
Multimedia Tools and Applications, 26, 345–363, 2005

An Integrated Framework for Semantic Annotation and Adaptation

M. BERTINI bertini@dsi.unifi.it
D.S.I., Universitá di Firenze, Italy

R. CUCCHIARA cucchiara.rita@unimore.it
D.I.I., Universitá di Modena e Reggio Emilia, Italy

A. DEL BIMBO delbimbo@dsi.unifi.it
D.S.I., Universitá di Firenze, Italy

A. PRATI prati.andrea@unimore.it
D.I.I., Universitá di Modena e Reggio Emilia, Italy

Abstract. Tools for the interpretation of significant events from video and video clip adaptation can effectively support automatic extraction and distribution of relevant content from video streams. In fact, adaptation can adjust meaningful content, previously detected and extracted, to the user/client capabilities and requirements. The integration of these two functions is increasingly important, due to the growing demand of multimedia data from remote clients with limited resources (PDAs, HCCs, Smart phones). In this paper we propose an unified framework for event-based and object-based semantic extraction from video and semantic on-line adaptation. Two cases of application, highlight detection and recognition from soccer videos and people behavior detection in domotic* applications, are analyzed and discussed.

Keywords: semantic annotation, semantic adaptation, semantic transcoding, video adaptation, event detection, motion segmentation, performance evaluation

1. Introduction

Adapting videos to the user's requirements and terminal constraints is commonly referred to as *transcoding* [6, 19, 26, 38] or as *video adaptation* [16]. Video adaptation assumes that video is available in a certain format that is changed to a different one which is better suited to the context where the video is used. Transcoding techniques can be classified into *intermedia* and *intramedia*. Intermedia transcoding assumes that the media type changes from source to destination (e.g. a video-to-text application in which a notification message is sent whenever an event is recognized in the video). In intramedia transcoding source and destination media are the same (e.g. a video-to-video or audio-to-audio) and transcoding can be either homogeneous or heterogeneous: in heterogeneous transcoding [13, 25], source and destination have different codes and information re-use is impossible.

Domotics is a neologism coming from the Latin word domus (home) and informatics.

As far as video is concerned, transcoding is often performed in the compressed domain to avoid decompression and re-compression. In fact, this process introduces quantization errors and is not efficient [18, 29, 31]. Typical direct video transcoding methods are *requantization*, *spatial resolution downscaling*, and *temporal resolution downscaling* or a combination of them [18, 25]. Requantization is performed by decoding the DCT coefficients and re-quantizing them to fit bandwidth requirement [15, 34]. Spatial resolution downscaling is useful to fit, for instance, the screen size capability of a PDA and it consists in generating a video with lower spatial resolution [12, 27, 36]. Temporal resolution downscaling reduces the video frame rate by skipping frames either in a predefined way or on the basis of the amount of changes in the motion vectors [6, 13, 14, 32]. Transcoding for MPEG video can use adaptive quantization as reported in [8, 11, 22, 23, 35]. In backward adaptative quantization [37], quantizers are updated based only on the previously quantized data which are available to both the encoder and the decoder and has the advantage of avoiding the transmission of additional information to the decoding end. In forward adaptive quantization [22], the encoder updates the quantizer by probing both current and future inputs. Since the encoder's decision is based on information that is not available to the decoder, additional information must be given to specify the due changes.

Video transcoding is very effective if the code change is driven by video content [20]. With *semantic adaptation*,[1] the most meaningful parts of the video may have different coding than others, so as to adapt video transmission to both user's requirements and device's capabilities. In this case direct transcoding techniques working on compressed domain are not effective, since higher level semantics must be extracted in the uncompressed image domain. As an example, in the transmission of a video of a soccer game, we can send good quality video only for the frames where interesting actions take place, or within the individual frames, provide high resolution sampling only for the most relevant parts (e.g. those in the surrounding of the players). Extracting such events in the compressed domain can be a challenging task.

Research in semantic transcoding mostly concentrated on the extraction and separate coding of meaningful objects rather than of meaningful events with both spatial and temporal extension. Smith et al. in [26] proposed image analysis processes for content-based image transcoding using image type (e.g. graphs or photos) and image purpose classes. The IBM's Video Semantic Summarization Systems described in [24] exploits MPEG-7 for semantic transcoding: semantic annotation is provided manually by human experts; the user specifies his/her request in terms of preference topics, topic ranking, query keywords, and time constraint; the system outputs a video summary. In [20], Nagao et al. employ a video annotation editor that is capable of scene change detection, speech recognition, and correlation of scenes with the text obtained from the speech recognition engine. In this way, semantic indexes for video-to-document or video translation and summarization are produced. In [33], Vetro et al. presented an object-based transcoding framework that uses dynamic programming or meta-data, for the allocation of bits among the multiple objects in the scene.

Sports video and surveillance video are two of the most interesting subjects of investigation for semantic annotation and transcoding. In sports video, extraction of the most significant highlights is important for broadcasters, in order to build meaningful resumes

with no human effort and make them available at any user's terminal. In surveillance video, semantic annotation and transcoding are motivated by the growing interest in recognizing a specific person, detecting suspicious people's behaviours for law enforcement, monitoring people's behaviour for surveillance and make remote-assistance possible. In order to provide effective semantic transcoding of sports and surveillance, it becomes therefore important to have the possibility of making annotations of video highlights and more in general of semantic events, automatically, and understanding significant entities within the individual frames.

Automatic detection and recognition of highlights in sport videos has been an active research topic in recent years. Typical events of tennis have been modeled and detected in [28]. In [21], shots of basketball game are classified into one of three categories (crowd cheer, scoreboard display, change of direction). Developing on this classification, basket events (e.g. goal events) are detected when the shot sequence displays certain visual patterns. In [17] MPEG motion vectors are used to detect events. In particular, they exploit the fact that fast imaged camera motion is observed with typical soccer events, such as shot on goal or free kick. In [10], the playfield is divided into several distinct zones. The framed zone is identified using patterns of the playfield lines which appear in the image. The ball position is also used to perform detection of shot on goal and corner kick events. More recently, in [7], Ekin et al. have performed event detection in soccer video using both shot sequence analysis and shot visual cues. In particular, they assume that the presence of highlights can be inferred from the occurrence of one or several slow motion shots and from the presence of shots where the referee and/or the goal box is framed. In [2] Assfalg et al. provided a general model based on a Finite State Machine, to represent the spatio–temporal behavior of the most important highlights in soccer and defined a limited number of observable cues whose combinations determine the transition from one state to the other of the models. Highlights were detected using a model checking engine.

Automatic semantic annotation of surveillance video has also received great attention. In [1], a Bayesian network was used to detect human actions: by tracking the movement of the head of the subject several typical actions were recognized. In [30] Hongeng et al. proposed a method for the recognition of events that employs finite state automata, referred to as "scenarios". In [9] Cupillard et al. proposed an approach for the recognition of the behaviours of isolated individuals or group of people or crowd in the context of visual surveillance of metro scenes.

