A Generic Virtual Content Insertion System Based on Visual Attention Analysis

Huiying Liu[1, 2, 3, 4], Shuqiang Jiang[1, 2], Qingming Huang[1, 2, 3, 4], Changsheng Xu[4, 5]
[1]Key Lab of Intelligent Information Processing, Chinese Academy of Sciences (CAS), China
[2]Institute of Computing Technology, CAS, Beijing 100190, China
[3]Graduate University of Chinese Academy of Sciences, Beijing 100049, China
[4]China-Singapore Institute of Digital Media
[5]National Lab of Pattern Recognition, Institute of Automation, CAS, Beijing 100190, China
Email: {hyliu, sqjiang, qmhuang}@jdl.ac.cn, csxu@nlpr.ia.ac.cn

ABSTRACT

This paper presents a generic Virtual Content Insertion (VCI) system based on visual attention analysis. VCI is an emerging application of video analysis and has been used in video augmentation and advertisement insertion. There are three critical issues for a VCI system: when (time), where (place) and how (method) to insert the Virtual Content (VC) into the video. Our system selects the insertion time and place by performing temporal and spatial attention analysis, which predicts the attention change along time and the attended region over space. In order to enable the inserted VC to be noticed by audience while not to interrupt the audience's viewing experience to the original content, the VC should be inserted at the time when the video content attracts much audience attention and at the place where attracts less. Dynamic insertion is performed by using Global Motion Estimation (GME) and affine transformation. Our VCI system is able to obtain an optimal balance between the notice of the VC by audience and disruption of viewing experience to the original content. Extensive subjective evaluations based on user study on the VCI result have verified the effectiveness of the system.

Categories and Subject Descriptors

I.4.9 [**IMAGE PROCESSING AND COMPUTER VISION**]: Applications; I.2.10 [**ARTIFICIAL INTELLIGENCE**]: Vision and Scene Understanding –*Perceptual reasoning*

General Terms

Algorithms, Experimentation

Keywords

Virtual content insertion, Visual attention

1. INTRODUCTION

With the development of digital media and communication, there is an explosive growth in the amount of multimedia information in our daily lives. This trend necessitates the development of content-based video analysis, indexing and retrieval technologies, and triggers new applications such as Virtual Content Insertion (VCI). VCI is an emerging application of video analysis and has been used in video augmentation [1] and advertisement insertion [2-7]. Video augmentation improves the audience's viewing experience to the original content. Advertisement insertion provides much more advertising opportunity to the advertisers. Examples of VCI in broadcast video can be seen in Figure 1, in which the virtual contents are inserted into the video by editors. However, this is a time costing and labor intensive job for the huge amount of video data. To tackle this problem, automatic VCI approaches and systems have been studied in the past years. The existing work mainly focuses on sports video advertisement insertion for its large amount of audience and huge commercial profit [2-6]. Another reason is that there is plenty of domain knowledge available which can be used in VCI to determine the insertion time [2] and place [3, 8], and to calibrate the camera [1, 5, 6]. However, the existing work lacks generality and is difficult to be extended to other video types. This motivates us to construct a generic VCI system which can be applied to various video types.

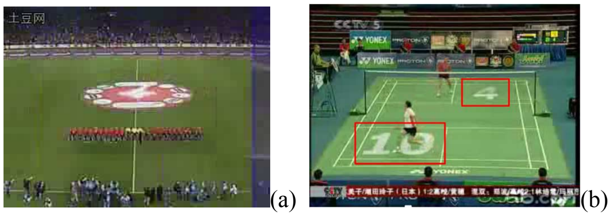


Figure 1. Examples of VCI. (a) A frame from soccer video; (b) A frame from a badminton game.

The general task of VCI is to make the inserted content more probable to be noticed by the audience and meanwhile not to interrupt the audience's viewing experience to the original content. To balance the two issues, the Virtual Content (VC) should be inserted at the time when the video attracts more audience attention [2] and at the region where attracts less [2, 4, 8]. Attention analysis, a hot research topic in recent years, can be used in this scenario. By performing attention analysis, Higher Attentive Shot (HAS) is detected as the insertion time and Lower Attention Region (LAR) is detected as the insertion place. Existing work about attention analysis is mainly for spatial attention. However temporal attention has rarely been researched. In this paper we propose a set of new methods to analyze visual attention, including both spatial and temporal attention.

Besides time and place, insertion method is also important for VCI system and may affect insertion result. Virtual content can be

MM'08, October 26–31, 2008, Vancouver, British Columbia, Canada.
Copyright 2008 ACM 978-1-60558-303-7/08/10...$5.00.

inserted into videos statically or dynamically. In static insertion VC flows over the original content (Figure 1 (a)). Dynamic insertion is less intrusive for fusing of the VC into the scene (Figure 1 (b)). To perform dynamic insertion in general video, a new method is proposed by using affine rectification and Global Motion Estimation (GME). This method needs only two pair of parallel lines, which are relatively easy to be obtained in most videos.

The novelty and contributions of our work can be summarized in the following aspects. 1) A generic virtual content insertion system is presented based on visual attention analysis for the first time. This system needs little domain knowledge and can be applied to general video. 2) HAS is detected as insertion time by using a new method of temporal attention analysis, which considers temporal context in analyzing the attention attracted by each frame/shot. 3) By using mosaic warping, we extend our previous work of LAR detection [8] to choose the appropriate place for dynamic insertion. Moreover, we propose a novel method for spatial attention analysis, which considers both saliency and novelty. Saliency is calculated by integrating contrast and information theory. Novelty is calculated as the distance between the prior and posterior data distribution. 4) Using affine transformation and GME, a novel method is proposed to perform dynamic insertion in general video.

The rest of the paper is organized as follows. In section 2, we review the related work on VCI and computational attention model. In section 3, the overview of the proposed system is illustrated. We present in section 4 the way to choose the insertion time, including the proposed temporal attention analysis method and HAS detection. The proposed spatial attention method and LAR detection are presented to choose the place for VCI in section 5. In section 6, the proposed dynamic insertion method is described. In section 7, experimental results on various video types are reported. We will conclude the paper with our future work in section 8.

