goals. Thus, image understanding is reduced to interaction between the levels. High-level description of content can be realized via low-level features while taking into account their relative or absolute spatial location in an image or applying artificial intelligence methods for their processing. For this purpose fuzzy production rules, variety of heuristics, cluster analysis, neural networks, diversity of filters and many more techniques enlisted by popularity in one of the authors' articles [8] can be used. One of such intelligent approaches imply assigning textual labels to different classes of images though construction of a semantic net based on tesaurus. Compliance of textual labels to images is defined by users who train the system. Such recognition algorithms imply searching for similarity measure within the semantic network by taking into consideration integrated visual features. However, because of inability to implement full-functional recognition, alike to human perception, the most commonly used convention employs so-called mid-level features that link semantic understanding with low-level concepts.

The more a priori information is available about initial data, the less number of features may be included for efficient analysis. Optimal feature set selection is a challenging task that requires some preliminary research. Aside from high correlation of some features, they may behave in a number of ways for different images. That is the reason why it is so hard to find a unique image processing algorithm that could cope with any application domain equally well. Considerable success has been reached in image understanding during the last few years. Despite of it, many issues still remain unsolved, and studies in this filed of computer vision continue [1]. Thus, it seems reasonable to think of any recognition problem as a link between the two aforementioned components, i.e. identification of meaningful areas in an image (segmentation) and their content interpretation.

3 Problem Statement and Legacy Techniques for Solution

Detection of changes in image sequences (or video frames) gives an opportunity to find out correlation between the integral parts of a sequence, find structural elements in video required for semantically meaningful temporal segmentation into separate scenes. In addition, it can be used for advertisement identification, estimation of repeats in a video, summarization and lossless compression of video. Recent video processing techniques that aim at change detection can be divided into groups as shown in Table 1.

Aside from the table above, there are methods that unite approaches from several groups, also there are methods that do not belong to any of the aforementioned

Table 1 Classification of generally applicable methods for change detection in image sequences

Method	Author (Year of Publication)
Color histogram difference	• B. Liang, W. Xiao, X. Liu (2012) • G. Liu, J. Zhao (2010) • S. Thakare (2012)
Statistics	• J. Almeida, N.J. Leite, R.S. Torres (2012) • S.S. Kanade, P.M. Patil (2013) • G.I. Rathod, D.A. Nikam (2013)
Clustering	• L. Li, X. Zhang, Y. Wang, W. Hu, P. Zhu (2008) • Z. Qu, L. Lin, T. Gao, Y. Wang (2013) • H. Zhou, A.H. Sadka, M.R. Swash, J. Azizi, U.A. Sadiq (2010)
Curve simplification	• S. Lim, D. Thalmann (2001) • K. Matsuda, K. Kondo (2004) • E. Bulut, T. Capin (2007)
Visual attention	• J. Peng, Q. Xiaolin (2010) • L.J. Lai, Y. Yi (2012) • Q.-G. Ji, Z.-H. Xie, Z.-D. Fang, Z.-M. Lu (2013)
Others	• M. Cooper, J. Foote (2002) • X. Yang, Z. Wei (2011) • D.P. Papadopoulos, V.S. Kalogeiton, S.A. Chatzichristofis, N. Papamarkos (2013)

groups (‘Others’ category), but the most part of them can be classified alike to this. Color histogram difference is the most intuitive approach to image comparison, though the results may lack from accuracy as mentioned in the previous sub-section. An example of color histograms for two consecutive video frames is shown in the figure below. Though these frames reside close to each other in the video sequence, their content differs extremely, but the histograms look quite similar in shape. All in all, histogram difference still remains one of the most popular techniques for change detection. Different kinds of statistics are also easy for calculation. This approach along also cannot boast of extraordinary precision, and it should be combined with something else.

Methods that appeared first in connection to change detection compared images using clustering algorithms. It is easily explainable as this artificial intelligent means emerged long time ago and their variety is enormous, starting from primitive k-means and KNN and up to modern fuzzy clustering solutions. A figure below illustrates an example of clustering video frames to check for their similarity. A number of images under analysis are separated into clusters by assuming some reasonable feature set discussed in the previous sub-section. Repetitions (or near duplicates) are located in one cluster in this case. Obviously that the closer the distance between cluster representatives, the better match is between the two images. A problem that arises here consists in a distance metric to choose as it should cope with the feature set in use.

Clustering techniques for video frame comparison started to be implemented from nearly the beginning of 2001, they remain in leading positions of popularity for now. The only difference between the earlier methods and contemporary ones is the computational complexity of the procedure that increases due to ever growing powers, seeking for near real-time processing. The main problem for clustering algorithms is traditionally connected with a priori needed number of clusters. This requires initial knowledge about the analyzed image sequence and user involvement in the procedure. Bayesian criterion overcomes this constraint, and clusters are selected automatically, however to decrease the number of computations, less parameters should be included for analysis, which will negatively influence the results.

Some authors argue that similarity matrix based on clustering results is a perfect match. In this relation, Chinese authors (X. Zeng, W. Hu, W. Liy, X. Zhang and B. Xu) proposed to assign several clusters to a single image (such kind of a fuzzy model) with an ability of one-to-one correspondence [9]. But their frame comparison model is too simplified that is inappropriate for a variety of initial data with different application domain. Aside from this, methods of cluster analysis perform badly for homogeneous data. Misclassification is observed very often.

One more approach related to clustering and classification, proposed by Xianfeng Yang and Qi Tian, deals with video repeats acquired by visual features. Recognition procedure of repeats differs here for a priori known and unknown cases. To recognize already known repeats, a set of feature vector is formed (color histogram, texture, etc.) based on video prototypes, and nearest neighbor classifier is used to recognize copies in video collections. The main attention in the work of Xianfeng Yang and Qi Tian is driven to presentation of these features and classifier training. The authors have made an attempt to perfect the results of recognition by incorporating discriminant analysis of sub-spaces. To decrease video volume and eliminate redundancy, frame extraction takes place each half a second. RGB color histogram is constructed for such extracted frames along with texture feature analysis. Nearest neighbor classifier offers effective means for video copy recognition. Closest prototype is specified for the analyzed image sequence when the distance is less than a threshold value, otherwise this sequence is related to another class with other prototype. In order to estimate error frequency and obtain optimal

