A Video Saliency Detection Model in Compressed Domain

Yuming Fang, Weisi Lin, *Senior Member, IEEE,* Zhenzhong Chen, *Member, IEEE,*
Chia-Ming Tsai, and Chia-Wen Lin, *Senior Member, IEEE*

***Abstract*—Saliency detection is widely used to extract regions of interest in images for various image processing applications. Recently, many saliency detection models have been proposed for video in uncompressed (pixel) domain. However, video over Internet is always stored in compressed domains, such as MPEG2, H.264, and MPEG4 Visual. In this paper, we propose a novel video saliency detection model based on feature contrast in compressed domain. Four types of features including luminance, color, texture, and motion are extracted from the discrete cosine transform coefficients and motion vectors in video bitstream. The static saliency map of unpredicted frames (I frames) is calculated on the basis of luminance, color, and texture features, while the motion saliency map of predicted frames (P and B frames) is computed by motion feature. A new fusion method is designed to combine the static saliency and motion saliency maps to get the final saliency map for each video frame. Due to the directly derived features in compressed domain, the proposed model can predict the salient regions efficiently for video frames. Experimental results on a public database show superior performance of the proposed video saliency detection model in compressed domain.**

***Index Terms*—Compressed domain, video saliency detection, visual attention.**

I. Introduction

SALIENCY detection models are widely used to extract regions of interest (ROIs) in images/frames for various image/video processing applications such as coding, classification, watermarking, transcoding, resizing. During the past ten years, various saliency detection models have been proposed for salient region detection for images [3], [4], [6], [7]. Compared with image saliency detection, video saliency detection algorithms have to calculate the motion saliency map since motion is an important factor to attract human beings' attention. Currently, several studies have tried to detect salient regions in video [5], [8], [9].

Existing saliency detection models mentioned above are implemented in uncompressed (pixel) domain. However, most video over Internet are typically stored in the compressed domain such as MPEG2, H.264, MPEG4 Visual. The compressed videos are widely used in various Internet-based multimedia applications since they can reduce the storage space and greatly increase the delivering speed for Internet users. Current video saliency detection models have to decompress the compressed video into the spatial domain for feature extraction. The full decompression process for video is not only time consuming but computation consuming as well. Therefore, the video saliency detection algorithm in compressed domain is much desired for various Internet-based multimedia applications.

Recently, the authors proposed a saliency detection model in the compressed domain for image retargeting [14], [38]. The saliency map is calculated based on the features extracted from the discrete cosine transform (DCT) coefficients of JPEG images. However, this model is designed for JPEG images and does not include the motion feature extraction. In addition, the encoding standards for video and images are greatly different. Therefore, the model in [38] cannot be used for video saliency detection. To the best of our knowledge, there is still no saliency detection model in the compressed domain for video. As an extension to our work in [38], in this paper, we propose a novel video saliency detection model in the compressed domain.

In this paper, the MPEG4 visual [advanced simple profile (ASP)] coded video are used for saliency detection experiments. Videos of other types, such as MPEG2, can be processed in a similar way. The YCrCb color space is used in the MPEG4 ASP video [10], where Y represents the luminance component, and the other two (Cr and Cb) are used to represent the chroma components. From the MPEG4 ASP standard, video frames are divided into 16×16 macroblocks for luminance component and 8×8 macroblocks for chroma components in 4:2:0 chroma subsampling [10]. Each 4:2:0 coded unit block consists of six 8×8 blocks: four 8×8 luminance blocks (luminance channel) and two 8×8 chrominance blocks (one is the Cr channel block while the other is the Cb channel block). The data in video bitstream are always processed by 8×8 blocks and, thus,

Manuscript received May 1, 2012; revised October 5, 2012 and March 8, 2013; accepted May 22, 2013. Date of publication July 16, 2013; date of current version January 3, 2014. This paper was recommended by Associate Editor T. Zhang.

Y. Fang is with the School of Information Technology, Jiangxi University of Finance and Economics, Nanchang, Jiangxi 330032, China (e-mail: fa0001ng@e.ntu.edu.sg).

W. Lin is with the School of Computer Engineering, Nanyang Technological University, Singapore 639798 (e-mail: wslin@ntu.edu.sg).

Z. Chen is with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, Hubei, China, 430079 (e-mail: zzchen@ieee.org).

C.-M. Tsai is with the Department of Computer Science and Information Engineering, National Chung Cheng University, Chiayi 30013, Taiwan (e-mail: tsaicm@cs.ccu.edu.tw).

C.-W. Lin is with the Department of Electrical Engineering, National Tsing Hua University, Hsinchu 30013, Taiwan (e-mail: cwlin@ee.nthu.edu.tw).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2013.2273613

we perform here the saliency detection in 8×8 block level. In this paper, the features of luminance, color, and texture are extracted from the DCT coefficients of the unpredicted frames (I frames), and these features are adopted for the static saliency map calculation for these unpredicted frames. Meanwhile, the motion feature is extracted from the motion vectors of the predicted frames (P and B frames), and this motion feature is adopted for the motion saliency map calculation of these predicted frames. A new fusion method is proposed to combine the static saliency and motion saliency maps to get the final saliency map for video frames. Experimental results show that the proposed saliency detection model in compressed domain outperforms other existing ones on a public database.

II. Related Work

Visual attention mechanism is an important characteristic in the human visual system (HVS). The selective attention may be stimulus-driven or goal-driven corresponding to bottom-up and top-down approaches in perception process [1], [27], [28]. Existing studies have explored visual attention mechanism from the various aspects such as psychology, biology, computer vision [1]–[3], [27]–[30]. In the 1980s, Treisman *et al.* [1] proposed the famous Feature Integration Theory (FIT). According to this theory, the early selective attention mechanism leads some image regions to be salient for their different features (including color, intensity, orientation, motion, and so on) from their surroundings [1], [32]. Meanwhile, Koch *et al.* proposed a neurophysiological model of visual attention in [31].

Recently, researchers in the area of computer vision have started to build computational models of visual attention for the emerging interest in the HVS. Itti *et al.* [3] proposed a saliency detection model based on the neuronal architecture of the primates' early visual system. The saliency map is obtained through the calculation of multiscale center-surround differences by using three features including intensity, color, and orientation. Harel *et al.* [15] proposed a graph-based saliency detection model based on the study in [3]. In this model, the saliency map is calculated based on two steps: forming activation maps on several features and the normalization of these feature maps [15]. Hou *et al.* [4] defined the concept of spectral residual to design a visual attention model. The spectral residual is computed based on the Fourier transform. Achanta *et al.* [6] tried to obtain more frequency information to get a better saliency measure. The difference of Gaussian (DoG) is used to extract the frequency information in that model [6]. Goferman *et al.* [7] designed a context-aware saliency detection model by including more context information in the final saliency map. The center-surround differences of patches are used for saliency detection.

Besides the saliency detection models for images, some video saliency detection models have also been proposed in this area. Guo *et al.* proposed a phase-based saliency detection model for video in [5]. This model obtains the saliency map through inverse Fourier transform on a constant amplitude and the original phase spectrum of input video frames based on the following features: intensity, color, and motion. Itti *et al.* [8] developed a model to detect the low-level surprising events in video; the surprising events are defined as the important information attracting human beings' attention in video. Zhai *et al.* [9] built a video saliency detection model by combining the spatial and temporal saliency maps. The color histograms of images are used for the spatial saliency detection, while the planar motion between images (estimated by applying RANSAC on point correspondences in the scene) is adopted for the temporal saliency detection [9]. In [16], the authors designed a dynamic visual attention model based on the rarity of features. The incremental coding length (ICL) is defined to measure the entropy gain of each feature for saliency calculation.