2. RELATED WORK

2.1 Related Work on VCI

An important application of VCI is advertisement insertion with several approaches. A direct method is to insert an advertisement segment into the video streams [7]. The time of insertion is chosen as the point where the content discontinuity is high and attractiveness is low to avoid interrupting the audience's viewing experience. However, this method prolongs the original video by adding extra content to the video stream. At the same time, inserting ads at low attractive point may result in negligence of the ads. Another choice is to insert the VC into each frame, which is a more challenging task for the choosing of insertion time, place and insertion method.

For time choosing, an important factor to consider is to make the VC noticeable to the audience. Therefore the VC is usually inserted into video highlights as they may attract more viewer attention [2]. Besides highlights, the consecutive frames with little camera motion can also be selected as candidates to hold the VC for a period of time [4]. Highlight extraction usually needs domain knowledge while the frames with little camera motion cannot ensure the frames to be attended. In this paper, temporal attention is performed to extract the shots which attract more audience attention as the candidates for virtual content insertion.

The insertion place can be detected by using domain knowledge. For example, static region, goalmouth, central circle and boundary line of soccer video were detected to identify suitable locations for VCI [3]. More generic approaches include Visual Relevance Measure (VRM) [2] and Lower Informative Region (LIR) [4]. Both approaches do not need any domain knowledge and can be used in all types of videos. However, in these approaches only information theory is considered. Visual attention, an important mechanism of human visual system, is neglected. To detect the place for VCI, a notation of Lower Attentive Region (LAR) is proposed and defined, from the cognition point of view, as a region of the video frame which attracts less audience's attention [8]. LAR is detected by using visual attention analysis. In this paper, we adopt the detection method in our previous work [8] and improve it with mosaic warping to detect LAR for dynamic insertion.

As presented in section 1, the VC can be inserted into the video statically or dynamically. Static insertion is relatively simple, while dynamic insertion needs VC adaptation to fuse the VC into the scene. VC adaptation can be done using predetermined landmarks [9]. Or, if the model of the scene is available, it can be used to estimate camera parameters [1] and to distort the inserted VC [5, 6, 10]. However, the above methods can't be applied to general video for the lack of prior knowledge of the scene. In this paper, dynamic insertion is performed using GME and affine rectification.

2.2 Related Work on Attention Analysis

Computational visual attention has been a hot research topic for years and has wide applications in artificial intelligence, computer vision and multimedia. Visual attention includes two critical parts, spatial attention and temporal attention. Spatial attention studies where the audience will pay attention to the image/frame. Temporal attention considers when audience will pay more attention to the video content. In this section, the related work on spatial and temporal attention analysis will be reviewed.

2.2.1 Spatial Attention

Existing work on spatial attention analysis can be classified into two categories, static attention analysis and spatio-temporal attention analysis, according to whether temporal information is taken into account.

Static Attention: Saliency is widely used in computational attention analysis to evaluate the standing out of an event over its spatial context. The existing works can be classified into contrast based method and information theory based method according to physiological basis. Contrast based attention analysis [11-18] takes the notion that the center-surround structure of receptive field provides HVS sensitivity to feature contrast. Information theory based methods adopt the premise that visual attention proceeds entirely by maximizing the information sampled from an image [19, 20]. Contrast and information sampling are two factors used to evaluate saliency in computational visual attention. However, to the best of our knowledge, these two factors have not been integrated. In this paper, we propose a new method which considers both contrast and information theory. We also propose a new way of contrast calculation which is more coherent with HVS and we adopt a simple while effective method to calculate information.

Spatio-temporal Attention: Recently, several approaches have been proposed to analyze video attention by extending static attention into spatial-temporal domain. Besides static features, motion, which attracts much human attention and plays an important role in video analysis, is an effective feature to detect attentive regions in spatio-temporal cues. In [11, 12, 21], motion saliency was calculated to evaluate motion attention. However, spatio-temporal attention is beyond motion. In video, an object standing out of its temporal background usually attracts much audience attention. Novelty, corresponding to the spatial definition of importance through saliency, is used to evaluate the importance of a temporal event. It is usually calculated by using information theory [22-24]. Both static novelty and motion novelty can be used in attention analysis. For the lack of stability of motion estimation, we neglect motion novelty and perform spatio-temporal attention analysis with static novelty and motion saliency.

2.2.2 *Temporal Attention*

The above mentioned work analyzes spatial attention while temporal attention is rarely studied. While watching a video, audience will pay different amount of attention to the video content at different time. In existing work, the attention attracted by each frame is calculated by simply summing up the saliency of the pixels of attentive region [11, 25]. Camera motion, which is often used by film producer to guide audience attention, is also used to analyze temporal attention [11]. However, the audience attention attracted by a frame is not only determined by its own content or camera motion, but also by the temporal context, i.e., the preceding ones since temporal dimension is single-line. Generally speaking, the more different a frame is to the preceding ones, the more attention it will attract. Based on this idea, we evaluate the attention of each frame with its contextual difference, which is also referred as novelty.

3. SYSTEM OVERVIEW

Figure 2 illustrates the framework of our system. The system is composed of four modules, video preprocessing, HAS detection, LAR detection and virtual content insertion. In video preprocessing module, the input video is segmented into shots using the shot boundary detection method proposed in [26]. Within a shot, the content is of spatial and temporal continuum. Thus inserting the virtual content into each shot makes the insertion integral and real.

The HAS detection module determines the insertion time. It calculates the attentive value of each shot by performing temporal attention analysis and chooses the attentive ones for VCI.

The LAR detection module chooses the insertion place. For the candidate shots, spatial attention analysis is first performed to generate the attention maps. And then LAR is detected as the region of lower attention value. For the purpose of static insertion, the attention maps are simply averaged to generate a shot attention map, from which the region of lower attention value is detected as static LAR. While for the purpose of dynamic insertion, the attention maps are warped into a mosaic one using homography matrices obtained from GME. The dynamic LAR detected from the mosaic attention map is a region of the scene.

