IEEE Transactions on Circuits and Systems for Video Technology, Vol. 15, No 10, pp. 1200-1209, October 2005

Semantic video analysis for adaptive content delivery and automatic description

Andrea Cavallaro, Olivier Steiger, Touradj Ebrahimi

***Abstract*— We present an encoding framework which exploits semantics for video content delivery. The video content is organized based on the idea of main content message. In the work reported in this paper, the main content message is extracted from the video data through semantic video analysis, an application-dependent process that separates relevant information from non relevant information. We use here semantic analysis and the corresponding content annotation under a new perspective: the results of the analysis are exploited for object-based encoders, such as MPEG-4, as well as for frame-based encoders, such as MPEG-1. Moreover, the use of MPEG-7 content descriptors in conjunction with the video is used for improving content visualization for narrow channels and devices with limited capabilities. Finally, we analyze and evaluate the impact of semantic video analysis in video encoding and show that the use of semantic video analysis prior to encoding sensibly reduces the bandwidth requirements compared to traditional encoders not only for an object-based encoder but also for a frame-based encoder.**

***Index Terms*— Video analysis, video encoding, object segmentation, metadata, MPEG.**

I. Introduction

The diffusion of network appliances such as cellular phones, personal digital assistants, and hand-held computers creates a new challenge for content delivery: how to adapt the media transmission to various device capabilities, network characteristics, and user preferences [1], [2], [3]. Each device is characterized by certain display capabilities and processing power. Moreover, such appliances are connected through different kinds of networks with diverse bandwidths. Finally, users with different preferences access the same multimedia content. Therefore there exists a need to personalize the way media content is delivered to the end user. In addition to the above, recent devices, such as digital radio receivers, and new applications, such as intelligent visual surveillance, require novel forms of video analysis for content adaptation and summarization. Digital radios allow for the display of additional information alongside the traditional audio stream to enrich the audio content. For instance, digital audio broadcasting (DAB) allocates 128 Kb/s to streaming audio, whereas 8Kb/s can be used to send additional information, such as visual data [4]. Moreover, the growth of video surveillance systems poses challenging problems for the automatic analysis, interpretation and indexing of video data as well as for selective content filtering for privacy preservation. Finally, the instantaneous indexing of video content is also an desirable feature for sports broadcasting [5].

To cope with these challenges, video content needs to be automatically analyzed and adapted to the needs of the specific application, to the capabilities of the connected terminal and network, and to the preferences of the user. Three main strategies for adaptive content delivery have been proposed throughout the literature, namely Info Pyramid, scalable coding and transcoding. The work presented in this paper aims to go beyond traditional adaptation techniques. We focus on semantic encoding by looking to exploit video analysis prior to encoding (Figure 1). Specifically, we use semantic video analysis to extract relevant areas of a video. These areas are encoded at a higher level of quality or summarized in textual form. The idea behind this approach is to organize the content so that a particular network or device does not inhibit the main content message. The main content message is dependent on the specific application. In particular, for applications such as video surveillance and sport video the main content message is defined based on motion information.

The contribution of this paper is twofold. On the one hand, a framework for adaptive video delivery is defined and implemented based on video objects and on their associated metadata. On the other hand, two new modalities of video delivery are proposed in such a framework. The first modality combines semantic analysis with a traditional frame-based video encoder. The second modality uses metadata to efficiently encode the main content message. In particular, the use of metadata enables not only to make the content more searchable, but also to improve visualization and to preserve privacy in video-based applications.

The paper is organized as follows. Section II is an overview of existing adaptation techniques. Section III presents the algorithm for extracting the main content message, the framework for adaptive video delivery and automatic description using semantic video analysis. Section IV discusses quality assessment issues, whereas experimental results are presented in Section V. Finally, Section VI concludes the paper.

II. Background

Three main approaches have been presented in the literature to provide adaptive content delivery, namely Info Pyramid, scalable coding and transcoding. Info Pyramid provides a general framework for managing and manipulating media objects [6], [7]. Info Pyramid manages different versions, or *variations*, of media objects with different modalities (e.g., video, image, text, and audio) and fidelities (summarized, compressed, and scaled variations). Moreover, it defines methods

A. Cavallaro is with the Multimedia and Vision Laboratory, Queen Mary, University of London (QMUL), E1 4NS London, United Kingdom.

O. Steiger and T. Ebrahimi are with the Signal Processing Institute, Swiss Federal Institute of Technology (EPFL), CH-1015 Lausanne, Switzerland.

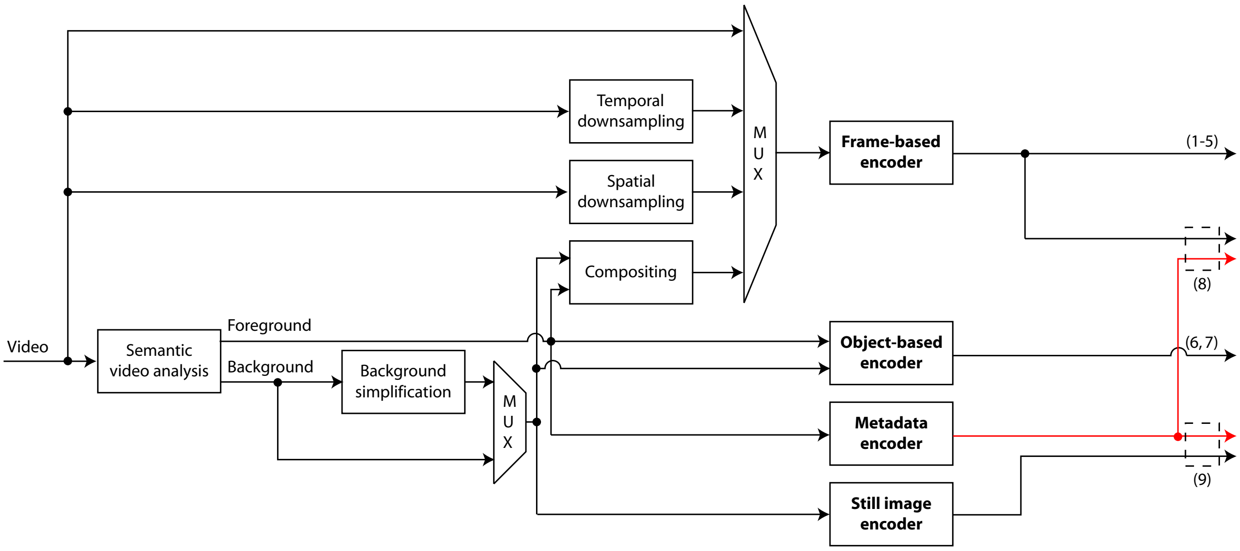


Fig. 1. Flow diagram of the proposed encoding framework based on semantic video analysis and description

for manipulating, translating, transcoding, and generating the content. When a client device requests a multimedia document, the server selects and delivers the most appropriate variation. The selection is made based on network characteristics and terminal capabilities, such as display size, frame rate, color depth and storage capacity.

