PAPER *Special Issue on Recent Advances in Circuits and Systems*

A Visual Attention based Region-of-Interest Determination Framework for Video Sequences*

Wen-Huang CHENG[†a)], **Wei-Ta CHU**[††b)], ***and*** **Ja-Ling WU**[†c)], ***Nonmembers***

SUMMARY This paper presents a framework for automatic video region-of-interest determination based on visual attention model. We view this work as a preliminary step towards the solution of high-level semantic video analysis. Facing such a challenging issue, in this work, a set of attempts on using video attention features and knowledge of computational media aesthetics are made. The three types of visual attention features we used are intensity, color, and motion. Referring to aesthetic principles, these features are combined according to camera motion types on the basis of a new proposed video analysis unit, frame-segment. We conduct subjective experiments on several kinds of video data and demonstrate the effectiveness of the proposed framework.

key words: *region-of-interest, video analysis, visual attention model, computational media aesthetics.*

1. Introduction

The rapid progress of the technology for multimedia production has contributed to the extensive use of multimedia, the explosive development of mobile communication, especially the ever-increasing importance of video communication, such as video phone and video-on-demand. Wide-ranging usage of video communications bring in several visible trends: 1) More and more end users have devices with diverse capability, such as Pocket PC and Smartphone. 2) As the types of networks, devices, and compression formats increase, interoperability among different systems and networks become more important. 3) There is too much redundant information in multimedia documents to be processed efficiently. In facing these challenges, one of the key technologies is region-of-interest (ROI) determination [1][2], which benefits in the applications of content adaptation, transcoding, and intelligent information management, etc. Moreover, it provides a practicable way for semantic level analysis without the need of fully understanding about the document's content.

In general, an ROI is a portion of a multimedia document that audiences show more interest in or pay more attention to than others. For the ease of explanation, we give a precise definition of an ROI, first. An ROI is a portion of a frame that contains the key concept or main subject of a visual scene and provides end users a more concise and informative representation of a document, e.g., the speaker should be one of the ROIs in a conference scene.

In the literature, schemes proposed for determining ROIs can be divided into two categories: *saliency-oriented* and *task-oriented*. The saliency-oriented scheme is to predict what will involuntarily attract our visual attention in a scene, and where to identify the interesting regions when the saliency information is given. According to psychological findings about the primate visual system and eye fixation, quite a few vision models for still images have been developed to simulate the cognitive mechanism of human beings. One well-known approach is based on Itti's visual attention model [3], in which several spatial visual features are combined into a single saliency map for representing local conspicuity in images. This model has been extensively studied in many fields and was shown to be robust in intelligent processing of digital images [4][5][6]. However, due to the ignorance of temporal aspects, its extension to moving pictures needs to be explored.

Some approaches for analyzing video attentions are then proposed. Ma *et al.* [7][8] presented user attention models for video skimming and summarization, which utilized more audio-visual features of semantics, for example, motion, speech, camera operation, and lexical information. In his paper, although the video features are shown to be effective in detecting temporal attentions, their interactions with spatial visual features are still unknown. Ho *et al.* [9] proposed a framework for video focus detection based on visual attention, which introduced a video-genre-based method for saliency map generation. That is, in different video categories, different parameter sets are elaborately optimized and accordingly assigned. The experiment shows impressive result, but the method is too highly domain-dependent to be extended for general purpose.

On the other hand, the task-oriented scheme is to determine where are relevant to a viewer's prede-

Manuscript received October 8, 2004.
Manuscript revised February 6, 2005.
Final manuscript received March 15, 2005.

[†]The author is with the Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei, Taiwan, 10617, R.O.C.

[††]The author is with the Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan, 10617, R.O.C.

a) E-mail: wisley@cmlab.csie.ntu.edu.tw
b) E-mail: wtchu@cmlab.csie.ntu.edu.tw
c) E-mail: wjl@cmlab.csie.ntu.edu.tw

*This work was partially supported by the CIET-NTU(MOE) and National Science Council of R.O.C. under the contract No. NSC93-2622-E-002-033, NSC93-2752-E-002-006-PAE and NSC93-2213-E-002-006.

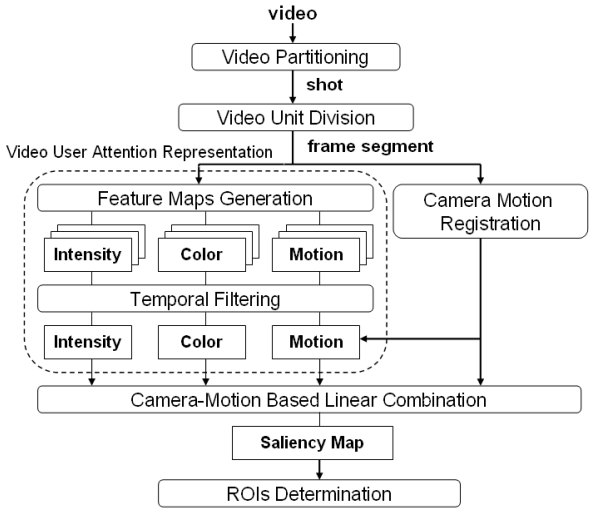


Fig. 1 Block diagram of the proposed framework for conducting video ROIs determination.

