18

User-aware Video Coding Based on Semantic Video Understanding and Enhancing

Yu-Tzu Lin[1] and Chia-Hu Chang[2]
[1]*National Chi Nan University*
[2]*National Taiwan University*
Taiwan

1. Introduction

Traditional video coding is devoted to represent the video data compactly by dealing with low-level features (e.g., color, motion, texture, etc.) of the video. However, with the insatiable demand of Internet and increased use of multimedia, the bit-rate control issues are more and more important. In order to achieve more efficient representation for coping with diverse network or devices, many researches devised scalable video coding schemes which adaptively change the bit-rate according to the available bandwidth or user requirements. However, most scalable video coding algorithms only consider low-level features of video content in frame-based format without utilizing semantic information, which lose the possibility of improving coding efficiency by employing semantic meaning of content. Therefore, it is valuable to investigate methodologies of semantic-level video coding to produce more compact and flexible coding results for various user preferences. Semantic analysis for video content will provide richer information about the content and then assist achieving higher compression rate with good visual quality. Besides the coding efficiency, various functions are required in current video services, such as manipulating, searching and interacting with semantic-level objects. To enhance the flexibility and interactivity for accessing and manipulating the video content adaptively for different users, user-aware functionalities based on semantic video analysis should be discussed. In this chapter, we will discuss the theory and practice of user-aware semantic video coding, focusing on the aspect of semantic manipulation and user adaptation of video, including the semantic analysis techniques, scalable coding, user attention model construction, and user-aware video coding, by considering requirements for different applications and giving an explanation about the methodologies for some example applications.

2. Semantic video coding

For instance, the background except the couple in the wedding video can be compressed with a higher rate than the area of the bride and groom because of its lower semantic importance (less interesting to humans). Many researches (Cheng et al., 2008; Bertini et al., 2006; Ng et al., 2010) investigate methodologies for analyzing the semantic meaning of the video. Within the MPEG-7 (ISO/IEC 15938) Multimedia Description Schemes specification, "event" is used in the Creation and Production description tools to describe the agents and

tools involved in creation process. The semantic event is also a fundamental concept in the Semantics Description tools where it is used to describe what is happening or being depicted in the actual content of the video object, which also plays a major role in MPEG's latest initiative, MPEG-21 (ISO/IEC 21000). Since the MPEG-21 standard (Vetro, 2004) highlights the importance of semantic video coding, more and more semantic video analysis approaches were devised for various applications. Fig. 1 shows the architecture of semantic video coding. The result of semantic analysis provides information for enhancing the spatial and temporal processing and then improves the coding efficiency.

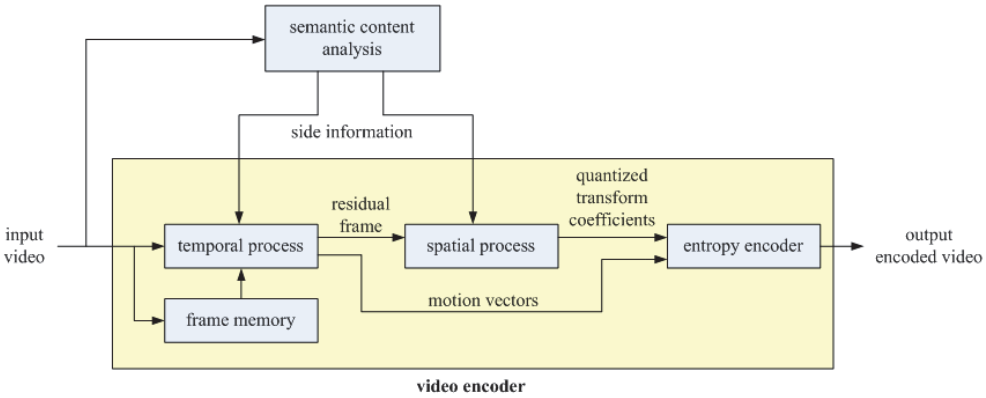
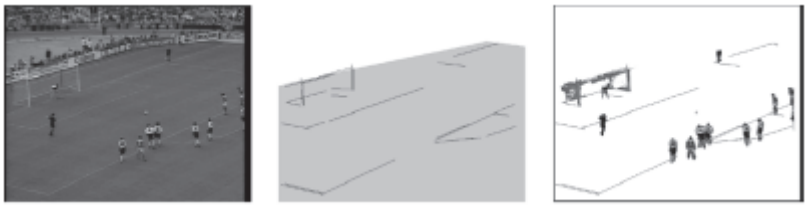


Fig. 1. The architecture of semantic video coding.

2.1 Object extraction and encoding

To mine semantic meaning from the video, low-level features have to be extracted at first, such as colour, texture, edge, shape, or motion features, to segment and describe objects with various descriptions, such as histogram, slope, graphs, and coefficients transformed to frequency-domain. By using the obtained low-level features of the objects, semantic rules can be applied to understand the video content by detecting meaningful events in the video. Object extraction is an essential procedure in semantic video analysis. The extracted objects are important basis for content event detection, and background except objects is the minor part of the video, which can be compressed with a higher rate while video coding. Fig. 2 (Bertini et al., 2006) shows one example of object extraction in the sports video.



(a) (b) (c)

Fig. 2. (a) Original frame; (b) playfield shape and playfield lines; (c) soccer players' blobs and playfield lines (Bertini et al., 2006).

However, in many applications, the objects in the video can not be easily found without segmenting the frame image at first. And the stability of the object-segment extraction is important for correctly detecting events in the video content. Before object extraction, the video frame has to be firstly segmented into several segments for locating candidates of the object-segment. Unfortunately, the pixel-based segmentation results of gray-level images are usually sensitive to the changes of the image pixels. Some researches (Lin et al., 2006) proposed reliable segmentation techniques called Geometric-Invariant Segmentation which is invariant to pixel changes. Even though the object moves, different frames will have the same segmentation result, so that the object extraction would be stable. Image pixels are firstly smoothed and binarized to reduce the noise possibly introduced in the edge detection step of the proposed segmentation algorithm. Instead of binarizing the image by a hard decision method, a fuzzy binarization approach was applied. A well-known segmentation method, Fuzzy Kohonen Clustering Network (FKCN) (Bezdek et al., 1992), was often applied to segment images. The comparison is provided in Fig. 3. After segmenting the video frame, the objects can be extracted according to criteria based on domain knowledge,

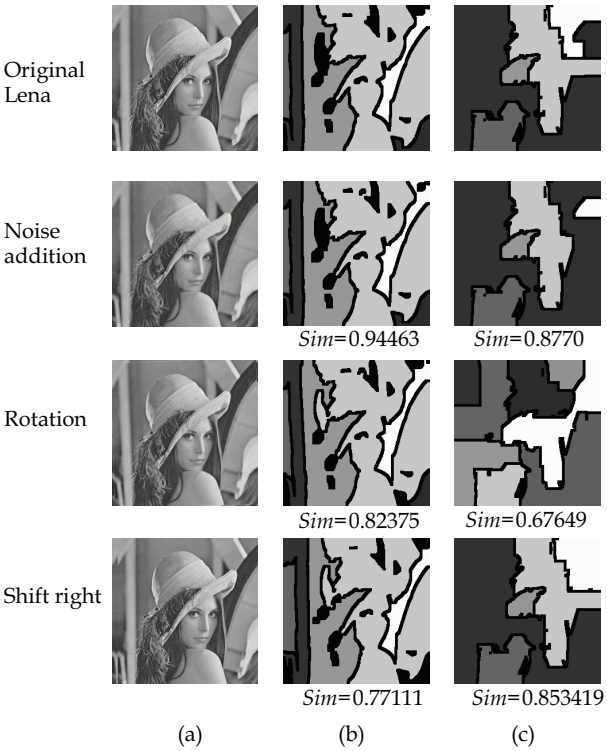


Fig. 3. The comparison between Geometric-Invariant Segmentation (GIS) and FKCN: Corresponding *Sim* values (the similarity between the segments of the manipulated image and original one) of GIS and FKCN after applying various attacks are listed below the images: (a) images manipulated by geometrical operations or signal processing, (b) the resulting images segmented by GIS, and (c) by FKCN. (Lin et al., 2006)

such as skin colour in the news video or playfield colour in the sports video. Bertini et al. (2006) segmented the playfield region from colour histogramming using grass colour information. There are many other segmentation algorithms and schemes proposed in the literature (Chen et al., 2005; Mezaris, et al., 2004; Borenstein & Ulman, 2008; Kokkinos et al., 2009).