In this paper, we propose an integrated framework for semantic annotation and transcoding of non-compressed video that applies transcoding on the basis of spatio-temporal cues that are extracted automatically from the video stream. Please note that, though transcoding is usually referred to techniques in which input video are coded, we actually work on uncompressed video. However, since our videos are taken by video sources that introduce some sort of compression (for network limited capacities), we can properly refer to it with the term "transcoding" (or *video adaptation*).

Meaningful highlights are modeled with Finite State Machines. Transcoding is applied to the frame sequence where the highlight is detected and to the part of each frame where the event takes place. Applications to sports and surveillance video are presented.

The paper is structured as follows. The system framework and the metric for performance evaluation that have been used are described in Section 2 and 3, respectively. The details on the algorithms used for annotation and transcoding are reported in Section 4 for soccer videos and in Section 5 for indoor surveillance. Experimental results of both cases are also presented. Conclusions are reported in Section 6.

2. The proposed framework

A *class of relevance* is defined as the set of meaningful elements in which the user is interested in and that the system is able to manage. The importance of classes of relevance is twofold. First, the set of classes defines an ontology of the scenario that must be recognized, annotated, and provided to the user. Secondly, the user can exploit the classes of relevance in order to define his/her preferences about the video content. In addition, it can be used for performance evaluation purposes, as reported in the next section. Actually, for our purposes, the set of classes of relevance includes all the *events* and *objects* of the scene that can be automatically identified and transcoded.

Formally, a *class of relevance* C is defined as a pair $C = <o_i, e_j>$, where o_i represents an object class and e_j is an event class, selected between the set of object classes O and event classes E detectable by the system:

$$O = \{o_1, o_2, \ldots, o_n\} \cup \{\tilde{o}\}; \qquad E = \{e_1, e_2, \ldots, e_m\} \cup \{\tilde{e}\}$$

The special class $\tilde{o}$ includes all the areas of the image that do not belong to user-defined classes (for example, the part outside the soccer playfield can be considered as $\tilde{o}$). Analogously, the event $\tilde{e}$ includes all the non interesting events or the case of no-event.

The scheme reported in figure 1 displays the process of semantic transcoding adopted in this research. The semantic annotation engine extracts from the raw video the meaningful

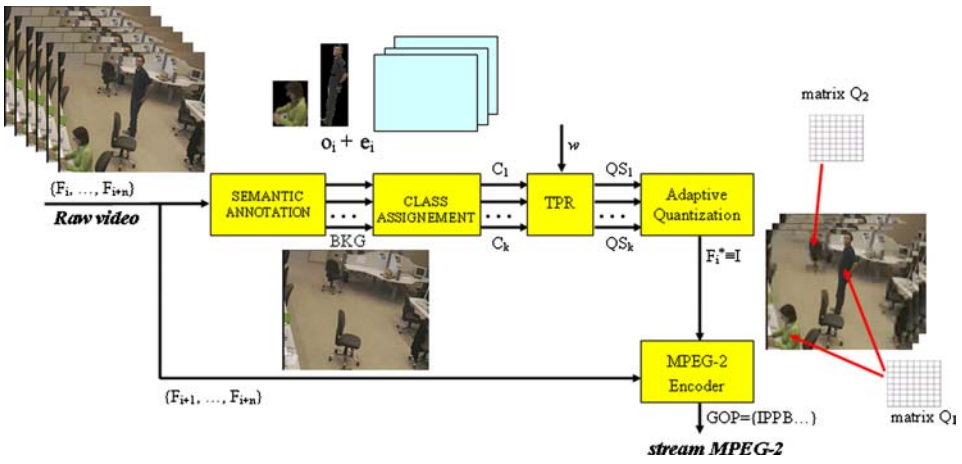

Figure 1. Scheme of the semantic annotation and transcoding (SAQ-MPEG) used.

objects (o_i) and the events (e_i). Then, objects and events are assigned to their class of relevance C_i. The TPR engine (Transcoding Policy Resolver) computes the quantization multipliers QS_i according to the user's defined relevance weights. By multiplying the QS_i with the MPEG-2 quantization matrix a quantization matrix for each class of relevance is obtained. Finally, a standard MPEG-2 encoder uses this coded frame as I frame and creates the GOP (Group Of Pictures) of the stream.

Automatic annotation performs the extraction of low-level features and their classification by means of high-level modules that are tailored on the specific application. For example, in the case of soccer videos, we partition the playfield into a number of different zones with slight overlapping and use the motion of the main camera as a cue for the description of the evolution of the play. Each event is modeled with a Finite State Machine, where key actions, defined in terms of the estimated cues, determine the transition from one state to the following. The event models are checked against the current observations, using a model checking algorithm. The objects of interest extracted are the playfield zones and the background. They are classified using Naïve Bayes classifiers. A short description of this subsystem is reported in Section 4, while interested readers may consult the detailed description provided in [3].

The semantic transcoding system can employ three different solutions. In the first solution (S-MJPEG), the extracted objects are encoded separately by considering the weights assigned to their class (S-MJPEG). Each object is sent in a separated image with the associated alpha-plane mask. Also the background is sent in a separated image (one every n frames, with n changing dynamically). At the client side, the decoder (non standard) superimposes the objects to the current background. Coding is therefore made *without any temporal prediction* but exploits the semantics to enhance the ratio quality/bandwidth [5, 6]. The second solution, called *semantic spatial transcoding* (SS-MJPEG), extends S-MJPEG. The frame size is adapted according to the user's display size, centering and resizing the most meaningful objects, and then compressed using a semantic coding as in the case of S-MJPEG. Thus, resolution and quality of meaningful parts of the video are preserved.

The third solution (SAQ-MPEG), employs the MPEG-2 standard and its capabilities of temporal prediction, to reduce the required bandwidth and produce a video that can be played by a standard decoder. The semantics extracted is used to drive the *adaptive quantization* of frame I in the MPEG stream (see figure 1). This results into standard MPEG stream, but with different compression, according to the image region that is under examination.

3. Performance evaluation metric

Performance evaluation of annotation and transcoding systems is typically based on a comparison with ground-truthed data obtained from manual annotation. In the case of annotation, comparison is made at object- or event-level by collecting errors or computing a confusion matrix with false positives and negatives. Instead, in the case of transcoding, the comparison is usually at pixel-level by computing figures, such as the PSNR (Peak

Signal-to-Noise Ratio), that evaluate the difference between original and distorted (adapted) images.

For semantic transcoding, since the user can access specific parts of the video he/she is interested in, a weighted version of PSNR is more appropriate as global measure of system performance.