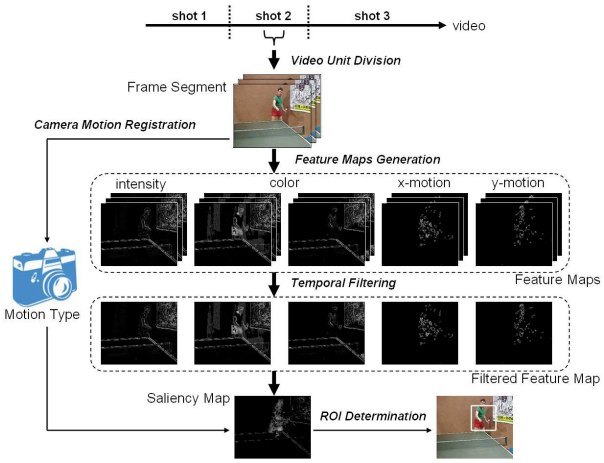


Fig. 2 An example of operations of the proposed framework. The input frame-segment is under static-with-object-motion camera type (defined in Section 3.4.2).

fined goal and what will voluntarily attract his attention when studying a scene. In this case, a salient location may be completely filtered out for its irrelevance of the viewer's goal. Navalpakkam *et al.* [10] propose an architecture to estimate the task-relevance of attended locations in a scene. The location-based relevant information is represented with a task-relevance map for each image. Recent researches by Cater *et al.* [11], Golenzer *et al.* [12], and Lin *et al.* [13] are also classified in this class. Generally, with an explicit description about the viewer's target, task-oriented schemes get better performance than those of saliency-oriented ones. However, the goal or tasks may not be always available in advance.

In summary, the problems associated with conventional video ROI determination, based on visual attention model, can roughly be divided into three categories. The first one is the lack or unsuitably treatment of temporal and motion information, and the second is that fixed or video-genre-based feature combining method seems to be problematic for practical use. Finally, little effort is put on integrating the advantages from both saliency-oriented and task-oriented schemes.

In this work, we consider the problem from the viewpoints of both visual attention model and computational media aesthetics [14][15]. Our goal is to develop a framework that can be used to determine the video ROI using computable visual features and general video-shooting principles. In this way, the superiorities of saliency and task oriented schemes are both integrated in our work. In addition to light and color, object motion is adopted as one visual feature in our attention model. Rather than a single frame, we choose a short video clip, i.e. a frame-segment, as the basic unit for conducting video analysis. A camera-motion assisted algorithm for combining visual features is developed and applied to the framework. We conduct lots of experiments on kinds of video data and demonstrate the effectiveness of the proposed framework in video ROI determination.

The rest of this paper is organized as follows. Section 2 presents the proposed framework for video ROI determination. Visual attention representation and camera motion utilization are described in Section 3. Section 4 discusses the dynamic ROIs determination from a saliency map. Section 5 shows experimental results, and Section 6 presents our concluding remarks.

2. An Overview of the Proposed Framework

The block diagram of the proposed framework is illustrated in Fig. 1. The input video is first segmented by a reliable shot boundary detection algorithm [16], which can correctly detect abrupt shot changes and gradual transitions. Further, each shot is partitioned into non-overlapped "frame-segment" (will be explained in Section 3). For each frame-segment, one camera motion type is registered. This camera motion information will be used to generate the saliency map later. Meanwhile, the corresponding feature maps generated from each of the feature models are computed. By taking account of the camera motion types, different kinds of feature maps are combined elaborately. Finally, the integrated saliency map is constructed. The video ROIs are then dynamically estimated according to the active area of saliency maps. An operational example of the proposed framework is illustrated in Fig. 2.

3. Visual Attention Representation

In this section, we discuss in detail how to represent visual attention in video sequences and propose a new video unit for ROI analysis. Subsequently, the relation-

ship between camera motion and visual attention is described. Based on our observations, a novel method for saliency map generation is presented.

3.1 Visual Attention Model

Visual attention refers to the ability of a viewer concentrating his attention on some visual objects or regions. Previous research showed that this physiological process could be modeled by the so-called visual attention model [3][8]. In our work, three types of video-oriented visual features (intensity, color, and motion) are adopted to model the visual attraction of videos by using the same idea.

3.1.1 Contrast Based Intensity and Color Feature Model

One of the most important ingredients of visual attention model is the contrast [17]. In psychology, perceptual experiments have shown that some color pairs, such as red-green and blue-yellow, possess high spatial and chromatic opposition. The same characteristics exist in high difference lighting or intensity pairs. Based on these observations, we include three contrast based feature models: intensity, red-green color contrast, and blue-yellow color contrast, into our visual attention representation module. The contrast maps are respectively defined as follows.

$$\mathcal{M}_{\mathcal{I}}(p) = \max_{p' \in w} |\mathcal{I}(p) - \mathcal{I}(p')|, \tag{1}$$

$$\mathcal{M}_{\mathcal{RG}}(p) = \max_{p' \in w} |(\mathcal{R}(p) - \mathcal{G}(p)) - (\mathcal{G}(p') - \mathcal{R}(p'))|, \tag{2}$$

$$\mathcal{M}_{\mathcal{BY}}(p) = \max_{p' \in w} |(\mathcal{B}(p) - \mathcal{Y}(p)) - (\mathcal{Y}(p') - \mathcal{B}(p'))|, \tag{3}$$

where $p = [x, y]^T$ is a position vector, w is a 3×3 window centered at p, and $\mathcal{I}$, $\mathcal{R}$, $\mathcal{G}$, $\mathcal{B}$, $\mathcal{Y}$ denote the intensity, red, green, blue, and yellow component value functions, respectively.

3.1.2 Motion Feature Model

The motion of objects plays an essential role in a video. It allows the video-maker to direct the audience's attention across the two-dimensional space of a frame [18]. In the proposed framework, two feature models: x-motion and y-motion, are used to represent the motion information of a video frame. The x-motion and the y-motion refer to the horizontal and the vertical movements of a specific pixel within a frame, respectively.