Another type of object extraction methods is to find objects using motion features. The highlight (the atomic entities of videos at semantic level) often has specific motions in the video rather than static, so it can be detected by analyzing the motion information. Some sport analysis algorithms (Li et al., 2010) estimate the motion vector to align the background. From the global motion analyzing result of two successive frames, the background can be accurately aligned (Fig. 4). This method also can be applied in moving background sports video. The player can be detected correctly in the video of diving game. Some researches (Papadopoulos et al., 2009) derived statistical approaches to determining the motion area. The kurtosis was used to localize active and static pixels in a video sequence to measure each pixel's s activity (Fig. 5).

(a) (b) (c)

Fig. 4. Result of global motion estimation: (a) -(b) two successive video frames, and (c) the detected background (Li et al., 2010).

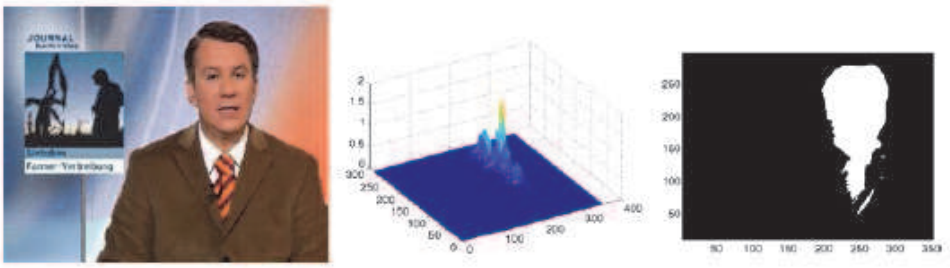


Fig. 5. One example of kurtosis field and activity area mask computation for a news video (Papadopoulos et al., 2009).

The object with higher semantic-level can be encoded by a structural description using low semantic-level objects. Xu et al. (2008) designed a hierarchical compositional model to represent the face, which makes the face representation more condensed and efficient for coding and recognition, as shown in Fig. 6. In another object-based video coding scheme (Wang et al., 2005), the high-level object is composed of the low-level shape and texture

information (Fig. 7). Another semantic video coding for videophone sequences (Zhang, 1998) used an adaptive face model using the deformable template to construct a 3D wireframe for the face (Fig. 8).

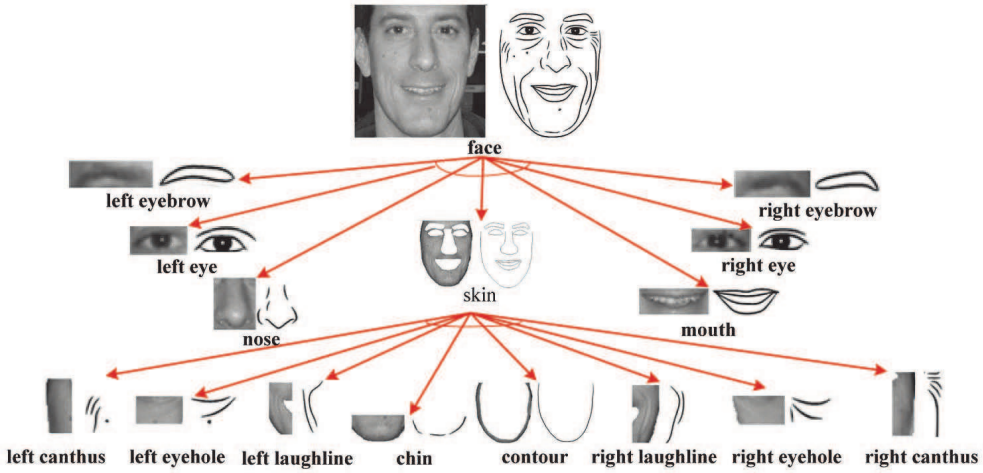


Fig. 6. The hierarchical compositional model for face representation (Xu et al., 2008).

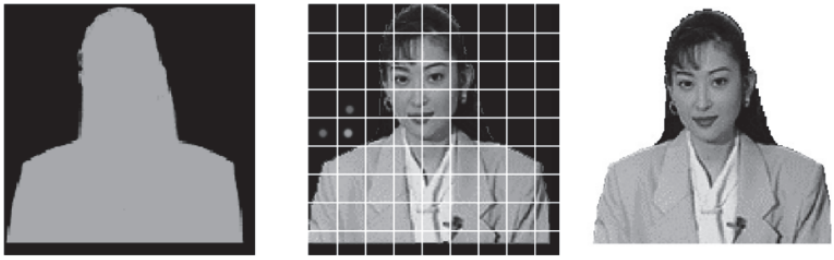


Fig. 7. Example of composition of shape and texture. (a) Shape. (b) Texture. (c) Composed object (Wang et al., 2005).

Fig. 8. The 3D wireframe of the face (Zhang, 1998).

2.2 Event detection and encoding

An event in the video content contains not only its spatial characteristics, but also particular features of the temporal order. Since it has even more semantic-level messages than objects, the structure of the video content should be analyzed based on more domain knowledge. A video can be represented as a multilayer structure, as illustrated in Fig. 9. Scenes, shots and frames are the units that can be found in video. A meaningful story is composed of several scenes. And a scene contains several shots which consist of the video frames that have been continuously recorded with a single camera operation. Shot change detection (Cotsaces et al., 2006; Koprinska & Carrato, 2001; Yuan et al., 2007) is to identify the shots of the video for the purpose of further video analyses and encoding.

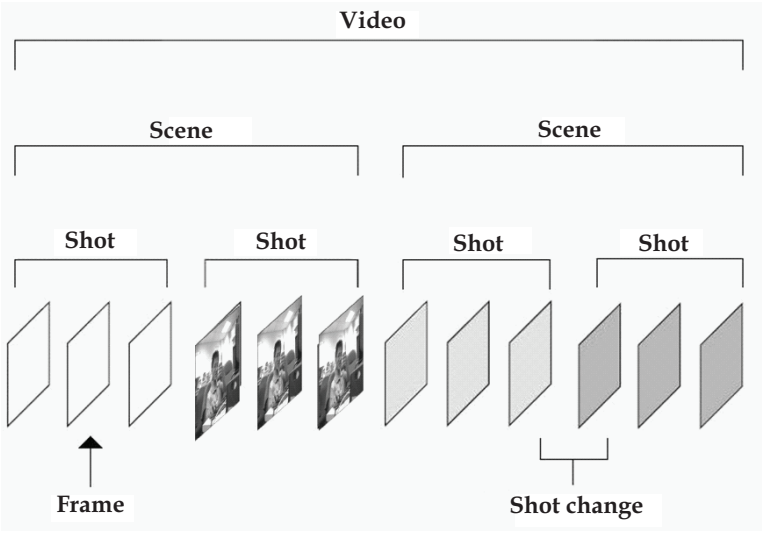


Fig. 9. The video structure.

Sports videos (Bertini et al., 2006; Li et al., 2010) were a frequently discussed application in semantic coding. Bertini et al. (2006) designed a automatic annotation scheme for soccer video based on MPEG-2 (ISO/IEC 13818), which performed event and event-object level compression by detecting camera motion, playfield zone, and players' position in the playfield (Fig. 2). And the players' position determines penalty kicks or free kicks. Further event detection for Forward launch, Shot on goal, Placed kicks, Attack action, and Counter attack are modelled with finite state machine constructed based on soccer rules.

Fig. 10. One example of four successive wedding events (Cheng et al., 2008).

Besides the finite sate machine, the Hidden Markov Model (Rabiner, 1989) is another common tool for modelling the content event. Based on the observation of wedding events, including speech/music types, applause activities, picture-taking activities, and leading roles, Cheng et al. (2008) exploited an HMM framework for segmenting wedding videos, which integrates the wedding event statistical models and the event transition model. Fig. 10 illustrates one example of four successive events, in which OP, WV, RE, WK represents officiant presenting, wedding vows, ring exchange, and wedding kiss, respectively.

3. User-aware semantic video coding

In a heterogeneous network/device environment, the video content should be adaptively encoded to satisfy different users' requirements. However, conventional user adaptive video coding approaches transform the video into bitstreams of various formats independently of the video content. The video is represented and compressed to the adaptive transmission rate and quality by only considering the physical environment, despite of user preferences. Therefore, besides the codec aspects of transmission and presentation constraints of the user's device or transmission capacity, understanding semantic components of the video content while coding, by analyzing the video content using semantic-based temporal or spatial features, could also be a major issue to help produce more condensed and meaningful video for different transmission requirements and user preferences. In this section, we will firstly introduce scalable video coding, then explain how user-aware semantic analysis and manipulations (including construction of user attention models, ROI extraction, and enhancing the interactivity of video coding) assist in improving the coding efficiency.