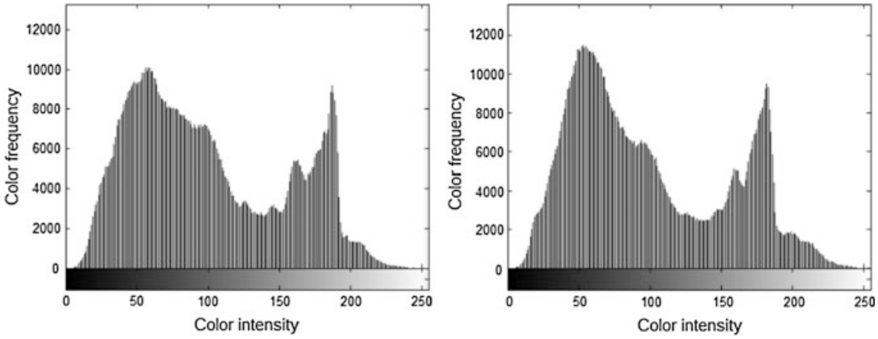

Fig. 1 Histogram difference approach illustration

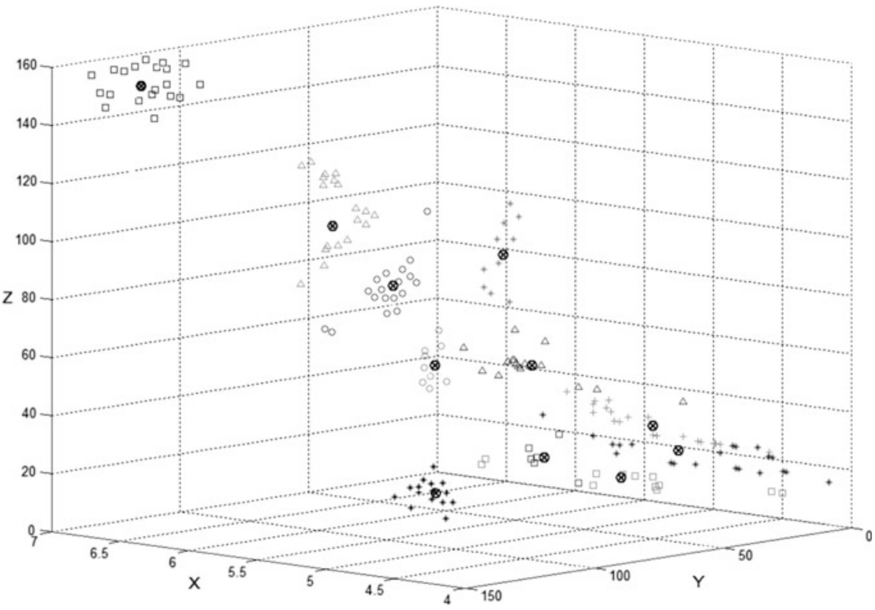

Fig. 2 Clustering approach illustration

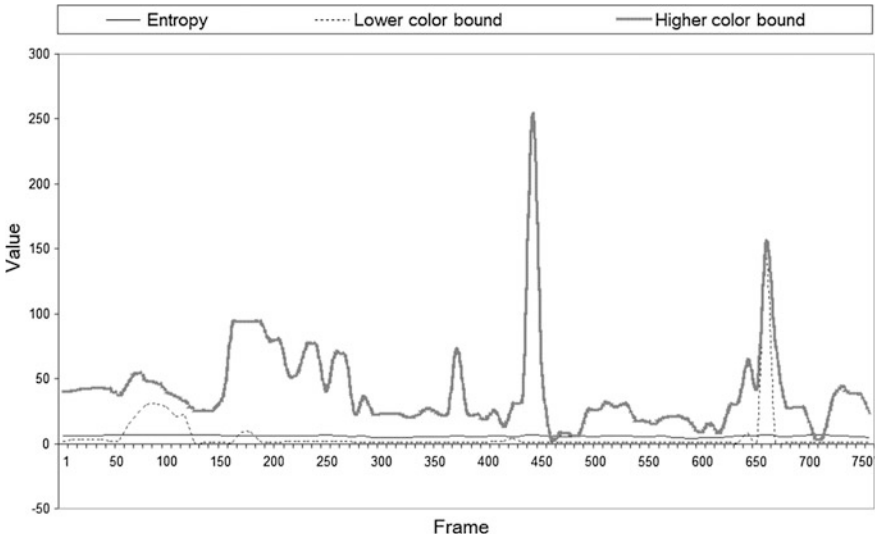

Fig. 3 Curve simplification approach illustration

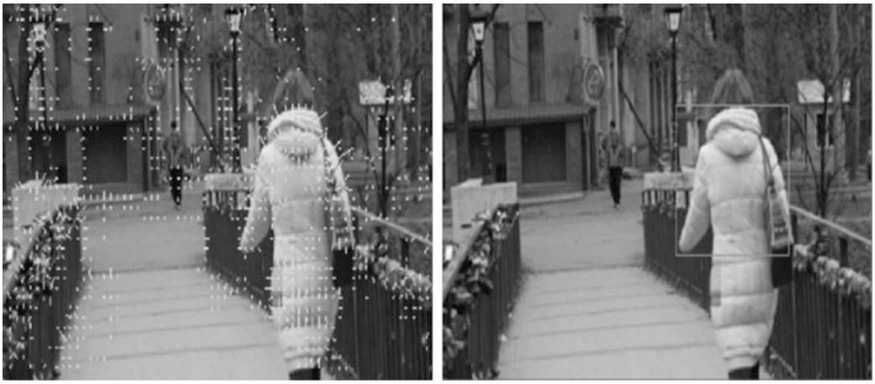

Fig. 4 Motion-based approach illustration

threshold for classification, there should be a notion about prototype feature vectors stored in a database.

As a rule, recognition of unknown repeats is implemented for collection from a single source. This problem is more complicated than the previous one as the content, image sequence length and camera location with an angle of shot are a priori unknown, that is actually true for all real-world applications. In addition, images from different collections may contain such contradictions as overlapping of headings and partial repeats. The aforementioned approach unites video segmentation, color fingerprinting, self-similarity analysis, two sequentially used detectors, LSH-indexing and self-learning. Results of implementing this approach were described for recognition of repeats in news video [10].

Another group of methods assume image sequence presentation by a curve constructed in multidimensional feature space. An example of three curves built on the basis of several features is shown in the figure below. For the purpose of clarity, all the three graphics are shown on one coordinate plane, though their physical meaning certainly differs from one another. Color changes are more explicit compared with texture, but the latter brings more precise information about exposure variation. Approximation of curve values leads to detection of slight changes, and significant changes are registered in local maximums and minimums of the curve. Color and texture features are primarily considered for such purposes (Figs. 1, 2 and 3).

Later, comparison methods based on motion emerged. As it was said before, optical flow algorithms are used most frequently. Figure 4 illustrates motion changes between consecutive frames of nearly the same content. Small lines depict motion shift of consecutive frames estimated by Horn-Schunck variational method. Motion-based solutions possess high computational complexity, and truly fine results are got only for significant motion. Though, this approach is a bit controversial as frames with small motion may also contain significant changes in content and, vice versa, frames with huge motion changes may be of similar content.

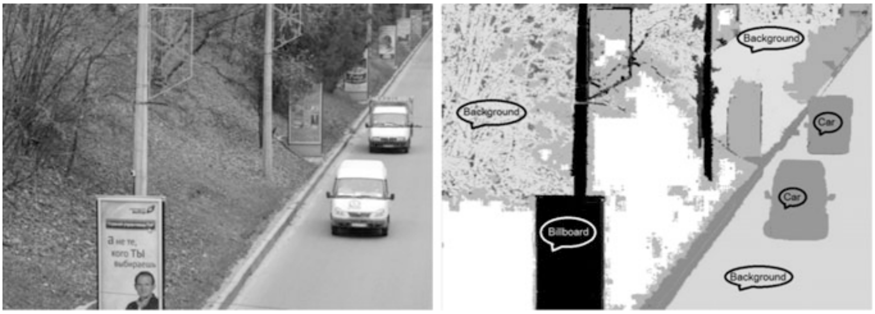

Fig. 5 Annotated saliency map of an image

Lighting conditions have direct influence on motion detection results, which often leads to wrong identification of motion change significance. For video sequences, it should be noted that the greatest motion is observed between scenes, i.e. on scene boundaries, or in high-textural frames occurring in the middle of a scene. The main difficulty lies in a necessity of motion sensitivity threshold specification. Each type of video/image content needs its own condition or constraint in relation to maximum and minimum motion.

Along with the aforementioned problems, this group of methods does not provide means for differentiation of motion significance, i.e. which motion to consider more significant and which is just a background motion. Because of this matter, many scientists merge motion analysis with visual features. Such a combination can look as follows: first, scene boundaries are detected by color change analysis, for example, and then, frames with significant changes are extracted from scene parts with decreased or increased speed of motion. If an application domain is a priori known, motion pattern can be constructed prior to performing the analysis [8]. The latter approach is convenient because it is not connected with threshold values, it increases processing time, though it is applicable for restricted types of image sequences while being linked with a predefined pattern.

In order to make the results semantically oriented, some researchers focus on addition of textual labels for objects. These labels are sometimes called annotations (see Fig. 5). In any case, users are involved in the procedure by accepting or declining perfect match of object labels. Unfortunately, assigning of textual labels does not guarantee that a new image will contain analogous material and the system will recognize it correctly. For arbitrary and unknown collection this is a big question that arises as all the changes in content cannot be predicted anyway. In spite of some achievements in semantically oriented concept implementation, correct machine-level interpretation still cannot be imagined without user involvement [9].

Visual attention model is another novel technique recently proposed for image comparison. It assumes generation of a saliency map or a curve of attention. An example of annotated saliency map can be seen in Fig. 5. Visual attention curve is

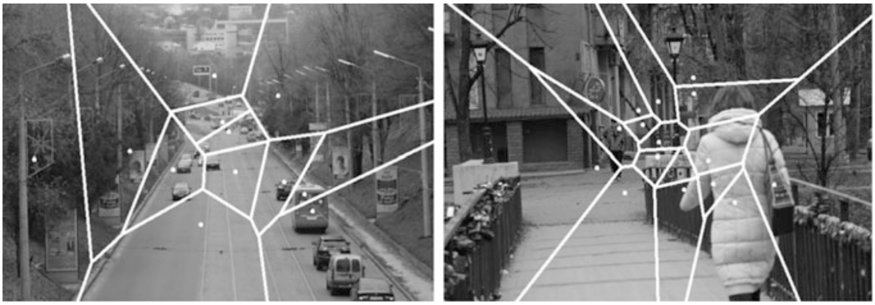

Fig. 6 Image presentation using Voronoi diagrams built on salient points

constructed on background and foreground changes, and it may look pretty much like a graphic shown in Fig. 3. Motion and intensity are mostly considered as low-level features for the model. Genetic algorithms, neural and immune networks are examples of some other intelligent options for frame comparison, though their authors honestly alert about additions and impromenents needed for reaching better quality [8].

Despite of a variety of available techniques, it is still hard to find a method with sufficient quality of image sequence content interpretation, that could cope with an arbitrary application domain. This is due to different sources of collections, image sizes and resolution, quality of shots, cameras in use, shot angles, noises, content characteristics and objective of frame extraction. None of existing techniques can cope with all the issues simultaneously. Solution to one problem usually uncovers another one. The following sub-section describes in detail an approach to change detection in an image sequence that is indifferent to data source. The authors tried to make an attempt of image presentation using Voronoi diagrams built on salient points which were found and reorganized by a number of features (see Fig. 6). Comparison of Voronoi diagrams using proposed metrics revealed some interesting properties that turned out the basis for similarity detection. The novelty of this method consists in using Voronoi tessellations as areas for image comparison and, thus, enabling detection of slight and significant changes.