If we consider a video as a frame sequence with spatial axes (x, y) and temporal axis t, the spatio-temporal slices are then a set of 2-dimensional frames along the t axis. The spatio-temporal slices can be further divided into horizontal slice with axes (x, t) and vertical slice with axes (y, t). To find the motion activity in the scene, the two-dimensional (2-D) structure tensor (ST) [19][20] of the slices is evaluated. Compared with other motion descriptors, the 2-D ST is adopted for that the coherence (or confidence) measure can also be estimated. The 2-D ST, J, is expressed as

$$J = \begin{bmatrix} J_{xx} & J_{xt} \\ J_{xt} & J_{tt} \end{bmatrix} = \begin{bmatrix} \sum_w H_x^2 & \sum_w H_x H_t \\ \sum_w H_x H_t & \sum_w H_t^2 \end{bmatrix}, \tag{4}$$

where w is the 3×3 support window. H_x and H_t are the partial derivatives of a horizontal slice along the spatial and temporal dimensions. Consequently, the local motion angle θ_x and its corresponding confidence measure (cm_x) can be computed as

$$\theta_x = \frac{1}{2} \tan^{-1} \frac{2J_{xt}}{J_{xx} - J_{tt}}, \tag{5}$$

and

$$cm_x = \frac{(J_{xx} - J_{tt})^2 + 4J_{xt}^2}{J_{xx} + J_{tt})^2}, \ 0 \le cm_x \le 1. \tag{6}$$

The vertical slice is processed in the same way to obtain the corresponding θ_y and cm_y. Finally, the x-motion and the y-motion maps are individually calculated as:

$$\mathcal{M}_{\mathcal{X}}(p) = \theta_x \times cm_x, \tag{7}$$

$$\mathcal{M}_{\mathcal{Y}}(p) = \theta_y \times cm_y, \tag{8}$$

where $p = [x, y]^T$ is, again, a position vector.

3.2 Frame-segment

In previous research, visual attention is modeled and determined mostly for only a single, at most, for two consecutive frames. The collection of determined regions of each independent single frame composes the final ROIs of a video sequence. However, based on our previous observations [9], we found that the single- or two-frame based approach only generates acceptable results for images but not for video ROI analysis. For example, the focus point may swiftly tremble due to a slight difference between two consecutive frames. This unpleasant phenomenon does not exist in viewers' attention. If the estimated ROIs are applied to other extended applications, such as scalable coding and content adaptation, the prescribed defect will cause significant deficiency in both bit rate and quality. Due to the fact that the content of a video would not change drastically in a short duration, we take a short video clip, called frame-segment, as the unit for conducting the video ROI analysis. The new defined frame-segment takes both spatial and temporal correlations into account and can suppress noises caused by sudden luminance change, such as flashlights. In our experiments, the length of a frame-segment is empirically set to 0.5 seconds or 15 frames. In the rest of this paper, we will use 0.5 seconds and 15 frames interchangeably to indicate the default length of a frame-segment.

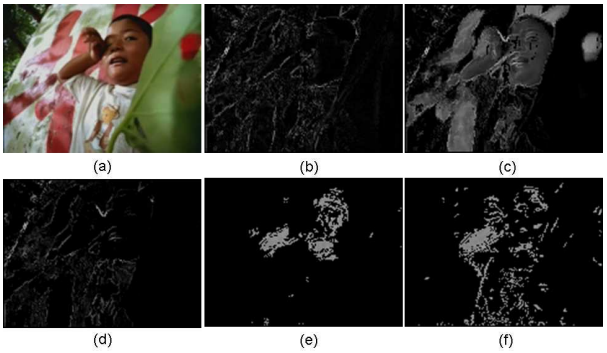


Fig. 3 Example of feature maps. (a) original video frame, (b) intensity, (c) red-green color, (d) blue-yellow color, (e) x-motion, and (f) y-motion feature maps.

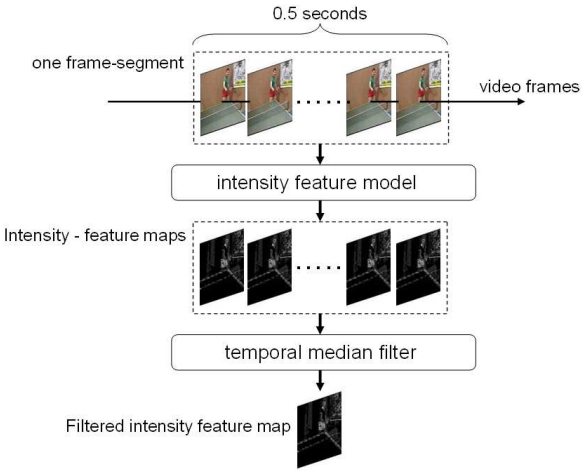


Fig. 4 The procedure of generating the filtered intensity feature map from a frame-segment.

3.3 Feature Map and Filtered Feature Map

For each video frame of the same frame-segment, the distributions of each of the features are calculated and constructed as five feature maps, as shown in Fig. 3. Therefore, each frame-segment has five map sets and each of them is composed of 15 feature maps belonging to a specific feature. A temporal median filter is then applied to each of the sets to find the corresponding filtered feature map. Fig. 4 shows an example for generating the intensity filtered feature map. Note that the temporal median filter plays an important role in the process. The average effect of filtering for frames within a segment can effectively suppress noises and sub-salient regions so that each filtered feature map represents the general characteristics of a specific feature in the frame-segment. In other words, both the spatial saliency distribution and the temporal saliency variation of a video are used to model the visual attraction of videos.

Fig. 5 Demonstrations of visual attention under left-pan camera motion. The t- to $(t + 2\Delta t)$-th video frames are captured in an interval of 0.5 seconds (i.e., $\Delta t = 0.5$ seconds) from a TV sports program. The white squares indicate the possible attentive regions in the first frame.

3.4 Camera Motion Based Saliency Map Generation

3.4.1 Relations between Camera Motion and Visual Attention

Nowadays, a large amount of videos are produced according to the principles of computational media aesthetics, especially the *expert-produced* videos [14]. From the viewpoint of video shooting, different camera motions have different impacts on the audience's reception. In other words, they influence the relative importance of each visual feature and reveal what and where the video-maker wants viewers to see. The idea has been extensively used in film or TV show productions. On the other hand, from the perspective of task-oriented gaze control, the phenomenon that directors purposely move their camera to control the audience's fixations appropriately serves as a high-level hint for integrating spatiotemporal visual features.