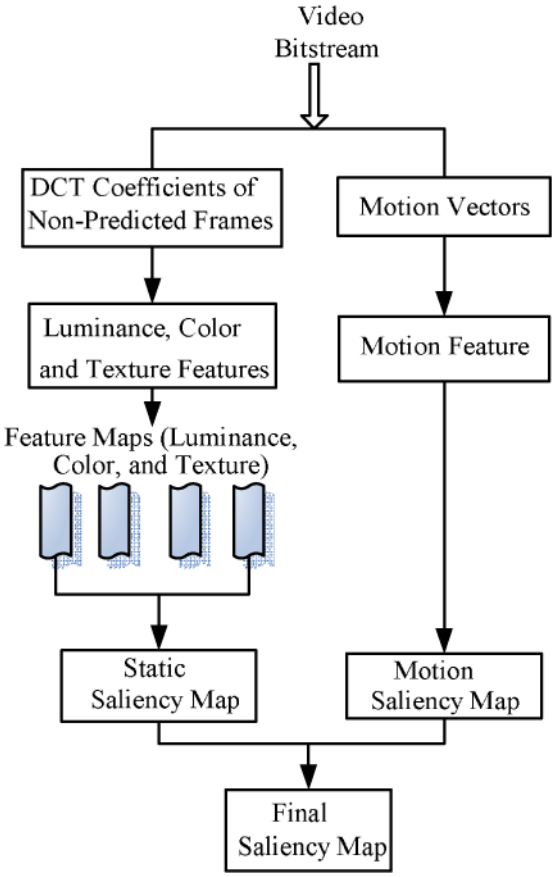

Fig. 1. Proposed framework.

All these saliency detection models mentioned above are implemented in uncompressed domain. As to these saliency detection models, coded videos have to be decompressed into spatial domain to extract features for saliency detection. In this paper, we propose a video saliency detection model in compressed domain. As mentioned in the first section, we calculate the saliency map of each video frame in 8×8 block level. The DCT coefficients of 8×8 image blocks in unpredicted frames (I frames) and motion vectors of 8×8 image blocks in predicted frames (P and B frames) are extracted from the video bitstream. The luminance, color, texture, and motion features of video frames are calculated from the DCT coefficients and motion vectors. Then, we calculate the static saliency map of unpredicted frames based on luminance, color, and

texture features. The motion saliency map of predicted frames are computed based on motion feature extracted from the motion vectors. Furthermore, we design a new fusion method to combine the static saliency and motion saliency maps to obtain the final saliency map for each video frame. Due to the directly derived features from the compressed domain, the proposed saliency detection model obtains promising results, as shown in the experimental section.

III. Proposed Framework

In this section, we describe the proposed model in detail. The proposed framework is depicted in Fig. 1. Firstly, three features including luminance, color, and texture are extracted from the video bitstream for unpredicted frames (I frames), and the motion feature is extracted from the motion vectors in video bitstream for predicted frames (P and B frames). Then, the static saliency map is obtained based on the features of luminance, color, and texture for unpredicted frames, while the motion saliency map is calculated on the basis of motion feature for predicted frames. Finally, the static saliency map and the motion saliency map are combined to get the final saliency map for each video frame. Here, we use MPEG4 ASP video to extract features in compressed domain. The features of other types of video such as MPEG2 video can be extracted in a similar way.

A. Feature Extraction From Video Bitstream

The proposed model uses DCT coefficients of unpredicted frames (I frames) to get luminance, color, and texture features, while motion vectors of predicted frames (P and B frames) are extracted to get motion feature for the motion saliency detection. Here, we do not use DCT coefficients of the predicted frames since these DCT coefficients represent the interpredicted block residue information. These residue information cannot be used to obtain the static saliency map by the proposed method. Meanwhile, there are no motion vectors for unpredicted frames in video bitstream. Therefore, in this paper, the DCT coefficients of unpredicted frames are used to calculate the static saliency map for these unpredicted frames, while the motion vectors of predicted frames are used to compute the motion saliency map for the predicted frames.

1) *DCT Coefficient and Motion Vector Extraction From Video Bitstream:* A natural video object is composed of a sequence (at different time points) of 2-D representations, which are referred to as video object planes (VOPs) [10], [17], [18]. The VOPs are coded using macroblocks by exploiting both temporal redundancies and spatial redundancies. Usually, a VOP consists of one or several video packets (slices) and each video packet is composed of an integer number of consecutive macroblocks. In each macroblock, the motion vectors and DCT coefficients are coded. The coded motion vector data are motion vector differences (predictively coded with respect to the neighboring motion vectors) after the variable length coding (VLC). The coded DCT coefficients are the 64 DCT coefficients encoded by zig-zag scanning and run-length-encoded, and the VLC.

The differential motion vector can be extracted from the coded motion vector data based on the VLC table. Then, it is added to a motion vector predictor component to form the real motion vector MV for predicted frames [10]. In a similar way, VLC tables of DCT coefficients are used to decode the coded DCT coefficients. The fixed length decoding is used to obtain the real DCT coefficients for video frames [10].

2) *Feature Calculation Based on DCT Coefficients:* In MPEG4 ASP video, DCT coefficients in one 8×8 block are composed of one DC coefficient and 63 AC coefficients. In each block, the DC coefficient is a measure of the average energy for the 8×8 block, while other 63 AC coefficients represent detailed frequency properties of this block. As mentioned above, YCrCb color space is used in MEPG4 video bitstream. In the YCrCb color space, the Y channel represents the luminance component, while Cr and Cb represent the chroma components. Thus, the DC coefficients in DCT blocks from the Y, Cr, and Cb channels are used to represent one luminance feature and two color features for 8×8 blocks as follows:

$$L = DC_Y \tag{1}$$

$$C_1 = DC_{C_r} \tag{2}$$

$$C_2 = DC_{C_b} \tag{3}$$

where L, C_1, and C_2 represent one luminance and two color features in each 8×8 DCT block, respectively, and DC_Y, DC_{C_r}, and DC_{C_b} are DC coefficients from the Y, Cr, and Cb components in each DCT block, respectively. It is noted that four 8×8 luminance blocks share two 8×8 chrominance blocks in the 4:2:0 chrominance format.

As mentioned above, AC coefficients include the detailed frequency information and existing studies have shown that AC coefficients can represent texture information for image blocks [11], [12]. In YCrCb color space, Cr and Cb components mainly include color information and little texture information is included in these two channels. Thus, only the AC coefficients in Y component are used to represent the texture information for images. In one DCT block, most of the energy is included in the first several low-frequency coefficients, which are in the left-upper corner of the block. The AC coefficients in the right-bottom corner of DCT blocks are equal to or close to zero and they are neglected during the quantization in coding process. In the progressive coding, the AC coefficients in one DCT block are ordered by zig-zag scanning. As the high-frequency AC coefficients include little energy for each DCT block, we use the first several AC coefficients to represent the texture feature of the DCT block. The existing study in [12] has shown that the first nine AC coefficients can represent most energy in each DCT block. Therefore, here, we use the first nine AC coefficients in each DCT block to represent the texture feature as follows:

$$T = \{AC_{01}, AC_{10}, AC_{20}, AC_{11}, ..., AC_{30}\}. \tag{4}$$

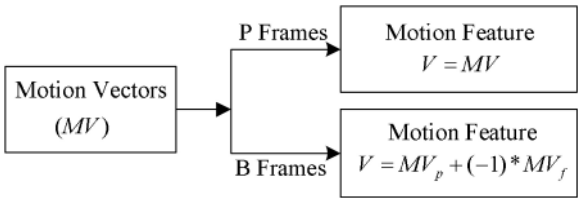

Fig. 2. Motion feature extraction.