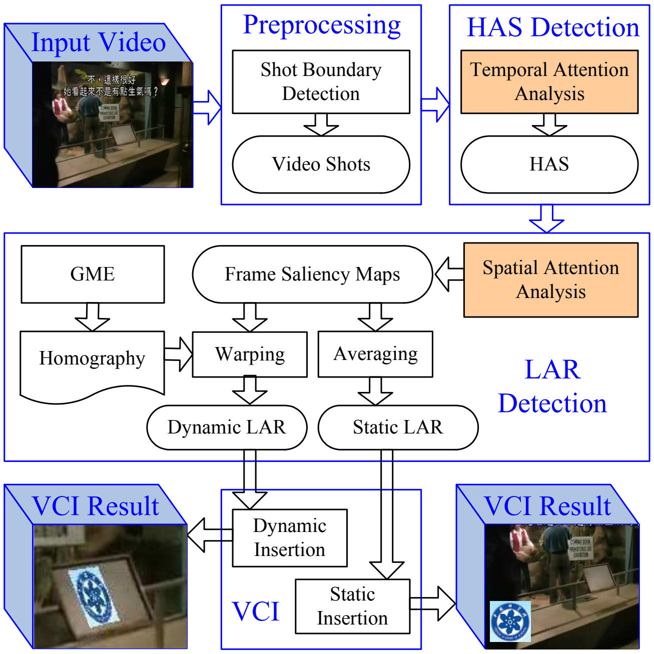


Figure 2. System Overview

The VCI module inserts the virtual content into the HAS at the LAR. For static insertion, the VC is resized to the LAR while for dynamic insertion, the camera parameters are needed to adapt the virtual content. If enough prior knowledge of the scene is available, camera calibration is performed to obtain camera parameters. Otherwise, two pairs of parallel lines around the LAR are detected to perform affine transformation, which provides the front view of the scene. The virtual content is resized to the front view and adapted with the homography matrices in the following frames.

4. HAS DETECTION

The VC inserted at the time when the video attracts more audience attention is more probable to be noticed and remembered by audience. In our work, temporal attention analysis is performed to evaluate the attention attracted by each shot. The shots of higher attention value are chosen as insertion time.

4.1 Temporal Attention Analysis

While watching a video, audiences pay different amount of attention to the video content at different time. Generally speaking, the frames different from the preceding ones attract more attention. Here we adopt again the notation of novelty to evaluate the attention of each frame as it is also along temporal dimension. In our work a frame's novelty is evaluated through its difference to its preceding ones. Let F_t be the *t th* frame of the video, its novelty is:

$$Nol\left(F_t\right)=\sum_{i=t-l}^{t-1} diff\left(F_t,F_i\right)w\left(F_t,F_i\right) \qquad (1)$$

where l is the length of context window which is set as 5 in our work. $diff\left(F_t, F_i\right)$ is the difference between the two frames, calculated through normalized color histogram:

$$diff\left(F_i, F_t\right) = 1 - \sum_{r=1}^{R}\sum_{g=1}^{G}\sum_{b=1}^{B}\min\left(H_i\left(r,g,b\right), H_t\left(r,g,b\right)\right) \tag{2}$$

where R , G and B are the numbers of histogram bins of R, G and B color channels. $w\left(F_t, F_i\right)$ is for the consideration that the nearer frames act more on the current one. It is a relative distance between two frames:

$$w\left(F_t, F_i\right) = \left(l - t + i + 1\right) \Big/ \sum_{j=1}^{l} j \tag{3}$$

An example of temporal attention result is illustrated in Figure 3. It is the result of a shot of “Adventure to the West”. Figure 3 (a) is the corresponding attention curve. It can be seen that the 16th frame, the red point on the curve, is of highest attention value. The frames from 12 to 17 are illustrated in Figure 3 (b). In frame 16, the sudden jumping out of Monkey King results in the high novelty of the frame.

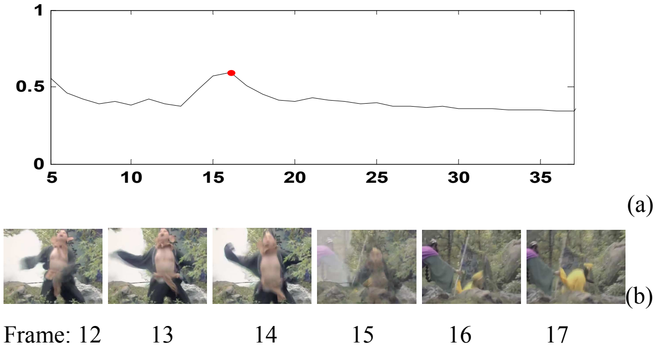


Figure 3. Temporal attention result. (a) Attention curve; (b) The frames of 12-17.

4.2 Shot Attention Evaluation

To choose insertion time for VCI, we evaluate the attention value of each shot instead of frame. Similar to that of frame, a shot’s attention value is determined by its novelty in temporal context. Shot novelty is calculated using the similar method with frame novelty. Let S_t be a shot of the video, its novelty is:

$$Nol\left(S_t\right) = \sum_{i=t-l}^{t-1} diff\left(S_i, S_t\right) w\left(S_i, S_t\right) \tag{4}$$

where the length of context window l is the number of shots included, which is also set as 5 in our work. The feature used here is normalized color histogram of shot, calculated by averaging the normalized frame histograms.

Different from a frame, a shot’s attention value is also determined by its length. Generally speaking, the longer a shot is, the more probable it is to be attended. In this regards, the attention value of a shot is $L_t \times Nol\left(S_t\right)$, where L_t is its length. An example of shot attention can be seen in Figure 4. Figure 4 (a) shows the video’s shot attention curve, from which it can be seen that the attention value of the 142th shot (the red star on the curve) is relatively high. The corresponding representative frames from 137th to 142th are illustrated in Figure 4 (b), from which it can be seen that the 142th shot is very different from the previous ones.

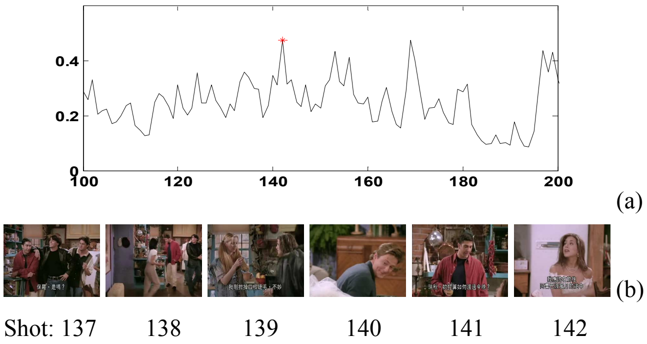


Figure 4. Shot attention result. (a) The shot attention curve of the video; (b) The representative frame of the shots 137-142.