Voronoi diagrams were chosen for several reasons for the purpose of spatial image segmentation. First of all, segmentation into real-life objects is not reasonable because there are too much of them when speaking about video. Secondly, video objects are very changeable in time. Today there is no method to detect and track objects without prior knowledge about them. The reason for such statement is constant improvement of such class of methods. Object overlapping, quick motion and dramatic changes in lighting conditions, non-stationary foreground and background, all of these may cause wrong object identification and classifying. Partitioning with Voronoi diagrams require much less computational resources than motion analysis of segmented objects. Voronoi diagrams have not been used earlier for the purpose of change detection, which is an undoubted interest for research from the point of application competitiveness of a new method [7].

4 Image Content Interpretation and Comparison

Comparison of images can be performed by machine-level analysis of salient points, local areas or the whole images. The latter is more robust and usually not applicable for precise content interpretation, it can be used as intermediary phase of analysis only. Lately, algorithms on salient points (also referred to as significant points, interesting points, meaningful points, key points, characteristic points, boundary points, corner points, sites, atoms, generators or generating points) have gained large popularity. In fact, when analyzing local areas (whether they are real objects or arbitrary tessellations of polygonal shape), their construction is usually marked or linked with those points of interest which posses some common properties. The points reside on object boundaries or other meaningful image areas. They are rigorously distinguished from surrounding locality by intensity, having invariance to geometric and radiometric distortions under potential movement [2].

Currently, many approaches to salient point detection have been developed: starting from complicated wavelets and up to pretty simple object corner detectors. In spite of a more precise output from wavelet transformations, corner detectors are used much often because of their relative computational simplicity. Scale-Invariant Feature Transform (SIFT) and its extensions (like Speeded-Up Robust Features, SURF) are commonly used in this relation. Though, Harris algorithm and its extensions are considered of higher correctness. FAST-detector (Features from Accelerated Segment Test) also analyses local intensity, but unlike the aforementioned techniques it operates in a real-time mode due to absence of differentiation procedures. However, the quality of the latter may degrade for some test collections. Other less popular detectors are: Susan, DoG, MSER, Hessian-Affine, GLOH, LBP, etc. Of course, the points only provide much less information than areas, but their utilization with respect to the areas guarantees effective image and video recognition [2, 11].

An obvious step of spatial image segmentation should be undertaken prior to visual content analysis. Partitioning of image space can be accomplished in a primitive manner of dividing it into equal rectangular areas [2, 3]. A more convenient and complicated approach consists in searching for background and foreground objects. The drawbacks of the latter segmentation arise with lighting condition changes, rotational and tangential movement of objects and their overlapping. Improper assignment of objects leads to poor recognition results. To avoid this problem, a novel approach to image change detection based on Voronoi diagrams is proposed [7, 8].

4.1 Construction of Voronoi Diagrams on Salient Image Points

Voronoi diagrams proposed to be considered as a basis for image structure presentation and interpretation are built upon salient points (they can be found using one of existing methods mentioned above). Primarily designed for geodesy, Voronoi diagrams have been lately started to be implemented in computer graphics for 3D modeling. However, this stochastic geometry method is still not distributed in image recognition despite of its apparent benefits [12]. Voronoi tessellations were mentioned for the first time by R. Descartes in 1644. Later, in 1850, they appeared in P.G.L. Dirichlet's works, and after that, at the beginning of 20th century, they were named after Russian mathematician G.F. Voronoi who has devoted his life to accumulation of knowledge about it. Aside from Voronoi diagrams or tessellations, several other names can be met in literature: Dirichlet cells or regions, Thiessen polygons, Wigner-Seitz cells. Contemporary scientists who studied Voronoi tessellations are: F. Aurenhammer, F. Preparata, M. Shamos, A. Okabe, B. Boots, K. Sugihara, S. Chiu, B.N. Delone and many others.

In order to provide formal definition for an arbitrary Voronoi region $v(p_i)$, define an image field of view as $D=[a,b]\times[c,d]$ with $a,b,c,d=const$. Let $\{p_1, p_2, \ldots, p_n\}$ be a set of salient points. Voronoi diagram is such a partitioning

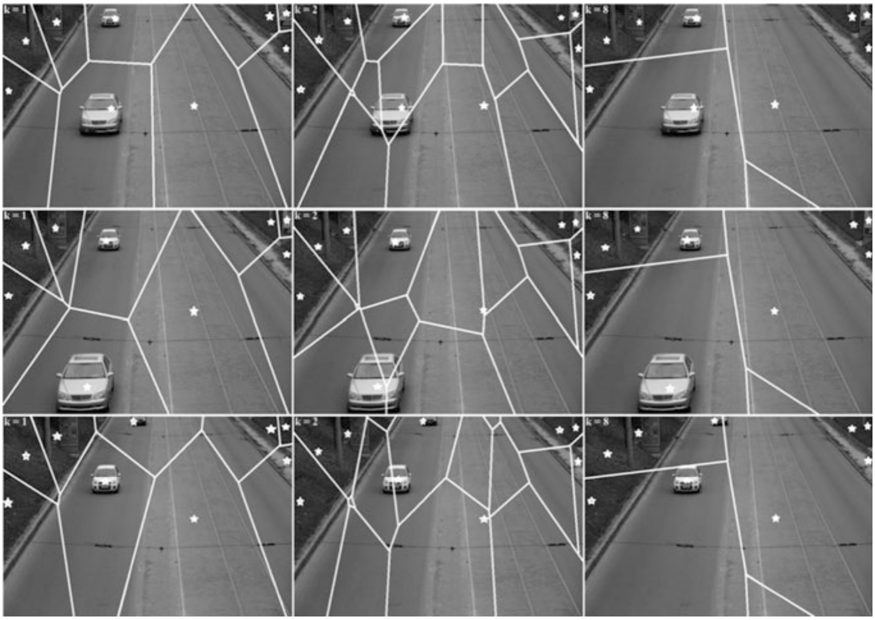

Fig. 7 Image segmentation with higher order Voronoi tessellations

of image field D into convex polygons $V = \{v(p_1) \cap D,\ v(p_2) \cap D,\ \ldots,\ v(p_n) \cap D\}$ that the following inequality is held for each region:

$$v(p_i) = \{z \in R^2 : d(z, p_i) \le d(z, p_j) \forall i \ne j\} \tag{4}$$

where $d(\circ, \circ)$ is a Euclidean metric. To put it differently, Voronoi region $v(p_i)$ linked with a point p_i is construction from a number of points Z, assuming that the distance from each of them to the corresponding salient point is less than or equals to the distance to any other non-corresponding salient point [15].

Content detailing with enhancement of tessellations can be performed either by increasing the number of initial salient points, or by construction of Voronoi diagrams of higher order (or generalized Voronoi diagrams). The latter is more sophisticated and generally turns out to be more computationally effective. Figure 7 shows partitioning of three video frames that reside close to each other in a video sequence. Each of the three images is segmented using Voronoi diagrams of the first, the second and the eighth order (from lest to right). Initially, nine salient points were detected for these frames. That is the reason why the eighth order is considered to be maximum possible in this case.

The increase in a number of segments for the second order is straightforward as well as the decrease in a number of segments for the highest possible order. From the figure above it is clear that the most stable to content change are Voronoi diagrams of the highest order. Frames provided here as an example were taken from a single scene where diagrams of the highest order remain practically unchanged. This is easily explainable from formal definition of generalized Voronoi diagrams [13].

Voronoi diagram of order k, $V^{(k)}$, built on the basis of n salient points in 2D space, is a plane partitioning into convex polygons, such that points z of each Voronoi region $v(p_i)^{(k)}$ have the same number of k closest salient points p_i linked with them. The aforementioned definition (4) of a Voronoi diagram is a special case of generalized Voronoi diagrams when $k = 1$.

To give definition of an arbitrary generalized Voronoi diagram, let $\{p_1,\ p_2,\ \ldots,\ p_m\}$ be the set of salient points, and $\{\{p_{1,1},\ \ldots, p_{1,k}\},\ \ldots,\ \{p_{l,1},\ \ldots,\ p_{l,k}\}\}$ be corresponding subsets of k closest salient points, then a convex Voronoi polygon $v(p_i)^{(k)}$ of order k, which is formed by salient points $\{p_{i,1}\ ,\ \ldots,\ p_{i,k}\}$, can be presented as follows:

$$\begin{aligned} v(p_i)^{(k)} \ = \ \{z \in \mathbb{R}^2 \colon \max\{d(z, p_{i,h}),\ p_{i,h} \in v(p_i)^{(k)}\} \le \\ \le \min\{d(z, p_{i,j}),\ p_{i,j} \in V^{(k)} \backslash v(p_i)^{(k)}\}\} \end{aligned} \tag{5}$$

Another way of putting it is that the distance between the farthest point of one Voronoi region to its corresponding salient points is closer or equals to the distance to any closest salient point of another region. An arbitrary Voronoi region of order k may contain from 0 to k salient points, i.e. a Voronoi region of order k may have no salient points [14, 15]. The increase in Voronoi diagram order leads to growth in a number of Voronoi regions. The undertaken experiments have proved that under gradual increase of order, the detailing starts fading when a definite higher order is reached (not waiting for the highest possible order that is equal to $k = n - 1$).