3) *Motion Feature Calculation Based on Motion Vectors:* In this paper, the extracted motion vectors from the video bitstream are used to calculate the motion feature for predicted frames. In MPEG4 ASP video, there are two types of predicted frames: P frames use motion compensated prediction from a past reference frame, while B frames are bidirectionally predictive-coded by using motion compensated prediction from a past and/or a future reference frame. As there is just one prediction direction (predicted from a past reference frame) for P frames, the original motion vector MV are used to represent the motion feature for P frames. As B frames might include two types of motion compensated prediction (the backward and forward prediction), we calculate the motion vectors for B frames as follows: assume the motion compensated prediction values from the past reference and the future reference frames are MV_p and MV_f, respectively, and the motion feature of B frames is obtained as follows:

$$V = MV_p + (-1) * MV_f. \tag{5}$$

The process of the motion feature extraction is demonstrated as Fig. 2. The motion feature of each DCT block in B frames is obtained from (5), while the original motion vector is used to represent the motion feature for each DCT block in P frames. Now, we can get luminance, color, and texture features for nonpredicted frames as L, C_1, C_2, and T. The motion feature of predicted frames can be obtained as V. We will describe how to use these features in compressed domain to calculate the saliency map for video frames.

B. Saliency Detection in Compressed Domain

Based on the above description, the luminance, color, and texture features (L, C_1, C_2, and T) for unpredicted frames can be extracted from the DCT coefficients. The motion feature (V) of predicted frames can be obtained from the motion vectors. In this paper, we use these features to calculate the saliency map for video frames in compressed domain.

1) *Static and Motion Saliency Map Calculation:* Existing studies have shown that observers will be attracted by the regions with different features from its surrounding when looking at a natural scene [1], [2]. The features which can be used to discriminate the image regions include intensity, color, motion, and so on. Based on the FIT [1], the center-surround differences of 8×8 DCT blocks are used to detect salient regions for images. The features of luminance, color, texture, and motion extracted from the DCT coefficients and motion vectors are used to calculate the DCT block differences for saliency detection in this paper.

It is commonly accepted that the HVS is highly space variant since there are different densities of cone photoreceptor cells in the retina of human eyes [13]. On the retina, the fovea has the highest density of cone photoreceptor cells, and thus the focus area is perceived at the highest resolution. The visual acuity decreases with the increasing eccentricity from the fixation areas [13]. This means that the HVS is more sensitive to the center-surround differences from the blocks with nearer distance compared with those from the farther blocks. Here, we use a Gaussian model to simulate this mechanism for weighting the center-surround differences among image blocks for saliency detection. The feature map of each video frame is calculated as follows:

$$S_i^k = \sum_{j \neq i} \alpha_{ij} D_{ij}^k \tag{6}$$

$$\alpha_{ij} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{d_{ij}^2}{2\sigma^2}} \tag{7}$$

where S_i^k indicates saliency value of the ith DCT block in the feature map with feature k; $k \in \{L, C_1, C_2, T, V\}$; σ is a parameter of the Gaussian distribution, d_{ij} is the Euclidean distance between DCT blocks i and j, D_{ij}^k is the feature differences between DCT blocks i and j with feature k.

As depicted in Fig. 1, the features of luminance, color, and texture are used to calculate the static saliency map for unpredicted frames. The luminance and color features only include one DC coefficient value from the luminance and color channels; thus, the feature differences of luminance and color among DCT blocks are represented as DC coefficient differences from the luminance and color channels. Since the texture feature is represented as a vector including nine AC coefficients, the Euclidean distance between the vectors is used to compute the texture difference between DCT blocks. The static saliency map S_s for unpredicted frames is calculated as linear combination of four feature maps from the luminance, color, and texture features (L, C_1, C_2, T) as follows:

$$S_s = \sum \beta_\theta N_\theta \tag{8}$$

where N is normalization operation; $\theta \in \{S^k\}$; β_θ is the parameter determining the weight for each feature map. In this paper, we set $\beta_\theta = 1/4$.

The motion feature V obtained from (5) is used to calculate the motion feature differences between DCT blocks. Here, the motion feature differences are computed as the Euclidean distance between motion features of DCT blocks. Therefore, the motion saliency map S_m can be obtained based on (6)–(7) for predicted frames. The final saliency map of video frames is calculated by combining the static saliency map S_s and motion saliency map S_m. We will describe how to compute the final saliency map as follows.

2) *Final Saliency Map Calculation:* Based on the above description, we can obtain the static saliency map for unpredicted video frames (I frames) and the motion saliency map for predicted video frames (P and B frames). The motion saliency map of unpredicted frames cannot be calculated since there are no motion vectors for unpredicted frames in video bitstream. However, there may be motion in unpredicted

frames (generally, there may be motion in all frames except the first I frame in video). Meanwhile, the static saliency map of predicted frames cannot be computed since the DCT coefficients of these predicted frames represent the DCT block residue information and cannot be used to calculate the static saliency map. Here, we use the static/motion saliency map of the previous unpredicted/predicted frames to replace that of the current predicted/unpredicted ones based on the implicit memory theory [24], [25]. Existing studies have shown that the focal attention and eye movements are guided by the recently attended locations and the implicit memory traces of context cueing [24]–[26]. These studies demonstrate that the previous attended targets will trigger the attention traces utilized in the following several fixations. Thus, human beings will continue focusing on the similar locations in future frames with these from the previous frames without much context change. Generally, the content in the consecutive video frames will not change greatly, and thus the saliency maps (static or motion) of the consecutive video frames are very similar in video. Therefore, we can use the static or motion saliency map of the previous video frames to represent that of the current ones.

As there is no motion saliency map for unpredicted frames, the motion saliency map of the previous predicted frame is adopted to represent that of the current unpredicted frame. Thus, the final saliency map for unpredicted frames (I frames) is calculated as follows:

$$S = f(S_s, S_{m_p}) \tag{9}$$

where S is the final saliency map of the current unpredicted frame, S_s is the static saliency map of the current unpredicted frame, S_{m_p} is the motion saliency map of the previous predicted frame, $f(S_1, S_2)$ is the fusion function to get the final saliency map from the saliency maps of S_1 and S_2.

Similarly, the static saliency map of the previous unpredicted frame is used to represent that of the current predicted frame, and thus the final saliency map of the predicted frames (P and B frames) is computed as follows:

$$S = f(S_{s_p}, S_m) \tag{10}$$

where S is the final saliency map of the current predicted frame, S_{s_p} is the static saliency map of the previous unpredicted frame, S_m is the motion saliency map of the current predicted frame, $f(S_1, S_2)$ is the fusion function to get the final saliency map from the saliency maps of S_1 and S_2.

According to (9) and (10), we can calculate the final saliency map for video frames based on the static saliency and motion saliency maps. We will describe the fusion method f [in (9) and (10)] for the static and motion saliency maps in the next section.

3) *Saliency Map Fusion:* Currently, there are many fusion methods for combining the static saliency and motion saliency maps into the final saliency map [33]. In this paper, we have tried several common fusion methods in [33] for combing the static saliency and motion saliency maps as follows.

1) Normalized and sum (NS): the most simple fusion method that normalizes the static saliency and motion saliency maps to the same dynamic range, and then sums these two maps to obtain the final saliency map as follows [33]:

$$S = \sum_i N(S_i) \tag{11}$$

where S is the final saliency map, N is the normalization operator, S_i is the static or motion saliency map, and $i \in \{1, 2\}$.

2) Normalization and maximum (NM): the fusion method that normalizes the static saliency map and motion saliency map to the same dynamic range, and then uses the maximum value as the final saliency value at each location

$$S = \max_i N(S_i) \tag{12}$$

where max is the maximum operator.