The following definition can be therefore used as in [6]:

$$WPSNR = 10 \log_{10} \left(\frac{V_{\mathrm{MAX}}^2}{WMSE} \right) \tag{1}$$

where V_{MAX} is the maximum (peak-to-peak) value of the signal to be measured and *WMSE* is the Weighted Mean Square Error, computed as:

$$WMSE = \sum_{k=1}^{N_{CL}} w_k \cdot MSE_k \tag{2}$$

where N_{CL} is the number of classes of relevance and with the *MSE* defined as:

$$MSE_k = \frac{1}{|C_k|} \sum_{p \in C_k} d^2(p) \tag{3}$$

where C_k is the set of the points belonging to the user-defined class k and $|C_k|$ its cardinality; $d(p)$ is a properly defined distance that measures the error between original and distorted images in the point p. As a distance, we use the *Euclidean distance* in the RGB color space (different color spaces can be used, but with similar comparison results). The weights w_k can be used to set the preferences for each class. For instance, in the case the user is particularly interested in the playfield and almost disregards other parts of the image, he/she can set the weights to 0.95 and 0.05, respectively.

Since the *WPSNR* is defined for a generic frame j the *PSNR* for the whole video can be defined as $PSNR = \frac{\sum_{j=1}^{NF} WPSNR_j}{NF}$ [5, 6], where *NF* is the number of frames. In our framework we have modified the previous metric to take into account the performance of the annotation engine (for each class of relevance) and the fact that the metric refers to events that have a finite temporal extension. Therefore we will have different *PSNR* for different frame sequences within the video stream.

In particular, the following measures must be taken into account:

$CR_k = (C/Overall)_k$ the ratio between the number of highlights correctly detected (C) over the total number of highlights for class k.

$FR_k = (F/Overall)_k$ the ratio between the number of falsely detected highlights (F) over the total number of highlights for class k.

$MR_k = (M/Overall)_k$ the ratio between the number of missed highlights (M) over the total number of highlights for class k.

Hence, from Eq. (2) we can define the following performance indexes for integrated semantic transcoding:

$$\begin{aligned} WMSE_{CR} &= \sum_{k=1}^{N_{CL}} w_k \cdot MSE_k \cdot CR_k \\ WMSE_{FR} &= \sum_{k=1}^{N_{CL}} w_k \cdot MSE_k \cdot FR_k \\ WMSE_{MR} &= \sum_{k=1}^{N_{CL}} w_k \cdot MSE_k \cdot MR_k \end{aligned} \tag{4}$$

that provide respectively measures of:

- Weighted Mean Square Error according to the recognition rate by the annotation engine for each class k.
- Weighted Mean Square Error due to falsely recognized highlights by the annotation engine for each class k (it results into some excess of bandwidth requirements).
- Weighted Mean Square Error due to missed highlights by the annotation engine for each class k (it results into some loss of quality in the transcoded data).

A global performance measure for a frame sequence is obtained by summing up these three figures over the individual frames:

$$WPSNR = 10 \log_{10} \left(\frac{V_{\text{MAX}}^2}{WMSE_{CR} + WMSE_{FR} + WMSE_{MR}} \right) \tag{5}$$

4. Semantic annotation and adaptation of soccer videos

Automatic annotation and adaptation of soccer video is becoming more and more important since telecommunication providers are ready to offer innovative services of live sport events on last generation cell phone or PDAs with very low bandwidth.

Inspection of videos showed that producers of soccer video use a main camera to follow the action of the game. The main camera is positioned along one of the long sides of the playing field. Due to the fact that the play always takes place next to the ball, it is fair to assume that the main camera follows the ball. Accordingly, ball motion is conveniently described in terms of pan/tilt camera motion, and estimated from the apparent image motion induced by the camera action. The validity of this assumption has been confirmed in the experiments carried out in the EU-ASSAVID project [2]. Identification of the part of the playing field currently framed and camera actions are the most significant features that can be used to describe and identify relevant game events. Recognition of the framed playfield zone is also useful for transcoding within the individual frames.

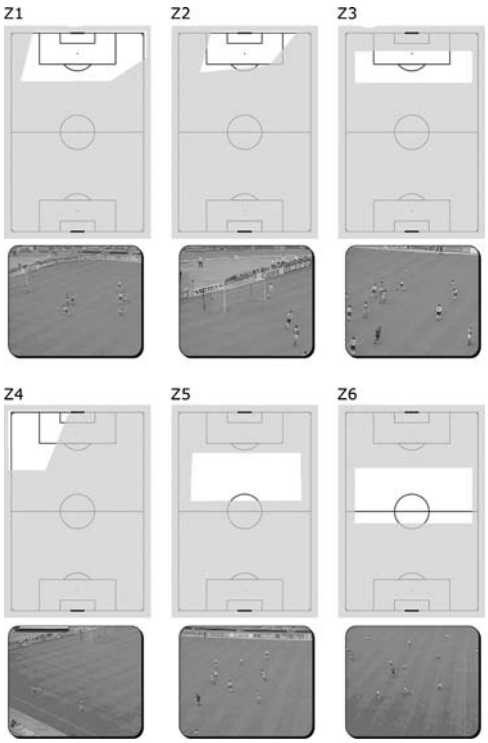

Figure 2. Playfield zones, Z7 to Z12 are symmetrical.

Events that are detected are *forward launch*, *shot on goal*, *turnover*, *placed kick* and *counterattack*. Placed kicks include penalty kicks, free kicks next to the goal box, and corner kicks. Counterattacks are a sequence of other basic events: a turnover followed by a forward launch.

The soccer playfield has been divided in 12 zones, 6 for each side (figure 2). These zones may overlap, and were chosen with support of a domain expert, so that passing from one to the other has a specific meaning in the way in which the play action evolves. The low-level features used to recognize the playfield zones are *playfield shape* and the *playfield lines*. The first one is extracted from color information, through color histogram analysis. This step is then followed by a processing chain of k-fill, flood fill and the erosion and dilation morphological operators. Then, the bitmap image is represented using a polygonal shape. The playfield lines are extracted from the edge map of the original image, and then creating a vectorial representation through a stick growing algorithm. Close and collinear segments are then merged, and length and color information is used to discard segments due to players. The classification of the playfield zone is performed according to a number of attributes derived from these low-level cues: playfield shapes descriptor, playfield line orientation descriptor, playfield size descriptor, playfield corner position, midfield line descriptor.

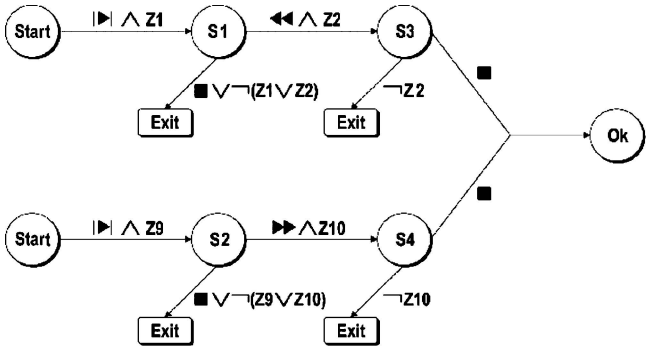

Figure 3. Shot model: the arcs report the camera motion and playfield zones needed for the state transition.