Fig. 5 gives a real example. If you take the first (or t-th) video frame as a still image, your eyes will freely scan the entire image and attracted by some noticeable regions, such as the scoreboard, screen texts, or players. However, if you take it as one of the video frames, that is, look at these frames rapidly in succession, you will find that your eyesight involuntarily moves left with the panning camera and is mostly attracted by horizontally moving objects. Your vision unconsciously follows the camera's track within the scene and the relative saliency of each region has accordingly been changed. Therefore, it is our belief that camera movement should be considered in the process of ROI determination.

3.4.2 Camera Motion Registration

In our work, seven camera motion types are registered:

Table 1 The ranges of the non-uniform bins used to quantize tensor histogram.

Bin #	Range
-2	$[-90^o, -45^o)$
-1	$[-45^o, -5^o)$
0	$[-5^o,\ 5^o]$
1	$(5^o,\ 45^o]$
2	$(45^o,\ 90^o]$

Table 2 Weights for filtered feature maps under different camera motion types.

	Zoom	L/R-Pan	U/D-Tilt	Static	Motion
Intensity	0.2	0.05	0.05	0.15	0.05
red-green	0.2	0.05	0.05	0.075	0.05
blue-yellow	0.2	0.05	0.05	0.075	0.05
X-motion	0.2	0.75	0.1	0.35	0.425
Y-motion	0.2	0.1	0.75	0.35	0.425

zoom, left-pan, right-pan, up-tilt, down-tilt, static-with-no-motion, and static-with-object-motion. The spatio-temporal slices based motion analysis techniques [19] are used to register the camera motion type of each frame-segment. We use two tensor histograms. One is for all horizontal slices and another is for all vertical slices, and they are denoted as $\mathcal{MH}$ and $\mathcal{MV}$, respectively. Within a frame-segment, all the local motions generated from the motion feature models are non-uniformly quantized into five bins Φ_i, i = -2 $\sim$ 2 (c.f. Table 1).

After constructing the two tensor histograms, a rule-based algorithm is applied to detect camera motion. We take two examples, say zoom and left-pan operations, to explain the detailed processes. For zoom, the tensor votings of positive-motion-angle bins and negative-motion-angle bins are approximately the same in both horizontal and vertical slices tensor histograms. That is,

$$\frac{\sum_{\Phi_i>0}\mathcal{MH}(\Phi_i)}{\sum_{\Phi_i<0}\mathcal{MH}(\Phi_i)} \approx 1 \; and \; \frac{\sum_{\Phi_i>0}\mathcal{MV}(\Phi_i)}{\sum_{\Phi_i<0}\mathcal{MV}(\Phi_i)} \approx 1. \quad (9)$$

For left-pan operation, the camera is moving fast toward left direction, so the detected right-direction motion would be much greater than the left-direction one. The value of right-direction motion, $\mathcal{MH}(\Phi_i)$, $\Phi_i > 0$, would be greater than a given camera motion threshold to ensure that the motion is induced by the camera itself. That is,

$$\sum_{\Phi_i>0}\mathcal{MH}(\Phi_i) > \kappa \; and \; \frac{\sum_{\Phi_i>0}\mathcal{MH}(\Phi_i)}{\sum_{\Phi_i<0}\mathcal{MH}(\Phi_i)} \gg 1, \quad (10)$$

where κ is the camera motion threshold. The other camera motion types can be decided following the similar way.

3.4.3 Saliency Map Generation

Weights of the filtered feature maps for combining the generic saliency map are decided according to the registered camera motion types (will be described in the next subsection). The generic saliency map is generated according to the following equation:

$$S(N) = \alpha_{c,1} \times FFM_1 + \cdots + \alpha_{c,n} \times FFM_n, \quad (11)$$

where $S(N)$ is the generated generic saliency map of a frame-segment with length N. FFM_i is the i-th filtered feature map of that segment, and $\alpha_{c,i}$ is the weight of the corresponding FFM_i under given camera motion type c. Table 2 shows the weights for various camera motion types and filtered feature maps used in our framework. These weights are defined elaborately to present characteristics of different camera motion types. For example, when camera panning occurs, the horizontal motion should be emphasized.

(a) (b)

Fig. 6 Examples of determined ROIs (dotted-line squares) for two different settings of feature weights. The manually marked ground truths are indicated by solid-line squares.

3.4.4 Procedure for Feature Weights Selection

As shown in Fig. 6, selection of appropriate feature weights is important in ROI determination. However, due to the large amount of candidate weights and their combinations, it's impossible to decide an appropriate combination of weights manually. On the other hand, it's also unpractical to do the selection through exhaustive search, because the weights selection depends highly on human's subjective perception. In this work, we exploit an generic procedure to sieve out some candidates from the weights combinations based on certain selection rule, first. Then the final decision is made by the end user. Since the procedure is applicable to all the adopted camera types, without loss of generality, we only describe the analysis of left-pan operation in the following.

First, a set of frame-segments $F = \{F_i,\ i = 1{\sim}T$ (e.g., $T = 50)\}$ are carefully chosen from various kinds of videos. Without loss of generality, one definite main subject is assumed to be contained in each F_i and have been manually marked as the ROI. These frame-segments with marked ROIs form the ground truth of our training benchmark. Let w_j, $j = 1{\sim}5$, be five random variables and each of them represents the feature weights of intensity, red-green contrast, blue-yellow contrast, x-motion and y-motion, respectively.

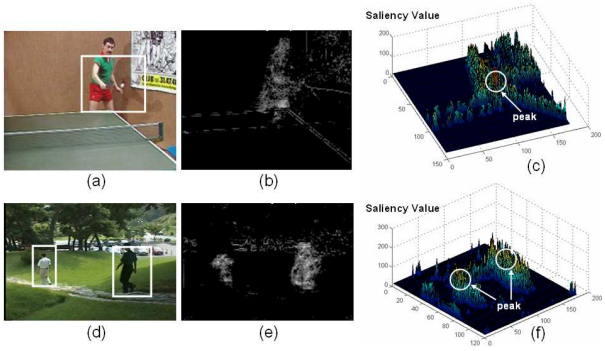


Fig. 7 Examples of frame-segments with (a) one and (d) two ROIs (indicated by the white squares); (b) and (e) are the corresponding saliency maps, and (c) and (f) are the 3-D profiles of the saliency maps of (a) and (d), respectively.