3.1 Scalable video coding

Scalable video coding is a technique to enable the encoding standard to encode the video into a set of bitstreams with different visual quality to satisfy the needs of different terminals/channels. As defined in MPEG-2, the bitstream is encoded into a base layer and a few enhancement layers, in which the enhancement layers add spatial, temporal, and/or SNR quality to the reconstructed base layer. Later, the fine granular scalability (FGS) is developed in the MPEG-4 (ISO/IEC 14496) Visual standard, which allows a much finer scaling of bits in the enhancement layer. Based on FGS provided in the MPEG-4, (Barrau, 2002) proposed both close-loop and open-loop solutions for the FGS transcoder, which reduce the bit-rate by cutting the enhancement information at know locations.(Qian et al., 2005) combined a scalable transcoder with space time block codes (STBBC) for an orthogonal frequency division multiplexing (OFDM) system to provide robust access to the pre-encoded high quality video server from mobile wireless terminals.

Heterogeneous transcoding converts the pre-compressed bitstream into another bitstream with different format. It is particularly important for the multimedia services which pre-encode the bitstream for storage and transmission. In (Siu et al., 2007), a transcoder from MPEG-2 to H.263 is proposed to convert a B-picture to a P-picture using the information of motion compensation in the DCT domain. Since one of the major differences between MPEG-2 and H.264/AVC is that MPEG-2 uses 8-tap DCT and H.264 uses 4-tap integer (DCT-like) transform (IT), (Shen, 2004; Chen et al., 2006a) designed fast DCT-to-IT algorithms to perform the MPEG-2-to-H.264 transcoding. Since wireless channels have lower bandwidth and higher error rate than wired channels, the error resilience transcoding

over wireless channel is more important. It is particularly useful in hostile environments, such as mobile networks and the Internet. There are many strategies to provide error resilience transcoding (Vetro et al., 2005): 1. Removing the spatial/temporal redundancy can help reduce the error propagation. 2. Group the coded data into several parts according to their importance to allow the unequal protection. 3. Add error-checking bits to the bitstream for robust decoding. 4. Embed additional information into the coded stream to enable the improved error concealment. (Chen et al., 2006b) designed a content-aware intra-refresh (CAIR) transcoding to improve efficiency of the intra-refresh allocation by avoiding the error propagation in the same prediction path.

Fig. 11 illustrates a video transcoder, which provides a seamless interaction between content creation and consumption, or among different channels/terminals. The format can be characterized by the bit-rate, frame rate, coding syntax, spatial resolution, or content (as shown in Fig. 11, in which RI, FI, CI are parameters of the input video, and RO, FO, CO are those of the output video).

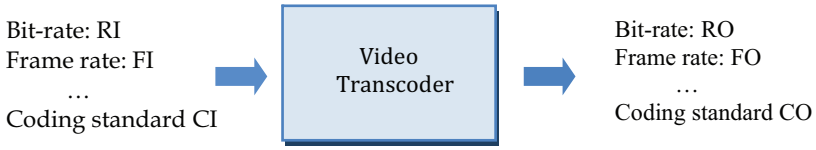


Fig. 11. The video transcoder.

Since video or image have much larger data sizes than other types of data, they have more needs for trascoding. For different applications, the requirements and techniques of the video/image trascoding are still quite different. A common requirement of transcoding is to reduce the complexity of the transcoder and the bit-rate of the data while preserving suitable content quality. It is especially an important issue for video streaming applications in both wired and wireless networks (Chen & Zakhor, 2005). To achieve a target bit-rate while maintaining consistent video/image quality and satisfying the required parameters, (e.g. bandwidth, delay, resolution, and memory constraints), there are various types of trancoding detailed as the following:

A. Frequency domain transcoding

Many video/image compression standards (e.g. JPEG, MPEG-2, MPEG-4, and H.264/AVC) carry out the residual coding in the DCT domain, which consists several steps: the run-length coding, quantization, and the motion compensation (MC). Consequently, many researches try to design DCT-based transcoders because the computational complexity will be much lower than in the pixel domain. (Kim et al., 2006) proposed a bit-rate adaptation method for streaming video in a QoS-based home gateway service, in which the input bitstream is partially decoded into the DCT domain first, then an adaptive motion mapping refinement and a DCT-based image downsizer are utilized to adapt the bit-rate. In (Assunco & Ghanbari, 1998), a drift-free transcoder working entirely in the frequency domain was proposed, in which a Lagrangian rate-distortion optimization was applied for bit reallocation to ensure the quality of the bitstream. Some literatures requantized the DCT coefficients to transcode the bitstream: (Werner, 1999) derived a cost function to estimate the quantizer so that the quantizer can achieve a larger SNR at the same bit-rate compared with the original quantizer used in MPEG-2. Besides, the MSE-based cost function and maximum

a posteriori used in this paper need minor additional complexity. Since the time complexity issue is significant in the real-time applications, (Seo et al., 2000) found an efficient requantization by using a piecewise linearly decreasing model.

B. Temporal resolution adaptation

Both spatial and temporal redundancies should be considered in a video compression algorithm. Besides the MC-based transcodings, temporal redundancies can also be removed by dropping some redundant frames while preserving the temporal smoothness of coded frames. Of course, temporal resolution adaptation is also one of the bit-rate control trascoding for video. In (Shu & Chau, 2005), some video frames are skipped by considering the motion change and reduces the jerky effect caused by undesired frame skipping. (Bonuccelli et al., 2005) designed a buffer-based temporal transcoding in a real-time mobile video application. Rather than dropping frames directly, Shu & Chau (2005) proposed a frame-layer bit allocation method to assign different number of bits for different frames.

C. Spatial resolution adaptation

Resizing is needed to adapt the spatial resolutions to devices with different display capabilities. Moreover, with the emergence of mobile devices and the desire for users to access video originally captured in a high spatial resolution, there is also a need to reduce the resolution for transmitting to and being displayed in such devices. (Shu & Chau, 2007) designed a two-stage structure for arbitrary resizing in DCT-based transcoding, in which some constraints are derived for anti-aliasing.

Although many studies (Lei & Georganas, 2003; Warabino et al., 2000; Elsharkawy et al., 2007) have investigated methodologies to solve the problems related to transcoding in the wireless environment, the rate control and error resilience for wireless applications are still challenging problems, especially for H.264/AVC, the more efficient but complex video standard.

3.2 User-aware semantic video analysis

As described in Section 3.1, different coding requirements should be satisfied for heterogeneous display resolution or communication abilities, which can give temporal, spatial, and quality scalability for the encoded bit stream. However, only considering low-level codec aspects produces limited efficiency gains. Semantic-level analysis will help design more feasible and flexible coding algorithms for different users' needs. To achieve this purpose, the user-aware attention model should be constructed for different applications: For real-time road traffic monitoring, content-based scalable coding (Ho et al., 2005) can help increase the compression rate. In the wireless environment, a temporal scalability scheme with background composition (Hung & Huang, 2003) was proposed in MPEG4. And effective bit-rate control can be achieved by considering the Region of Interest (ROI) (Grois et al., 2010). As shown in Fig. 12, the ROI is used as the baselayer of H.264/AVC standard, which can provide various resolution and bit-rates for different users' needs. Table 1 presents the bit-rate savings when exploiting this method. For lecture videos, a learner-focused model can be designed to reduce the network traffic in case of the real-time streaming video (Lin et al., 2009, 2010a). Fig. 13 illustrates one example of lecture video coding, in which a lot of lecture video frames are compressed into one lecture slide with teaching focus and the temporal redundancy is reduced based on semantic-level analysis.

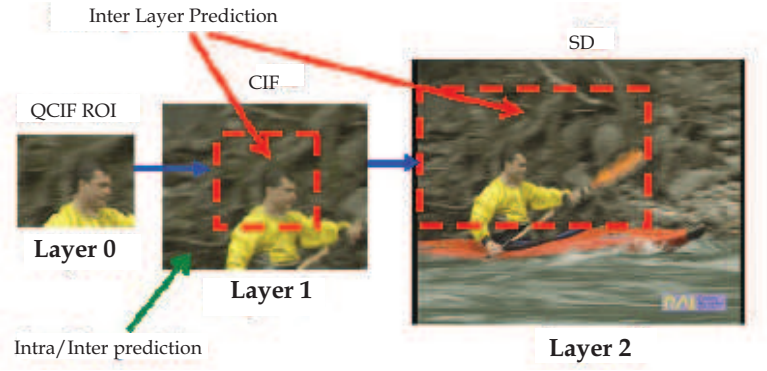


Fig. 12. Example of the ROI dynamic adjustment and scalability for mobile devices with different spatial resolution (Grois et al., 2010).