5. LAR DETECTION

Insertion place is a critical issue of VCI as improper insertion may result in intrusion of the original content. In this section we detect LAR as the insertion place of VCI by analyzing spatial attention. To ensure the result to be consistent with audience vision, we also propose a novel method of spatial attention analysis.

5.1 Spatial Attention Analysis

The proposed spatial attention analysis method consists of static saliency, motion saliency and static novelty, as illustrated in Figure 5.

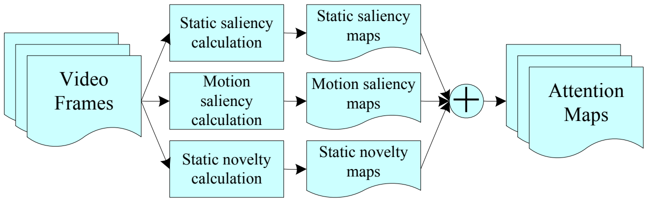


Figure 5. Framework of attention analysis

5.1.1 Saliency Calculation

Contrast and information are two main factors used in computational attention model. Contrast is an event’s standing out in its local context. Information is its importance in global background. While watching a scene, the two factors work together and determine the final attentive point. In our work, we integrate contrast and information and calculate saliency as their product:

$$Sal\left(x,y\right) = Con\left(x,y\right) ID\left(x,y\right) \tag{5}$$

where $Con\left(x,y\right)$ and $ID\left(x,y\right)$ are the contrast and information density of point $\left(x,y\right)$ respectively, normalized to $\left[0,1\right]$. The details of contrast and information density calculation will be presented in the following sections.

5.1.1.1 Contrast Calculation

As presented in Section 2, the center-surround structure of the receptive field provides HVS the sensitivity to contrast. The term receptive field refers to the specific receptors that feed into a given cell in the nervous system, with one or more synapses intervening. The structure of receptive field is verified to be an ellipse with its main axis 20° to the horizon and is modeled with Difference of Gaussian (DoG) [27]. For simplicity we adopt an isotropic model:

$$DoG(x,y)=\frac{1}{2\pi\sigma^2}\exp\left(-\frac{x^2+y^2}{2\sigma^2}\right)-\frac{1}{2\pi\lambda^2\sigma^2}\exp\left(-\frac{x^2+y^2}{2\lambda^2\sigma^2}\right) \quad (6)$$

where σ^2E and $\lambda^2\sigma^2E$ ($\lambda>1$, E is identity matrix) are the covariance matrices of the two Gaussians. The receptive field's center is:

$$center=\{(x,y)\mid DoG(x,y)\geq 0\}=\left\{(x,y)\mid (x^2+y^2)\leq\frac{4\sigma^2\ln\lambda}{1-1/\lambda^2}\right\} \quad (7)$$

The signal received by the cell is the convolution of the signal intensity $f(x,y)$ and $DoG(x,y)$:

$$s(x,y)=\iint f(u,v)DoG(u-x,v-y)dudv \quad (8)$$

Using the above idea, we evaluate feature contrast of an image I . Substituting the signal intensity in (8) with relative intensity, i.e., distance between image features, gives the contrast of a point:

$$Con(x,y)=-\sum\nolimits_{(u,v)\in I}d\left(f(x,y),f(u,v)\right)\times DoG(u-x,v-y) \quad (9)$$

where $d\left(f(x,y),f(u,v)\right)$ is the distance between two features and DoG is of discrete form.

In (9) there are two parameters, λ and σ , to be determined. The parameters can be set to satisfy:

$$\sum\nolimits_{(x,y)\in center}DoG(x,y)=\varepsilon \quad (10)$$

where ε is a predefined constant between 0 and 1. From (7) and (10) we obtain:

$$-\left(\lambda^2\right)^{-\frac{\lambda^2}{\lambda^2-1}}+\left(\lambda^2\right)^{-\frac{1}{\lambda^2-1}}=\varepsilon \quad (11)$$

In our work we set $\varepsilon=1/2$ with the result of $\lambda\approx 2.0984$. Let R be the center's radius (R is different from the radius of [28]), then from (7) we obtain $\sigma=0.5104R$.

While watching a scene, the size of receptive field changes adaptively according to the content of the scene [28]. While watching an image, the size of the image patch corresponds to the center of the receptive field, referred to as perceptive unit, also changes adaptively for the regulation of the eyeball's focal length. The choosing of perceptive unit will be presented in section 5.1.2.

5.1.1.2 Information Density Calculation

In existing work [20], information is calculated by computing a joint likelihood measure based on local statistics, which requires estimation of a $3\times M\times N$ dimensional probability density function for a local window size of $M\times N$ in RGB space. In the above method, the window size is predefined and keeps unchanged when analyzing an image. However, in our work, the perceptive unit changes adaptively to the context. So information density is adopted instead.

We firstly make a simplification that the pixels are independent to each other. This simplification has two advantages. One is that the information contained by each pixel can be calculated by using image histogram. Let H be the normalized color histogram of the image, the information contained by a pixel of color (r,g,b) is $-\log\left(H(r,g,b)\right)$. The other is that the information contained by each perceptive unit is simply the sum of the information contained by its pixels. Information density is calculated as follow:

$$ID(x,y)=\sum\nolimits_{(u,v)\in center}I(u,v)\times DoG(u-x,v-y) \quad (12)$$

where $I(u,v)$ is the amount of information contained in pixel (u,v) . Since $\sum\nolimits_{(u,v)\in center}DoG(u-x,v-y)$ is constant, $ID(x,y)$ provides us a good representation of information density.