Twelve independent Naıve Bayes classifiers have been used to classify the playfield zones shown in the video. Each highlight is modeled with a Finite State Machine, where key events, defined in terms of the playfield zone currently framed and motion of the ball, determine the transition from one state to the following. For instance, the forward launch requires that the ball moves quickly from midfield toward the goal box. This approach was chosen over other methods, such as HMMs, due to the fact that it does not need a very large training set, to cope with all the possible occurrences of an highlight, and it is easier to deal with the temporal duration of an event, that is necessary to define a class of relevance. The highlight models are checked against the current observations, using a model checking algorithm. figure 3 shows the FSM for the shot on goal. The model is composed of 4 states: $Start$, OK, and two other states for each side of the playfield. Logical symbols (and, or, not) are used to combine visual cues extracted from the video stream. For example the sentence "play is around the left goal box" is modeled by expression written above the arc connecting $S1$ to $S2$. This expression is made by two constraints, one related to the motion of the play (direction toward left and fast motion), and the other one (Z_2) related to the current framed playfield zone (which is the part of the playfield surrounding the goal box, as shown in figure 2).

In conclusion, in the case of soccer videos, our system is able to extract at least the following object and event classes:

$$O = \{Z_1, \ldots, Z_{12}\} \cup \{\tilde{o}\}; \qquad E = \{FL, SG, TO, PK, CO\} \cup \{\tilde{e}\}$$

where $Z_1 \ldots Z_{12}$ are the twelve playfield zones detected; FL is forward launch, SG is shot on goal, TO is turnover, PK is placed kick, CO is counterattack.

It must be noted that the events belonging to the E class extend over a certain amount of time, from the beginning to the end of the action, and are not punctual. The system has been tested on about one hour of soccer videos, including 85 sequences, selected from 15 European competitions. Table 1 reports the results: it can be noticed that, for most of the highlights, correct detection is over 90%. False detection of shots on goal is due to attack actions near the goal box. The model for turnovers was designed so as to achieve a low missed detection, even if this results in a high false detection rate. This was done since they

Table 1. Highlight classification results.

	Detected	Correct	Missed	False
Forward launch	36	32	1	4
Shot on goal	18	14	1	4
Turnover	20	10	3	10
Counterattack	3	3	1	0
Placed kick	13	13	2	0

are used to detect counterattacks, that are a composition of turnovers followed by forward launches. Therefore this false detection is not a serious problem, since false turnover can be discarded after checking for forward launches.

Transcoding can be applied to objects and events detected by the annotation engine. For example suppose that the classes of relevance defined by the user are the followings: $C_1 = <^*, \tilde{e}> \vee <\tilde{o}, \tilde{e}>$, $C_2 = <\tilde{o},^* >$, $C_3 = <^*, e_x>$, and $C_4 = <^*, e_y>$, with $e_x \in \{FL, TO, CO\}$, $e_y \in \{SG, PK\}$ and $*$ represents the case of "any event" or "any object".

A user will be probably more interested in class C_4 to see the most relevant actions at the best quality, at least in the playfield zone. Besides, the user does not want to waste bandwidth with actions of no interest, thus class C_2 will be less relevant to him/her. Possible values of the weights could be $w = \{w_1, w_2, w_3, w_4\} = \{0.005, 0.005, 0.20, 0.79\}$. As a consequence of this definition, the transcoding system will heavily compress the classes C_1 and C_2 (or even not sending data in the case of not relevant events), by using an average quality in the C_3 case, and by preserving as much of the quality as possible for the class C_4.

To evaluate the performance of the semantic transcoding proposed, this has been compared with a standard "syntactic" transcoding based on JPEG and MPEG-2. S-MJPEG and SAQ-MPEG "semantic" transcoding methods have been used. The figures reported in the following display the performance expressed in terms of $PSNR$ and $WPSNR$. Figure 4 shows the graph of the PSNR of two different compression levels of standard JPEG compared with our S-MJPEG. During this video some events occur and objects (as playfield zones or $\tilde{o}$) are detected. From frame 12 to 26 a false detection of FL occurs, that is encoded using the C_3 class of relevance, and thus resulting in a small waste of bandwidth. No missed detection occurs.

The S-MJPEG code operates a different video adaptation depending on the classes of relevance. The corresponding average bandwidth is reported in the caption. It is possible to note that the bandwidth occupation of our method (S-MJPEG) is comparable to that of JPEG at high compression level ($C = 80$). Using a lower compression level (JPEG $C = 20$) increases significantly the PSNR but the bandwidth required is about 3 times higher than in the other cases, that is impractical for slow connection devices, such as PDAs or cell phones. In figure 4, it can be observed that S-MJPEG obtains similar results as JPEG with $C = 80$ for the classes C_2 and C_3, and, obviously, lower performance for the class C_1 (frames from 26 to 87 or from 202 to 250) that is of no interest for the user and is therefore heavily degraded. On the other hand, in the case of interesting events

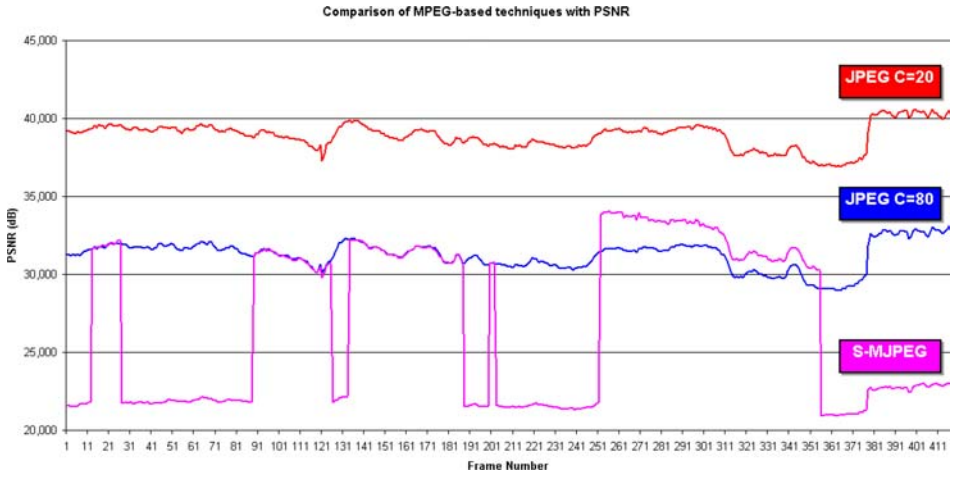

Figure 4. Comparison of JPEG-based techniques with our S-MJPEG with standard PSNR. Average bandwidth occupations are 1128.63 kbps for JPEG $C = 20$, 423.46 kbps for JPEG $C = 80$, and 432.11 kbps for our S-MJPEG.