Quantization Parameters	Four Layers (640x360, and three HD layers)		Eight Layers (two CIF layers, three SD layers, and three HD layers)		Bit-Rate Savings (%)
	PSRN	Bit-Rate	PSNR	Bit-Rate	
32	34.48	2566.15	34.49	3237	20.73
34	33.93	1730.21	33.93	2359	26.66
36	33.27	1170.01	33.27	1759	33.48

Table 1. The bit-rate savings when using ROI adaptive scalable video coding (Grois et al., 2010).

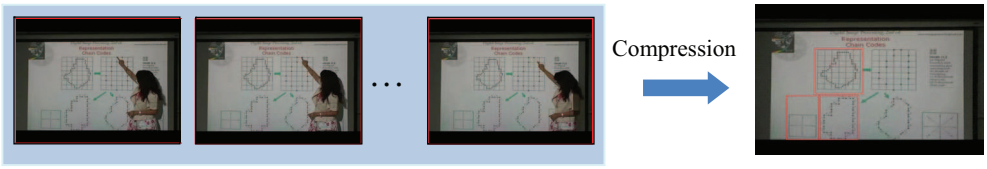


Fig. 13. Lecture video coding: (a) the lecture video frames are compressed into (b) the lecture slide with detected teaching focuses (Lin et al., 2009).

In the following, we will introduce user-aware semantic understanding techniques for videos by extracting and analyzing user-aware visual/aural features, including the analysis of expression, gesture, emotion, motion, and event detection, for the purpose of enhancing the video coding.

Fig. 14 illustrates one example of user-aware video analysis schemes, in which the learner-focused attention model was constructed and provided for enhancing the video lecture representation (Lin et al., 2010b).

Visual analysis can be used to understand the video semantically by merely finding low-level features (color, texture, pixel histogram, etc.) or further extracting semantic-level features (gesture, expression, action, etc.) from low-level visual features, which will be introduced by providing examples in the following.

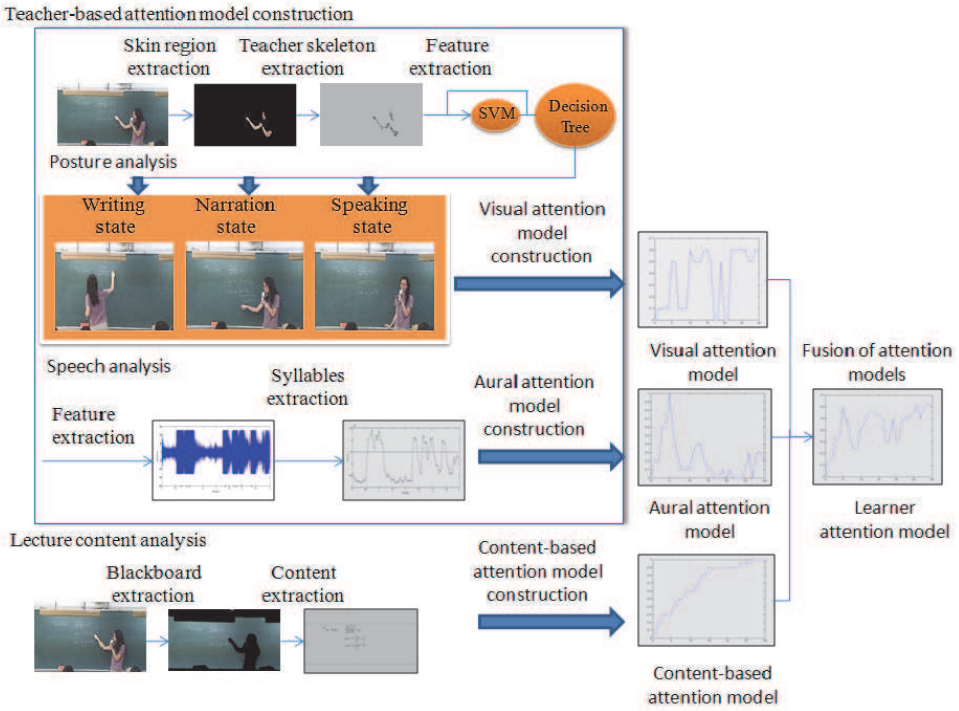


Fig. 14. One example of the user-aware video analysis scheme for constructing a learner-focused attention model. (Lin et al., Sept. 2010b)

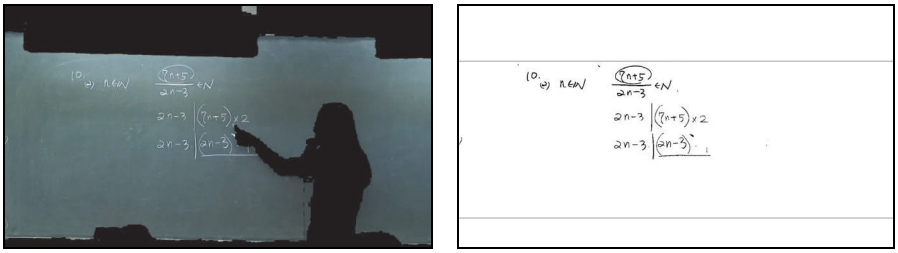
3.2.1 User-aware visual analysis

Low-level features can be extracted by directed computig the visual characteristics in the spatical or frequency domain, which contains no semantic meaning for humans at first glance. In the adaptive video learning system proposed in (Lin et al., 2010b), the importance of the lecture content are decided by analyzing the extracted lecture content and also the instructor's behavior. In lecture content, color features are used to couting the chalk pixels. The blackboard region is at first obtained by extracting the regions of the blackboard colour and merging them(Fig. 15 (a)). After deciding the blackboard region, the set of chalk pixels P_{chalk} can be computed as

$$P_{chalk} = \bigcup \left\{ x \middle| I(x) > I_{cp} \right\}, \tag{1}$$

where $I(x)$ is the luminance of pixel x and I_{cp} is the luminance threshold. Fig. 15 shows one example of lecture content extraction.

The chalk text or figures written on the blackboard by the lecturer are undoubtedly the most important part that lecturers want students to pay attention to. It is obvious that the more there is lecture content (chalk handwriting or figures), the more revealed semantics are in the lecture video. Therefore, the attention values are evaluated by extracting the lecture content on the blackboard and analyzing the content fluctuation in lecture videos.



(a) (b)

Fig. 15. Lecture content extraction: (a) the blackboard region, and (b) the extracted lecture content.

Another important low-level feature for videos is motion extracted in the pixel or compression domain. Many sophisticated motion estimation algorithms has been developed in the literature, for examples, the optical flow in (Beauchemin & Barron, 1995) and the feature tracking in (Shi & Tomasi, 1994). However, they often have high computational complexity because the operations are executed in the pixel domain and the estimated motions are accurate. In the work of (Chang et al., 2010a), accurate motion estimation is not needed, so the motion information can be directly extracted from the motion vectors of a compressed video. Since the process is done directly in the compression domain, the induced complexity is very low. Therefore, the motions in each video frame can be efficiently obtained.

As mentioned in Section 3.1.1, object extraction is an important process before deciding video events. In many user-aware applications, human detection is the major work while extracting objects. In the lecture video application, the lecturer should be detected for further analysis. In the work of Lin et al. (2009), the human area was extracted by detecting the moving object in the video frames, which was carried out by finding the eigenregions in the frames. That is, the moving objects can be distinguished from the still objects by methods of classification. The PCA (Principal Component Analysis)-based approach is used in this paper, which is detailed in the following. Three successive frames F_{i-1}, F_i, and F_{i+1} are firstly transformed into a matrix $X=[F_{i-1}\ F_i\ F_{i+1}]$, then the covariance C is computed as $C=X^TX$. Finally, each frame is aligned with the first two principle vectors (which are the eigenvectors of C associated with the two largest eigenvalues. Thus, the area with higher values represents that with higher variances, i.e., the moving object.

After the eigenregion is extracted, the produced image (Fig. 16 (d)) is binarized and applied by morphological operators to fill and smooth the region in order to obtain a more stable mask. Fig. 16 shows one example of moving object detection, in which (a), (b), and (c) are three successive frames, (d) is the corresponding eigenregion, (e) is the binarized one, and (f) is the final mask after morphological operations.