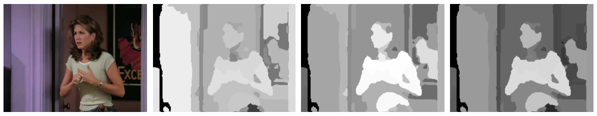
5.1.2 Static Saliency

The choosing of perceptive unit differentiates the existing attention analysis methods. The perceptive unit can be chosen as pixel [12, 13], image block [11, 14], region [15-17] or object [18]. A pixel/block contains little perceptive information. Comparatively, an object contains much perceptive information but is difficult to be obtained because object detection is still an open problem in the area of computer vision. In color images, an object is composed of one or more regions. In other words, a region is a unit between a pixel/block and an object. It contains more perceptive information than a pixel/block and can be obtained by image segmentation, which is much easier than object detection. So, in our work we adopt region as perceptive unit. This choice also enables the proposed method to analyze visual attention at multi-scales for the adaptive size of region.

Since our purpose is to obtain the image patches which can be used as perceptive unit, image segmentation is simplified by performing color quantization using K-Means. Then the neighboring pixels of same color are regarded as a region. From equation (5), (8) and (12) the saliency of each region can be calculated by using equation (13):

$$Sal(k)=\log p\left(f(k)\right)\sum_{i=1}^{K}d\left(f(k),f(i)\right)\times G_k(i,k) \quad (13)$$

where K is the total number of regions in the image. G_k is the DoG function of region k , of which the radius is the same with the region. $f(k)$ is the feature of region k and $p\left(f(k)\right)$ is its probability calculated by using the color quantization result. $d\left(f(k),f(i)\right)$ is the distance between features. It is evaluated in our work by using Gaussian distance. An example of static saliency calculation is illustrated in Figure 6. Figure 6 (b) and (c) show the contrast map and information map. Figure 6 (d) shows the saliency map.



(a) (b) (c) (d)

Figure 6. Spatial attention result. (a) Input image; (b) Contrast map; (c) Information map; (d) Saliency map

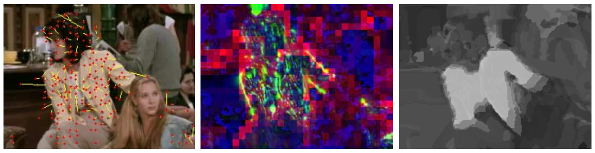
5.1.3 Motion Saliency

Motion is widely used to detect attentive region in spatio-temporal cues [11, 12, 21-23]. Motion vector can be obtained by several methods such as optical flow. However, a critical issue is that motion estimation under moving camera is still a challenging problem and the motion vector obtained is not so reliable. In this paper, a cone-shaped Motion Vector Space (MVS) [29] is adopted to alleviate the negative impact caused by camera motion. This method transforms the MVS to HSV color space as follow:

$$Angle \rightarrow H; Magnitude \rightarrow S; Texture \rightarrow V \qquad (14)$$

where the motion magnitude and the texture are normalized to $[0, 255]$. The selection of texture as value, which follows the intuition that a high-textured region produces a more reliable motion vector, provides this method a significant advantage that when the motion vector is not reliable for camera motion, the V component can still provide a good presentation of the frame.

After transforming the MVS to HSV color space, motion saliency can be calculated using the segmentation result of section 5.1.2 and equation (13). An example of MVS and motion attention is illustrated in Figure 7. Figure 7 (b) shows the result HSV image, in which the motion vector is presented intuitively. Figure 7 (c) is the corresponding motion saliency map.



(a) (b) (c)

Figure 7. Motion attention result. (a) Input image with MVF; (b) HSV image; (c) Motion saliency map

5.1.4 Static Novelty

Besides motion saliency, novelty, an event's standing out of its temporal background, also affects audience attention. Itti [22, 23] measured novelty by using information theory. The information carried by data is measured as the difference between prior and posterior distributions over the set of all models. KL divergence is used to calculate the difference. We also adopt an information theory based method to evaluate novelty in videos. Similar to Itti's work, we calculate the distance between the prior and the posterior distributions as the novelty of each event. Different from Itti's work, we model the original feature of the video instead of the center-surround feature maps. Supposing that M_{t-1} and M_t are data models at $t-1$ and t respectively, the novelty at t is calculated by using KL distance:

$$Nol(t) = KL(M_{t-1}, M_t) = \int_{X} M_t(x) \log \frac{M_t(x)}{M_{t-1}(x)} dx \qquad (15)$$

We adopt Gaussian distribution to model the data at each position. At time t the data is presented as:

$$M_t \sim N(\mu_t, \sigma_t^2) \qquad (16)$$

where $\mu_t = \sum_{i=1}^{t} x_i \Big/ t$, $\sigma_t^2 = \sum_{i=1}^{t} (x_i - \mu_t)^2 \Big/ (t-1)$.

However, there is a problem that the data accumulation with time may decrease the model's sensitivity to data change. To avoid this problem, the model is reset at the beginning of each shot. An example of novelty attention result is illustrated in Figure 8, which shows three frames (7th, 10th and 19th) of a shot from "Adventure to the West" and their corresponding novelty maps. In the 7th frame, the moving of Monkey King (the region with red bounding box) results in the high novelty of the corresponding region. In the 10th frame, the sudden appearance of the monster (the region with yellow bounding box) is of highest novelty. However, the monster is stationary, while Monkey King keeps moving. Therefore, with time going, the novelty of the monster decreases while the novelty of Monkey King increases, as in the 19th frame.

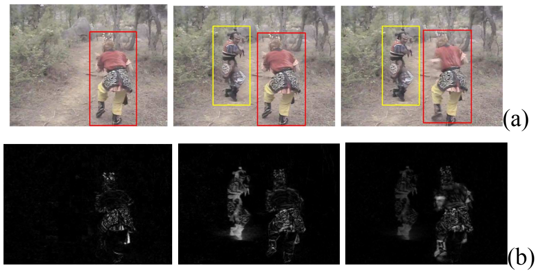


Figure 8. Novelty Calculation result. (a) Original frames; (b) Corresponding novelty maps.

5.1.5 Map Fusion

After attention analysis, we obtain three saliency/novelty maps including static saliency map (M_S), motion saliency map (M_M) and novelty map (M_N). Suitable fusion of the maps produces the final attention map. Map fusion can be performed with linear [14] and nonlinear [11] methods. In our work we adopt linear method for simplicity and with adaptive coefficients to fit different types of videos.