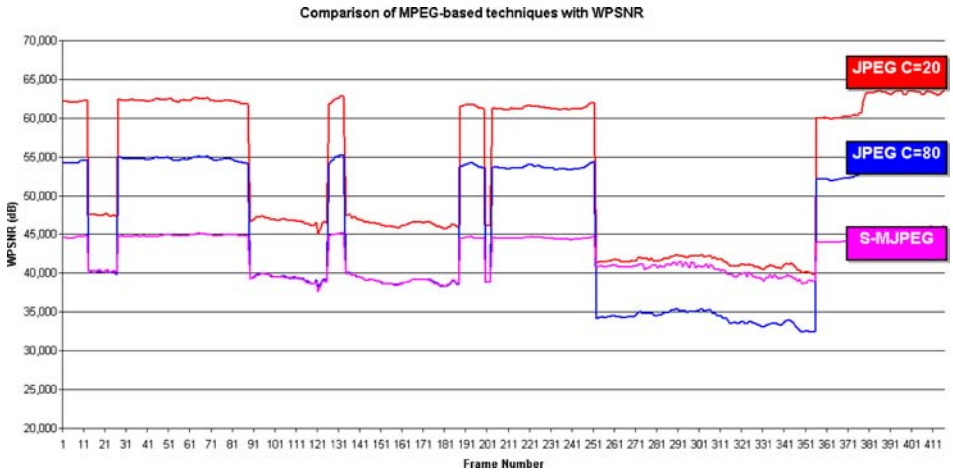

Figure 5. Comparison of JPEG-based techniques with our S-MJPEG with weighted PSNR. Bandwidth occupations are the same as above.

(class C_4) S-MJPEG outperforms JPEG of comparable bandwidth (frames from 251 to 354).

A similar behaviour can be observed in figure 5 where WPSNR is used. In this case, S-MJPEG has almost the same quality of the low compression JPEG (with $C = 20$). This is due to the fact that in the frames from 251 to 354 we detected both objects of class C_4 (the playfield in the case of event "shot at goal") and of the class C_2 (outside the playfield). Besides we should consider that the frames in the test sequences used have backgrounds that change from one frame to the other. This fact is the worst case situation for the S-MJPEG, that

send background objects every frame. Visual results of S-MJPEG are reported as example in figure 6: figure 6(a) reports a frame associated with class C_1, (b) associated with classes C_2 or C_3, figures (c) and (d) one image in the case of interesting events, thus associated with class C_4.

Comparison with standard transcoding based on MPEG shows even more interesting results. In this case SAQ-MPEG method has been tested, with similar bandwidth allocation. Figure 7 presents PSNR of MPEG-compressed video and the PSNR of SAQ-MPEG for each class of relevance. Weights were taken as $w = \{0.005, 0.005, 0.20, 0.79\}$. It is possible to notice that for the classes C_1 and C_2 (the less relevant ones) the standard method outperforms SAQ-MPEG. Instead, for relevant classes (C_3 and C_4) the SAQ-MPEG (and therefore the use of adaptive quantization) provides higher quality, particularly for the most relevant events (C_4). Similar results are obtained with weighted PSNR as shown in figure 8. It must be noted that the video examples used were particularly suited to standard prediction techniques due to the fact that large parts of the frames have uniform green color.

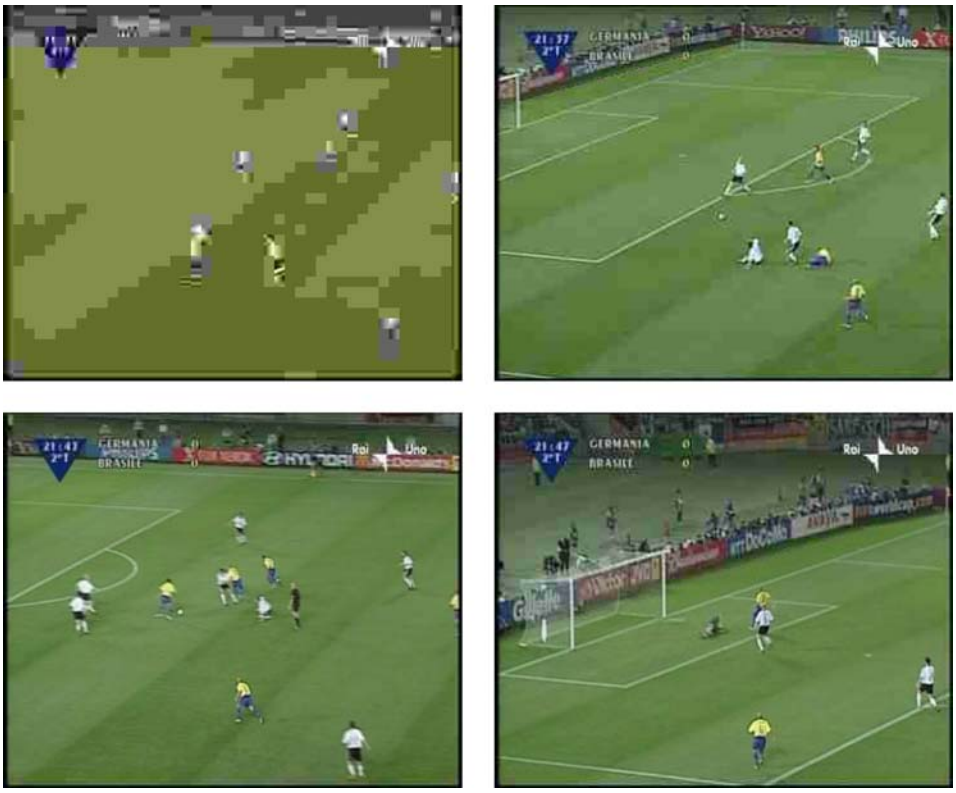

Figure 6. Examples of the S-MJPEG transcoding results on the soccer video: (a) frame with no interesting events (class C_1), (b) frame with low interest objects and/or events (classes C_2 and C_3), and (c) and (d) two examples of class with high-interest objects and events (class C_4).