Note that $\sum_{j=1}^{5} w_j = 1$. Then, all possible combinations of the weight vector $(w_1, w_2, \ldots, w_5)$ are generated according to a predefined deviation step size, say 0.005. As shown in Fig. 6, if a weight vector correctly reflects the relative importance of each feature, the region size and location of a determined ROI will highly match with those of the ground truth. That is, the overlapped area of the dotted-line and solid-line squares will nearly equal to their joint region. Based on the observations, we define a simple but well-defined fitness value e_w for each weight vector w to test whether it is a possible candidate as follows.

$$e_w = \frac{1}{T} \sum_{F_i \in F} \frac{|DR_{w,i} \cap GR_{w,i}|}{|DR_{w,i} \cup GR_{w,i}|} > \epsilon, \tag{12}$$

where $DR_{w,i}$ and $GR_{w,i}$ are pixel sets of the determined and ground truth ROIs in F_i, respectively. ϵ is a dynamic threshold used by end users to control the amount of obtained candidates to a manageable number. For example, the value of ϵ is increased as to reduce the number of candidates. Note that the fitness value e_w is mainly used for preprocessing the weight enumeration, but not for picking up the best result. Finally, an appropriate weight vector among the candidates is selected by end users according to their subjective judgement. To avoid bias, three end users are invited to joint the test and the decision is made by majority vote. It is our belief that the proposed procedure successfully integrates the advantages of both computer search and manual selection. Moreover, according to our simulation results, this approach promises the reliability of selected feature weights well.

4. Video ROI Determination

Although the saliency maps have showed the ability to characterize the visual attraction of a video, the generated ROIs still have the probability of failure in capturing the essence if they are not determined properly according to these maps. In this work, the appropriate position and size of an ROI are determined by the regular moments [21]. Since there may be multiple key objects in a frame-segment, a method for dynamically determining the number of ROIs will then be presented.

1. Take the first sample x_1 as the representative of the first cluster:
   $z_1 = x_1$, where z_1 is the center of the first cluster.
2. Take the next sample and compute its distance $d_i(x, z_i)$ to all the existing clusters, and choose the minimum of d_i: $\min d_i$.
   (a) Assign x to z_i if $\min d_i \le \theta_\tau$, $0 \le \theta \le 1$, where τ is the membership boundary for a specified cluster, and update the center of this cluster.
   (b) A new cluster with center x is created if $\min d_i > \tau$.
   (c) No decision will be made if $\theta_\tau < \min d_i \le \tau$.
   In this case, sample x is in the intermediate region.
3. Repeat *step 2* until all samples have been checked once. Calculate the variances of all the clusters.
4. If the current variance is the same as that of the previous iteration, the clustering process is converged, go to *step 5*. Otherwise, go to *step 2* for further iteration.

If the ratio of samples in the intermediate region is larger than a threshold, adjust θ and τ and go to *step 2* again. Otherwise, the process ends.

Fig. 8 An Euclidean-distance based clustering algorithm.

4.1 Saliency Weighted Regular Moment

Saliency weighted regular moments [21] are effective tools for calculating the center coordinate of a set of weighted data points. They are adopted in our work to determine the centroid of each ROI. Let

$$m_{pq} = \sum_{x=1}^{M} \sum_{y=1}^{N} x^p y^q s(x, y), \ p, q = 0, 1, 2, ..., \tag{13}$$

where M, N are the dimensions of the saliency map and $s(x, y)$ is the saliency value function corresponding to the pixel (x, y). In the saliency map, the centroid (x, y) of an ROI is given by $(\overline{x}, \overline{y}) = (m_{10}/m_{00}, m_{01}/m_{00})$. On the other hand, based on our observations, the region size of each ROI is proportional to the active area of the saliency map. A saliency weighted invariant [21] is defined to measure the variation of computed centroid as follows. Let

$$\eta_{pq} = \frac{\sum_{x=1}^{M} \sum_{y=1}^{N} (x - \overline{x})^p (y - \overline{y})^q s(x, y)}{m_{00}}, \tag{14}$$

and the region size is set as $(k\sqrt{\eta_{20}}) \times (k\sqrt{\eta_{02}})$, where $k = 2$.

When determining the ROIs, we observe that if those saliency points are clustered around a concentrated area, they generate a small ROI. It implies that an obvious attentive region exists. Contrarily, if those saliency points are scattered across the saliency map, it implies that there is no obvious attentive region and the region size will be very large. In this case, even one

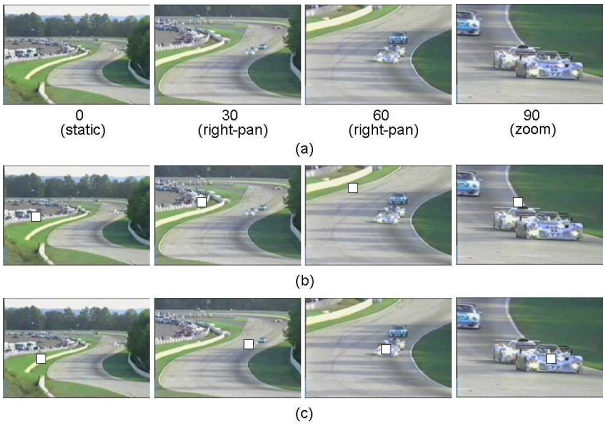


Fig. 9 Detection results of ROI centroid for the sequence "car-racing". (a) Original video frames with frame index and camera motion type, (b) ROI centroid detected by conventional approach, and (c) ROI centroid detected by the proposed framework. (The white square indicates the location of detected ROI centroid.)