As opposed to Info Pyramid, scalable coding processes multimedia content only once. Lower qualities, lower spatial and temporal resolutions of the same content are then obtained by truncating certain layers or bits from the original stream [8]. Basic modes of video scalability include quality scalability, spatial scalability, temporal scalability, and frequency scalability. Combinations of these basic modes are also possible. *Quality scalability* is defined as the representation of a video sequence with varying accuracies in the color patterns. This is typically obtained by quantizing the color values with increasingly finer quantization step sizes. *Spatial scalability* is the representation of the same video in varying spatial resolutions. *Temporal scalability* is the representation of the same video at varying temporal resolutions or frame rates. *Frequency scalability* includes different frequency components in each layer, with the base layer containing low-frequency components and the other layers containing increasingly high-frequency components. Such decomposition can be achieved via frequency transforms like the DCT or wavelet transforms. Finally, the basic scalability schemes can be combined to reach *fine-granularity scalability*, such as in MPEG–4 FGS [9]. The various scalable coding methods introduced previously perform the same operation over the entire video frame. In object-based temporal scalability (OTS), the frame rate of foreground objects is enhanced so that it has a smoother motion than the background.

Video transcoding is the process of converting a compressed video signal into another compressed signal with different properties. Early solutions to video transcoding determine the output format based on network and appliance constraints, independently of the semantics in the content (content-blind transcoding). Content-blind transcoding strategies include spatial resolution reduction, temporal resolution reduction, and bit-rate reduction [10]. Recent transcoding techniques make use of semantics to minimize the degradation of important image regions [11], [12]. In [13], optimal quantization parameters and frame skip are determined for each video object individually. The bit-rate budget for each object is allocated by a difficulty hint, a weight indicating the relative encoding complexity of each object. Frame skip is controlled by a shape hint, which measures the difference between two consecutive shapes to determine whether an object can be temporally downsampled without visible composition problems. Key objects are selected based on motion activity and on bit complexity.

The transcoding strategies described thus far are referred to as *intramedia transcoding* strategies and do not change the media nature of the input signal. On the other hand, *intermedia transcoding*, or *transmoding*, is the process of converting the media input into another media format. Examples of intermedia transcoding include speech-to-text [14] and video-to-text [15] translation. Both the intramedia and the intermedia adaptation concepts are used in this paper for video encoding, as described in the following section.

III. Adaptive content delivery and description using semantics

The proposed framework for adaptive video delivery and automatic description uses video content analysis and semantic pre-filtering prior to encoding (Figure 1) in order to improve the perceived content quality and to provide additional functionalities, such as privacy preservation and automatic video indexing. Semantic video analysis and semantic encoding are described next.

A. Semantic video analysis

Semantic video analysis is used to extract the main content message from the video. The semantics to be included in the analysis is dependent on the specific application. In the following, we discuss possible semantics and, in particular, we describe the use of motion as semantics.

Semantic video analysis refers to a human abstraction and uses *a priori* information to translate the semantics into rules. The rules are then applied through an algorithm. Examples of semantic video analysis based on *a priori* information are template matching, extraction of captions and text, face detection, and moving object segmentation. Template matching is used to implement the semantics when the shape objects we want to segment is known *a priori*. In this case, which includes in particular the detection of captions and text, the extraction method searches for specific object features in terms of geometry. For segmenting faces of people, color-based segmentation can be used [16]. The face detection task consists in finding the pixels whose spectral characteristics lie in a specific region in the chromaticity diagram. For extracting moving objects, *motion information* can be used as semantics. Several applications, such as sport broadcasting and video surveillance, deal with segmenting moving objects.

A typical tool used to tackle the problem of object segmentation based on motion is change detection. Different change detection techniques can be employed for moving camera and static camera conditions. If the camera moves, change detection aims at recognizing coherent and incoherent moving areas. The former correspond to background areas, the latter to video objects. If the camera is static, the goal of change detection is to recognize moving objects (foreground) and the static background. The semantic analysis we use addresses the static camera problem and is applicable in the case of a moving camera after global motion compensation. The change detector decides whether in each pixel position the foreground signal corresponding to an object is present. This decision is taken by thresholding the frame difference between the current frame and a frame representing the background. The frame representing the background is dynamically generated based on temporal information [17]. The thresholding aims at discarding the effect of the camera noise after frame differencing. A locally adaptive threshold, $\tau(i,j)$, is used that models the noise statistics and applies a significance test. To this end, we want to determine the probability that frame difference at a given position (i,j) is due to noise, and not to other causes. Let us suppose that there is no moving object in the frame difference. We refer to this hypothesis as the *null hypothesis*, H_0. Let $g(i,j)$ be the sum of the absolute values of the frame difference in an observation window of q pixels around (i,j). Moreover, let us assume that the camera noise is additive and follows a Gaussian distribution with variance σ. Given H_0, the conditional pdf of the frame difference follows a χ^2_q distribution with q degrees of freedom defined by

$$f\big(g(i,j)|H_0\big) = \frac{1}{2^{q/2}\sigma^q\Gamma(q/2)} g(i,j)^{(q-2)/2} e^{-g(i,j)^2/2\sigma^2}, \tag{1}$$

where $\Gamma(\cdot)$ is the Gamma function, that can be evaluated as $\Gamma(x+1) = x\Gamma(x)$, and $\Gamma(1/2) = \sqrt{\pi}$. To obtain a good trade-off between robustness to noise and accuracy in the detection we choose $q = 25$ (5×5 window centered in (i,j)). It is now possible to derive the significance test as

$$P\{g(i,j) \geq \tau(i,j)|H_0\} = \frac{\Gamma(q/2, g(i,j)^2/2\sigma^2)}{\Gamma(q/2)}. \tag{2}$$

When this probability is smaller than a certain significance level, α, we consider that H_0 is not satisfied at the pixel position (i,j). Therefore we label that pixel as belonging to a moving object. The significance level α is a stable parameter that does not need manual tuning along a sequence or for different sequences. Experimental results indicate that valid values fall in the range from 10^{-2} to 10^{-6}.

The change detection process produces the segmentation of the moving objects from the background (Figure 2) and, coupled with video object tracking [18], enables the subsequent extraction of object metadata, as described in the following section.

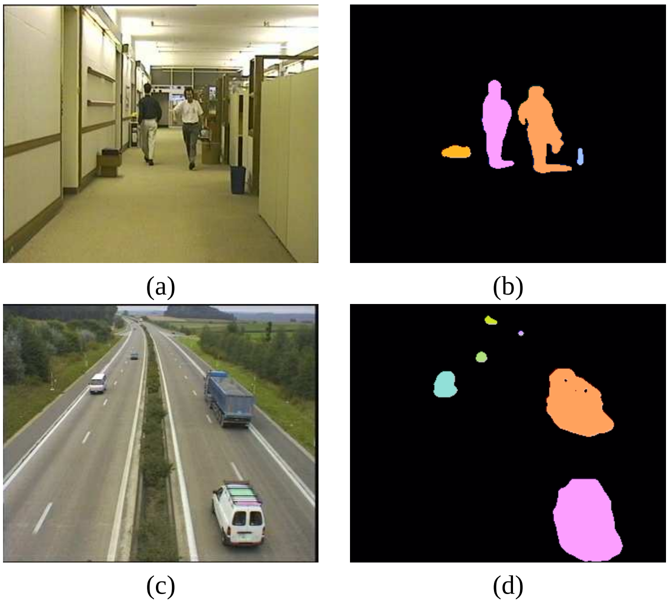


Fig. 2. Example of semantic video analysis results for the test sequences (a) *Hall Monitor* and (c) *Highway*: (b) separation of foreground and background for *Hall Monitor*. (d) separation of foreground and background for *Highway*. The background is color-coded in black

B. Semantic encoding

The decomposition computed with semantic video analysis is used with an object-based encoder as well as with a traditional frame-based encoder. We will refer to the former case as *object-based encoding* and to the latter as *frame-based encoding*. Furthermore, metadata are used to efficiently encode relevant information and to enhance relevant part of a low-quality coded video. These approaches are referred to as *metadata-based encoding* and *metadata-enhanced encoding*, respectively. Relevant examples of the modalities presented in this section are illustrated in Figure 3. The analysis and evaluation of the different approaches in terms of results and bandwidth requirements are presented in Section V.