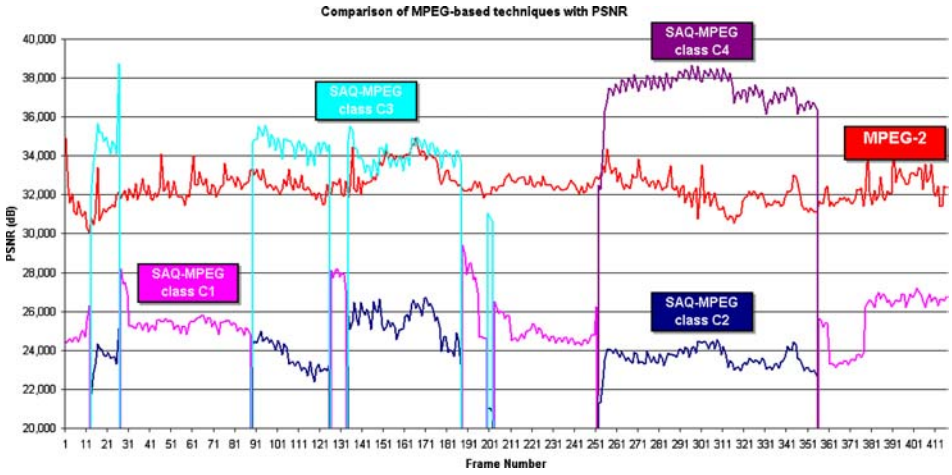

Figure 7. Comparison of MPEG-based techniques with our SAQ-MPEG with standard PSNR. Bandwidth occupations are 233.88 kbps for MPEG-2 and 253.25 kbps for SAQ-MPEG total.

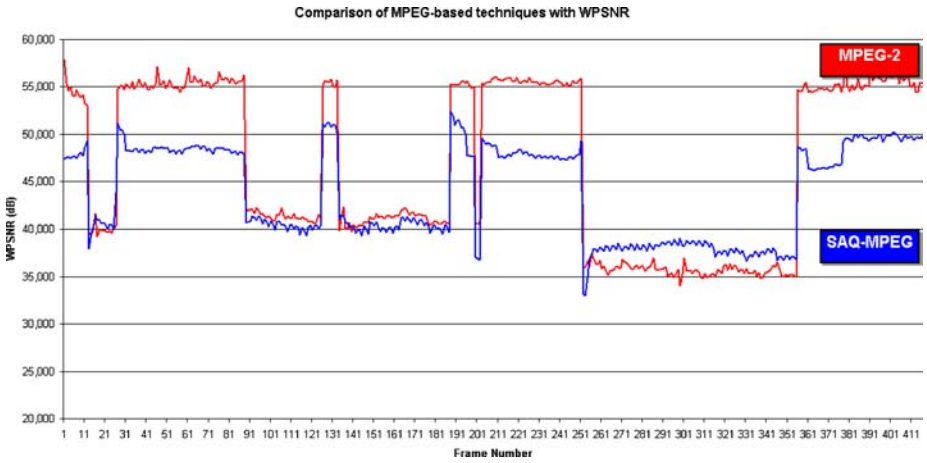

Figure 8. Comparison of MPEG-based techniques with our SAQ-MPEG with weighted PSNR. Bandwidths occupations are the same as above.

5. Semantic annotation and adaptation for indoor surveillance for domotic applications

The reference scenario used in in semantic annotation and adaptation of surveillance video is referred to a domotic application in which tele-presence and tele-viewing are essential for the safety of disabled people. In this scenario the staff taking care of a disabled person can use a PDA to monitor his/her condition continuously, or when an event occurs. Since PDAs have limited resources and (typically) a low Internet connection event

annotation is performed, followed by downscaling transcoding to adapt to the display capability.

Objects are extracted from the scene acquired with a static camera, by background suppression. A statistical, knowledge-based updating process is used to react quickly to scene changes as presented in [4]. Moving objects are classified into two distinct classes, namely "people" and "others". People objects are recognized according to a number of features that take into account shape and color. In particular the ratio between height and width of the silhouette is used as a first cue. The presence of a person face is detected using both color and edge data: a generalized Hough transform is used to check for the presence of elliptical patterns, and the ellipsis that has the color histogram most similar to a face model histogram is selected and tracked along time. The posture of the person is recognized considering head and feet positions and a-priori knowledge of the 3D scene geometry. Features that are obtained from the sequence analysis are to input to FSM (similarly to the case of soccer video highlight detection) that detects whether the person is actually "walking", "sitting", "falling down", or "laying down". Our experiments showed no misdetection and no false detection of these events over 2 hour video sequences used for test.

The objects and the events that the annotation engine detects are the following:

$$O = \{FF, FP, FO\} \cup \{\tilde{o}\}$$
$$E = \{PW, PS, PF, PL\} \cup \{\tilde{e}\}$$

where FF is Foreground Faces, FP is Foreground People (the person full body), FO is other Foreground Objects, and the background ($\tilde{o}$). As events we use PW to indicate a person that is walking, PS for a person sitting, PF for one falling down (as transition between a state of standing or sitting and one of laying down), and PL for a person laying down on the floor.

Three classes can be identified: $C_1 = <FP, PL>$, $C_2 = <FP, e_x>$ (with $e_x \in E, e_x \neq \{PL\}$), and $C_3 = <\tilde{o}, *>$. Possible weights are $w = \{w_1, w_2, w_3\} = \{0.7, 0.2, 0.1\}$, i.e. the user is very interested in having the image of the fallen person (C_1) at the best quality possible, while in other situations the image of the person (C_2) can be sent with lower quality.

In this scenario, since live video is concerned, SAQ-MPEG can not be used, being necessary a forward prediction (for B-type frame) that are not available as far as no delay in the transmission is admitted. Therefore experiments will compare semantic spatial transcoding method (SS-MJPEG, that use adaptive transcoding) against S-MJPEG (that only encodes objects and events, with no frame size adaptation), the spatial resolution downscaling classical method (SRD, that uses fixed transcoding) and the standard JPEG based method.

In figure 9 frame *9a* and *b* show results of the SS-MJPEG transcoding method. Frame *9c* and *d* present the results obtained with the spatial resolution downscaling with fixed transcoding. It is worth notice that spatial resolution downscaling still displays the whole scene but with much lower quality than SS-MJPEG.

Table 2 reports a quantitative analysis of the performance of SS-MJPEG with respect to S-MJPEG, SRD and JPEG, in terms of PSNR, WPSNR and required bandwidth. WPSNR

Table 2. Numerical Results of the Comparison (original bandwidth = 24304.22 kbps).

Transcoding policy	PSNR	WPSNR $w_i = [0.9, 0.1, 0]$	WPSNR $w_i = [0.9, 0.1]$	Req. bandwidth (kbps)
SS-MJPEG	33.98	31.32	32.23	181.74
S-MJPEG	31.37	29.98	25.58	118.32
SRD (C = 20)	26.89	24.85	26.11	494.34
JPEG (C = 20)	41.09	38.05	38.42	1342.42
JPEG (C = 80)	30.35	27.54	28.05	511.28

Figure 9. Frames with semantic spatial transcoding (a) and (b) or spatial transcoding (c) and (d).

has been analysed for two different cases where they are considered: (a) person's face (FF), person's full body (FP) and background $\tilde{o}$ (in this case weights $w = \{0.9, 0.1, 0\}$ are assigned); and (b) person's full body and background only (in this case weights $w = \{0.9, 0.1\}$ are assigned).