3) Normalization and product (NP): the fusion method that normalizes the static saliency map and motion saliency map to the same dynamic range, and then products the static saliency and motion saliency maps for the final saliency map

$$S = \prod_i N(S_i). \tag{13}$$

These three methods are the common fusion methods of the research area. However, there is no spatial competition between the static saliency map and the motion saliency map with the above fusion methods. In these fusion methods, the static saliency and motion saliency maps are considered with the same weighting whatever the differences between these two maps. To address the drawbacks with these fusion methods, we propose a new fusion method of parameterized normalization, sum and product (PNSP) based on the characteristics of the static saliency map and the motion saliency map. The final saliency map from the proposed fusion method PNSP for video frames is calculated as follows:

$$S = \gamma_1 S_s + \gamma_2 S_m + \gamma_3 S_s S_m \tag{14}$$

where S is the final saliency map for video frames, S_s is the static saliency map, S_m is the motion saliency map, γ_1, γ_2, and γ_3 are the parameters to determine the weighting of each component.

The weighting parameters in (14) are determined by the characteristic of the static saliency and motion saliency maps. A good saliency map should include the small compact salient regions rather than the spread salient points. If the salient regions in the motion saliency map are small and compact, the motion contrast of this frame can be considered strong. In this case, we can use a large weighting parameter γ_2 to weight the motion saliency map. Thus, the motion saliency map contributes much to the final saliency map. Similarly, a large weighting parameter γ_1 should be used to weight the static saliency map if the feature contrast is also strong in the static saliency map. As we can see, the weighting parameter γ_3 is used to measure the importance of these regions which both static saliency and motion saliency maps detect as salient. Here, we set $\gamma_3 = (\gamma_1 + \gamma_2)/2$.

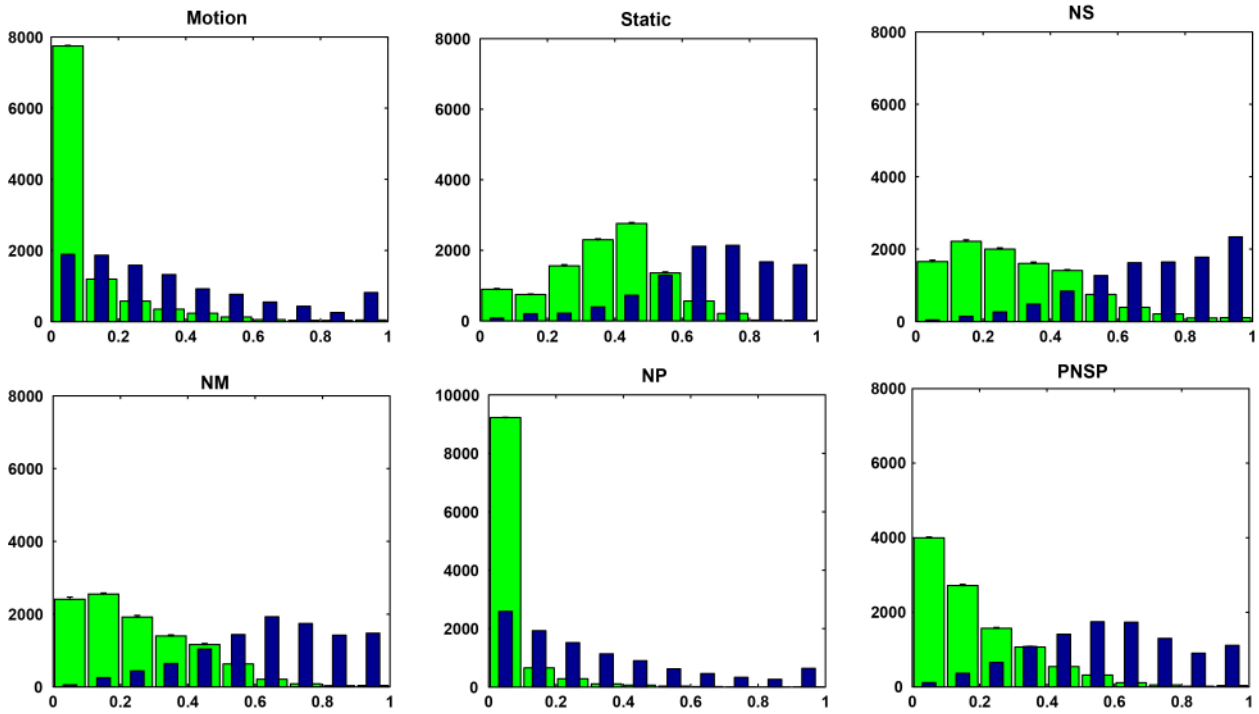

Fig. 3. Saliency distributions at human saccade locations (narrow blue bars) and random locations (wide green bars) from different algorithms. The x- and y-axis represent the predicted saliency values from different models, and the number of locations with the corresponding salient values, respectively.

The spatial variance is adopted to measure the feature contrast in the static saliency map or motion saliency map. Given a static saliency map or motion saliency map S_k ($S_k \in \{S_s, S_m\}$), the Otsu's thresholding method [34] is used to binarize it and the spatial variance can be calculated as follows:

$$v_k = \frac{\sum_{(i,j)} \sqrt{(i - S_{i,e})^2 + (j - S_{j,e})^2} * S_k(i, j)}{\sum_{(i,j) S_k(i,j)}} \tag{15}$$

$$S_{i,e} = \frac{\sum_{(i,j)} i * S_k(i, j)}{\sum_{(i,j)} S_k(i, j)} \tag{16}$$

$$S_{j,e} = \frac{\sum_{(i,j)} j * S_k(i, j)}{\sum_{(i,j)} S_k(i, j)} \tag{17}$$

where v_k is the spatial variance of the saliency map S_k, $S_k(i, j)$ is the saliency value at the location (i, j) in the saliency map S_k, $S_{i,e}$ and $S_{j,e}$ represent the spatial expectation values of the salient regions in the horizontal and vertical directions, respectively.

The spatial variance of the static saliency map or motion saliency map can be calculated from (15) to (17). The spatial variance values of the saliency maps containing small and compact salient regions are smaller than those of the saliency maps containing spread salient points. Therefore, we set $\gamma_k = 1/v_k$ (we use the normalized γ_k as the weighting parameter determining the weighting of the static saliency map or motion saliency map ($\gamma_k \in \{\gamma_1, \gamma_2\}$)). Thus, the final saliency map for each video frame can be calculated based on (14)–(17). We will present the performance of this fusion method compared with the existing ones (11)–(13) in the experiment section below.

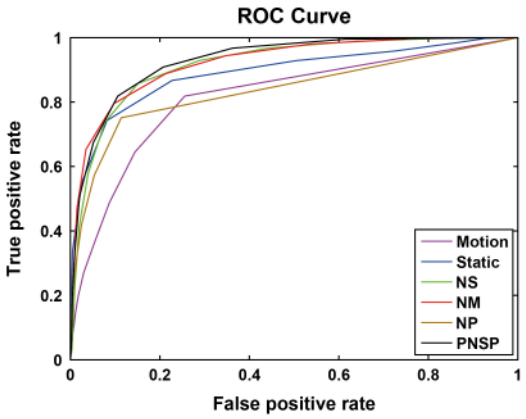

Fig. 4. ROC curves for the static saliency map, the motion saliency map, and the final saliency maps from different fusion methods.

IV. Experiments

In this section, we evaluate the performance of the proposed model based on a public video database [8]. This database includes 50 videoclips totaling over 25 min and their saccade data from the eight observers. It includes various types of video such as outdoor video in daytime and nighttime, sports video, news television broadcast, video games, etc. The human saccade data was recorded by a 240-Hz infrared-video-based eye tracker. In this paper, we have performed two experiments to evaluate the performance of the proposed model: the first one is conducted to demonstrate why we have to combine the static saliency and motion saliency maps to obtain the final saliency map for each video frame; the other one is performed

TABLE I
AUC RESULTS FROM THE STATIC SALIENCY MAP

Models	Motion	Static	NS	NM	NP	PNSP
KL Distance	0.846	1.529	1.596	1.761	1.174	1.828
AUC	0.817	0.89	0.918	0.922	0.837	0.93

The motion saliency map and the final saliency maps from different fusion methods.

to compare the performance of the proposed video saliency detection model with other existing ones.