1) Object-based encoding: With object-based encoding, the encoder needs to support the coding of individual video objects (e.g., MPEG–4 object-based). Each video object is assigned to

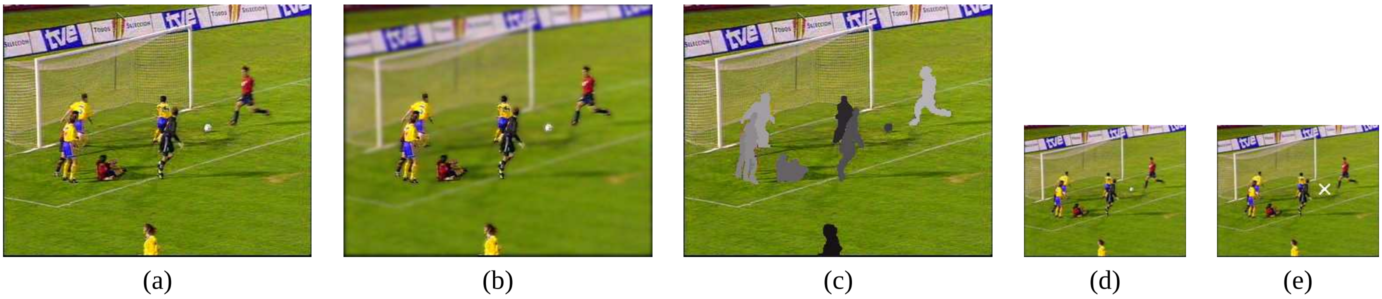


Fig. 3. Examples of encoding modalities. (a) Sample frame from the sequence *Soccer*; (b) Semantic frame-based encoding: the background is selectively lowpass-filtered prior to encoding; (c) Metadata-based encoding: object shapes are superimposed on the background; (d) Spatial resolution reduction; (e) Metadata-enhanced encoding: metadata are used to enhance relevant portions of a video

a distinct object class, according to its importance in the scene. The encoding quality can be set depending on the object class: the higher the relevance, the higher the encoding quality. One advantage of this approach is the possibility of controlling the sequencing of objects: video objects may be encoded with different degree of compression, thus allowing better granularity for the areas in the video that are of more interest to the viewer. Moreover, objects may be decoded in their order of priority, and the relevant content can be viewed without having to reconstruct the entire image (network limitations). Another advantage is the possibility of using a simplified background (appliance limitation), so as to enhance the relevant objects. Using a simplified background aims at taking advantage of the task-oriented behavior of the human visual system for improving compression ratios. Recent work on foveation [19] demonstrated that using nonlinear integration of low-level visual cues mimicking the processing in primate occipital and posterior parietal cortex allows one to sensibly increase compression ratios. Moreover, the work reported in [20] demonstrated that an overall increase in image quality can be obtained when the increase in quality of the relevant areas of an image more than compensates for the decrease in quality of the image background.

2) Semantic frame-based encoding: The semantic frame-based encoding mode exploits semantics in a traditional frame-based encoding framework (e.g., MPEG–1). The use of the decomposition of the scene into meaningful objects prior to encoding, referred here as *semantic pre-filtering*, helps support low bandwidth transmission. The areas belonging to the foreground class, or semantic objects, are used as region of interest. The areas not included in the region of interest may either be eliminated, that is set to a constant value, or lowered in importance by using a low-pass filter. The latter solution simplifies the information in the background, while still retaining essential contextual information. An example of this solution is reported in Fig. 4 (a). On the other hand, filtering the entire image inhibits the main content message Fig. 4 (b). Another way to take into account less relevant portions of an image before coding is to take advantage of the specifics of the coding algorithm. In the case of block-based coding, each background macroblock can be replaced by its DC value. Semantic frame-based encoding mimics the way humans perceive visual information and allows for a reduction of information to be coded.

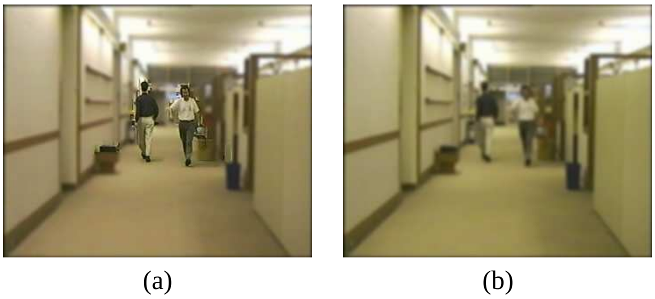


Fig. 4. (a) Selective lowpass-filtering simplifies the information in the background, while still retaining essential contextual information; (b) filtering the entire image inhibits the main content message

3) Metadata-based and metadata-enhanced encoding: A further processing of the video content is performed to cope with limited device or network capabilities as well as to automatically generate metadata. Such processing transforms the foreground objects extracted through semantic analysis into quantitative descriptors and enables video annotation. Video annotation is desirable for applications such as video surveillance, where terabytes of data are produced and need to be searched quickly. Moreover, the descriptors can be transmitted instead of the video content itself and superimposed by the terminal on a still background. For example, an object identifier and a shape descriptor are used in [21]. The object identifier is a unique numerical identifier describing the spatial location of each object in the scene. The shape descriptor is used to represent the shape of an object, ranging from a bounding box to a polygonal representation with a different number of vertices (Figure 3(c)). This approach is useful to preserve privacy in video surveillance applications as well as to reduce bandwidth requirements under critical network conditions. A progressive representation is used where the number of vertices corresponding to the best resolution is computed, and any number of vertices smaller that this maximum can be used according to the requirements of the application. In addition to the above, other features such as color and texture descriptors may be added in the description process. The choice of these additional features depends on the application at hand.

In addition to the above, the descriptors can be transmitted along with the video itself and used for rendering the video content. This solution, consisting in a mixture of video-based and text-based modalities, is here referred toe as *metadata-enhanced encoding*. Using metadata-enhanced encoding, content descriptors help enhance parts of the video that are hidden or difficult to perceive due to heavy compression. In this case, the video itself is the background and the descriptors highlight relevant portions of the data. One example is the ball in a football match for transmission to a PDA or a mobile phone, as shown in Figure 3(e).