It is important to notice that the threshold value for detailing reduction varies under different number of salient points. Despite the amount of Voronoi regions depends on planar location of salient points, experiments have shown that this amount differs by the value of 20–30 % for diagrams of the same order and the same number of salient points. Detailing changes under parabola: first, it increases, and then fades smoothly reaching its threshold value close to the highest order. Thus, generalized Voronoi diagrams constitute an excellent tool for image detailing when they are going to be compared with each other [14]. Voronoi diagrams of any order are not intended for the purposes of robust shaping of object borders. Object presence in one or another Voronoi region testifies fine distribution of image pixels from the point of color and texture only. It also testifies stability of such a partitioning under condition of initial number and location of salient points [13].

4.2 Similarity Metrics for Voronoi Diagram Comparison

In order to compare video frames, local image regions, feature sets or anything else related to processing of visual information, different metrics are used for the most part. These metrics or distances show proximity of essences under analysis. However, the ‘distance’, ‘similarity’ or ‘proximity measure’ are reversed concepts from the terminological point of view. One may argue that the distance between two arbitrary sets is the difference between them. Conventionally, a metric is considered to be any function that satisfies conditions of reflexivity, symmetry and triangle inequality [16].

$$\begin{aligned}
&(1)\ \rho(B'(z), B''(z)) = 0 \Leftrightarrow B'(z) = B''(z),\\
&(2)\ \rho(B'(z), B''(z)) = \rho(B''(z), B'(z)),\\
&(3)\ \rho(B'(z), B'''(z)) \le \rho(B'(z), B''(z)) + \rho(B''(z), B'''(z))
\end{aligned} \qquad (6)$$

where $\rho(\circ, \circ)$ is the similarity metric between images $B'(z)$, $B''(z)$, $B'''(z)$.

As a rule, image comparison is made via widely applicable Manhattan distance, Hausdorff distance, Mahalanobis distance, Euclidean or Squared Euclidean distance. The authors of the chapter propose using special metrics that assume color, texture and region shape properties. Sometimes linear combinations of metrics or measures are used (they are called aggregation), weight coefficients may also be assigned to values obtained as a result of application of a metric or aggregation. Statistical measures implemented for these purposes do not provide high-quality output. However, they may be used for data processing in a real time mode and for solution of limited application domain problems. To get an ability of image content comparison that is presented by Voronoi tessellations, principally novel metrics should be developed as comparison of Voronoi diagrams have not been performed earlier by researchers. As it has been already mentioned, this gives an opportunity of finding maximal match between images by making video frame content presentation more stable.

An attempt of Voronoi diagram match has been made by Japanese scientist Yukio Sadahiro who has tried to compare them not by means of similarity metrics, but by areal properties and statistical measures [17]. He has proposed using different methods of visual and quantitative analysis, including χ^2 criterion, Kappa index and their extensions, area and perimeter of Voronoi regions, variance and standard deviation, center of mass concept, etc. For the purpose of comparison of Japanese division systems into administrative regions, Y. Sadahiro has proposed to use a density measure of detailing and hierarchical relationships of overlapping, partial overlapping and inclusion. However, areal methods sometimes have dual meaning for image and video processing applications because objects can be shot with different scale. Different objects in images or video frames may have the same size. Thus, areal properties only cannot be trusted for object recognition and change detection. Some other properties should be taken into consideration as well: spatial location, texture, color, shape, motion, which are the main attributes utilized in content-based image and video retrieval systems.

To compare Voronoi tessellations constructed for two arbitrary images $B^{'}(z)$ and $B^{''}(z)$ with salient points $\{p^{'}_1, p^{'}_2, \ldots, p^{'}_n\}$ and $\{p^{''}_1, p^{''}_2, \ldots, p^{''}_m\}$ respectively, take advantage of the following metric $\rho_1(V^{'}, V^{''})$ [16]:

$$\rho_1(V^{'}, V^{''}) = \sum_{i=1}^{n}\sum_{j=1}^{m} \mathrm{card}(v(p^{'}_i)\Delta v(p^{''}_j))\ \mathrm{card}(v(p^{'}_i) \cap v(p^{''}_j)) \qquad (7)$$

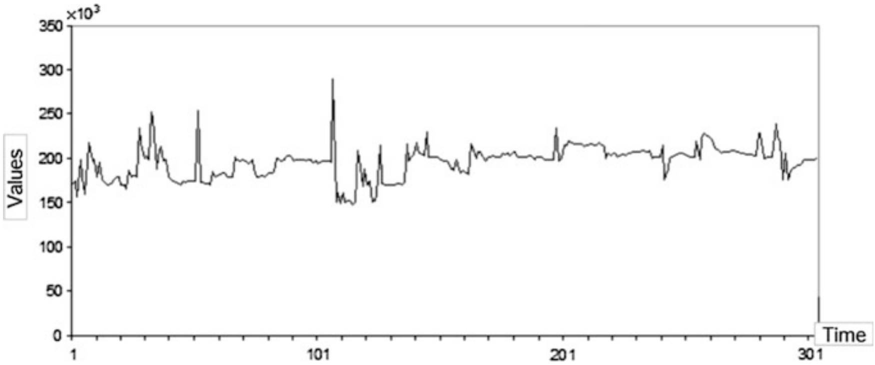

Fig. 8 Voronoi region shape comparison using symmetrical difference

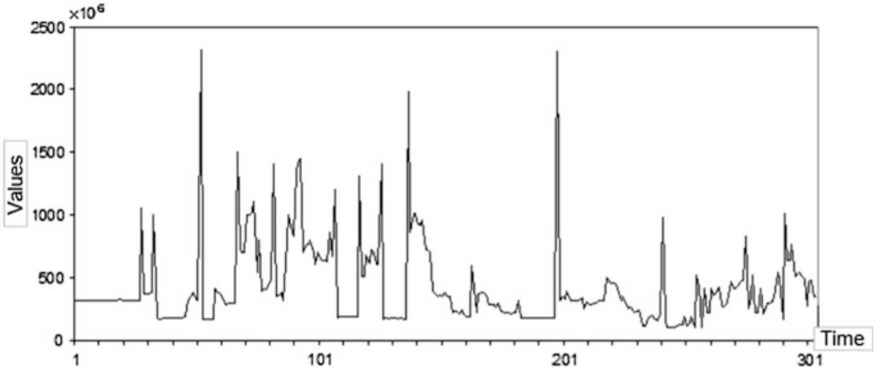

Fig. 9 Region shape comparison using symmetrical difference multiplied by region intersection

where $v(p_i^{'})\Delta v(p_j^{''}) = (v(p_i^{'})\backslash v(p_j^{''})) \cup (v(p_j^{''})\backslash v(p_i^{'}))$ is symmetrical difference that define the number of pixels which makes the difference between the regions $v(p_i^{'})$ and $v(p_j^{''})$.

The value obtained from the symmetrical difference does not depend on region size, i.e. two spatially large areas with three points of mismatch (or difference between them) will be treated the same way in terms of distance as two spatially small areas with three points of mismatch. Multiplication of symmetrical difference by intersection of areas leads to a huge scatter of values for different and similar images. Implementation of symmetrical difference only does not guarantee such a huge scatter plot, which is easily seen from the two figures below. Figure 8 illustrates graphic of symmetrical difference values calculated for Voronoi regions of consecutive frames taken from news video fragment “factories_512 kb.mp4” from Internet Archive open-source test collection. Figure 9 illustrates graphic of symmetrical difference values multiplied by Voronoi region intersection calculated for the same video frames.

The range of values, obtained as a result of symmetrical difference multiplication by Voronoi region intersection for the news video under analysis, equals to $2,3 \times 10^9$, while the values of symmetrical difference vary from 148×10^3 to 289×10^3, which is 10^5 times less. In addition, graphic changes observed for symmetrical difference values much less correlate with changes in video content than the graphic values obtained from Eq. (7). That is why it is hard to make any decisions concerning image changes by taking into consideration values of symmetrical difference only. While assuming region intersection, the resulting value from Eq. (7) turns out to be much greater for regions of different shape, and most multipliers equal to 0 for regions of almost the same shape due to less number of uncommon intersections, that leads to huge scatter of values.