A. Evaluation Method

We use the similar measurement method as the study [8], [19] to evaluate the performance of the proposed model. The performance of the video saliency detection models is measured by comparing the response values at saccade locations and random locations in the saliency map. Here, we calculate the salient value at a human saccade location as the maximum value over a circular aperture around the saccade location in the saliency map. A high salient value at the human saccade location means that the saliency detection model can predict the salient locations exactly. Similarly, the salient value at a randomly chosen location is calculated as the maximum value over a circular aperture around the randomly chosen location.

Generally, an efficient video saliency detection model would have high response values at saccade locations and have no response at most randomly chosen locations. Here, we first calculate the saliency distributions at human saccade locations and random locations with ten bins for saliency values over the saliency map. Figs. 3 and 6 show the saliency distributions at human saccade locations and random locations in the saliency map from different algorithms. The x-axis represents the saliency value bins from 0 to 1 with 0.1 interval, while y-axis represents the number of human saccade locations or random locations with different saliency value bins. To evaluate the performance of saliency detection models, Kullback–Leibler (KL) distance [20], [21] is used to measure the similarity between these two distributions as follows [19]:

$$KL(H, R) = \frac{1}{2}\left(\sum_n h_n log\frac{h_n}{r_n} + \sum_n r_n log\frac{r_n}{h_n}\right) \tag{18}$$

where H and R are saliency distributions at human saccade locations and random locations with probability density functions h_n and r_n, respectively, and n is the saliency value bins ($n \in \{1, 2, 3, ..., 10\}$).

As mentioned above, a good video saliency detection model can get high salient values at human saccade locations and low salient values (even zero) at most randomly chosen locations in the saliency map. In this case, the saliency distributions at human saccade locations and random locations are greatly different, and thus the KL distance between these saliency distributions is large. Therefore, the video saliency detection model with a higher KL distance can more easily discriminate human saccade locations from the random locations in video frames, and thus a better performance in saliency detection for video [8].

In addition, we use receiver operating characteristic (ROC) curve [22], [23] to evaluate the performance of the proposed model. The ROC curve is a graphical plot of the true positive rate (TPR) versus the false positive rate (FPR) for a binary classifier system with varied discrimination thresholds [22], [23], as shown in Figs. 4 and 7. Here, we first normalize the saliency values in the range [0, 1], and then use the thresholds from 0 to 1 with the interval 0.1 to get the ROC curves for saliency detection models. The saliency distribution at human saccade locations and random locations are used as the test set and discrimination set, respectively. For each threshold, the TPR is calculated as the percentage of the number of human saccade locations with salient values larger than this threshold over the total number of human saccade locations; the FPR is computed as the percentage of the number of random locations with salient values larger than this threshold over the total number of random locations. The overall quantitative evaluation with the ROC curve is the area under ROC curve (AUC). The larger the AUC is, the better the saliency predication for the saliency detection model is for video frames.

B. Experiment 1

In this subsection, we compare the performance of the static saliency map, the motion saliency map and the combined saliency maps to demonstrate the importance of the combination of these two saliency maps for the final saliency map calculation for video frames. In addition, we compare the performance of the final saliency map from the proposed fusion method with those from the other existing ones to demonstrate the advantages of the proposed fusion method.

Fig. 3 shows the saliency distributions of the human saccade locations and random locations for the static saliency map, the motion saliency map, and the final saliency maps from different fusion methods. From the first graph (motion) in Fig. 3, the amount of lower saliency values (0–0.2) at random locations is large. Thus, the saliency values at most random locations are low. Although the saliency values at most human saccade locations are higher than those at random locations, the amount of the low saliency values (0–0.5) at human saccade locations is still large. From the second graph (Static) in Fig. 3, the saliency values at most human saccade locations and random locations in the static saliency map are larger than the corresponding ones from the motion saliency map. In addition, compared with the motion saliency map, the saliency distribution at human saccade locations is more different from that at random saccade locations in the static saliency map, as shown in Fig. 4 and Table I. From Fig. 4 and Table I, we can see that the KL distance and AUC values from the static saliency map are larger than those from the motion saliency map. Thus, the performance of the static saliency map is better than that of the motion saliency map.

Figs. 3 and 4, and Table I also present the experimental results of the final saliency maps from different fusion methods for the static saliency and motion saliency maps. Based on the statistic results in Table I, the fusion methods of NS, NM, and PNSP can obviously obtain better performance than the static saliency and motion saliency maps, while the NP

Fig. 5. Comparison for different saliency maps: first column: the original images with human saccade locations marked with 64-radius circles, second column: the motion saliency map, third column: the static saliency map, final column: the final saliency map from the proposed fusion method.

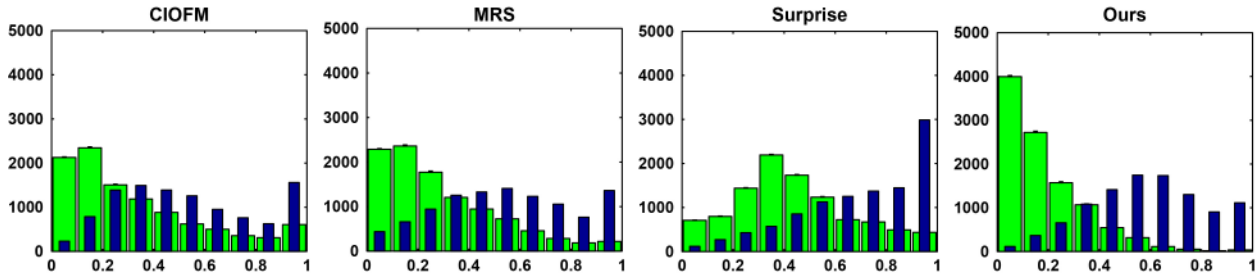

Fig. 6. Saliency distributions at human saccade locations (narrow blue bars) and random locations (wide green bars) from different models. x-axis and y-axis represent the predicted saliency values from different models, and the number of locations with the corresponding salient values, respectively.

fusion method cannot get good performance. The disadvantage of NP fusion method can be demonstrated by experimental results in Fig. 3. From this figure, we can see that the saliency distributions at human saccade locations and random saccade locations are more similar, compared with other fusion methods (NS, NM, and PNSP). From Table I, although the KL distance and AUC values from the fusion methods of NS, NM, and the proposed one (PNSP) are all larger than those from the static and motion saliency map, the KL distance and AUC values from the proposed fusion method (PNSP) are larger than those from the NS and NM fusion methods. This demonstrates the proposed fusion algorithm PNSP can obtain better performance than other ones (NS, NP, and NM).

In Table I, we can see that the KL distance and AUC values of the final saliency map from the proposed fusion method are much larger than the static saliency and motion saliency maps. The good performance of the proposed method demonstrates the importance of the fusion method in (14). Additionally, we present some comparison samples for the static saliency map, the motion saliency map and the final saliency map from the proposed fusion method in Fig. 5. In Fig. 5, the human saccade locations in original images and saliency maps are marked with circles. The saliency values at human saccade locations are calculated as the maximum value within a circle in the saliency map [8], [19]. From Fig. 5, we can see that the saliency value at the human saccade location is largest in each final saliency map from the proposed fusion method. On the contrary, in the motion saliency map and static saliency

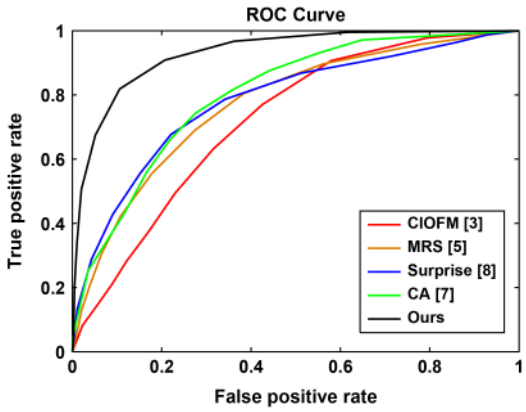

Fig. 7. ROC curves for different video saliency detection models in comparison.