The classical JPEG transcoding method at low compression rates ($C = 20$) has the highest bandwidth requirements, the quality of the compressed images is obviously very high. On the other hand, JPEG with higher compression rate, while reducing the required bandwidth, presents bad performance for quality. SRD has bandwidth requirement similar to JPEG at $C = 80$, but with lower quality. S-MJPEG and SS-MJPEG both show better performance in bandwidth requirements and, moreover, they achieve a good quality performance, as soon as referred to the correct weights in the WPSNR. It is shown that whenever the transcoding applies a higher compression to the background and lower to the person, and the WPSNR is measured viceversa by applying a larger relevance (weight) to the background, the result for S-MJPEG falls down (fourth column of Table 2. SS-MJPEG appears to outperform all the other methods as to the quality of the compressed video. It should be considered that although the quality of the compressed video could not, in some cases, be perceived at a visual inspection, it actually prevents any further processing of the video stream.

6. Conclusions

In this paper we have proposed an unified framework for event-based and object-based semantic extraction from video and semantic on-line adaptation. Two cases of application, highlight detection and recognition from soccer videos and people behavior detection in domotic applications, were analysed and discussed.

Results have shown that semantic transcoding performance is dependent on the performance of the annotation engine, but provides high performance both in terms of bandwidth and quality for recognition rate from 90–100% of the annotation engine.

Acknowledgments

This work is partially funded by the "Domotics for disability" by Fondazione CRM, the FIRB/WP1 "Perf" project, and by European Commission under grant IST-1999-13082 (`http://viplab.dsi.unifi.it/ASSAVID`)

Note

1. Many researchers refer to this technique as *semantic transcoding*.

References

1. J.K. Aggarwal and A. Madabhushi, "A bayesian approach to human activity recognition," in Proc. of the Second IEEE International Workshop on Visual Surveillance (CVPR workshop), Fort Collins, CO (USA), June 1999, pp. 25–30.
2. J. Assfalg, M. Bertini, C. Colombo, A. Del Bimbo, and W. Nunziati, "Automatic interpretation of soccer video for highlights extraction and annotation," in Proceeedings of the ACM Symposium on Applied Computing, March 2003, pp. 769–773.
3. J. Assfalg, M. Bertini, C. Colombo, A. Del Bimbo, and W. Nunziati, "Semantic annotation of soccer videos: Automatic highlights identification," Computer Vision and Image Understanding, Vol. 92, No. 2/3, pp. 285–305, 2003.
4. R. Cucchiara, C. Grana, M. Piccardi, and A. Prati, "Detecting moving objects, ghosts and shadows in video streams," in press on IEEE Transcations on Pattern Analysis and Machine Intelligence, 2003.
5. R. Cucchiara, C. Grana, and A. Prati, "Semantic transcoding for live video server," in Proceedings of ACM Multimedia 2002 Conference, December 2002, pp. 223–226.
6. R. Cucchiara, C. Grana, and A. Prati, "Semantic video transcoding using classes of relevance," International Journal of Image and Graphics, Vol. 3, No. 1, pp. 145–169, 2003.
7. A. Ekin, A. Murat Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization," IEEE Transactions on Image Processing, 2003 (to appear).
8. D. Farin, M. Ksemann, P.H.N. de With, and W. Effelsberg, "Rate-distortion optimal adaptive quantization and coefficient thresholding for MPEG coding," in 23rd Symposium on Information Theory in the Benelux, May 2002.
9. F. Brémond, F. Cupillard, and M. Thonnat, "Behaviour recognition for individuals, groups of people and crowd," in IEEE Proc. of the IDSS Symposium—Intelligent Distributed Surveillance Systems, London (UK), February 2003.
10. Y. Gong, L.T. Sin, C.H. Chuan, H. Zhang, and M. Sakauchi, "Automatic parsing of tv soccer programs," in Proceedings of IEEE Int'l Conference on Multimedia Computing and Systems, 1995, pp. 15–18.

11. C.A. Gonzales and E. Viscito, "Motion video adaptive quantization in the transform domain, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 1, No. 4, pp. 374–378, 1991.
12. M.R. Hashemi, L. Winger, and S. Panchanathan, "Compressed domain motion vector resampling for downscaling of MPEG video, in Proceedings of IEEE Int'l Conference on Image Processing, Vol. 4, pp. 276–279, 1999.
13. K.-L. Huang, Y.-S. Tung, J.-L. Wu, P.-K. Hsiao, and H.-S. Chen, "A frame-based mpeg characteristics extraction tool and its application in video transcoding, IEEE Transcations on Consumer Electronics, Vol. 48, No. 3, pp. 522–532, 2002.
14. J. Hwang, T. Wu, and C. Lin, "Dynamic frame-skipping in video transcoding," in Proceedings of the IEEE Second Workshop on Multimedia Signal Processing, 1998, pp. 616–621.
15. G. Keesman, R. Hellinghuizen, Fokke Hoeksema, and Geert Heideman, "Transcoding of MPEG bitstreams," Signal Processing: Image Communication, Vol. 8, No. 6, pp. 481–500, 1996.
16. J.-G. Kim, Y. Wang, and S.-F. Chang, "Content-adaptive utility-based video adaptation," in Proceedings of IEEE Int'l Conference on Multimedia Computing and Expo, 2003.
17. R. Leonardi and P. Migliorati, "Semantic indexing of multimedia documents," IEEE Multimedia, Vol. 9, No. 2, pp. 44–51, 2002.
18. Y. Liang and Y-P. Tan, "A new content-based hybrid video transcoding method," in Proceedings of IEEE Int'l Conference on Image Processing, Vol. 1, 2001, pp. 429–432.
19. R. Mohan, J.R. Smith, and C. Li, "Adapting multimedia internet content for universal access," IEEE Transactions on Multimedia, Vol. 1, No. 1, pp. 104–114, 1999.
20. K. Nagao, Y. Shirai, and K. Squire, "Semantic annotation and transcoding: Making web content more accessible," IEEE Multimedia, Vol. 8, No. 2, pp. 69–81, 2001.
21. S. Nepal, U. Srinivasan, and G. Reynolds, "Automatic detection of 'goal' segments in basketball videos," in Proceedings of ACM Multimedia, 2001, pp. 261–269.
22. A. Ortega and K. Ramchandran, "Forward-adaptive quantization with optimal overhead cost for image and video coding with applications to MPEG video coders," in SPIE Digital Video Compression, February 1995.
23. K. Ramchandran and M. Vetterli, "Rate-distortion optimal fast thresholding with complete JPEG/MPEG decoder compatibility," IEEE Transactions on Image Processing, Vol. 3, No. 5, pp. 700–704, 1994.
24. IBM research. http://www.research.ibm.com/MediaStar/VideoSystem.html.
25. T. Shanableh and M. Ghanbari, "Heterogeneous video transcoding to lower spatio-temporal resolution and different encoding formats," IEEE Transactions on Multimedia, Vol. 2, No. 2, pp. 101–110, 2000.
26. J.R. Smith, R. Mohan, and C. Li, "Content-based transcoding of images in the internet," in Proceedings of IEEE Int'l Conference on Image Processing, October 1998, Vol. 3, pp. 7–11.
27. J. Song and B.-L. Yeo, "Fast extraction of spatially reduced image sequences from MPEG-2 compressed video," IEEE Transactions on Circuits and Systems for Video Technology, Vol. 9, No. 7, pp. 1100–1114, 1999.
28. G. Sudhir, J.C.M. Lee, and A.K. Jain, "Automatic classification of tennis video for high-level content-based retrieval," in Proceedings of Int'l Workshop on Content-based Access of Image and Video Databases, 1998.
29. H. Sun, A. Vetro, J. Bao, and T. Poon, "A new approach for memory-efficient atv decoding," IEEE Transcations on Consumer Electronics, Vol. 43, No. 3, pp. 517–525, 1997.
30. F. Brémond S. Hongeng and R. Nevatia, "Representation and optimal recognition of human activities," in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition CVPR00, South Carolina (USA), June 2000.
31. A. Vetro, C. Chrisopoulos, and H. Sun, "Video transcoding architectures and techniques: An overview," IEEE Signal Processing Magazine, Vol. 20, No. 2, pp. 18–29, 2003.
32. A. Vetro and H. Sun, "Encoding and transcoding multiple video-objects with variable temporal resolution," in Proceedings of Intern. Symposium of Circuit and Systems, May 2001.
33. A. Vetro, H. Sun, and Y. Wang, "Object-based transcoding for adaptable video content delivery," IEEE Transactions on Circuits and Systems for Video Technology, Vol. 11, No. 3, pp. 387–401, 2001.
34. O. Werner, "Requantization for transcoding of MPEG-2 bit streams," IEEE Transactions on Image Processing, Vol. 8, No. 2, pp. 179–191, February 1999.
35. P.H. Westerink, R. Rajagopalan, and C.A. Gonzales, "Two-pass MPEG-2 variable-bitrate encoding," IBM Journal of Research and Developement, Vol. 43, No. 4, July 1999.