IV. Quality assessment

Perceptual video quality assessment is a difficult task already when dealing with traditional coders [22]. When dealing with object-based coders, the task becomes even more challenging. For this reason, we use a combination of subjective and objective evaluation techniques to compare the performance of the different encoding modalities. Subjective evaluation includes the visual comparison of frames and frame details. This analysis is performed at different bitrates and at different frame resolutions. Objective evaluation includes temporal signal-to-noise ratio analysis and the analysis of rate-distortion curves.

A. Semantic peak signal-to-noise ratio

Traditional peak signal-to-noise ratio (PSNR) analysis uniformly weights the contribution of each pixel in an image when computing the mean squared error (MSE). This analysis gives the same importance to relevant as well as less relevant areas of an image. To account for the way humans perceive visual information, different areas of an image, or object classes, should be considered [11]. We take into account object classes through a distortion measure, the *semantic mean squared error*, SMSE, defined as:

$$\text{SMSE} = \sum_{k=1}^{N} w_k \cdot \text{MSE}_k, \tag{3}$$

where N is the number of object classes and w_k the weight of class k. Class weights are chosen depending on the semantics, with $w_k \geq 0, \forall k = 1, \dots, N$ and $\sum_{i=1}^{N} w_k = 1$. The mean squared error of each class, MSE_k, can be written as

$$\text{MSE}_k = \frac{1}{|C_k|} \sum_{(i,j)\in C_k} d^2(i,j), \tag{4}$$

where C_k is the set of pixels belonging to the object class k and $|C_k|$ is its cardinality. The class membership of each pixel (i,j) is defined by semantic video analysis. The error $d(i,j)$ between the original image I_O and the distorted image I_D in Eq.(4) is the pixel-wise color distance. The color distance is computed in the 1976 CIE *Lab* color space in order to consider perceptually uniform color distances with the Euclidean norm and is expressed as:

$$d(i,j) = \sqrt{\left(\Delta I^L(i,j)\right)^2 + \left(\Delta I^a(i,j)\right)^2 + \left(\Delta I^b(i,j)\right)^2}, \tag{5}$$

with $\Delta I^L(i,j) = I_O^L(i,j) - I_D^L(i,j)$, $\Delta I^a(i,j) = I_O^a(i,j) - I_D^a(i,j)$, and $\Delta I^b(i,j) = I_O^b(i,j) - I_D^b(i,j)$. The final quality evaluation metric, the *semantic peak signal-to-noise ratio*, SPSNR, is the following:

$$\text{SPSNR} = 10 \log_{10}\left(\frac{V_M^2}{\text{SMSE}}\right), \tag{6}$$

where V_M is the maximum peak-to-peak value of the color range. When the object classes are foreground and background, then $N = 2$ in Eq (3). If we denote with w_f the foreground weight, then SPSNR $\equiv$ PSNR when $w_f = 0.5$. The larger w_f, the more important the contribution of the foreground. When $w_f = 1$, then the foreground only is considered in the evaluation of the peak signal-to-noise ratio.

An illustration of the impact of w_f in the distortion measure is shown in in Fig. 5. The figure presents a comparison of the average SPSNR of the sequence *Hall Monitor* for the different encoding modalities described in Section III-B as function of w_f. The value of w_f to be used is estimated as described in the next section.

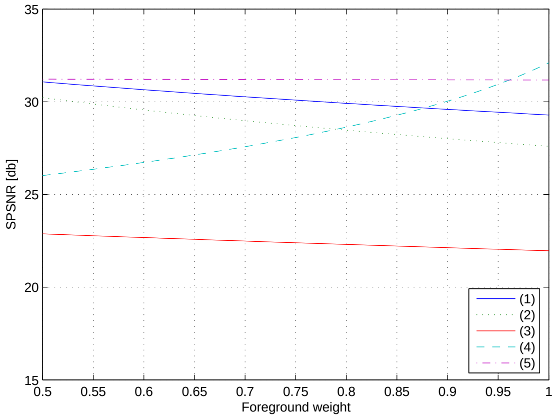


Fig. 5. Illustration of the impact of w_f in the distortion measure: average SPSNR vs. foreground weight for *Hall monitor* sequence. The five labels correspond to the following sequence types: (1) coded original; (2) temporally down-sampled; (3) spatially down-sampled; (4) static background; (5) simplified background. Content-blind coding methods (1)-(3) decrease their performance when the foreground is given more importance. Methods based on semantic, (4) and (5), increase their performance when the foreground is given more importance

B. Determination of the foreground weight

Subjective performance evaluation experiments have been performed to estimate the foreground weight leading to the closest match between SPSNR prediction and human judgment. Twenty non-expert observers of different ages and backgrounds have been presented a series of video sequences according to ITU-T Recommendation P.910, *Absolute Category Rating* [23]. The evaluation has been carried out using the MPEG–4 test sequences *Akiyo*, *Hall Monitor*, *Children*, and *Coastguard*. Video sequences have been generated using the encoding strategies described in Section III-B, at different bitrates, and rated by the observers on a scale ranging from 0 (bad) to 100 (excellent). This range of values was presented to the observers in a training phase.

The foreground weight, w_f, is determined for each test sequence by maximizing the Pearson correlation [24] between

SPSNR and subjective results. The results are summarized in Table I. For the sequence *Akiyo*, where the foreground covers a large area of each frame and the background is simple, the the observers focused mostly on foreground, thus leading to a value of $w_{\mathrm{f}} = 0.97$. For *Hall Monitor*, whose background is more complex and objects are smaller, the foreground attracted slightly more the attention than the background ($w_{\mathrm{f}} = 0.55$). The sequence *Children* has a very complex and colored background that attracted the observers' attention, thus resulting in foreground and background being equally weighted ($w_{\mathrm{f}} = 0.5$). The sequence *Coastguard* contains camera motion that prevented the observer from focusing on background steadily, even though the background is quite complex. In this case, the resulting foreground weight is $w_{\mathrm{f}} = 0.7$. In general, results confirm that large moving objects and complex background tend to attract user's attention. Based on the data collected with subjective experiments, it is possible to predict the foreground weight based on the following formula:

$$w_f = \alpha \cdot r + (\delta - \beta \cdot r)\sigma_b + \gamma \cdot v + \delta, \qquad (7)$$

where r represents the portion of the image occupied by foreground pixels, expressed as $r = |C_f|/(|C_f| + |C_b|)$, with $|C_f|$ and $|C_b|$ representing the number of foreground and background pixels, respectively. The background complexity is taken into account with σ_b, the standard deviation of the luminance of background pixels. The presence of camera motion is considered with the term v: $v = 1$ for moving camera, and $v = 0$ otherwise. α, β, γ, and δ are constants whose values are determined based on the results of the subjective experiments and are the following: $\alpha = 5.7$, $\beta = 0.108$, $\gamma = 0.2$ and $\delta = 0.01$. The final value of w_f is the average of the foreground weights over the sequence.