Low-level feature can provide limited information for human perception, for example, the DCT coefficients could not be understood well by humans, even though these features play an important role in pattern recognition. Therefore, semantic understanding for videos can be improved by extracting semantic-level features like gestures, expression, actions, etc. For instance, the posture of the lecturer will generally change with the delivered lecture content or the situation in lecture presentation. For example, when teaching the math problems, the lecturer may firstly write the lecture content on the blackboard and shows their back to the

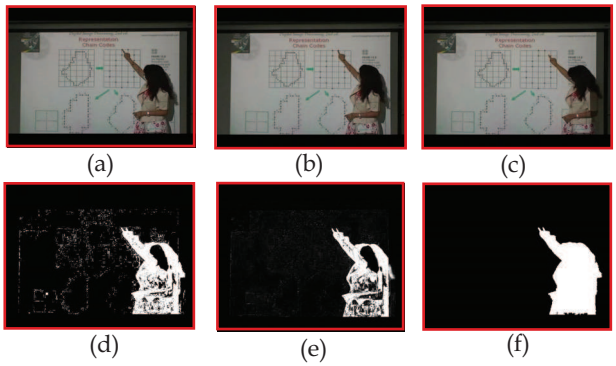


Fig. 16. Human detection: (a), (b), and (c) are successive frames, (d) Eigenregion (lighter area), (e) binarized eigenregion, and (f) the resulting mask. (Lin et al., Sept. 2010b)

students. Next, he starts to narrate the written equations and moves sideways to avoid occluding the content which students should focus their gazes on. After writing the complete lecture content, the lecturer will face the students to further explain the details. All of the lecturing statuses and postures mentioned above will repeatedly occur with alternative random order in a course presentation. Different states represent different presentation states and also different semantics. Therefore, the lecturing states can be decided according to the changes of the lecturer's posture. In (Lin et al., 2010b), the skeleton of the lecturer is extracted to represent the posture and then the lecturing states are identified by using the SVM approach. The regions of the head and hands are detected by using the skin-color features. The lecturer's skeleton is then constructed by considering the relations between the positions of head and hands.

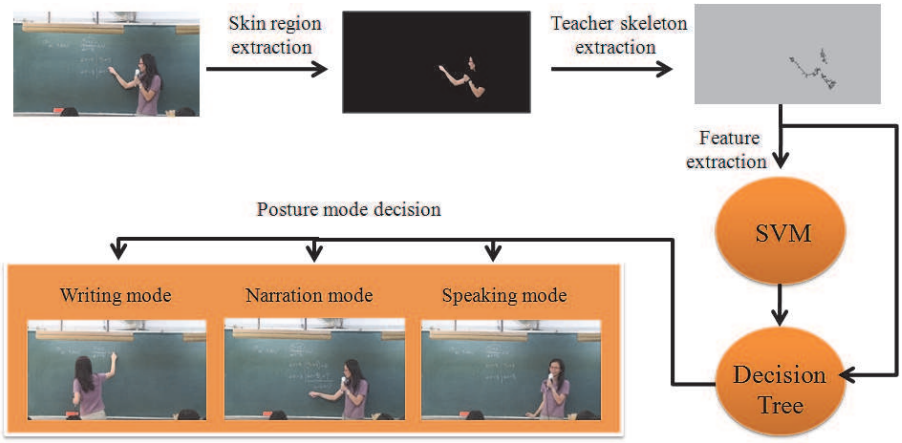


Fig. 17. Analysis of lecturer's posture.

After constructing skeletons, several features derived from the skeleton are used for posture discrimination to estimate the lecturing state (Fig. 17), including the distance between end points of the skeleton, the joint angle of the skeleton, and the orientation of the joint angle.

Then features mentioned above are used to train a SVM classifie, so that the other defined lecturing states can be identified.

3.2.2 User-aware aural analysis

Aural information in multimedia contents is also an important stimulus to attract viewers and should be utilized to affect the inserted virtual content. Compared to visual saliency analysis, researches on aural saliency analysis are rare. In (Ma et al., 2005), an aural attention modeling method, taking aural signal, as well as speech and music into account, was proposed to incorporate with the visual attention models for benefiting video summarization. Intuitively, a sound with loud volume or sudden change usually grabs human's attention no matter what they are looking at. If the volume of sound keeps low, even if a special sound effect or music is played, the aural stimulus will easily be ignored or be treated as the environmental noise. In other words, loudness of aural information is a primary and critical factor to influence human perception and can be used to model aural saliency. Similar to the ideas stated in (Ma et al., 2005), the sound is considered as a salient stimulus in terms of aural signal if the following situation occurs: loudness of sound at a specific time unit averages higher than the ones within a historical period which human continued listening so far, especially with peaks.

Based on the observations and assumptions, the aural saliency response $AR(T_h, T)$, is defined at a time unit T and within a duration T_h, to quantify the salient strength of the sound. That is,

$$AR(T_h, T) = \frac{E_{avr}(T)}{\hat{E}_{avr}(T_h)} \cdot \frac{E_{peak}(T)}{\hat{E}_{peak}(T_h)}, \tag{2}$$

where $E_{avr}(T)$ and $E_{peak}(T)$ are the *average sound energy* and the *sound energy peak* in the period T, respectively.

After analyzing the aural saliency of the video, an *AS feature sequence* is generated which describes the aural saliency response with the range [0, 1] at each time unit T, as shown in Fig. 18.

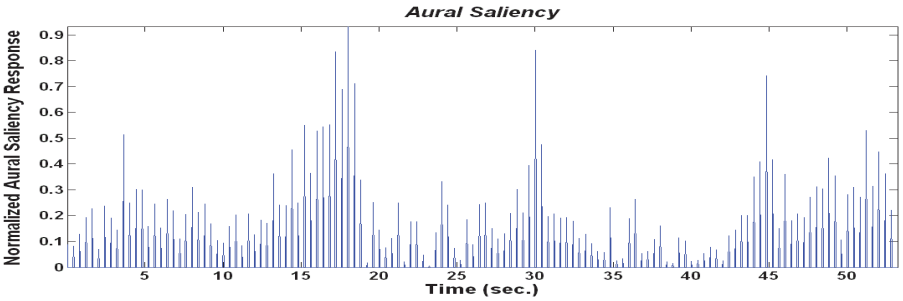


Fig. 18. The normalized aural saliency response of the audio segment.

In (Lin et al., 2010b), besides gesture and posture, making sounds or changing tones is another way that lecturers usually used to grab students' attentions while narrating the lecture content. Therefore, aural information of lecturers is an important cue to estimate the attentions for lecture videos. Since more words are spoken by lecturers within a period may

imply more semantics are conveyed or delivered in such duration, the aural attention can be modeled based on the lecturer's speech speed. Generally, each Chinese character corresponds to at least one syllable, so we can analyze the syllables by extracting the envelope (Iked et al., 2005) of audio samples to estimate the lecturer's speech speed as (3).

$$W_{syllable}(t) = \begin{cases} 1 \ , \ if \ \ e(t) > C_w Max\{e(T)\} \\ 0 \ , \ otherwise \end{cases}, \tag{3}$$

where $W_{syllables}(t)$ represents whether it is a syllable ending at time t, $e(t)$ is the envelope size at time t, T is a time period, and c_w is a threshold.

3.3 ROI estimation

ROI could be considered as one of the semantic scalability in spatial dimension. The virtual content should be inserted at suitable spatial and temporal location, which is often an area attractive to humans, that is, the ROI region. While considering the human perception and viewing experience, a compelling multimedia content is usually created by artfully manipulating the salience of visual and aural stimulus. Therefore, attractive regions or objects are usually utilized to direct and grab viewers' attention and play an important role in multimedia contents. Algorithms of both patial and temporal ROI estimation will be discussed in this section.

In order to automatically identify such attractive information in visual contents, a great deal of research efforts on estimating and modeling the visual attention in human perception have proliferated for years. The systematic investigations about the relationships between the vision perceived by humans and attentions are provided in (Chun & Wolfe, 2001; Itti & Koch, 2001; Chen et al., 2003; Ma et al., 2005; Liu et al., 2007; Zhang et al., 2009). Itti et al. (2001) presented a frameworkfor a computational and neurobiological understanding of visual attention modeling. Ma et al.(2005) proposed a generic framework of user attention model by fusing several visual and aural features and applied it to video summarization. As for practical applications, numerous visual attention models were explored to adapt images (Chen et al., 2003; Liu et al., 2007) and videos (Cheng et al., 2007) for improving viewing experience on the devices with small displays. Zhang et al. (2009) proposed a distortion-weighing spatiotemporal visual attention model to extract the attention regions from the distorted videos. Instead of directly computing a bounding contour for attractive regions or objects, most approaches construct a saliency map to represent the attention strength or attractiveness of each pixel or image block in visual contents. The value of a saliency map is normalized to [0, 255] and the brighter pixel means higher salience. Several fusion methods for integrating each of the developed visual feature models have been developed and discussed in (Dymitr & Bogdans, 2000). Different fusion methods are designed for different visual attention models and applications. The goal of this module is to be able to provide a flexible mechanism to detect various ROIs as the targets, which the inserted ads can interact with, according to the users' requirements. For this purpose, Chang et al. (2009) utilize linear combinations for fusion, so that users can flexibly set each weight of corresponding feature salience maps. The ROI saliency map, which is denoted as S_{ROI} , is computed as

$$S_{ROI} = \sum_{i=1}^{n} w_i \times F_i, \tag{4}$$

where F_i is the i-th feature map of that frame, and w_i is the i-th weight of the corresponding i-th feature map F_i with the constraints of $w_i \geq 0$, and $\sum_{i=1}^{n} w_i = 1$. The ROI can be easily derived by evaluating the center of gravity and the ranging variance on the basis of the saliency map.