To present Voronoi regions in terms of salient points only, and rewrite the above formula, the following concepts should be marked out. Because of the fact that Voronoi regions are constructed based on perpendicular bisectors between the

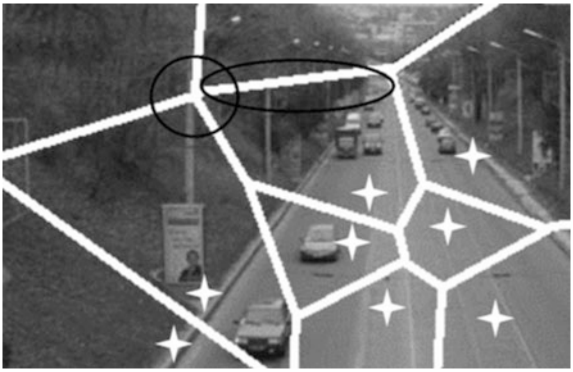

Fig. 10 Edge of a Voronoi polygon and its vertex

adjacent salient points p_i and p_λ [15], consider Voronoi region $v(p_i)$ as an intersection of half-planes that connect adjacent salient points:

$$v(p_i) = \bigcap_{\lambda \in [1;\psi]} H(p_i,\ p_\lambda) \tag{8}$$

where ψ is a number of salient points p_λ adjacent to p_i.

The following properties are held for adjacent salient points p_i and p_λ:

1. $\exists\, \gamma_\psi,\ d(p_i, \gamma_\psi) = d(p_\lambda, \gamma_\psi)$ where γ_ψ is a vertex of a Voronoi polygon;

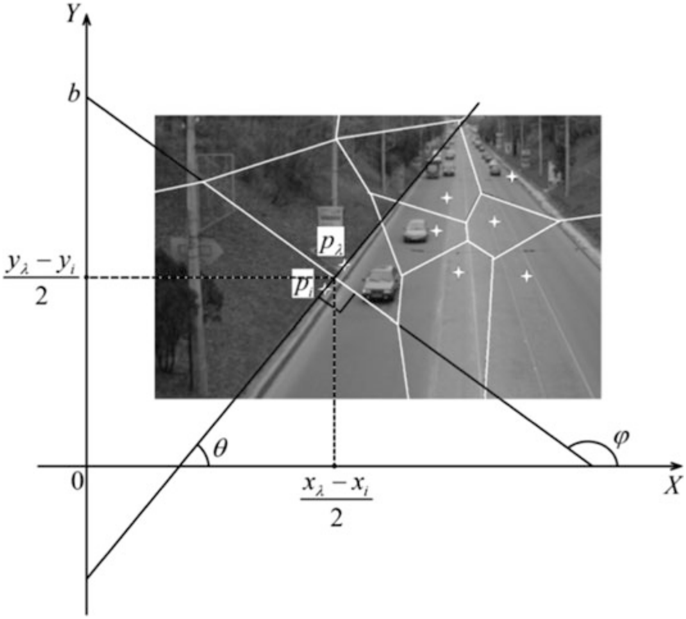

Fig. 11 Slope angle of the line containing perpendicular bisector between the adjacent salient points

2. $(\frac{x_\lambda - x_i}{2}; \frac{y_\lambda - y_i}{2}) \in v(p_i), v(p_\lambda)$ where $p_i(x_i; y_i)$, $p_\lambda(x_\lambda; y_\lambda)$;
3. $v(p_i)\bigcap v(p_\lambda) \neq \varnothing$ where $v(p_i)\bigcap v(p_\lambda)$ is an edge of a Voronoi polygon or its vertex (see Fig. 10).

However, salient points are adjacent only in case their corresponding regions have non-degenerated boundary (Voronoi polygon edge, i.e. more than just a single point). Figure 10 shows an example of an edge of a Voronoi polygon and its vertex obtained as a result of intersection of lines containing perpendicular bisectors between the adjacent salient points.

To present Voronoi regions in terms of salient points, an equation for perpendicular bisectors between the adjacent salient points $p_i^{'}$ and $p_\lambda^{'}$ of one image $B^{'}(z)$ and adjacent salient points $p_j^{''}$ and $p_\lambda^{''}$ of another image $B^{''}(z)$ should be written. Equation of a straight line that connects the two points with coordinates $p_i^{'}(x_i; y_i)$ and $p_\lambda^{'}(x_\lambda, y_\lambda)$ looks as follows:

$$\frac{y - y_i}{y_\lambda - y_i} = \frac{x - x_i}{x_\lambda - x_i} \text{ or } (y_i - y_\lambda)x + (x_\lambda - x_i)y + (x_i y_\lambda - x_\lambda y_i) = 0\ . \tag{9}$$

From the formula above, the slope angle θ of the line that connects two adjacent points can be calculated as $\mathrm{tg}\theta = -\frac{y_i - y_\lambda}{x_\lambda - x_i}$. From the Fig. 11, $\mathrm{tg}\varphi = \mathrm{tg}(\theta + \frac{\pi}{2})$, that is why the slope angle φ for the perpendicular bisector can be expressed by $\mathrm{tg}\varphi = -\mathrm{ctg}\theta$. Thus, $\mathrm{tg}\varphi = \frac{x_\lambda - x_i}{y_i - y_\lambda}$.

Knowing the slope angle φ and coordinates $(\frac{x_\lambda - x_i}{2}; \frac{y_\lambda - y_i}{2})$ of the point on a line which produces half-planes, it is not hard to find out an equation for a straight line, containing perpendicular bisector between the adjacent salient points. Assuming that point coordinates should satisfy straight-line equation with angular coefficient $y = kx + b$, $k = \mathrm{tg}\varphi$, define the point of intersection of the perpendicular bisector with the ordinate axis:

$$b = \frac{y_\lambda - y_i}{2} - \mathrm{tg}\varphi \times \frac{x_\lambda - x_i}{2}, \text{ i.e. } b = \frac{y_\lambda - y_i}{2} - \frac{x_\lambda - x_i}{y_i - y_\lambda} \times \frac{x_\lambda - x_i}{2}.$$

Consider equation of the straight line, containing perpendicular bisector between the adjacent salient points:

$$y = \frac{x_\lambda - x_i}{y_i - y_\lambda} x + \frac{y_\lambda - y_i}{2} - \frac{x_\lambda - x_i}{y_i - y_\lambda} \times \frac{x_\lambda - x_i}{2}$$

or

$$y = \frac{x_\lambda - x_i}{y_i - y_\lambda}(x - \frac{x_\lambda - x_i}{2}) + \frac{y_\lambda - y_i}{2} \tag{10}$$

Intersection of lines, containing perpendicular bisectors between the adjacent salient points, generates Voronoi regions in a form of convex polygons. To find out coordinates of polygon vertices, point coordinates of such line intersections should

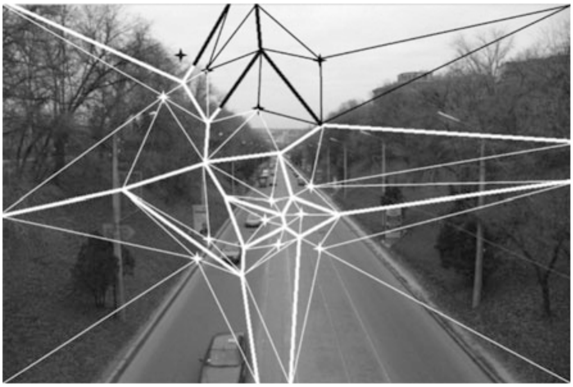

Fig. 12 Partitioning of polygons into triangles needed for area calculation

be determined: $\bigcap\limits_{\lambda \in [1;\psi]} \frac{x'_\lambda - x'_i}{y'_i - y'_\lambda}\left(x - \frac{x'_\lambda - x'_i}{2}\right) + \frac{y'_\lambda - y'_i}{2}$. With a knowledge of the salient point coordinates $(x_i; y_i)$ and polygon vertex coordinates $(x_{\gamma_\psi}; y_{\gamma_\psi})$, area of each polygon can be found using Heron's formula by dividing the polygons into triangles. A triangle vertex will reside in a salient point, and a triangle base will be an edge of a Voronoi polygon (a side between the two vertices with already known coordinates) or a segment of a field of view, bounding an image or a frame, which coordinates are known as well. With this in mind, if a salient point resides within a triangle which base is presented by a segment of an image field of view, then area of such a triangle is computed by assuming its vertex and the base segment, without taking the salient point into consideration (see Fig. 12). Thus, area of each Voronoi region can be calculated to compare two images in future.