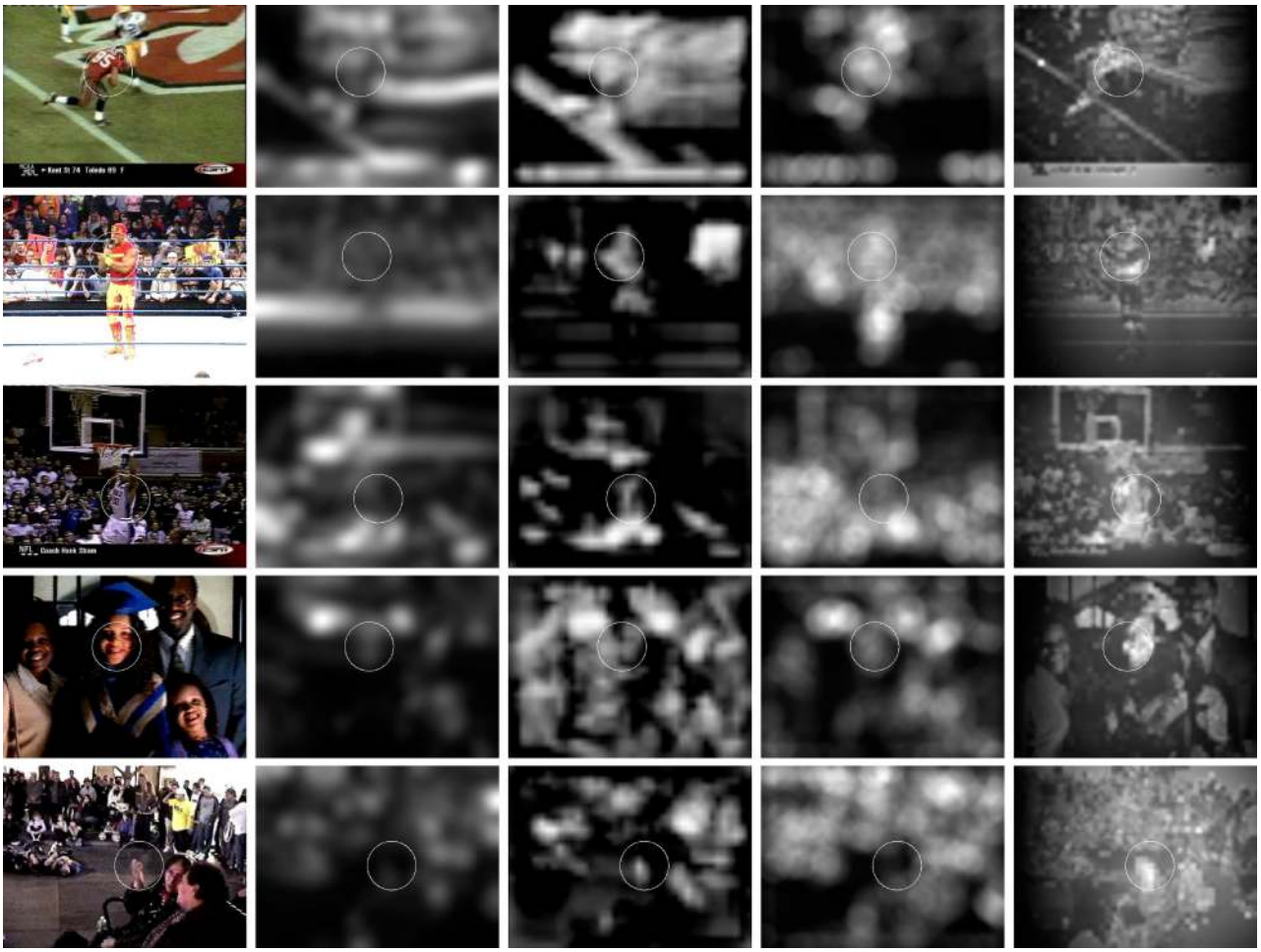

Fig. 8. Comparison of saliency maps from different models. First column: original images with fixation data. Second to the last column: saliency maps from CIOFM [3], Surprise [8], MRS [5], and the proposed model.

TABLE II
COMPARISON RESULTS FROM DIFFERENT VIDEO SALIENCY DETECTION MODELS

Models	CIOFM [3]	MRS [5]	Surprise [8]	CA [7]	Ours
KL Dist.	0.402	0.529	0.593	0.76	1.828
AUC	0.720	0.771	0.782	0.802	0.93

map, the saliency value at human saccade locations may not be the largest value or there may be the largest values at other locations except the human saccade locations.

C. Experiment 2

In this paper, three existing video saliency detection models are used to do the comparison: CIOFM [3], Surprise [8], and MRS [5].

The experimental results from different existing models are shown in Figs. 6–8 and Table II. In Fig. 7 and Table II, we present the video saliency results from an image saliency detection model [7] since there are various image saliency detection models proposed in this area and the model [7] is proposed recently. However, we do not show the results with [7] in Figs. 6 and 8 for the consideration of the paper length. Fig. 6 shows the saliency distributions at the human saccade locations and random locations from different saliency detection models. Fig. 7 and Table II show the AUC and KL distance values for different saliency detection models. From Table II, we can see that the KL distance and AUC values of CA [7], MRS [5], and Surprise [8] are larger than those of CIOFM [3]. This demonstrates that the performance of these three models is better than the one [3]. From this table, the KL distance and AUC values of the proposed model is the largest, and thus the performance of the proposed model is the best among these compared models. Among these models, the CA [7] only considers the static saliency detection. The MRS [5] is a phase-based saliency detection model. In this model, the Fourier transform is first used to obtain the amplitude and phase. Then, the final saliency map for the image/frame is obtained by the inverse Fourier transform based on the user-defined constant amplitude and the original phase. It obtains the saliency map for images/frames mainly by considering the global contrast. CIOFM [3] and Surprise [8] mainly calculate the local contrast for computing the saliency map for images/frames. On the contrast, in the proposed model, the saliency value of each DCT block is calculated from the center-surround differences between this DCT block and all other DCT blocks in the image/frame. Thus, the proposed model considers both local and global contrast

TABLE III
COMPARISON RESULTS OF THE COMPUTATIONAL COMPLEXITY FROM DIFFERENT VIDEO SALIENCY DETECTION MODELS

Models	CIOFM [3] (Matlab)	MRS [5] (Matlab)	Surprise [8] (C++)	CA [7] (Matlab)	Ours (Matlab)
Time(s)	0.11	0.05	0.25	55.7	1.73

for saliency calculation and it achieves better performance than others.

Fig. 8 provides some comparison samples of saliency maps from different saliency detection models. In the first column of Fig. 8, the human saccade locations are indicated by circles. As mentioned previously, the saliency value of the saccade location is calculated by the maximum value within this circle in the saliency map from the saliency detection model. From this figure, we can see that the saliency values at human saccade locations from the model CIOFM [3] are very low, some of them even approaching zero (the second column in Fig. 8). From the third column of Fig. 8, we can see that the saliency values at human saccade locations are not the largest ones in the saliency map from the model Surprise [8]. The saliency values at many other locations are larger than those at the human saccade locations from the results of the model Surprise [8]. Similarly, the saliency values at human saccade locations from the model [5] are low, as shown from the fourth column in Fig. 8. On the contrary, the saliency values at fixation points from the proposed model are almost the largest values in the saliency map. This demonstrates that the saliency map from the proposed model can predict more accurate human saccade locations compared with other existing ones.