In addition to semantic weight, Table I provides information about *accuracy*, *monotonicity* and *consistency* of the SPSNR metric. Accuracy is given by Pearson linear correlation coefficient r_p, monotonicity by Spearman rank-order correlation coefficient r_s, and consistency by outliers ratio r_o [24]. Pearson correlation of PSNR, $r_p(0.5)$, is given for comparison. Pearson correlation r_p and Spearman correlation r_s are close to 1 for all sequences. Thus, accuracy and monotonicity of SPSNR are high. Outliers ratio r_o is around 10%, thus consistency of the metric is good as well. Note that using semantics improves accuracy by up to 8% (*Akiyo*), as compared to PSNR.

TABLE I

Foreground weight and SPSNR accuracy

	Akiyo	*Hall monitor*	*Children*	*Coastguard*
$\mathbf{w_f}$	0.97	0.55	0.50	0.7
$\mathbf{r_p(w_f)}$	0.95	0.90	0.95	0.92
$\mathbf{r_p(0.5)}$	0.87	0.89	0.95	0.90
$\mathbf{r_s(w_f)}$	0.90	0.84	0.95	0.93
$\mathbf{r_o(w_f)}$	0.10	0.11	0.07	0.07

V. Experimental results

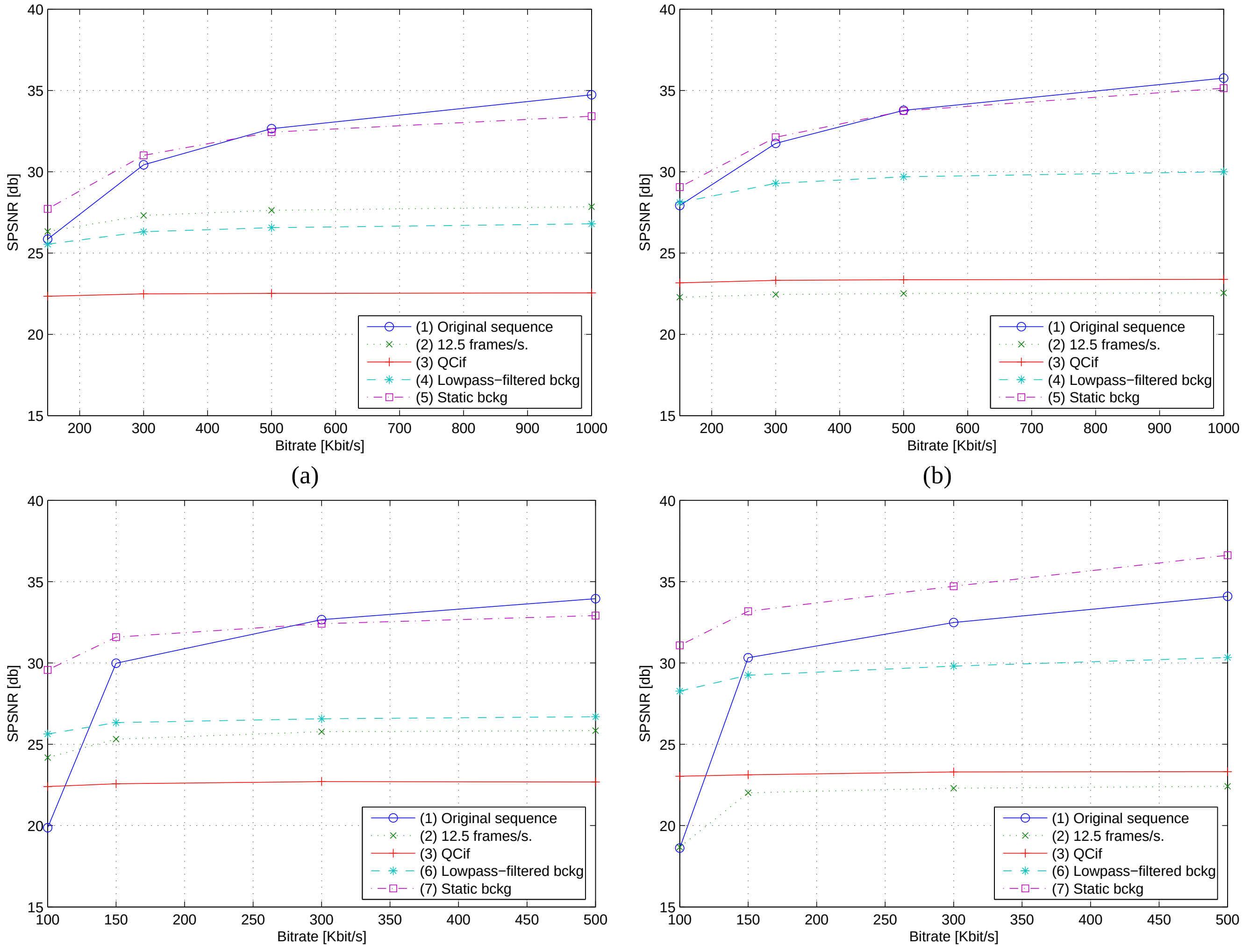
In this section, experimental results of the proposed semantic video encoding and annotation framework with standard test sequences are presented. The results illustrate the impact of semantic analysis on the encoding performance of frame-based as well as object-based coders and demonstrate the use of the proposed approach for advanced applications, such as privacy preservation in video surveillance. Sample results are shown from the MPEG–4 test sequence *Hall Monitor* and from the MPEG–7 test sequence *Highway*. Both sequences are in CIF format at 25 Hz. The modalities under analysis are: (1) coded original sequence; (2) temporal resolution reduction (from 25 frames/s. to 12.5 frames/s.); (3) spatial resolution reduction (from CIF to QCIF); (4,6) video objects composited with static background; (5,7) video objects composited with simplified background. The background is simplified using a Gaussian $9x9$ low-pass filter with $\mu = 0$ and $\sigma = 2$.

The following coders have been used in the encoding process: (i) TMPGEnc 2.521.58.169 using constant bitrate (CBR) rate control for frame-based MPEG–1; (ii) MoMuSys MPEG-4 VM reference software version 1.0 using VM5+ global rate control for object-based MPEG–4; (iii) Expway MPEG-7 BiM Payload encoder/decoder version 02/11/07 for MPEG–7 metadata; (iv) Kakadu JPEG2000 codec version 4.2 for JPEG200 still images. The value of the foreground weight used in the objective evaluation is $w_f = 0.55$ for *Hall Monitor*, as determined with the subjective experiments, and $w_f = 0.53$ for *Highway*, computed using Eq. (7) with $r = 0.07, \sigma_b = 48, v = 0$.

Figure 6 shows the rate-distortion diagrams for the test sequences. The average SPSNR for five encoding modalities is plotted against the encoding bitrate. Figures 6 (a) and (b) show the rate-distortion diagrams for MPEG–1 at bitrates between 150 Kbit/s and 1000 Kbit/s. At low bitrates (150-300 Kbit/s), semantic encoding with static background (4) leads to a larger SPSNR than any of the content-blind methods (1-3). This is because inter-coded static background blocks do not produce residue and most of the available bitrate can be allocated to foreground objects. In Figures 6 (c) and (d), foreground and background are encoded in two separate streams using object-based MPEG–4 at bitrates between 100 Kbit/s and 500 Kbit/s. Here semantic analysis is used in all five modalities. It possible to notice that quality is improved at low bitrates by low-pass filtering the background or using a still frame representing the background.