Areal image similarity does not assume spatial location of partitions, so differently adjusted regions of the same area will have the same value of granularity. That is why color and texture information should be taken into account as well, along with shape characteristic of Voronoi regions. Other features may also be included in final computations under condition of less weight assignment. Equation (7), containing the metric for comparison of Voronoi region shape, can be presented in terms of salient point coordinates using Eq. (8), containing expression of a Voronoi region by half-planes, and Eq. (10) with the straight-line equation, containing perpendicular bisectors between the salient points:

$$\begin{aligned}\rho_1(V', V'') = \sum_{i=1}^{n}\sum_{j=1}^{m} \operatorname{card}\Bigg(& H\Bigg(\bigcap_{\lambda\in[1;\psi]} \frac{x'_\lambda - x'_i}{y'_i - y'_\lambda}\Big(x - \frac{x'_\lambda - x'_i}{2}\Big) + \frac{y'_\lambda - y'_i}{2}\Bigg)\Delta \\ & \Delta H\Bigg(\bigcap_{\lambda\in[1;\psi]} \frac{x''_\lambda - x''_j}{y''_j - y''_\lambda}\Big(x - \frac{x''_\lambda - x''_j}{2}\Big) + \frac{y''_\lambda - y''_j}{2}\Bigg)\Bigg)\times \\ & \times \operatorname{card}\Bigg(H\Bigg(\bigcap_{\lambda\in[1;\psi]} \frac{x'_\lambda - x'_i}{y'_i - y'_\lambda}\Big(x - \frac{x'_\lambda - x'_i}{2}\Big) + \frac{y'_\lambda - y'_i}{2}\Bigg)\bigcap \\ & \bigcap H\Bigg(\bigcap_{\lambda\in[1;\psi]} \frac{x''_\lambda - x''_j}{y''_j - y''_\lambda}\Big(x - \frac{x''_\lambda - x''_j}{2}\Big) + \frac{y''_\lambda - y''_j}{2}\Bigg)\Bigg).\end{aligned} \tag{11}$$

The above similarity metric shows how two Voronoi diagrams match each other in terms of the region shape. To consider color and texture features, two additional metrics ($\rho_2(B'(z),\ B''(z))$ and $\rho_3(B'(z),\ B''(z))$ respectively) should be defined for the common parts of partitions:

$$\rho_2(B'(z),\ B''(z)) = \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{x_q}\sum_{y_u}{}_{(x_q, y_u)\in(v(p'_i)\cap v(p''_j))}(B'(x_q, y_u) - B''(x_q, y_u))^2 \tag{12}$$

where $B'(x_q, y_u)$ is intensity value of a pixel from the region $(v(p'_i)\cap v(p''_j)$, and

$$\rho_3(B'(z),\ B''(z)) = \sum_{i=1}^{n}\sum_{j=1}^{m}{}_{(v(p'_i),\ v(p''_j))\supseteq(v(p'_i)\cap v(p''_j))}\left|E(v(p'_i)) - E(v(p'_j)\right| \tag{13}$$

where $E(v(p'_i))$ is an entropy value for the region $v(p'_i)$.

Similarity measures of color and texture ensure assumption of corresponding changes in Voronoi regions of images being analyzed. Squared Euclidean distance guarantees increased weight for remote (in terms of color) objects. Manhattan distance was chosen for texture similarity analysis because entropy value calculation is made for each region in whole, and each value is presented by a single floating number (one number per a Voronoi region), whereas color similarity is estimated in each pixel present in both images being analyzed. By analogy, metrics (12) and (13) may also be presented in terms of salient point coordinates, instead of Voronoi regions, much like Eq. (11). Thus, three non-normalized estimations are got. Further, normalization of Eqs. (7), (12) and (13) can be made to obtain values ranging from 0 to 1. Transformation of the above metrics to a limited form imply usage of a function named range compander:

$$\rho'(B'(z),\ B''(z)) = \frac{\rho(B'(z),\ B''(z))}{1 + \rho(B'(z),\ B''(z))} \tag{14}$$

Combination of this function with a metric still leads to a metric that satisfies reflexivity, symmetry and triangle inequality rules. Because linear combination of metrics will give a metric, the following resulting metric can be used:

$$\widehat{\rho}(B'(z),\ B''(z)) = \alpha_1\rho_1' + \alpha_2\rho_2' + \alpha_3\rho_3' , \quad \sum_{\gamma=1}^{\gamma=3} \alpha_\gamma = 1 , \quad \alpha_\gamma \geq 0 \tag{15}$$

where $\widehat{\rho}(B'(z),\ B''(z))$ shows image similarity and α_γ indicates importance of each feature in use [7].

4.3 Procedure of Image Change Detection

To compare content from several images, analysis of image sequence homogeneity is proposed to be performed at the beginning, prior to any other processing. This will give understanding on a threshold value to use for content change detection. For this purpose, entropy values computed for each image may come in hand. High entropy testifies huge scatter of pixel values, whereas low entropy says about pixel

Table 2 Miniatures of frames with corresponding entropy values

Range of Entropy Values	Frame Miniatures			
3,25 – 4,99				
5 – 5,99				
6 – 6,99				
7 – 7,25				
7,26 – 7,5				
7,51 – 7,75				

Table 3 Classification of generally applicable methods for change detection in image sequences

5:4	4:3	16:10	16:9
SXGA (1280 * 1024)	QVGA (320 * 240)	CGA (320 * 200)	WVGA (854 * 480)
QSXGA (2560 * 2048)	VGA (640 * 480)	WSXGA + (1680 * 1050)	HD 720 (1280 * 720)
	PAL (768 * 576)	WUXGA (1920 * 1200)	HD 1080 (1920 * 1080)
	SVGA (800 * 600)	WQXGA (2560 * 1600)	
	XGA (1024 * 768)		
	SXGA + (1400 * 1050)		
	UXGA (1600 * 1200)		
	QXGA (2048 * 1536)		

homogeneity and detail homogeneity as a consequence. Thus, entropy value shows the amount of details present in a local image area for which it is computed. Image with no objects, without any texture (totally black, white or gray frame), will have 0 entropy value. It is interesting to note that any other fill color (from light yellow to dark blue) will lead to entropy of 1.58, however, addition of any details to a white background will influence appearance of tenth parts of entropy value only. For real photos and video frames, some texture will be present for sure, and entropy value usually varies in a range from 3.25 to 7.75. When entropy is less than 5, light texture is present with small number of details. More values usually reside in the range from 7 to 7.75.

When speaking about sequential video frames, difference of entropy for a tenth part probably means that the object of tracking has not been changed and simply moved in space. Though, a new scene may also be started, as the same value of entropy may be computed for different frame content. Because of such ambiguity, entropy only cannot be used for correct interpretation of content changes, but it can be utilized for less precise homogeneity estimation of the whole visual sequence. As

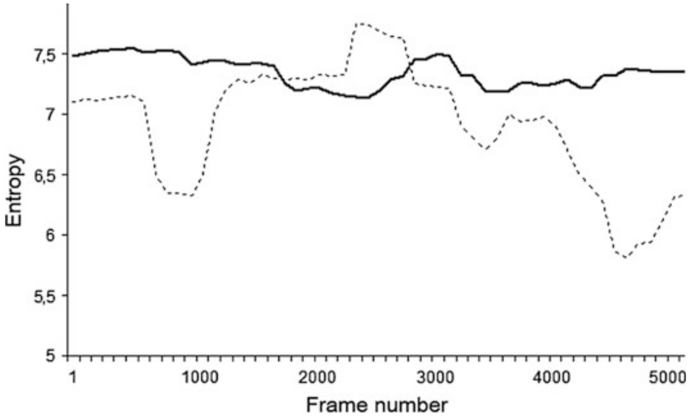

Fig. 13 Graphics of entropy change for homogeneous and heterogeneous video content

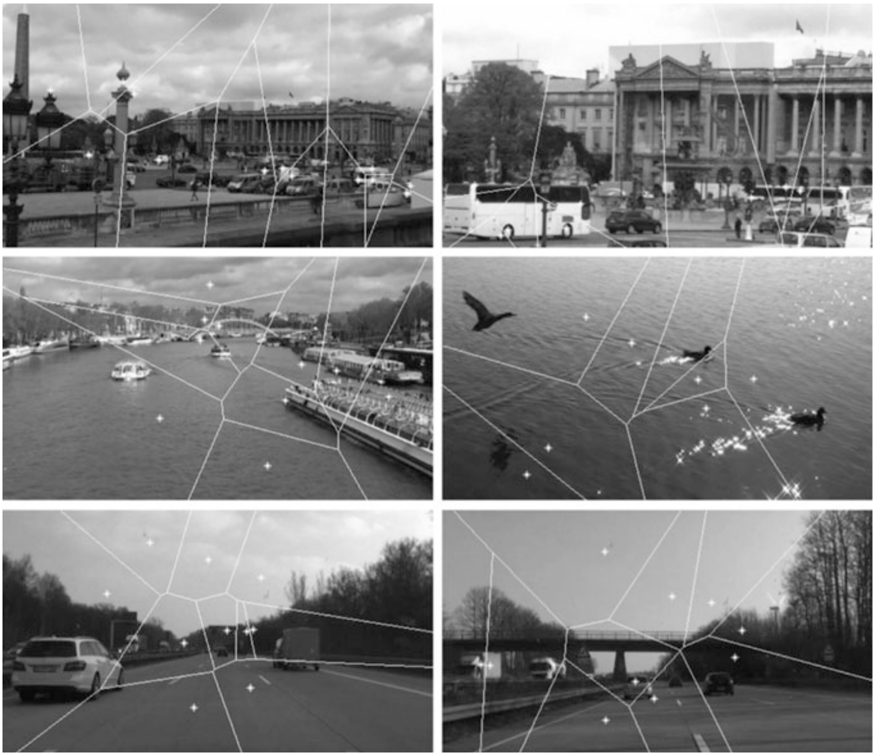

Fig. 14 Video frames extracted with slight and significant changes in content

a proof of the above statements, Table 2 shows entropy values for miniatures of frames form four video sequences (Table 3).