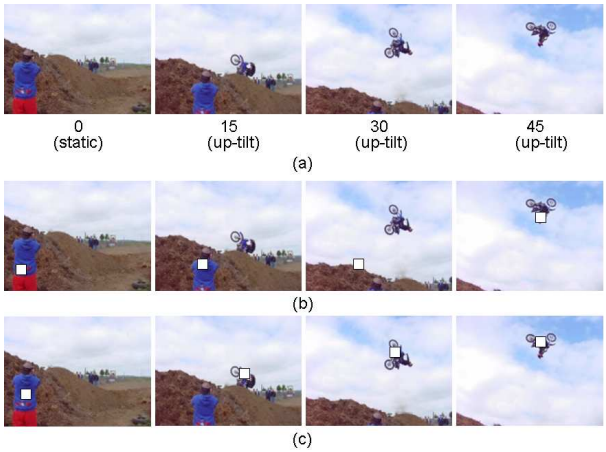


Fig. 10 Detection results of ROI centroid for the sequence "acrobatics". (a) Original video frames with frame index and camera motion type, (b) ROI centroid detected by conventional approach, and (c) ROI centroid detected by the proposed framework. (The white square indicates the location of detected ROI centroid.)

ROI is claimed, actually, it implies no apparent region attracts the audience's attention.

4.2 Dynamic Determination of ROIs

Sometimes, there are more than one ROI in a frame-segment. For example, in a distance view of a tennis game, two players may form two different ROIs. We devised a method to resolve this problem explicitly. In a saliency map, each ROI usually consists of a set of saliency values peaked at the center of its 3-D profiles. For example, if a frame-segment has two ROIs (e.g. there are two separate moving persons in Fig. 7(d)), its saliency map usually has two separate peaked sets, as shown in Fig. 7(f). We assume that the saliency value ranges from 0 to R (in this work, R is 255). In each saliency map, if a pixel's saliency value is greater than a predefined threshold, it is added to the peak set (PS). The pixels in the PS are further clustered according to an Euclidean-distance based algorithm (cf. Fig. 8), and are then divided into several disjoint subsets. That is,

$$PS = \bigcup_{i=1}^{n} PS_i, \ where \ PS_i \cap PS_j = \phi \ if \ i \neq j. \quad (15)$$

In this way, a saliency map is divided into n regions, and each region corresponds to a peak subset PS_i. One ROI is declared for each region. With this scheme, the number of ROIs can be determined dynamically and automatically for each frame-segment. However, normally, the number of ROIs in a frame will be no more than three. If there are more than three key objects in a frame, the viewer may be confused and lose his focus [14].

5. Experimental Results

Although the problem of ROI determination has been extensively studied, there is no standard method to evaluate the corresponding performance. The difficulties arise from the strong subjectivity of human perception. To verify the effectiveness of the proposed framework, we conduct some experiments and compare the so-obtained results with those of a conventional approach [3]. Subsequently, we have carried out a user study experiment to take human factors into account. Note that, for fair competition, orientation feature is replaced by motion feature in the conventional approach [3]. In addition, the experimental data are all *expert-produced* sequences and are chosen from three categories of videos: TV shows, sports programs, and TV commercials.

Two fundamental elements are used to define an ROI. One is the centroid and another is the region size. Fig. 9 illustrates an example of ROI centroid detection. As shown in the 0^{th} frame, the result is consistent for both frameworks. In this case, there is no obvious camera movement so that the proposed framework acts like a conventional visual attention model. In the following frames, when camera motion type is available, the proposed framework shows its superior performance to that of the conventional one. In Fig. 9(c), the detected centroid location rapidly moves along the camera direction and follows the racing cars. Even for the vague-viewed cars (cf. the 60^{th} frame), they are still detected with the help of camera-motion based method. Further, as shown in the 90^{th} frame, under the "zoom-in" operation the car in the upper side of the frame is reasonably ignored. On the contrary, in Fig. 9(b) the detected

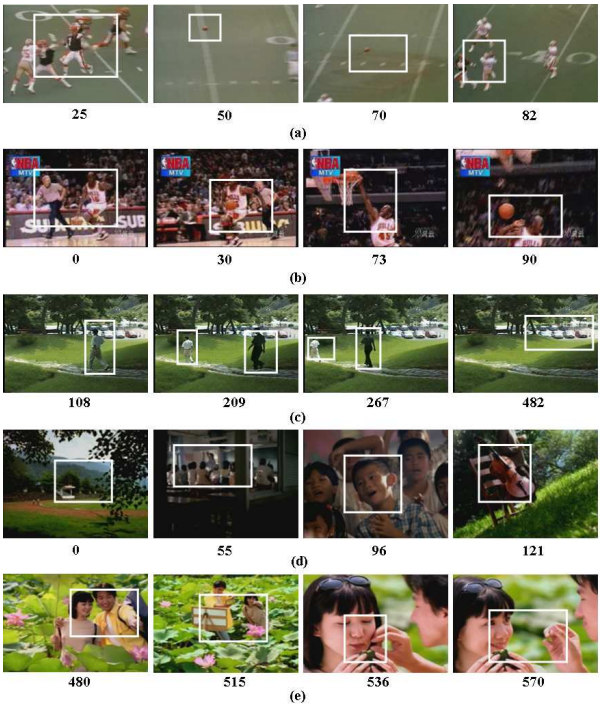


Fig. 11 Sample results of ROI determination for the sequences (a) "football", (b) "Jordan", (c) "walking", (d) "commercial-1", and (e) "commercial-2". (The number below each frame is the time index in the corresponding video sequence.)

Table 3 Comparison of the user study between (a) conventional approach and (b) the proposed framework.