A human visual system (HVS) has been introduced for finding ROIs in many researches. In (Lee et al., 2006) and (Kankanhalli & Ramakrishnan, 1998), an HVS was used to improve the quality of a watermarked image. In (Geisler & Perry, 1998), an HVS was used to skip bits without influencing the visual perceptibility of video encoding applications. It also can be applied to build the user-attentive model proposed in (Cox et al., 1997) for deciding the ROI. In (Lin et al., 2010b), the user-attentive model is constructed based on the graylevel and texture features of the image. Regions with mid-gray levels will have a high score for selection because regions with very high or low gray levels are less noticeable to human beings. In addition, the strongly textured segments will have low scores. The distances to the image center are also considered because human beings often focus on the area near the center of an image.

Besides the spatial ROIs, temporal ROIs should also be considered for removing temporal redundancy. The temporal ROI is the video clip which is attractive to humans. The curve derived from the user attention model can be used to determine the temporal ROIs, which have higher values of user attention function.

3.4 Interactivity of video coding

Some video coding standards, such as the MPEG-4, allow developing algorithms of audio-visual coding for not only high compression, but also interactivity and universal accessibility of the video content. In addition to the traditional "frame"-based functionalities of the MPEG-1 and MPEG-2 standards, the MPEG-4 video coding algorithm will also support access and manipulation of "objects" within video scenes. The "content-based" video functionality is to encode the sequence in a way that will allow the separate decoding and reconstruction of the objects using the concept of Visual Objects (VOs), and to allow the manipulation of the original scene by simple operations on the bit stream. The properties of objects are described in the bit stream of each object layer. As illustrated in Fig. 19, a video scene can be encoded into several Visual Object Planes (VOPs), wich can be manipulated by simple operations.

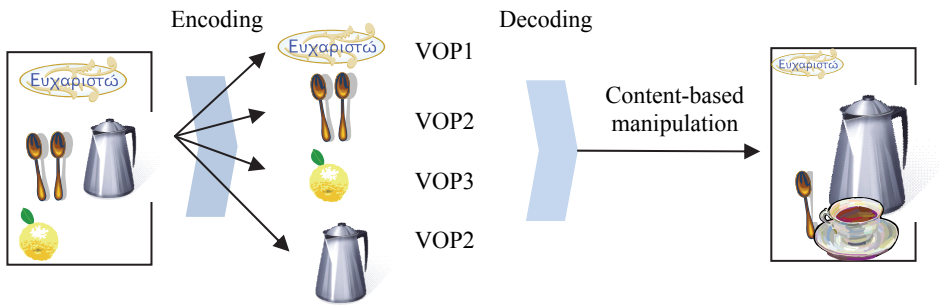


Fig. 19. Composition and manipulation of MPEG-4 videos.

To enhance the interactivity of the video content for user adaptive applications, Chang et al. (2010a, 2010b) presented an interactive virtual content insertion architecture which can insert virtual contents into videos with evolved animations according to predefined behaviors emulating the characteristics of evolutionary biology. The videos are considered not only as carriers of message conveyed by the virtual content but also the environment in which the lifelike virtual contents live. Thus, the inserted virtual content will be affected by the videos to trigger a series of artificial evolutions and evolve its appearances and behaviors while interacting with video contents. By inserting virtual contents into videos through the system, additional entertaining storylines can be easily created and the videos will be turned into visually appealing ones. The above mentioned concept is illustrated in Fig. 20.

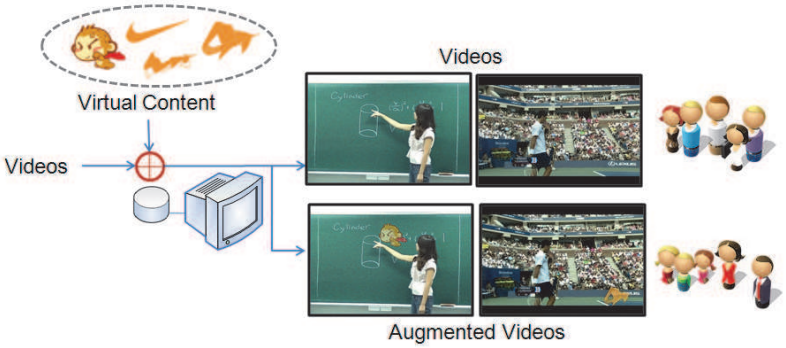


Fig. 20. The augmented videos can be served by using techniques in interactive virtual content insertion to enrich the original videos.

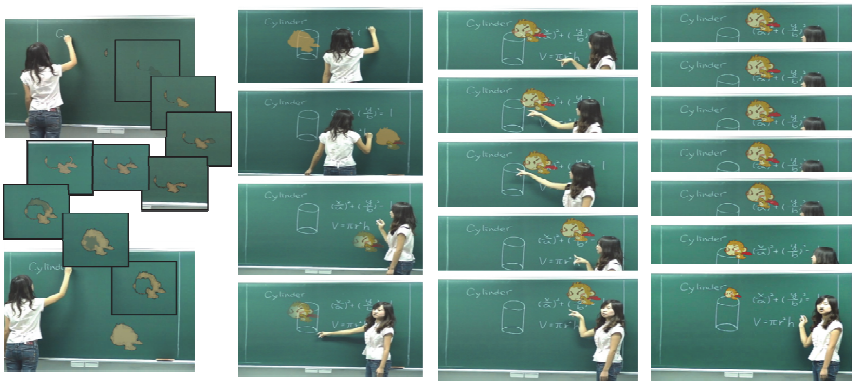


Fig. 21. Snapshots of sample results of the virtual learning e-Partner. The e-Partner can evolve according to the lecturer's teaching behavior in the lecture video and assist in pointing out or enhancing the important lecture content.

In (Chang et al., 2010b), a virtual learning e-partner scheme was presented. The e-partner is assigned the ability to seek for the salient object, which is detected by finding the ROIs based on the algorithms described in Section 3.3, and is simulated to obtain the color and

texture by absorbing the energy of the salient object. At last, the e-partner owns the ability to dance with the music or show the astonished expression while perceiving loud sound. Besides, the e-partner would interact with the moving salient object in an intelligent manner. The e-partner would either tend to imitate the behavior of the moving salient object, or moves to the salient object for further interactions. With the extracted feature space of the lecture videos and the behavior modeling of the e-partner, the proposed system automatically generates impressive animations with an evolution way on a virtual layer. Finally, the virtual layer, in which the e-partner is animated on, is integrated with the video layer. Fig. 21 shows sample results of the virtual learning e-partner, in which the e-partner can evolve according to the lecturer's teaching behavior in the lecture video and assist in pointing out or enhancing the important part of lecture content.

To support the interactivity of video coding, many researches (Naman & Taubman, 2007; Ng et al., 2010; Wang et al., 2005; Tran et al., 2004) propsed content-based scalable coding schemes, most of which applied the concept of VOPs of MPEG-4. Fig. 22 shows a generic architecture of semantic scalable video coding based on the multilayer structure.

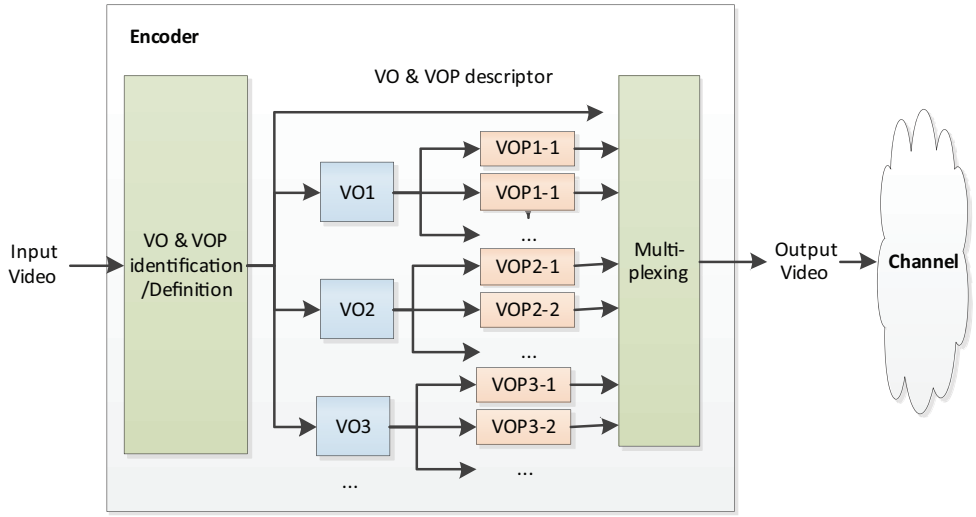


Fig. 22. The architechre of object-based scalable video coding.