Figure 7 shows a sample frame from each test sequence coded with MPEG–1 at 150 Kbit/s with and without semantic pre-filtering. Figure 8 shows magnified excepts of both test sequences coded with MPEG–1 at 150 Kbit/s. Figure 8 (top) shows the person that carries a monitor in *Hall monitor*. The amount of coding artifacts is notably reduced by semantic pre-filtering ((d) and (e)). In particular, the person's mouth and the monitor are visible in (e), whereas they are corrupted by coding artifacts in the non-semantic modalities. Similar observations can be made for Figure 8 (bottom), which shows a blue truck entering the scene at the beginning of the *Highway* sequence. Coding artifacts are less disturbing on the object in (d) and (e) than in (a)-(c). Moreover, the front-left wheel of the truck is only visible with semantic pre-filtering ((d) and (e)).

Next, we evaluate the cost of sending metadata for metadata-based and metadata-enhanced encoding. Table II shows the



(a) (b)

(c) (d)

Fig. 6. Rate-distortion diagrams. (a) *Hall monitor*, MPEG–1; (b) *Highway*, MPEG–1; (c) *Hall monitor*, MPEG–4 object-based; (d) *Highway*, MPEG–4 object-based

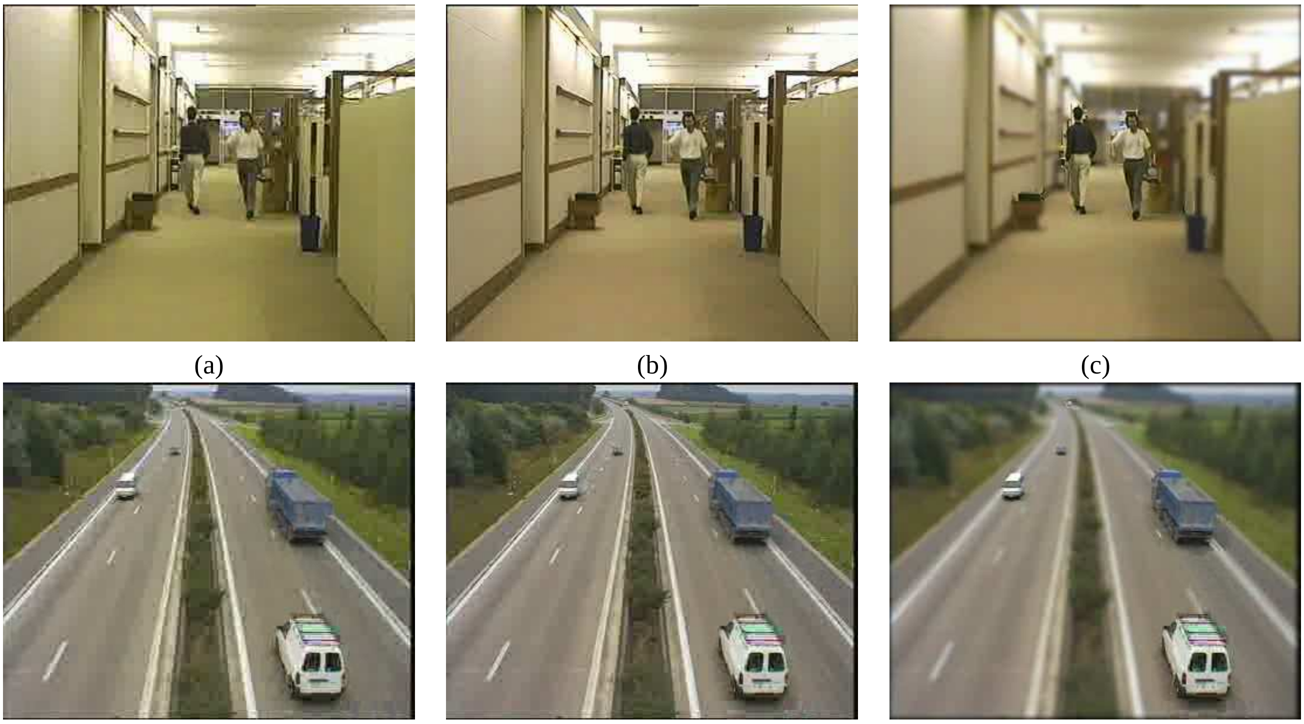
bitrate required by three types of description for *Hall Monitor* and *Highway* using MPEG–7 binary format (BiM). MPEG–7 binary format is used for sending summary information to terminals with limited capabilities and to enhance heavily compressed videos. The descriptions are represented by the spatial locators of the foreground objects, their bounding boxes, and an approximation of their shape with 20-sided polygons, respectively. The metadata size increases with the description complexity and with the number of objects in the scene (*Hall Monitor* vs. *Highway*). The cost for metadata-enhanced encoding can be further reduced by sending the description of critical objects only. In addition to the above, metadata-enhanced encoding is used for privacy preservation in video surveillance. Figure 9 shows an example of different level of information hiding obtained using object descriptors for the sequence *Hall Monitor*. A surveillance operator can be shown different video types, ranging from the full appearance of the objects (Figure 9 (a)) to the visualization of a position locator that allows the operator to derive statistics about number of objects, their behavior and position without disclosing their identity (Figure 9 (d)). Intermediate levels of visualization include the approximation of object shapes that hides the identity of the subjects captured by the surveillance camera, while allowing to derive information about their size and form (Figure 9 (b)), and the bounding box (Figure 9 (c)). The encoding cost associated with this additional functionality added to a surveillance system is 21 Kbit/s for the spatial locator, 59 Kbit/s for the bounding box and 89 Kbit/s for the polygonal shape. The choice of the description to be used depends on the trade-off between privacy and the monitoring task at hand.

TABLE II
AVERAGE BITRATE OF MPEG–7 BIM SEQUENCE DESCRIPTION

DESCRIPTION	Spatial locator	Bounding box	Polygon shape
Hall monitor	21 Kbit/s	59 Kbit/s	89 Kbit/s
Highway	26 Kbit/s	66 Kbit/s	98 Kbit/s

VI. CONCLUSIONS

We presented a content-based video encoding framework which is based on semantic analysis. Semantic analysis enables



(a) (b) (c)

Fig. 7. Frame 190 of *Hall monitor* (top) and frame 44 of *Highway* (bottom) coded with MPEG–1 at 150 Kbit/s using different modalities: (a) coded original sequence; (b) static background; (c) simplified background

the decomposition of a video into meaningful objects. Using this decomposition, the encoder may adapt its behavior to code relevant and non relevant objects differently. Three modalities of video delivery have been discussed, analyzed, and compared using standard encoders. The first exploits semantics in traditional frame-based encoding. Semantically pre-filtering the video prior to coding leads to significant improvements in video compression efficiency in terms of bandwidth requirements as well as visual quality at low bitrates. The second modality uses metadata to efficiently encode relevant information. Object descriptors are generated for content retrieval as well as used for coding at very low bit-rates or for devices with limited capabilities. The third modality combines video and metadata for visualization. Metadata are used for content enhancement at low bitrates and for preserving privacy in video surveillance applications.