36. C. Yim and M.A. Isnardi, "An efficient method for dct-domain image resizing with mixed field/frame-mode macroblocks," IEEE Transactions on Circuits and Systems for Video Technology, Vol. 9, No. 5, pp. 696–700, 1999.
37. Y. Yoo and A. Ortega, "Adaptive quantization without side information using svq and tcq," in 29th Asilomar Conference on Signals, Systems, and Computers, November 1995.
38. Y. Yu and C.W. Chen, "SNR scalable transcoding for video over wireless channels," in Proceedings of the Wireless Communications and Networking Conference (WCNC), 2000, Vol. 3, pp. 1396–1402.

Marco Bertini has a research grant and carries out his research activity at the Department of Systems and Informatics at the University of Florence, Italy. He received a M.S. in electronic engineering from the University of Florence in 1999, and Ph.D. in 2004. His main research interest is content-based indexing and retrieval of videos. He is author of more than 25 papers in international conference proceedings and journals, and is a reviewer for international journals on multimedia and pattern recognition.

Rita Cucchiara (Laurea Ingegneria Elettronica, 1989; Ph.D. in Computer Engineering, University of Bologna, Italy 1993). She is currently Full Professor in Computer Engineering at the University of Modena and Reggio Emilia (Italy). She was formerly Assistant Professor ('93–'98) at the University of Ferrara, Italy and Associate Professor ('98–'04) at the University of Modena and Reggio Emilia, Italy. She is currently in the Faculty staff of Computer Engenering where has in charges the courses of Computer Architectures and Computer Vision.

Her current interests include pattern recognition, video analysis and computer vision for video surveillance, domotics, medical imaging, and computer architecture for managing image and multimedia data.

Rita Cucchiara is author and co-author of more than 100 papers in international journals, and conference proceedings. She currently serves as reviewer for many international journals in computer vision and computer architecture (e.g. IEEE Trans. on PAMI, IEEE Trans. on Circuit and Systems, Trans. on SMC, Trans. on Vehicular Technology, Trans. on Medical Imaging, Image and Vision Computing, Journal of System architecture, IEEE Concurrency). She participated at scientific committees of the outstanding international conferences in computer vision and multimedia (CVPR, ICME, ICPR, . . .) and symposia and organized special tracks in computer architecture for vision and image processing for traffic control. She is in the editorial board of Multimedia Tools and Applications journal. She is member of GIRPR (Italian chapter of Int. Assoc. of Pattern Recognition), AixIA (Ital. Assoc. Of Artificial Intelligence), ACM and IEEE Computer Society.

Alberto Del Bimbo is Full Professor of Computer Engineering at the Università di Firenze, Italy. Since 1998 he is the Director of the Master in Multimedia of the Università di Firenze. At the present time, he is Deputy Rector of the Università di Firenze, in charge of Research and Innovation Transfer. His scientific interests are Pattern Recognition, Image Databases, Multimedia and Human Computer Interaction. Prof. Del Bimbo is the author of over 170 publications in the most distinguished international journals and conference proceedings. He is the author of the "Visual Information Retrieval" monography on content-based retrieval from image and video databases edited by Morgan Kaufman. He is Member of IEEE (Institute of Electrical and Electronic Engineers) and Fellow of IAPR (International Association for Pattern Recognition). He is presently Associate Editor of Pattern Recognition, Journal of Visual Languages and Computing, Multimedia Tools and Applications Journal, Pattern Analysis and Applications, IEEE Transactions on Multimedia, and IEEE Transactions on Pattern Analysis and Machine Intelligence. He was the Guest Editor of several special issues on Image databases in highly respected journals.

Andrea Prati (Laurea in Computer Engineering, 1998; PhD in Computer Engineering, University of Modena and Reggio Emilia, 2002). He is currently an assistant professor at the University of Modena and Reggio Emilia (Italy), Faculty of Engineering, Dipartimento di Scienze e Metodi dell'Ingegneria, Reggio Emilia. During last year of his PhD studies, he has spent six months as visiting scholar at the Computer Vision and Robotics Research (CVRR) lab at University of California, San Diego (UCSD), USA, working on a research project for traffic monitoring and management through computer vision. His research interests are mainly on motion detection and analysis, shadow removal techniques, video transcoding and analysis, computer architecture for multimedia and high performance video servers, video-surveillance and domotics. He is author of more than 60 papers in international and national conference proceedings and leading journals and he serves as reviewer for many international journals in computer vision and computer architecture. He is a member of IEEE, ACM and IAPR.