		GOOD (%)	ACCEPTABLE (%)	FAILED (%)
Data Set I	(a)	69	18	13
	(b)	82	15	3
Data Set II	(a)	71	20	9
	(b)	73	21	6
Avg.	(a)	70	19	11
	(b)	78	18	4

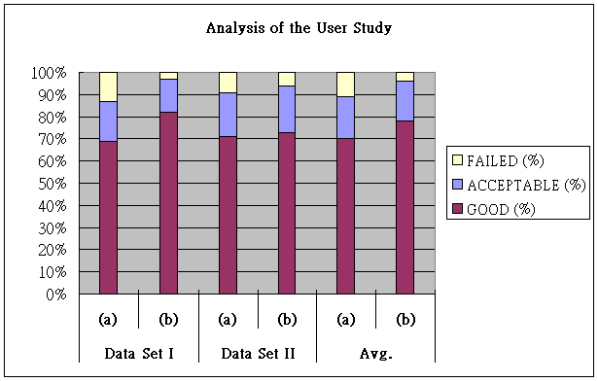


Fig. 12 Statistical comparison of the user study between (a) conventional approach and (b) the proposed framework.

centroid is continually drawn to the range around the road-fence. It is due to the inflexibility of the conventional approach. From the viewpoint of a single image, the road-fence is indeed the most salient object in the scene, but the right-pan camera operation has revealed the director's actual intention: "to fast turn right and concentrate on the leading cars." Without taking the camera motion and temporal information into account, it is hard to capture the actual subjects in such a scene. Similar cases are also exhibited in Fig. 10. Based on the experimental results and associated comparisons, it demonstrates that the proposed framework provides a more reliable and robust channel for ROI centroid detection.

Examples of ROI determination are shown in Fig. 11. The video sequences of Fig. 11(a) and (b) are classified as sports programs. Without fully understanding of the game rules, the subjects can still be correctly identified by the proposed framework. The 50^{th} frame of Fig. 11(a) illustrates an excellent example. The football almost disappeared in the image, but the detected ROI successfully captures it according to the information of rapid camera panning. Fig. 11(c) demonstrates the accuracy of the proposed dynamic algorithm in deciding the number of ROIs. Through the whole clip, the number of ROIs appropriately changes according to the video content. Finally, Fig. 11(d) and (e) show two results of TV commercials. At the first glance, the determined ROIs in the 536^{th} and the 570^{th} frames of Fig. 11(e) seem to be unsatisfactory. However, object motion is one of the important feature adopted in our attention model. As a result, the hand's fast moving highly influences the determined ROIs.

To further evaluate the performance of the proposed framework, a subjective experiment is designed. Currently, we have two testing data sets. Data-Set-I contains videos from TV shows, films, and commercials. Data-Set-II contains clips from various sports programs. Each data set includes about 15 sequences with determined ROIs, and the total length of them is approximately 60 minutes. Then, twenty observers were invited to participate in the user study. To verify the connection between determined ROIs and human's perception, observers are requested to assign one subjective comment for each test sequence. Three comments, GOOD, ACCEPTABLE, and FAILED, are adopted in our experiments. In order to avoid bias, observers are separated into two groups: one for the proposed framework and another for the conventional approach.

The statistical results of the experiment are listed in both Table 3 and Fig. 12. Obviously, the proposed framework outperforms the conventional approach in all cases. For Data-Set-I, the improvement are even higher than 15%. Although conventional approach also provides reasonable ACCEPTABLE rate, its FAILED percentage are almost 4 times higher than that of the proposed framework. For TV show and commercial se-

quences, the scene compositions are usually very complex and contain various special effects like stage lighting. By exploiting the camera motion information, the proposed framework doesn't easily misdetect the "false" salient regions. However, the conventional approach is unaware of that and always misses the true key subject. On the other hand, for Data-Set-II, the performances of both frameworks are more consistent. In quite a few sport videos, the key subjects are exactly the most salient objects with high color or intensity contrast (e.g., Fig. 11(b)), so conventional approach also achieves satisfactory performance. However, based on computational media aesthetics, the proposed framework generally provides more robust performance than that of the conventional approach. In summary, it is reasonable that the overall performance of Data-Set-I is better than that of Data-Set-II, because the videos in Data-Set-I are produced more reliably according to the principles of computational media aesthetics than those of the Data-Set-II. The clips in Data-Set-II are sport related videos, and more semantic or game related rules are needed to facilitate the accuracy of ROI determination. In conclusion, the experimental results show that most of the observers (more than 95%) feel comfortable with the estimated video ROIs, which demonstrates the effectiveness of the proposed framework.

6. Conclusion

This paper presents an automatic video ROI determination framework based on visual attention model, which provides an alternative way towards high-level semantic video analysis. The main contribution of this work is the investigation of a video-oriented fusion scheme for integrating visual features to facilitate the ROI determination. Both visual attention model and computational media aesthetics are considered in the scheme. Experimental results show that the proposed framework is effective in video ROI determination. This work is very useful for a variety of vision systems and video content analysis. It is noticeable that the computational complexity of the proposed framework will be dominated by two tasks: feature map generation and feature weights selection. Fortunately, with the recent advances in circuits and systems, lots of feature extraction devices [22] have been developed. For example, intensity and color related circuits can be found in a digital camera, while motion feature extraction circuits are reachable in MPEG encoder related products, such as digital camcoders and DVD recorders. Similarly, several advanced circuits designed for neural network and/or support vector machine (SVM) circuits [22] are available for handling the issue of feature weights selection. In other words, with the aid of recent progress in circuits and systems, the proposed framework for ROI determination is expected to be applicable to various video applications in real time. Our future work will focus on the development of more semantic-level attention features and on the investigation of video-oriented aesthetic principles and film grammars.

References

[1] N. Dimitrova, H.-J. Zhang, B. Shahraray, I. Sezan, T. Huang, and A. Zakhor, "Applications of video-content analysis and retrieval," IEEE Multimedia, vol. 9, pp. 42-55, July-Sept. 2002.