In the specific implementation discussed in Section V, the semantics is defined by motion. Given the modularity of the proposed encoding framework other semantics can also be used in the analysis step. Examples are face detection and text segmentation.

The quality metric used in this work is a promising first step towards measuring the quality taking semantics into account. Future work includes the study and definition of a perceptual metric that accounts for user satisfaction, depending on the application and the user preferences. To this end, an object of interest metric, such as that used in [3], will be an important building block of the overall quality metric. This quality metric will be used to automatically select the best encoding technique that maximizes user experience.

References

[1] P. van Beek, J. Smith, T. Ebrahimi, T. Suzuki, and J. Askelof, “Metadata-driven multimedia access,” *IEEE Signal Processing Magazine*, pp. 40–52, March 2003.

[2] R. Mohan, J. Smith, and C.-S. Li, “Adapting multimedia internet content for universal access,” *IEEE Transactions on Multimedia*, vol. 1, no. 1, pp. 104–114, 1999.

[3] 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.

[4] I. P. Duncumb, P. F. Gadd, G. Wu, and J. L. Alty, “Visual radio: Should we paint pictures with words, or pictures?,” Tech. Rep., Loughborough University, UK, 2004.

[5] Gopal S. Pingali, Agata Opalach, Yves D. Jean, and Ingrid B. Carlbom, “Instantly indexed multimedia databases of real world events,” *IEEE Transactions on Multimedia*, vol. 4, no. 2, pp. 269–282, June 2002.

[6] Chung-Sheng Li, Rakesh Mohan, and John R. Smith, “Multimedia content description in the Info Pyramid,” in *Proc. of the IEEE International Conference on Accoustics, Speech and Signal Processing*, May 1998, pp. 171–178.

[7] John R. Smith, Rakesh Mohan, and Chung-Sheng Lj, “Scalable multimedia delivery for pervasive computing,” in *Proc. of the ACM conference on Multimedia*, Oct.-Nov. 1999, vol. 1, pp. 131–140.

[8] Yao Wang, Jörn Ostermann, and Ya-Qin Zhang, *Video Processing and Communications*, Prentice Hall, 1 edition, 2001.

[9] W. Li, “Overview of fine granularity scalability in MPEG–4 video standard,” *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 11, no. 3, pp. 301–317, March 2001.

[10] A. Vetro, C. Christopoulos, and H. Sun, “Video transcoding architectures and techniques: an overview,” *IEEE Signal Processing Magazine*, vol. 20, no. 2, pp. 18–29, March 2003.

[11] R. Cucchiara, C. Grana, and A. Prati, “Semantic transcoding for live video server,” in *Proceedings of ACM Multimedia*, Juan–Les–Pins (France), December 2002, pp. 223–226.

[12] Andrea Cavallaro, Olivier Steiger, and Touradj Ebrahimi, “Semantic segmentation and description for video transcoding,” in *Proceedings of the IEEE International Conference on Multimedia and Expo*, July 2003, vol. 3, pp. 597–600.

[13] A. Vetro, H. Sun, and Y. Wang, “Object-based transcoding for adaptable

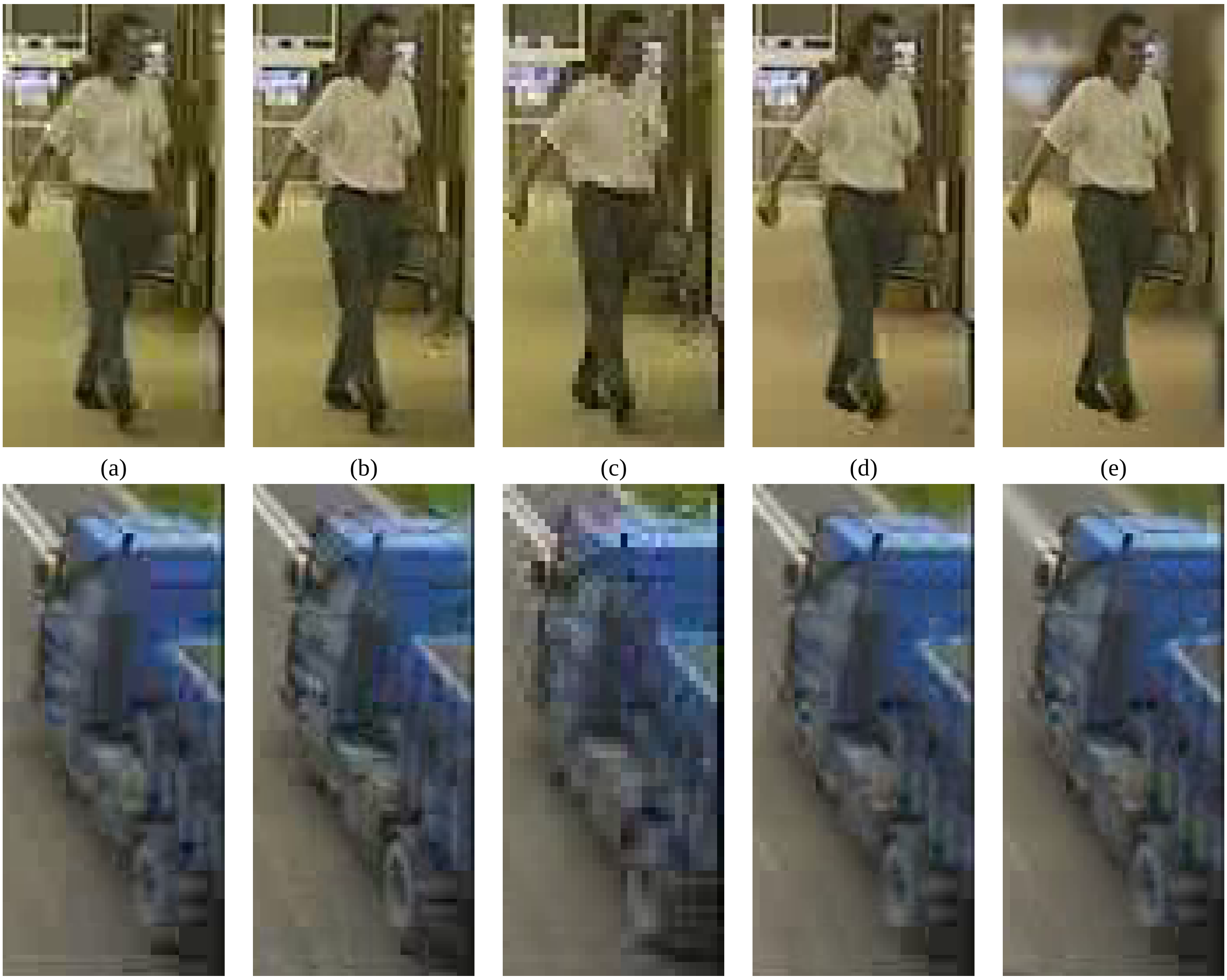


Fig. 8. Details of frame 280 of *Hall monitor* (top) and frame 16 of *Highway* (bottom). The sequences are encoded with MPEG–1 at 150 Kbit/s using different encoding modalities: (a) coded original sequence; (b) temporal resolution reduction; (c) spatial resolution reduction; (d) static background; (e) simplified background