Frame numbers in the upper left corner indicate one of the four video sequences, they were taken from. Thus, the range of entropy for the first video sequence varies from 3.25 to 7.74. This video fragment is characterized by the greatest scene change among the analyzed materials. The range of entropy for the second video sequence equals to 4.12–6.59. Here, the content is of high heterogeneity as well, but it was shot in the nighttime, and the dark background narrows the range of visible details. The third material is characterized by homogeneity of changes, all the objects are shot under the same conditions, and they are very similar to each other, entropy range varies from 7.1 to 7.53. The fourth video has several scenes in it, entropy range here is quite diverse (from 5.85 to 7.75), simply not all the frames from the video have been included to the table.

Figure 13 illustrates graphics of entropy change for homogeneous and heterogeneous content from the third and the fourth video sequences included to the table above. For the demonstration purposes, fragments lasting nearly three minutes are taken from both videos, containing 29 frames per second. From the figure below it is obvious that entropy values computed for one of the videos change slightly,

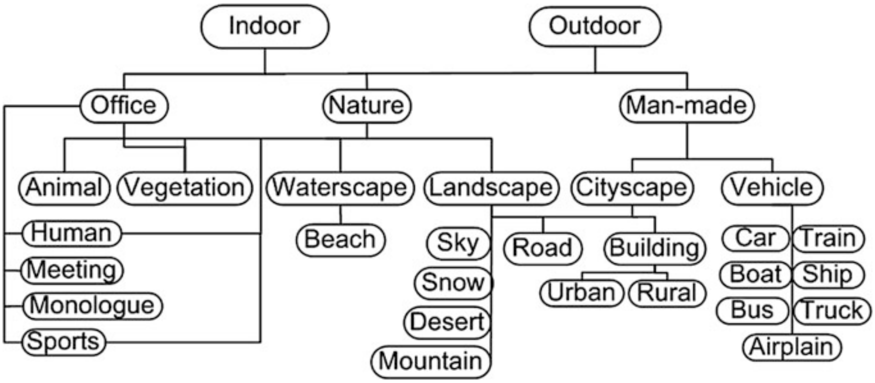

Fig. 15 TRECVid test video categories

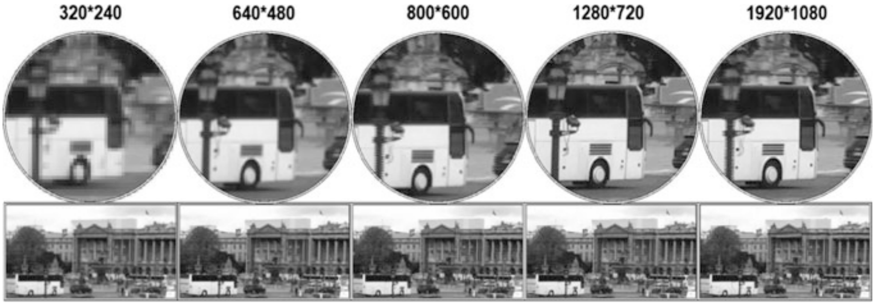

Fig. 16 Example of image information detailing enhanced with size and resolution

whereas for another one they change significantly. These research results form a background for initial analysis of image changes (Figs. 14, 15 and 16).

The whole procedure of matching Voronoi diagrams for image change detection is described below.

Step 1 Determine homogeneity of an image sequence. Calculate texture variance for all images in a sequence and according to it, set a threshold value by assuming the following rule:

$$Threshold = \begin{cases} \frac{1}{4}\ , & \frac{1}{K-1}\sum_{k=1}^{k=K}\left(E(B_k(z)) - \frac{1}{K}\sum_{k=1}^{k=K}E(B_k(z))\right)^2 \to \infty\ , \\ \frac{1}{2}\ , & \frac{1}{K-1}\sum_{k=1}^{k=K}\left(E(B_k(z)) - \frac{1}{K}\sum_{k=1}^{k=K}E(B_k(z))\right)^2 \to \frac{1}{K}\sum_{k=1}^{k=K}E(B_k(z)), \\ \frac{3}{4}, & \frac{1}{K-1}\sum_{k=1}^{k=K}\left(E(B_k(z)) - \frac{1}{K}\sum_{k=1}^{k=K}E(B_k(z))\right)^2 \to 0\ , \end{cases} \qquad (16)$$

where K indicates the total number of images/frames in a sequence, and $E(B_k(z))$ is the entropy value for k-th image/frame. The threshold value is set as follows: it should be less than $\frac{1}{4}$ for images/frames with heterogeneous content and a variety of scenes, it should be increased to $\frac{3}{4}$ for images/frames with homogeneous content and a small number of scenes (or even a single scene), otherwise it is set to $\frac{1}{2}$.

Step 2 Take the first $(B_k(z))$ and the second $(B_{k+1}(z))$ images/frames for comparison. Set $k = 1$.

Step 3 Compare Voronoi tessellations frame-by-frame to find significant changes in content. According to Eq. (15) calculate $\widehat{\rho}(B_k(z),\ B_{k+1}(z))$ for the two images/frames. If $\widehat{\rho}(B_k(z),\ B_{k+1}(z))$ turns out less than the preset threshold, then both images/frames $B_k(z)$ and $B_{k+1}(z)$ are considered containing significant changes (they may be reassigned as $B_r^*(z)$ and $B_{r+1}^*(z)$), and the procedure is continued from the next (4th) step. Otherwise, the procedure is continued from the 5th step.

Step 4 Reassign images/frames as $B_k(z) = B_{k+1}(z)$, $B_{k+1}(z) = B_{k+2}(z)$ and go to step 6.

Step 5 Remain $B_k(z) = B_k(z)$ and set $B_{k+1}(z) = B_{k+2}(z)$.

Step 6 Repeat step 3 until $B_{k+1}(z) \leq K$.

Step 7 Extract images/frames with significant changes by assuming scene boundaries to eliminate redundancy. Do not take into account similar frames from a single scene, if they are extracted, as they usually have much in common. Scene boundaries can be found through histogram difference, time series analysis or any other currently available technique [7]. Figure below illustrates frames of similar and different content, extracted from tourist's video. Frames with similar content reside horizontally.

Application of specifically designed similarity metrics (with orientation to low-level features in use) ensures high quality of image change detection. The procedure can be enhanced by utilization of weighted feature estimates, as in this case a linear combination only is used with the same weight coefficients. Additional features may also enhance the procedure. Determination of each feature importance is the prime direction of future research.

5 Test Collections, Processing Evaluation and Development Trends

Each application domain has its own specificity, but before operating with real-world industrial, medical or other video tracking objects, specialized test collection is needed. However, there is no universal video collection designed for such kind of purposes. That is why researchers choose video test samples by their own will, providing full descriptions in a form of video name, length, content type, number of frames/shots/scenes and the source it was taken. Publicly available collections from CERN Document Server (particle physics laboratory), Open Video Project (http://www.open-video.org), Movie Content Analysis Project (http://pi4.informatik.uni-mannheim.de/pi4.data/content/projects/moca/index.html), and Internet Archive (http://archive.org) are the main sources of video test samples. Sometimes commercials and self-made high-definition videos are taken into consideration as publicly available collections usually lack from high resolution. Custom video is also used for testing on specific application domain, such as bank office or street pedestrian tracking. The following figure illustrates commonly used categories of video samples that can be derived from datasets used for testing purposes by TRECVid community. Along with traditional video sources, clips from digital video libraries (DVL) can also be chosen: Informedia DVL (http://www.informedia.cs.cmu.edu), Consumer DVL (http://www.cdvl.org/), Hemispheric Institute DVL (http://hidvl.nyu.edu/), Harvard-Smithsonian Center for Astrophysics DVL (http://hsdvl.org/), etc. Materials from some digital video libraries cannot be downloaded, but they can be permitted for research usage under request. Popular human action recognition video datasets are: KTH, Weizmann, IXMAS, UCF50, HMDB51. Complete list of less famous data sets are available at http://www.datasets.visionbib.com/info-index.html and http://www.cvpapers.com/datasets.html [8].