5. Conclusion

With the rapid development of information technology, access to internet service and use of multimedia has been increasing in recent years. Since the network bandwidth is limited, it is important to investigate approaches to control the bit-rate adaptively for various requirements of different transmission capacity, devices, or user preferences. Moreover, in order to increase the flexitility and interactivitly for accessing and manipulating the video content, semantic-level analysis should be considered to achieve user-aware functionalities. In this chapter, we have introduced theory and practice of user-aware semantic video coding., includng concepts and techniques of salable video coding, transcoding, semantic analysis, and semantic coding. In addition, related methods for user adaptive video coding,

containing ROI estimation, virtual content insertion, and object-based video coding, are also discussed.

6. Acknowledgement

This work was partially supported by the National Science Council and the Ministry of Education of China under contact no. NSC 99-2511-S-260-001-MY2.

7. References

Assuncao, P. A. A. & Ghanbari, M. (Dec. 1998). A Frequency-domain Video Transcoder for Dynamic Bitrate Reduction of MPEG-2 Bit Streams, *IEEE Trans. Circuits Syst. Video Technol.*, Vol. 8, No. 8, pp. 953-967.

Barrau, E. (2002). MPEG Video Transcoding to A Fine-granular Scalable Format, *IEEE ICIP*, pp. I717-720,.

Beauchemin, S. S. & Barron, J. L. (1995). The computation of optical flow, *ACM Computing Surveys*, pp. 433-467.

Bertini, M.; Cucchiara, R. ; Del Bimbo, A. & Prati, A. (June 2006). Semantic Adaptation of Sports Video with User-centred Performance Analysis, *IEEE Transactions on Multimedia,* Vol. 8, No. 3, pp. 433-443.

Bezdek, J. C. ; Tsao, E. C.-L. ; & Pal, N. R. (1992). Fuzzy Kohonen Clustering Networks, *IEEE ICFS 1992*, pp.1035–1043, 1992.

Bonuccelli, M. A.; Lonetti, F. & Martelli, F. (2005). Temporal Transcoding for Mobile Video Communication, *IEEE Proc. Mobile and Ubiquitous Systems: Networking and Services*.

Borenstein, E. & Ullman, S. (December 2008). Combined Top-Down/Bottom-Up Segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 30, No. 12, pp. 2109-2125.

Chang, C. H.; Chiang, M. C. & Wu, J. L. (2009). Evolving virtual contents with interactions in videos, *Proceedings of the ACM International Workshop on Interactive Multimedia for Consumer Electronics* (*IMCE*), pp. 97-104.

Chang, C. H.; Hsieh, K. Y.; Chiang, M. C. & Wu, J. L. (Oct. 2010a). Virtual spotlighted advertising for tennis videos, *Journal of Visual Communication and Image Representation* (*JVCIR*), Vol. 21, No. 7, pp. 595-612.

Chang, C. H.; Lin, Y. T. & Wu, J. L. (April 2010b). Adaptive video learning by the interactive e-partner, *Proceedings of the 3rd IEEE International Conference on Digital Game and Intelligent Toy Enhanced Learning* (*DIGITEL'10*), Kaohsiung, Taiwan, pp. 207-209.

Chen, G.; Lin, S. & Zhang, Y. (2006a). A Fast Coefficients Convertion Method for the Transform Domain MPEG-2 to H.264 Transcoding, *International Conference on Digital Telecommunications* (*ICDT'06*).

Chen, C. M.; Chen, Y. C. & Lin, C. W. (2006b). Error-Resilience Transcoding Using Content-Aware Intra-Refresh Based on Profit Tracing, *IEEE ISCAS*, pp.5283-5286.

Chen, J.; Pappas, T. N.; Mojsilović A. & Rogowitz, B. E. (October 2005). Adaptive perceptual color-texture image segmentation, *IEEE Transactions on Image Process*, Vol. 14, No. 10, pp.1524-1536.

Chen, L.; Xie, X.; Fan, X.; Ma, X.; Zhang, H. & Zhou, H. (October 2003). A visual attention model for adapting images on small displays, Multimedia Systems, Vol. 9, No. 4, pp. 353-364.

Chen, M. H. & Zakhor, A. (2005). Rate Control for Streaming Video over Wireless, *IEEE Wireless Communications*, Vol. 12, No. 4.

Cheng, W. H.; Chuang, Y. Y .; Lin, Y. T.; Hsieh, C. C.; Fang, S. Y.; Chen, B. Y. & Wu, J. L. (November 2008). Semantic Analysis for Automatic Event Recognition and Segmentation of Wedding Ceremony Videos, *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 18, No. 11, pp. 1639-1650.

Cheng, W. H.; Wang, C. H. & Wu, J. L. (2007). Video adaptation for small display based on content recomposition, *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 1, pp. 43-58.

Chun, M. M. & Wolfe, J. M. (2001). *Visual attention in Blackwell handbook of perception*, USA: Wiley-Blackwell, Ch. 9, pp. 272-310.

Cotsaces, C. ; Nikolaidis, N. & Pitas, I. (2006). Video Shot Detection and Condensed Representation a review, *IEEE Signal Processing Magazine*, Vol. 23, No. 2, pp. 28-37.

Cox, I.; Kilian, J.; Leighton, F. & Shamoon, T. (Dec. 1997). Secure Spectrum Watermarking for Multimedia, IEEE Trans. on Image Processing, Vol. 6, No. 12, pp. 1673-1687.

Dymitr, R. & Bogdan, G. (2000). An overview of classifier fusion methods, Computing and Information Systems, Vol. 7, No. 1, pp. 1-10.

Elsharkawy, M. I.; Aly, & Elemandy, H. (2007). Secure Scalable Video Transcoding over Wireless Network, *IEEE Intelligent Computer Communication and Processing*, pp.287-292.

Geisler, W. S. & Perry, J. S. (1998). A Real-Time Foveated Multiresolution System for Low-Bandwidth Video Communication, *SPIE proceedings: Human Vision and Electronic Imaging* (*VCIP'98*), pp.294-305.

Grois, D.; Kaminsky, E. & Hadar, O. (Oct. 2010). ROI Adaptive Scalable Video Coding for Limited Bandwidth Wireless Networks, *IFIP Wireless Days*.

Ho, W. K.-H.; Cheuk, W.-K. & Lun, D. P.-K. (August 2005). Content-based Scalable H.263 Video Coding for Road Traffic Monitoring, *IEEE transaction on multimedia*, Vol. 7, No. 4, pp. 615-623.

Hung, B. G. & Huang, C. L. (December 2003). Content-Based FGS Coding Mode Determination for Video Streaming Over Wireless Networks, *IEEE Journal on Selected Areas in Communications*, Vol.21, No. 10, pp. 1595-1603.

Ikeda, O. (2005). Estimation of speaking speed for faster face detection in video-footage, *Proceedings of the International Conference on Multimedia and Expo.*, pp. 442-445.

Itti, L. & Koch, C. (March 2001). Computational modelling of visual attention, *Nature reviews*, Neuroscience, Vol. 2, No. 3, pp. 194-203.

Kankanhalli, M. S. & Ramakrishnan, K. R. (1998). Content Based Watermarking of Images, *ACM Multimedia'98*, pp.61-70.

Kim, J. W.; Kwon, G. R.; Kim, N. H.; Morales, A. & Ko, S. J. (Feb. 2006). Efficient Video Transcoding Technique for QoS-Based Home Gateway Service, *IEEE Trans. Consumer Electronics*, Vol. 52, No. 1, pp. 129-137.

Kokkinos, I.; Evangelopoulos, G. & Maragos, P. (January 2009). Texture analysis and segmentation using modulation features, generative models, and weighted curve evolution, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (*T-PAMI*), Vol. 31, No. 1, pp. 142-157.

Koprinska, I. & Carrato, S. (2001). Temporal Video Segmentation: A Survey, *Signal Processing: Image Communication*, Vol. 16, pp. 477-500.

Lee, Y. H.; Kim, H. & Lee, H. K. (March 2006). Robust image watermarking using local invariant features, Optical Engineering, SPIE 2006,Vol. 45, No. 3.