Considering that our goal is to detect Region of Interest (ROI) or LAR from the saliency/novelty maps, we model this progress as binary classification and use the variance between classes to determine the fusion coefficients. M_S is first classified with the method of maximum variance between classes. Let Var_S be the result variance between classes:

$$Var_S = \left(n_1 (\mu_1 - \mu)^2 + n_2 (\mu_2 - \mu)^2 \right) \Big/ n \qquad (17)$$

where n_1 and n_2 are the number of samples of the two classes, n is the total number of samples. μ_1 and μ_2 are the means of two classes, μ is the mean of all the samples. Let Var_M and Var_N be the variances between classes of M_M and M_N respectively. Then the fusion weight for static saliency maps is

$$w_S = Var_S \Big/ \left(Var_S + Var_M + Var_N \right) \qquad (18)$$

The weights for motion saliency map and novelty map are similarly calculated. The final attention map is:

$$AM = w_S M_S + w_M M_M + w_N M_N \qquad (19)$$

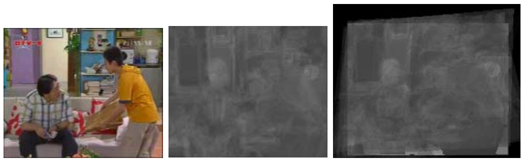
5.2 LAR Detection

LAR is defined as the region that attracts less audience attention [8]. It is detected by using spatial attention analysis. Corresponding to the insertion methods, the LAR is also classified into static LAR and dynamic LAR. A static LAR is a region of the same position on each frame, while a dynamic LAR is a region of the same position in the scene. To detect LAR, the frame attention maps of a shot are fused into an integrated one. For static LAR, simple averaging is enough. While for dynamic LAR it is much more complex for the existence of camera motion. So mosaic image stabilization is adopted to present the appearance of a locked down camera.

The homography matrices between consecutive frames are obtained by using Global Motion Estimation (GME) [30]. Let $H_{t,t-1}$ be the homography between frame t and $t-1$. Then for a point P_t in frame t, its corresponding point in frame $t-1$ is $P_{t-1} = H_{t,t-1}P_t$. Here P_t and P_{t-1} are the aligned coordinates of the points. Then for a point P_M in the mosaic attention map, its attention value is calculated as:

$$MAM\left(P_M\right) = \sum_{t=1}^{L} w\left(t\right) AM_t\left(P_t\right) \tag{20}$$

where $P_t = \prod_{k=1}^{t} H^{-1}_{k,k-1} P_M$ is P_M 's corresponding point in frame t. L is the shot's length. $w\left(t\right)$ is the weight of the t *th* frame and can be set as $1/L$. It can also be determined by the frame's attention value from temporal attention analysis. The higher its attention value is, the bigger its weight is. The regions of lower attention value on the mosaic attention map are chosen as the candidate LAR. Figure 9 shows an example of LAR detection. In the shot, the camera is moving and the mosaic attention map is shown in Figure 9(c). The corresponding static attention map, which is obtained by simple averaging, is shown in Figure 9(b).



(a) (b) (c)

Figure 9. Example of mosaic saliency map. (a) Key frame; (b) Shot saliency map; (c) Mosaic saliency map;

6. VIRTUAL CONTENT INSERTION

As presented in Section 1, VC can be inserted into video statically or dynamically. In dynamic insertion the content is merged into the background. Generally speaking, static insertion applies to video of complex scene, for example natural scene, and dynamic insertion applies to man-made scenes such as sports video and indoor video. Dynamic insertion needs to adapt the virtual content to camera parameters.

The existing work performs dynamic insertion in sports video as the structure of the playfield can be used to calibrate camera. However, for general video there is too little prior knowledge to perform camera calibration. In this paper a novel method is proposed to insert VC with as little as scene information by using affine transformation. To insert the virtual content into the scene with reality, affine rectification is first performed to obtain the front view of the scene. The virtual content is inserted into the front view and then adapted to the following frames using the affine matrix and the homography matrices. In our work, the affine matrix is obtained by using two pairs of parallel lines. The method is briefly presented here and the details can be referred in [31].

Suppose l_1 and l_2 are a pair of lines in the frame, corresponding to a pair of spatial parallel lines L_1 and L_2.

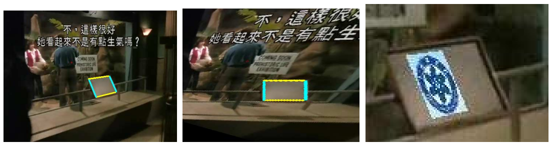
$$l_i = a_i x + b_i y + c_i, i = 1,2 \tag{21}$$

If l_1 and l_2 are not parallel in the image plane, the vanishing point is v_1 with aligned coordinate $V_1 = \left(x_1 \quad y_1 \quad f\right)^T$. If l_1 and l_2 are parallel in the image plane, the vanishing point is infinite and its aligned coordinate can be written as $V_1 = \left(a_1 \quad -b_1 \quad 0\right)^T$. The normalized spatial direction of the lines L_1 and L_2 is $V_1/\left|V_1\right|$, represented as $r_1 = V_1/\left|V_1\right|$. Using the other pair of parallel lines we can obtain $r_2 = V_2/\left|V_2\right|$. Then the normal of the spatial plane is calculated as $r_3 = r_1 \times r_2$. From the three vectors we obtain the affine matrix $A = \left(r_1 \quad r_2 \quad r_3\right)$. Then for a point of aligned coordinate P on the original image, its corresponding pixel on the affined image is $P_A = P_0 - A^{-1}P_0 + \alpha A^{-1}P$. Here the coefficient α is set to make $P_A\left(3\right)$ equal to f. The focal length f can be obtained using camera calibration or be set to 1 for simplification. P_0 is the point kept unchanged in affine rectification.

For a point P_A on the front view, its position on the t *th* frame is calculated as follow:

$$P_t = \prod_{k=0}^{t-1} H_{k,k+1} \times \left(AP_0 + P_0 - AP_A\right) \tag{22}$$

where A is the affine matrix, $H_{k,k+1} = H^{-1}_{k+1,k}$ is the homography matrix. An example of dynamic insertion is illustrated in Figure10. Figure 10 (a) is the original frame with two pair of parallel lines marked as yellow and green respectively. (b) is the affined result in which the front view of the marked region is obtained. In (c) the VC is inserted into the frame by using the affine matrix.