Furthermore, we have done the experiments for video saliency detection with different compression ratios from different quantization parameter (QP) values and lengths of group of pictures (GOP). The comparison results are shown in Figs. 9 and 10. In these figures, GOP8, GOP16, and GOP32 represent saliency results from the video sequences with GOP values as 8, 16, and 32 (with QP value as 8); QP2, QP8, QP16, QP32, QP48, and QP64 represent the saliency results from the video sequences with QP values as 2, 8, 16, 32, 48, and 64, respectively (with GOP value as 12). From these figures, we can see that the saliency performance does not change greatly with different QP and GOP values. Thus, the video compression will not greatly change the video saliency results. A recent research study has also shown that video coding impairments would not disturb the visual attention regions from the observers [35]. Therefore, the normal video compression will not change the saliency results from the proposed model and observers.

For the computational complexity, we have conducted the experiment on a subset of the database among these comparison models. With the same computer environment (CPU: Intel(R) Xeon 3.2 GHz, memory: 6.00 GB), the average computational time for each video frame is shown in Table III. From Table III, we can see that the computational time of CA [7] is the highest among these models, while the time cost of MRS [5] is the lowest. MRS [5] calculates the saliency map mainly when FT and FT operation is fast. Thus, its computational complexity is low. Compared with other models, the computational time of the proposed model is modest. Of course, we can still further improve the computational complexity of the proposed model by optimizing the code (such as implementation by C/C++). In the proposed method, we only need to extract motion vectors for predicted frames (P and B frames) and DCT coefficients for unpredicted frames (I frames), respectively, in entropy decoding as features for saliency analysis, instead of completely decoding the whole frame. As reported in [36] and [37], the time cost of the operations after entropy decoding for MPEG4 decoder is up to 90% of the decoding time for a GOP. Therefore, by skipping unnecessary operations, the overall decoding time cost for MPEG4 applications we could save is up to 90%. Additionally, the performance of the proposed model is much better than other existing models in uncompressed domain, as shown in the experiments.

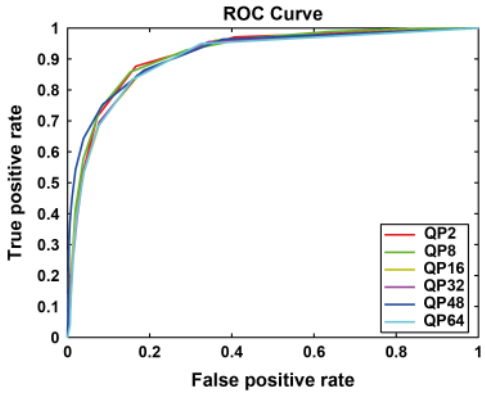

Fig. 9. ROC curves from different QP values in comparison.

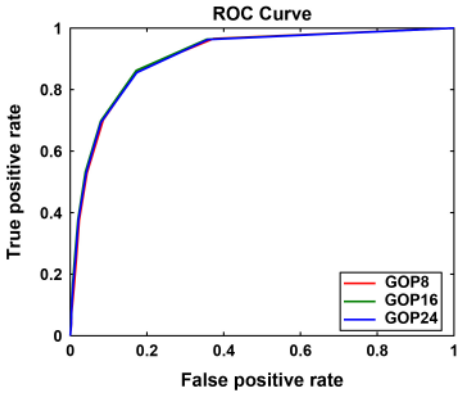

Fig. 10. ROC curves from different GOP values in comparison.

V. CONCLUSION

In this paper, we have proposed a video saliency detection model in compressed domain based on four types of features: luminance, color, texture, and motion. These four types of features are extracted from the DCT coefficients and motion vectors in video bitstream. The static saliency map is calculated from the features of luminance, color, and

texture, while the motion saliency map is computed from the motion feature. A new fusion method has been designed to combine the static saliency and motion saliency maps to get the final saliency map for video frames. Experimental results based on a public database show that the proposed video saliency detection model outperforms the relevant existing ones.

It it noted that existing video saliency detection models are implemented in uncompressed domain. Compared with the video saliency detection in uncompressed domain, the proposed video saliency detection model in compressed domain can be used more conveniently in the Internet-based multimedia applications such as video retargeting, video quality assessment. Therefore, the proposed saliency detection model in compressed domain is significant in this research area. As the next step of the paper, we will explore various multimedia applications of the proposed video saliency detection model in compressed domain.

REFERENCES

[1] A. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[2] J. Wolfe, "Guided search 2.0: A revised model of guided search," *Psychonomic Bull. Rev.*, vol. 1, no. 2, pp. 202–238, 1994.

[3] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[4] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE Int. Conf. Comput. Vision Pattern Recognit.*, Jun. 2007, pp. 1–8.

[5] C. Guo and L. Zhang, "A novel multi-resolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 185–198, Jan. 2010.

[6] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Int. Conf. Comput. Vision Pattern Recognit.*, Jun. 2009, pp. 1597–1604.

[7] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," in *Proc. IEEE Int. Conf. Comput. Vision Pattern Recognit.*, Jun. 2010, pp. 2376–2383.

[8] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," *Adv. Neural Inform. Process. Syst.*, vol. 46, nos. 8–9, pp. 1194–1209, Apr. 2006.

[9] Y. Zhai and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," in *Proc. ACM Int. Conf. Multimedia*, 2006, pp. 815–824.

[10] ISO/IEC JTC1, *Information technology—Coding of audio-visual objects Part2: Visual*, ISO/IEC 14492-2, (MPEG-4 Visual), Version 1: Apr. 1999, Version 2: Feb. 2000, Version 3: May 2004.

[11] Y. L. Huang and R. F. Chang, "Texture features for DCT-coded image retrieval and classification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 6. 1999, pp. 3013–3016.

[12] C. Theoharatos, V. K. Pothos, N. A. Laskaris, G. Economou, and S. Fotopoulos, "Multivariate image similarity in the compressed domain using statistical graph matching," *Pattern Recognit.*, vol. 39, pp. 1892–1904, Oct. 2006.

[13] B. A. Wandell, *Foundations of Vision*. Springfield, MA, USA: Sinauer Associates, 1995.

[14] Y. Fang, Z. Chen, W. Lin, and C. Lin, "Saliency-based image retargeting in the compressed domain," in *Proc. ACM MM*, 2011, pp. 1049–1052.

[15] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," *Adv. Neural Inform. Process. Syst.*, vol. 19, pp. 545–552, MIT Press, 2007.

[16] X. Hou and L. Zhang, "Dynamic visual attention: Searching for coding length increments," *Adv. Neural Inform. Process. Syst.*, vol. 21, pp. 681–688, MIT Press, 2008.

[17] T. Ebrahimi and C. Home, "MPEG-4 natural video coding: An overview," *Signal Process. Image Commun.*, vol. 15, no. 4, pp. 365–385, 2000.

[18] R. Talluri, "Error-resilient video coding in the ISO MPEG-4 standard," *IEEE Commun. Mag.*, vol. 36, no. 6, pp. 112–119, Jun. 1998.

[19] R. J. Peters and L. Itti, "Applying computational tools to predict gaze direction in interactive visual environments," *ACM Trans. Appl. Perception*, vol. 5, no. 2, pp. 1–19, 2008.

[20] S. Kullback, *Information Theory and Statistics.* New York, NY, USA: Wiley, 1959.

[21] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Stat.*, vol. 22, no. 1, pp. 79–86, 1951.

[22] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. San Diego, CA, USA: Academic Press, 1990.

[23] C. W. Therrien, *Decision Estimation and Classication: An Introduction to Pattern Recognition and Related Topics*. New York, NY, USA: Wiley, 1989.

[24] M. M. Chun and K. Nakayama, "On the functional role of implicit visual memory for the adaptive deployment of attention across scenes," *Visual Cognition*, vol. 7, nos. 1–3, pp. 65–81, 2000.