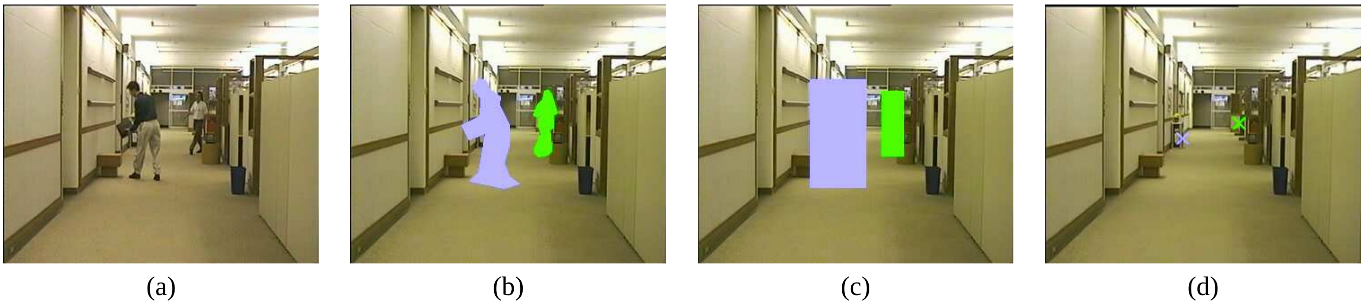


Fig. 9. Example of use of the proposed encoding framework for privacy preservation in an indoor surveillance application. Four different methods are shown representing different privacy levels. a) Video objects; (b) Object shape; (c) Bounding box; (d) Object position. The method can also be used to adapt the video delivery to the channel capacity and the terminal characteristics.

video content delivery," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 11, no. 3, pp. 387–401, March 2001.

[14] N. Morgan and H. Bourlard, "Continuous speech recognition," *IEEE Signal Processing Magazine*, vol. 12, no. 3, pp. 24–42, May 1995.

[15] K. Jung, K.I. Kim, and A.K. Jain, "Text information extraction in images and video: a survey," *Pattern Recognition*, vol. 37, no. 5, pp. 977–997, May 2004.

[16] R.L. Hsu, M.Abdel-Mottaleb, and A. Jain, "Face detection on color images," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 696–706, 2002.

[17] A. Cavallaro and T. Ebrahimi, "Video object extraction based on adaptive background and statistical change detection," in *Proceedings of SPIE*

Electronic Imaging - Visual Communications and Image Processing, San Jose, California, USA, 2001, pp. 465–475.

[18] A. Cavallaro, O. Steiger, and T. Ebrahimi, “Multiple object tracking in complex scenes,” in *Proceedings of ACM Multimedia*, Juan–Les–Pins (France), December 2002, pp. 523–532.

[19] L. Itti, “Automatic foveation for video compression using a neurobiological model of visual attention,” *IEEE Trans. on Image Processing*, vol. 13, no. 10, pp. 1304–1318, 2004.

[20] A.P. Bradley and F.W.M. Stentiford, “Visual attention for region of interest coding in jpeg 2000,” *Journal of Visual Communication and Image Representation*, vol. 14, pp. 232–250, 2003.

[21] O. Steiger, A. Cavallaro, and T. Ebrahimi, “MPEG-7 description of generic video objects for scene reconstruction,” in *Proceedings of SPIE Electronic Imaging*, San Jose, California, USA, January 2002, pp. 223–226.

[22] S. Olsson, M. Stroppiana, and J. Baina, “Objective methods for assessment of video quality : state of the art,” *IEEE Transactions on Broadcasting*, vol. 43, no. 4, pp. 487–495, 1997.

[23] ITU, “Subjective video quality assessment methods for multimedia applications,” Tech. Rep. P.910, ITU-T Recommandation, September 1999.

[24] David Freedman, Robert Pisani, and Roger Purves, *Statistics*, W.W. Norton & Company, 3 edition, 1997.

[25] Yap-Peng Tan, Yongqing Liang, and Haiwei Sun, “On the methods and performances of rational downsizing video transcoding,” *Signal Processing: Image Communications*, vol. 19, pp. 47–65, 2004.

[26] Anthony Vetro, Toshihiko Hata, Naoki Kuwahara, Hari Kalva, and Shun-Ichi Sekiguchi, “Complexity-quality analysis of transcoding architectures for reduced spatial resolution,” in *IEEE Transactions on Consumer Electronics*, August 2002, pp. 515–521.

[27] P. A. A. Assunção and M. Ghanbari, “A frequency-domain video transcoder for dynamic bit-rate reduction of MPEG–2 bit streams,” *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 8, no. 8, pp. 953–967, Dec. 1998.

[28] T. Shanableh and M. Ghanbari, “Heterogeneous video transcoding to lower spatio-temporal resolutions and different encoding formats,” *IEEE Trans. on Multimedia*, vol. 2, no. 2, pp. 101–110, June 2000.

[29] Y. Liang and Y.-P. Tan, “A new content-based hybrid video transcoding method,” in *Proc. of IEEE Int. Conf. on Image Processing 2001*, Oct. 2001, vol. 1, pp. 429–432.

[30] A. Cavallaro, E. Salvador, and T. Ebrahimi, “Shadow detection in image sequences,” in *Proc. of IEE Conference on Visual Media Production*, London, UK, 2004, pp. 165–174.

[31] A. Vetro and H. Sun, “Media conversions to support mobile users,” in *Proc. of IEEE Canadian Conf. on Electrical and Computer Engineering, CCECE 2001*, May 2001, vol. 1, pp. 607–612.

[32] R. Mohan, J.R. Smith, and C.-S. Li, “Adapting multimedia internet content for universal access,” *IEEE Trans. on Multimedia*, vol. 1, no. 1, pp. 104–114, March 1999.

[33] H. Sun, A. Vetro, and K. Asai, “Resource adaptation based on MPEG–21 usage environment description,” in *Proc. IEEE Int. Symposium on Circuits and Systems*, May 2003, vol. 2, pp. 536–539.

[34] ISO/IEC, “Information technology – generic coding of moving pictures and associated audio information: Video, 2nd ed.,” Tech. Rep. ISO/IEC FDIS 13818-2:2000, ISO/IEC JTC 1/SC29/WG11, 2000.

[35] ISO/IEC, “Information technology – coding of audio-visual objects – part 2 visual–amendment 2: Streaming video profiles,” Tech. Rep. ISO/IEC FDIS 14496-2:2001, ISO/IEC JTC 1/SC29/WG11, 2001.

[36] Julien Bourgeois, Emmanuel Mory, and François Spies, “Video transmission adaptation on mobile devices,” *Journal of System Architecture*, vol. 49, pp. 475–484, 2003.

[37] Surya Nepal and Uma Srinivasan, “Dave: A system for quality driven adaptive video delivery,” *Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval*, pp. 223–230, 2003.

[38] Niklas Björk and Charilaos Christopoulos, “Video transcoding for universal multimedia access,” in *ACM Multimedia Workshop*, 2000, pp. 75–79.