(a) (b) (c)

Figure 10. Example of dynamic insertion. (a) Original frame; (b) Affine result of (a); (c) VCI result.

7. EXPERIMENTAL RESULTS

Both visual attention analysis and virtual content insertion are subjective tasks and there is not a widely accepted method to evaluate the performance. For such work user study is an effective way to evaluate the performance subjectively. So in this paper we also adopt user study to evaluate our system. We invited 16 users to evaluate our experimental results. The users are postgraduate students aging from 22 to 30 years old. They are all audience of the testing videos and familiar with the VC used in the experiment.

7.1 Data Set

Testing video: We applied our system on different types of videos to verify its generality. The dataset, detailed in Table 1, includes situation comedy, outdoor TV play series and interview video. The situation comedy includes American soap opera “Friends” (Friends.rmvb), and Chinese TV play series “There are Children at Home” (Children.rmvb). The outdoor TV is “Adventure to the West” (Adventure.rmvb). “A Date with Lu Yu” (LuYu.rmvb) is the interview video adopted. Among the datasets, situation comedy and interview video are indoor video and we perform either static or dynamic VCI on them. Outdoor teleplay is in natural scene and so static VCI is more suitable. The detail of the testing videos is given in Table 1. The total length of testing video is 72:50, consisting of 800 shots.

Table 1. Testing videos

No.	Video	Genre	Shot	Time
1	Friends.rmvb	situation comedy	200	11:25
2	Children.rmvb	situation comedy	200	14:48
3	LuYu.rmvb	Interview	200	20:49
4	Adventure.rmvb	Outdoor teleplay	200	25:48
Sum	----	----	800	72:50

Virtual Content: We choose 60 famous brands as VC including car brand such as Benz, noshery brand such as Mcdonald's, and other brands with which users are familiar. We adopt these brands for the consideration that it should be easier for the users to choose the ones he/she has noticed in the video. Another reason is to eliminate the noticing difference caused by the brands themselves. In the experiment, each shot is inserted, at LAR, with a unique VC randomly chosen from the VC database.

7.2 HAS Evaluation

We first test the proposed temporal attention and HAS detection method. The users are invited to watch the result videos. Each time, a video segment consisting of 50 shots is played. After playing, the users are shown to a set of brands, which includes the ones inserted into the video. The users are required to choose the brands he/she has noticed in the video. The noticing rate of each brand (i.e., each shot, since each shot corresponds to a unique brand), is calculated as n_i/N, where n_i is the number of users who identified the brand. N is the total number of users. Then a noticing curve is obtained. The attention curve and noticing curve of a segment of “Adventure to the West” are illustrated in Figure 11 (a). It can be seen intuitively that the attention curve is consistent with the noticing curve, meaning that the result of temporal attention analysis is accordant with audience attention.

To evaluate the result of temporal attention analysis quantitatively, we define its consistency as the similarity between attention curve and noticing curve:

$$cos = \sum\nolimits_k \min\left(AC(k), NC(k)\right) \qquad (23)$$

where AC and NC are normalized attention curve and noticing curve, respectively. The average consistency values of the 4 video clips are 0.75, 0.79, 0.84 and 0.82, respectively. Besides consistency, we also studied the relationship between noticing rate and attention value. We divide the attention value into 10 equal intervals and calculate their average noticing rate. The result is shown in Figure 11 (b). It can be seen that the noticing rate increases with shot attention value. From the result, it can be concluded that the proposed temporal attention method and the HAS detection method are effective and consistent with audience attention.

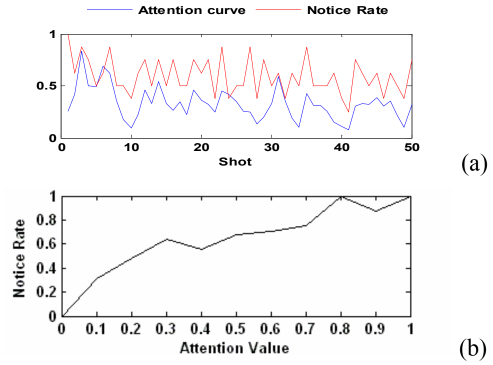


Figure 11. HAS evaluation result. (a) Attention curve and noticing curve; (b) Noticing rate to attention value.

7.3 LAR Evaluation

To evaluate the proposed spatial attention and LAR detection method, we invited the users to evaluate the brands he/she has noticed in the video. The users are required to give each brand an assessment, GOOD, NEUTRAL or BAD, according to whether the inserted brand has interfered with their viewing experience on the original video clip. For a video, we calculate the rate of GOOD as follow:

$$Rg = \sum_{i=1}^{n} Ng_i \Big/ \sum_{i=1}^{n} N_i \qquad (24)$$

where n is the total number of users and $n = 16$. N_i is the number of brands identified by user i and Ng_i is the number of brands assessed as GOOD by the same user. The rates of NEUTRAL (Rn) and BAD (Rb) are calculated similarly. The result is shown in Table 2, from which it can be seen that the average rate of “GOOD” is around 70 percent and the rate of “BAD” is below 10 percent. The variance is used to evaluate the differences of the collected data. The low variances of Rg , Rn and Rb means that the difference between the data is not significant. The result is relatively reliable under limited data set.

The promising result verifies the effectiveness of the proposed spatial attention analysis and the LAR detection methods. Some of the static insertion result is illustrated in Figure 12 and dynamic insertion will be presented independently in section 7.4. Figure 12 (a) and (b) show examples of successful result while (c) and (d) show examples of failed result. The reason for failure is that the static feature in our experiment is only *RGB* value. Sometimes the color of human faces is not distinct from the background and the regions are detected as LAR and occluded by the inserted content. This problem can be solved by integration of more features such as texture and shape. Moreover, semantic features such as human face and human body, can also be used to improve the performance of our system.

Table 2. Evaluation result on VCI

Video	Good(*Rg*)	Neutral(*Rn*)	Bad(*Rb*)
1	72.25	19.13	8.62
2	70.87	23.13	6.00
3	66.25	25.00	8.75
4	70.38	25.62	4.00
Mean/Var	69.94/6.67	23.22/8.55	6.84/5.20

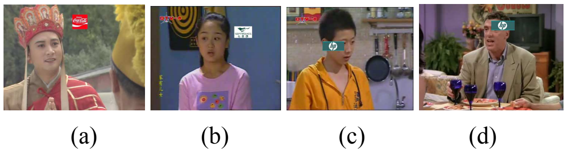


Figure 12. Examples of static VCI. (a, b) Successful results; (c, d) Failure results.