[25] M. F. Land and S. Furneaux, "The knowledge base of the oculomotor system," *Philosoph. Trans. Roy. Soc. London B, Biol. Sci.*, vol. 352, pp. 1231–1239, Aug. 1997.

[26] R. A. Rensink, "Seeing, sensing, and scrutinizing," *Vision Res.*, vol. 40, no. 10-12, pp. 1469–1487, 2000.

[27] J. M. Wolfe, S. J. Butcher, and M. Hyle, "Changing your mind: On the contributions of top-down and bottom-up guidance in visual search for feature singletons," *J. Experimental Psychol. Human Perception Performance*, vol. 29, no. 2, pp. 483–502, 2003.

[28] J. Braun and D. Sagi, "Vision outside the focus of attention," *Perception Psychophysics*, vol. 48, no. 1, pp. 45–48, 1990.

[29] Z. Lu, W. Lin, X. Yang, E. Ong, and S. Yao, "Modeling visual attention's modulatory aftereffects on visual sensitivity and quality evaluation," *IEEE Trans. Image Process.*, vol. 14, no. 11, pp. 1928–1942, Nov. 2005.

[30] H. Pashler, *The psychology of attention*. Cambridge, MA, USA: MIT Press, 1997.

[31] C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," *Human Neurobiology*, vol. 4, no. 4, pp. 219–227, 1985.

[32] A. Treisman and S. Gormican, "Feature analysis in early vision: Evidence from search asymmetries," *Psychology Rev.*, vol. 95, no. 1, pp. 15–48, 1988.

[33] C. Chamaret, J. C. Chevet, and O. Le Meur, "Spatio-temporal combination of saliency maps and eye-tracking assessment of different strategies," in *Proc. Int. Conf. Image Process.*, 2010, pp. 1377–1387.

[34] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. 9, no. 1, pp. 62–66, Jan. 1979.

[35] O. Le Meur, A. Ninassi, P. Le Callet, and D. Barba, "Do video coding impairments disturb the visual attention deployment?" *Signal Process. Image Commun.*, vol. 25, no. 8, pp. 597–609, 2010.

[36] F. M. Porikli, "Real-time video object segmentation for MPEG-encoded video sequences," in *Proc. SPIE Conf. Real-Time Imagining VIII*, vol. 5297. 2004, pp. 195–203.

[37] F. Cavalli, R. Cucchiara, M. Piccardi, and A. Prati, "Performance analysis of MPEG-4 decoder and encoder," in *Proc. Int. Symp. Video/Image Process. Multimedia Commun.*, 2002, pp. 227–231.

[38] Y. Fang, Z. Chen, W. Lin, and C.-W. Lin, "Saliency detection in the compressed domain for adaptive image retargeting," *IEEE Trans. Image Process.*, vol. 21, no. 9, pp. 3888–3901, Sep. 2012.

Yuming Fang received the B.E. degree from Sichuan University, Chengdu, China, the M.S. degree from the Beijing University of Technology, Beijing, China, and the Ph.D. degree from Nanyang Technological University, Singapore.

He is a faculty member with the School of Information Technology, Jiangxi University of Finance and Economics, Nanchang, China. From 2011 to 2012, he was a visiting scholar with National Tsinghua University, Hsinchu, Taiwan, and he was with University of Waterloo, Waterloo, ON, Canada in 2012. His research interests include visual attention modeling, visual quality assessment, image retargeting, computer vision, 3-D image/video processing.

Weisi Lin (M'92–SM'98) received the B.Sc. and M.Sc. degrees from Zhongshan University, Guangzhou, China, and the Ph.D. degree in computer vision from Kings College, London, U.K.

He was the Laboratory Head, Visual Processing, and the Acting Department Manager, Media Processing, with the Institute for Infocomm Research, Singapore. He is currently an Associate Professor at the School of Computer Engineering, Nanyang Technological University, Singapore. He has published over 240 referred papers in international journals and conferences. His research interests include image processing, perceptual modeling, video compression, and multimedia communication.

Dr. Lin is on the editorial boards of the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE SIGNAL PROCESSING LETTERS, and the *Journal of Visual Communication and Image Representation*. He is a Chartered Engineer (U.K.), a fellow of the Institution of Engineering Technology, and an Honorary fellow, Singapore Institute of Engineering Technologists. He has been the lead Technical Program Chair for the Pacific-Rim Conference on Multimedia in 2012 and the Technical Program Chair in the IEEE International Conference on Multimedia and Expo in 2013. He has been elected as a Distinguished Lecturer of the Asia Pacific Signal and Information Processing Association in 2012 to 2013.

Zhenzhong Chen (M'07) received the B.Eng. degree from Huazhong University of Science and Technology, Wuhan, China, and the Ph.D. degree from the Chinese University of Hong Kong, Shatin, China, both in electrical engineering.

He is currently a Professor at Wuhan University (WHU). Before joining WHU, he worked at Media Tek USA Inc. San Jose, CA, USA; Nanyang Technological University (NTU), Singapore; and National Institute for Research in Computer Science and Control (INRIA), France. His current research interests include visual perception, image processing, video processing, and multimedia communications.

Dr. Chen is a Voting member of the IEEE Multimedia Communications Technical Committee (MMTC), an Invited member of the IEEE MMTC Interest Group of Quality of Experience for Multimedia Communications (QoEIG), from 2010 to 2012. He has served as a Guest Editor of special issues for the IEEE MMTC E-LETTER and *Journal of Visual Communication and Image Representation*. He has co-organized several special sessions at international conferences, including the IEEE International Conference on Image Processing 2010, IEEE International Consulting Management and Engineering 2010, and Packet Video 2010, and has served as a technical program committee member of the IEEE ICC, GLOBECOM, CCNC, ICME, etc. He was a recipient of the CUHK Faculty Outstanding Ph.D. Thesis Award, Microsoft Fellowship, and ERCIM Alain Bensoussan Fellowship. He is a member of SPIE.

Chia-Ming Tsai received the B.S. degree from Feng Chia University, Taichung, Taiwan, in 2003, and the M.S. degree from National Chung-Cheng University, Chiayi, Taiwan, in 2005, both in computer science and information engineering. Since 2005, he has been pursuing the Ph.D. degree with the Department of Computer Science and Information Engineering, National Chung-Cheng University.

He was a Software Engineer with CyberLink, Inc., Taipei, Taiwan, from 2012 to 2013. His research interests include video coding and video content adaptation.

Chia-Wen Lin (S'94–M'00–SM'04) received the Ph.D. degree in electrical engineering from National Tsing Hua University (NTHU), Hsinchu, Taiwan, in 2000.

He is an associate professor with the Department of Electrical Engineering and the Institute of Communications Engineering, NTHU. He was with the Department of Computer Science and Information Engineering, National Chung Cheng University, Chiayi, Taiwan, from 2000 to 2007. Prior to joining academia, he was with the Information and Communications Research Laboratories, Industrial Technology Research Institute, Hsinchu, Taiwan, from 1992 to 2000. His current research interests include video content analysis and video networking.

Dr. Lin is an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE MULTIMEDIA, and the *Journal of Visual Communication and Image Representation*. He is also an Area Editor of the *EURASIP Signal Processing: Image Communication*. He has been the Chair of the Multimedia Systems and Applications Technical Committee since September 2013. He served as the Technical Program Co-Chair of the IEEE International Conference on Multimedia and Expo (ICME) in 2010, and Special Session Co-Chair of the IEEE International Consulting Management and Engineering in 2009. He was a recipient of the 2001 Ph.D. Thesis Awards presented by the Ministry of Education, Taiwan, and also of the Young Faculty Awards presented by CCU in 2005 and the Young Investigator Awards presented by National Science Council, Taiwan, in 2006. His paper won the Young Investigator Award presented by VCIP 2005.