[2] C. M. Privitera and L. W. Stark, "Algorithms for defining visual regions-of-interest: comparison with eye fixations," IEEE Trans. Pattern Anal. Machine Intell., vol. 22, no. 9, pp. 970-982, Sept. 2000.

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

[4] Y.-F. Ma and H.-J. Zhang, "Contrast-based image attention analysis by using fuzzy growing," in ACM Multimedia Conf., 2003, pp. 374-381.

[5] H. Liu, X. Xie, W.-Y. Ma, and H.-J. Zhang, "Automatic browsing of large pictures on mobile devices," in ACM Multimedia Conf., 2003, pp. 148-155.

[6] G. Deco and J. Zihl, "A neurodynamical model of visual attention: feedback enhancement of spatial resolution in a hierarchical system, " Journal of Computational Neuroscience, vol. 10, pp. 231-253, 2001.

[7] Y.-F. Ma and H.-J. Zhang, "A model of motion attention for video skimming," in ICIP'02, 2002, pp. 129-132.

[8] Y.-F. Ma, L. Lu, H.-J. Zhang, and M. Li, "A user attention model for video summarization," in ACM Multimedia Conf., 2002, pp. 533-542.

[9] C.-C. Ho, W.-H. Cheng, T.-J. Pan, and J.-L. Wu, "A user-attention based focus detection framework and its applications," in Proc. of Pacific-Rim Conference on Multimeida, Singapore, 2003.

[10] V. Navalpakkam and L. Itti, "A goal oriented attention guidance model," Lecture Notes Comput. Sci., vol. 2525, pp. 453-164, Nov. 2002.

[11] K. Cater, A. G. Chalmers, and P. Ledda, "Selective Quality Rendering by exploiting human inattentional blindness: looking but not seeing," in Proc. of Symposium on Virtual Reality Software and Technology, 2002, pp. 17-24.

[12] J. Golenzer, C. Viard-Gaudin, and P. Lallican, "Finding regions of interest in document images by planar HMM," in IEEE International Conference on Pattern Recognition, 2002, pp.145-148.

[13] H. Lin, J. Si, and G. P. Abousleman, "Knowledge-based hierarchical region-of-interest detection," in IEEE ICASSP'02, 2002, pp. 3628-3631.

[14] H. Zettl, Sight, Sound, Motion: Applied Media Aesthetics. Belmont, CA: Wadsworth, 1990.

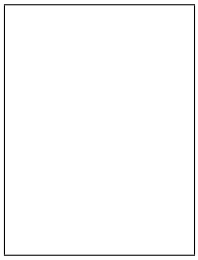
[15] C. Dorai and S. Venkatesh, Media Computing: Computational Media Aesthetics. Norwell, MA: Kluwer, 2002.

[16] W.-T. Chu, W.-H. Cheng, S.-F. He, C.-W. Wang, and J.-L. Wu, "A unified framework using spatial color descriptor and motion-based post refinement for shot boundary detection," in Proc. of Pacific-Rim Conference on Multimedia, Tokyo, Japan, 2004.

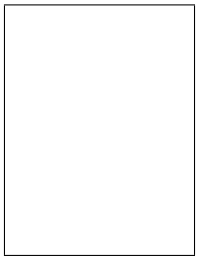
[17] S. Engel, X. Zhang, and B. Wandell, "Colour tuning in human visual cortex measured with functional magnetic resonance imaging," Nature, vol. 388, no. 6,637, pp. 68-71, July 1997.

[18] D. Bordwell and K. Thompson, Film Art: An Introduction. New York: McGraw-Hill, 2001.

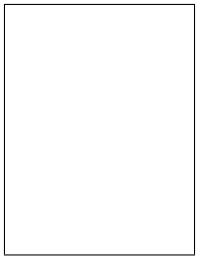
[19] C.-W. Ngo, T.-C. Pong, and H.-J. Zhang, "Motion analysis and segmentation through spatio-temporal slices processing," IEEE Trans. Image Processing, vol. 12, no. 3, pp. 341-355, March 2003.
[20] B. Jahne, Spatio-Temporal Image Processing: Theory and Scientific Applications. New York: Springer-Verlag, 1991.
[21] R. Paramesan, P. Ramaswamy, and S. Omatu, "Regular moments for symmetric images," IEE Electronics Letters, vol. 34, no. 15, pp. 1481-1482, July 1998.
[22] P.-C. Tseng, Y.-C. Chang, Y.-W. Huang, H.-C. Fang, C.-T. Huang, and L.-G. Chen, "Advances in hardware architectures for image and video coding - a survey," Proceedings of the IEEE, vol. 93, no. 1, January 2005.



Wen-Huang Cheng received the B.S. and M.S. degrees in computer science and information engineering from National Taiwan University, Taipei, Taiwan, R.O.C., in 2002 and 2004, respectively, where he is currently pursuing the Ph.D. degree in the Graduate Institute of Networking and Multimedia. His research interest includes multimedia data management and analysis.



Wei-Ta Chu received the B.S. and M.S. degrees in Computer Science and Information Engineering from National Chi Nan University in Nantou, Taiwan, in 2000 and 2002. He is currently pursuing his Ph.D. degree in the Department of Computer Science and Information Engineering, National Taiwan University, Taiwan. As he is in the Communication and Multimedia Laboratory, his research interests include digital content analysis, multimedia indexing, digital signal process, and pattern recognition.



Ja-Ling Wu received the B.S. degree in electronic engineering from TamKang University, Tamshoei, Taiwan, R.O.C., in 1979, and the M.S. and Ph.D degrees in electrical engineering from Tatung Institue of Technology, Taipei, Taiwan, in 1981 and 1986. Since 1987, he has been with the Department of Computer Science and Information Engineering, National Taiwan University, where he is presently a Professor. He has published more than 200 journal and conference papers. His research interests include algorithm design for DSP, data compression, digital watermarking and multimedia systems. Prof. Wu was the recipient of the Excellent Research Award from NSC, Taiwan, in 1999, 2001 and 2004.