Lei, Z. & Georganas, N. D. (2003). Video Transcoding Gateway for Wireless Video Access, *IEEE CCGEI*, pp.1775-1778.

Li, H.; Tang, J. ; Wu, S.; Zhang, Y. & Lin, S. (Mar. 2010). Automatic detection and analysis of player action in moving background sports video sequences, *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 20, No. 3, pp.351-364.

Lin, Y. T. ; Wu, J. L. & Kao, Y. F. (March 2006). Geometric-invariant image watermarking by object-oriented embedding, *International Journal of Computer Science and Network Security*, Vol. 6, No. 3A, pp. 169-180.

Lin, Y. T.; Yen, B. J.; Chang, C. C.; Yang, H. F. & Lee, G. C. (Dec. 2009). Indexing and teaching focus mining of lecture videos, *Proceedings of 11th IEEE International Symposium on Multimedia* (*ISM'09*), San Diego, California, USA, pp. 681-686.

Lin, Y. T.; Yen, B. J.; Chang, C. C.; Yang, H. F.; Lee, G. C. & Lin, Y. C. (2010a). Content-based indexing and teaching focus mining for lecture videos, *Interactive Technology and Smart Education*, Vol. 7, No. 3, pp.131-153.

Lin, Y. T.; Tsai, H. Y.; Chang, C. H. & Lee, G. C. (Sept. 2010b). Learning-focused structuring for blackboard lecture videos, *Proceedings of the 4th IEEE International Conference on Semantic Computing* (*ICSC'10*), Carnegie Mellon University, Pittsburgh, PA, USA.

Liu, H.; Jiang, S.; Huang, Q.; Xu, C. & Gao, W. (2007). Region-based visual attention analysis with its application in image browsing on small displays, *Proceedings of the 15th International Conference on Multimedia* (*MULTIMEDIA*), pp. 305-308.

Ma, Y.; Hua, X.; Lu, L. & Zhang, H. (2005). A generic framework of user attention model and its application in video summarization, *IEEE Transactions on Multimedia* (*T-MM*), vol. 7, no. 5, pp. 907-919.

Mezaris, V. ; Kompatsiaris, I. & Strintzis, M. G. (June 2004). Video object segmentation using Bayes-based temporal tracking and trajectory-based region merging, *IEEE Transactions on Circuits and Systems for Video Technology* (*T-CSVT*), Vol. 14, No. 6, pp. 782-795.

Naman, A. T. & Taubman, D. (Jan. 2007). JPEG2000-Based Scalable Interactive Video (JSIV), *IEEE Transactions on Image Processing*, Vol. 6, No. 1, pp. 1-16.

Ng, K. T.; Shing, Q. W.; Chan, C. & Shum, H. Y. (April 2010). Object-Based Coding for Plenoptic Videos, *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 20, No. 4, pp. 548-562.

Papadopoulos, G. H.; Briassouli, A.; Mezaris, V.; Kompatsiaris, I. & Strintzis, M. G. (Oct. 2009). Statistical motion information extraction and representation for semantic video analysis, *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 19, No. 10, pp.1513-1528.

Qian, T.; Sun, J.; Xie, R.; Su, P.; Wang, J.& Yang, X. (2005). Scalable Transcoding for Video Transmission over Space-time OFDM Systems," *IEEE SIPS*.

Rabiner, L. R. (February 1989). A tutorial on hidden Markov models and selected applications in speech recognition, *Proceedings of the IEEE*, Vol. 77, No. 2, pp. 257-285.

Seo, K. D.; Lee, S. H.; Kim, J. K. & Koh, J. S. (Nov. 2000). Rate Control Algorithm for Fast Bit-rate Conversion Transcoding, *IEEE Trans. Consumer Electronics*, Vol. 46, No. 4.

Shen, B. (2004). From 8-Tap DCT to 4-Tap Integer-Transform for MPEG to H.264/AVC Transcoding, *IEEE ICIP*.

Shi, J. & Tomasi, C. (1994). Good features to track, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), pp. 539-600.

Shu, H. & Chau, L.P. (Feb. 2007). A Resizing Algorithm with Two-stage Realization for DCT-based Transcoding, *IEEE Trans. Circuits and Systems for Video Technology*, Vol. 17, No. 2, pp. 248-253.

Shu, H. & Chau, L.P. (May 2005). Frame-skipping Transcoding with Motion Change H. Shu and L.P. Chau. "Frame Layer Bit Allocation for Video Transcoding, *IEEE International Symposium on Circuits and Systems*, ISCAS2005, Kobe, Japan.

Siu, W.C.; Chan, Y.L. & Fung, K.T. (Oct. 2007). On Transcoding a B-Frame to a P-Frame in the Compressed Domain, *IEEE Trans. Multimedia*, Vol. 9, No. 6, pp. 1093-1101.

Tran. S. M.; Lajos, K.; Balazs, E.; Fazekas, K. & Csaba S. (June 2004). A Survey on The Interactivity Feature of MPEG-4, *46th International Symposium Electronics in Marine*, Zadar, Croatia, pp. 30-38.

Vetro, A. (Jan.-Mar. 2004). MPEG-21 digital item adaptation: enabling universal multimedia access, *IEEE Multimedia*, Vol. 11, No. 1, pp. 84 – 87.

Vetro, A.; Xin, J. & Sun, H. (Aug. 2005). Error Resilience Video Transcoding for Wireless Communications, *IEEE Wireless Communications*, pp. 14-21.

Wang , H.; Schuster, G. M. & Katsaggelos, A. K. (2005). Rate-Distortion Optimal Bit Allocation for Object-Based Video Coding, IEEE Trans. Circuits and System for Video Technology, Vol. 15, No. 9, pp. 1113-1123.

Warabino, T.; Ota, S.; Morikawa, D. & Ohashi, M. (Oct. 2000). Video Transcoding Proxy for 3Gwireless Mobile Internet Access, *IEEE Communication Magazine*, pp. 66-71.

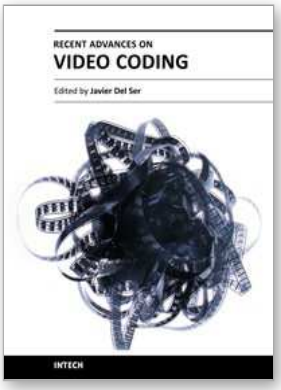
Werner, O. (Feb. 1999). Requantization for Transcoding of MPEG-2 Intrframes, IEEE Trans. Image Processing, Vol. 8, No. 2, pp. 179-191.

Xu, Z.; Chen, H.; Zhu, S. C. & Luo, J. (June 2008). A Hierarchical Compositional Model for Face Representation and Sketching, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 30, No. 6, pp. 955-969.

Yuan, J. ; Wang, H. ; Xiao, L. ; Zheng, W.; Lin, J., Li, F. & Zhang, B. (2007). A Formal Study of Shot Boundary Detection, *IEEE Trans. on Circuits and Systems for Video Technology*, Vol. 17, No. 2, pp.168-186.

Zhang, H.; Tian, X. & Chen, Y. (2009). A distortion-weighing spatiotemporal visual attention model for video analysis, *Proceedings of the International Congress on Image and Signal Processing,* pp.1-4.

Zhang, L. (Oct. 1998). Automatic adaptation of a face model using action units for semantic coding of videophone sequences, *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 8, No. 6, pp. 781-795.



Recent Advances on Video Coding

Edited by Dr. Javier Del Ser Lorente

ISBN 978-953-307-181-7
Hard cover, 398 pages
Publisher InTech
Published online 24, June, 2011
Published in print edition June, 2011

This book is intended to attract the attention of practitioners and researchers from industry and academia interested in challenging paradigms of multimedia video coding, with an emphasis on recent technical developments, cross-disciplinary tools and implementations. Given its instructional purpose, the book also overviews recently published video coding standards such as H.264/AVC and SVC from a simulational standpoint. Novel rate control schemes and cross-disciplinary tools for the optimization of diverse aspects related to video coding are also addressed in detail, along with implementation architectures specially tailored for video processing and encoding. The book concludes by exposing new advances in semantic video coding. In summary: this book serves as a technically sounding start point for early-stage researchers and developers willing to join leading-edge research on video coding, processing and multimedia transmission.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Yu-Tzu Lin and Chia-Hu Chang (2011). User-aware Video Coding Based on Semantic Video Understanding and Enhancing, Recent Advances on Video Coding, Dr. Javier Del Ser Lorente (Ed.), ISBN: 978-953-307-181-7, InTech, Available from: http://www.intechopen.com/books/recent-advances-on-video-coding/user-aware-video-coding-based-on-semantic-video-understanding-and-enhancing

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821