All the open-source materials are usually provided of small size and low resolution. However, by reducing these image characteristics, small details are poured together with the background, and by enhancing these characteristics, detailing level of image information increases (see figure below). Thus, testing any image processing method with different size and resolution is one of the key point. The most frequently used video standards are considered PAL (720 * 576), HD (1280 * 720) and Full HD (1920 * 1080). The following table shows a list of currently available video standards which are reasonable to be used for testing purposes.

Someone may think that three to four test samples of different (or even the same) content type is enough, others perform testing of their proprietary methods on 20–100 movies lasting for more than an hour. Image and video genre has a great impact on visual features in use. That is the reason why to obtain truthful results, testing on different data types should be held: news, sports, cartoons, documentaries, talk-shows, if only the designed method is not oriented for a restricted application domain with homogeneous content [14].

Along the above considerations, test videos should include camera motion, zoom and lighting condition change because some video processing methods (especially those ones which are designed for tracking) could not cope with such changes. As it was already mentioned, any image and video processing method should be checked for quality and performance on different size and resolution data, and when speaking about image sequences and videos, these data should include huge and small inter-frame difference. It was noticed that methods showing sufficiently fine results for the most part of test collections, very often they fail with face recognition and processing of textual information that overlap in an image [8].

The accuracy of image changes detected is a subjective matter because respondents are involved in each evaluation that is almost always performed by precision (amount of found samples that turned out relevant) and recall (amount of found relevant samples from all the relevant ones). The benefit from using two measures simultaneously is obvious. In many cases one of the measures turns out to be more important. Some respondents do not appreciate obtaining any misclassified results, they want to get the least number of outputs, but all of them should be relevant (high precision). On the contrary, other respondents are interested in high recall, they treat low precision tolerantly. Precision and recall contradict with each other. Recall can always be increased to '1' with very low precision by simply returning all the values, no matter whether they are right or wrong. Recall does not decrease with increase in a number of wrong outputs (precision is decreased in such a case). Generally speaking, sufficient balance between these two measures should be reached. F-measure is an opportunity of doing so [18].

Along with traditional precision and recall measures, sometimes also percentage of wrong detected essences (e.g. significant changes) to missed ones can be computed. While the previous estimations are obtained from the information provided by the respondents with some supplementary calculations, criterions of informativeness and enjoyability simply depict users' feedback. Any method evaluation can be held using Fidelity measure and compression ratio. Actually, the estimation can be made in numerical, graphical or other form of comparison of results obtained after implementation of different image processing algorithms by showing benefits and drawbacks of the proposed one. Researchers and developers should check validity, performance and quality of image/video processing. It is considered that minimum 20 respondents are needed for reliable estimation. Expert estimates can be absolute and comparative. The first type implies presence of some viability scale while the second one assumes ranking the results or methods in terms of quality. The easiest way to express absolute expert estimates is to calculate an average value [14].

When using any clustering technique, in order to check whether clusters are well separated, cluster validity measure is used. This measure provides numerical information on distances within and between clusters. In most part of statistical software, ability to compute average/maximum/minimum distances within each cluster is realized, along with the sum of distances calculation ability, Euclidean distances, squared Euclidean distances, etc. Matlab silhouette plot visually demonstrates distances within and between clusters. Plot value of '+1' indicates that observations from different clusters are maximally remote from each other; plot value that resides near '0' indicates that these observations are too close to each other, and clusters could be assigned in a wrong manner; when a plot value tend to '−1', the observations are most probably related to clusters in a wrong manner.

Theoretical and practical results presented in this chapter facilitate indexing and archiving, summarization and annotating, searching and cataloguing large image sequences. In near future, works will be most probably concentrated on explorative approaches based on the middle level as a linking unit between low-level visual features and high-level concepts, and they have to meet the requirements for real-time processing in both performance and effectiveness [8]. The proposed approach to image change detection is unique and can be implemented in a range of applications aside from the intentional one.

6 Conclusion

An attempt towards high-level description of moving objects has shown that only specific motion patterns can be recognized by a machine. Current level of evolution does not permit implementation of semantically universal systems which could cope with a variety of content types [8]. Problem of machine-level video content comparison on frame-by-frame basis, i.e. image change detection, emerged several decades before. This chapter provides an overview on intelligent approaches utilized worldwide for solution of the aforementioned problem. Novel method proposed by the authors in this connection is also described in detail. It differs from existing ones by reliability and increased performance of change detection due to utilization of higher order Voronoi diagrams which assume all the advantages of operation with salient points.

Image sequence presentation using Voronoi diagrams with further their comparison provides means for machine interpretation of how objects are moved in space and time. Similar content in sequential video frames turned out to have identical diagrams, whereas the proposed metrics for diagram comparison ensure recognition of slight and significant changes. Voronoi diagrams of higher order simplify detailing image content compared with increasing the number of initial salient points [14]. Some other useful properties of higher order Voronoi diagrams have been revealed in relation to image change detection.

Evaluation of image and video processing is an integral part of any novel method development. Complexity of such evaluation lies in subjectivity of estimations provided by respondents. By analysing commonly applicable measures, several estimators have been marked out to test validity and performance of visual processing. The most frequently used measures of precision and recall can be combined in a form of Dice coefficient to obtain reasonable balance between them. Importance of performing tests on different image genres, high and low size and resolution, has been grounded. As long as clustering algorithms are widely applied in image and video processing, cluster validity measures have been also discussed. Datasets of open-source test video samples have been mentioned, among which TRECVid collection, Internet Archive, Movie Content Analysis Project and Open Video Project have gained greatest popularity. Less often used video libraries have been listed as well.

References

1. Sonka, M., Hlavac, V., Boyle, R.: Image Processing, Analysis, and Machine Vision, International Student Edition. Thomson, Toronto (2007)
2. Bezdek, J.C., Keller, J., Krisnapuram, R., Pal, N.R.: Fuzzy Models and Algorithms For Pattern Recognition and Image Processing. Springer, NY (2005)
3. Haralick, R.M., Shanmugam, K., Dinstein, I.: Textural features for image classification. IEEE Trans. Syst. Man Cybern. **3**(6), 610–621 (1973)

4. Deselaers, T., Weyand, T., Ney, H.: Image retrieval and annotation using maximum entropy. Evaluation of multilingual and multi-modal information retrieval. Lect. Notes Comput. Sci. **4730**, 725–734 (2007)
5. http://vision.middlebury.edu
6. Baker, S., Scharstein, D., Lewis, J.P., Roth, S., Black, M.J., Szeliski, R.: A database and evaluation methodology for optical flow. Int. J. Comput. Vis. **92**(1), 1–31 (2011)
7. Mikhnova, O., Vlasenko, N.: Key frame partition matching for video summarization. Int. J. Inf. Models Anal. **2**(2), 145–152 (2013)
8. Mashtalir, S., Mikhnova, O.: Key frame extraction from video: framework and advances. Int. J. Comput. Vis. Image Process. **4**(2), 67–78 (2014)
9. Mikhnova, O.: A template-based approach to key frame extraction from video. In: Proceedings of International Scientific and Technical Internet Conference on Computer Graphics and Image Recognition, pp. 120–127. VNTU, Vinnytsia (2012)
10. Yang, X., Tian, Q.: Video repeat recognition and mining by visual features. In: Schonfeld, D., Shan, C., Tao, D., Wang, L. (eds.) Video Search and Mining. Studies in Computational Intelligence, vol. 287, pp. 305–326. Springer, Berlin (2010)
11. Lee, W.-T., Chen, H.-T.: Histogram-based interest point detectors. Comput. Vis. Pattern Recogn. 1590–1596 (2009)
12. Ledoux, H., Gold, C.M.: Modelling three-dimensional geoscientific fields with the Voronoi diagram and its dual. Int. J. Geogr. Inf. Sci. **22**(5), 547–574 (2008)
13. Mikhnova, O.D.: Analiz videodannyh na osnove diagramm Voronogo razlichnogo poryadka. Zbirnyk naukovyh prats HUPS **1**(38), 142–145 (2014)
14. Mashtalir, S.V., Mikhnova, O.D.: Stabilization of key frame descriptions with higher order Voronoi diagram. Bionics Intell. **1**, 68–72 (2013)
15. Okabe, A., Boots, B., Sugihara, K., Chiu, S.N.: Spatial Tessellations: Concepts and Applications of Voronoi Diagrams. Wiley, Chichester (2000)
16. Mashtalir, V., Mikhnova, O., Shlyakhov, V., Yegorova, E.: A novel metric on partitions for image segmentation. In: Proceedings of International conference on Video and Signal Based Surveillance, pp. 1–6. IEEE CS, Washington (2006)
17. Sadahiro, Y.: Analysis of the relationship among spatial tessellations. J. Geogr. Syst. **13**(4), 373–391 (2011)
18. Manning, C.D., Raghavan, P., Schutze, H.: Introduction to Information Retrieval. Cambridge University Press, Cambridge (2008)