7.4 Evaluation of Dynamic Insertion

Based on the common sense that in man-made scene there are usually many straight lines, we use the line number to choose the candidate shots for dynamic VCI. We first detect lines using Hough transform on the first frame of each shot. If enough lines are detected, the shot is chosen as candidate. Then from the mosaic saliency map, the LAR is detected and the lines around the LAR are used in affine transformation.

Six results of dynamic VCI are evaluated. In the results there are camera motion and scene deformation caused by view angle. The users are requested to score the results based on the following criteria. (1) Is the result's deformation consistent with the scene? (2) Does the inserted VC follow the camera motion? (3) To what degree the user is satisfied with the result? The scores are scaled from 1 to 5 to represent the satisfactory degree with 1 being not satisfying at all and 5 being very satisfying. Each result's mean and variance are illustrated in Figure 13. All the variances are below 1, showing that the user feedback is reliable. Among the results, five means are above 3 with two of them above 4, which verify the effectiveness of the proposed method. Some of the examples are illustrated in Figure 14, in which an example of satisfying result and an example of failed result are shown in the first and second row respectively. In the failed result, GME provides an unreliable output which makes the VC departure from its original position.

The experimental results on various types of video demonstrate that the proposed system applies to general video and works effectively. The HAS and LAR evaluations show that the application of temporal and spatial attention analysis in VCI can increase the probability of the inserted content to be noticed by audience and decrease the damage to the original content. However, the system can still be improved in the following two aspects. Firstly, integrating of more features can improve the performance of the system. Secondly, better GME result can improve the result of dynamic insertion.

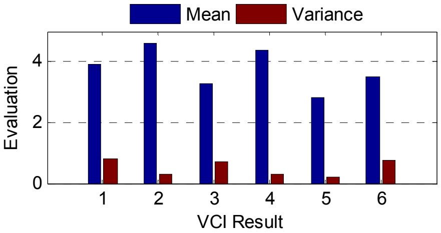


Figure 13. The mean and variance of the user study result.

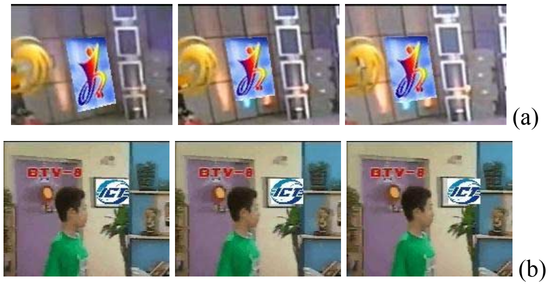


Figure 14. Examples of dynamic VCI. (a) Example of satisfying result; (b) Example of failure result.

8. CONCLUSIONS

Based on visual attention analysis, a generic VCI system is presented in this paper. The system determines insertion time by detecting HAS with temporal attention analysis, and determines insertion place by detecting LAR with spatial attention analysis. By inserting VC into the attentive shots at LAR our system balances between the notice of the VC by audience and disruption of viewing experience to the original content. Another problem of VCI is the insertion method. In the proposed system, dynamic insertion is performed by using affine transform and GME. This method needs little prior knowledge and can be implemented by two pairs of parallel lines. The performance of the system has been tested on several types of videos, including situation comedy, interviewing TV shows and outdoor TV play series. The experimental results have verified the effectiveness of the system.

However, the system can still be improved. In fact, virtual content itself also plays a role in the effect of VCI. It should be coherent with its spatial and temporal context to improve the insertion effect and reduce intrusion [7]. This will be one of our next-step works in the future. Besides, in our work, the static feature is only *RGB* value. We believe that integrating of more features can improve the performance of the system. In our future work, we will take shape perception into account to analyze visual attention.

9. ACKNOWLEDGEMENT

This work was supported in part by National Natural Science Foundation of China under Grant 60702035 and 60773136, in part by National Hi-Tech Development Program (863 Program) of China under Grant 2006AA01Z117 and Science100 Plan of Chinese Academy of Sciences: 99T3002T03. The authors also gratefully acknowledge the support of K. C. Wong Education Foundation, Hong Kong.

10. REFERENCES

[1] X. Yu, X. Yan, T. T. P. Chi and L. F. Cheong, “Inserting 3D Projected Virtual Content into Broadcast Tennis Video”, *Proceedings of the 14th ACM international conference on Multimedia*, pp: 619-622, 2006.

[2] K. Wan, C. Xu, “Automatic Content Placement in Sports Highlights”, *IEEE International Conference on Multimedia & Expo*, pp: 1893-1896, 2006.

[3] C. Xu, K. W. Wan, S. H. Bui, Q. Tian, “Implanting Virtual Advertisement into Broadcast Soccer Video”, *Pacific-Rim Conference on Multimedia*, pp: 264-271, 2004.

[4] Y. Li, K. Wah Wan, X. Yan, C. Xu, “Real Time Advertisement Insertion in Baseball Video Based on Advertisement Effect”, *Proceedings of the 11th ACM international conference on Multimedia*, pp: 343-346, 2005.

[5] M. Tamir, et al., “Method and Apparatus for Automatic Electronic Replacement of Billboards in a Video Image”, US Patent 6,292,227, 2001.

[6] S. Deshpande, et al., “Method and Apparatus for Including Virtual ADs in Video Presentations”, US Patent, 7,158,666, 2007.

[7] T. Mei, X-S. Hua, L. Yang, S. Li, “VideoSense-Towards Effective Online Video Advertising”, *16th ACM International Conference on Multimedia*, pp: 1075-1084, 2007.

[8] H. Liu, S. Jiang, Q. Huang and C. Xu. “Lower Attentive Region Detection for Virtual Content Insertion”, *IEEE International Conference on Multimedia & Expo*, 2008.

[9] S. Das, et al., “Method of Tracking Scene Motion for Live Video Insertion Systems”, US Patent 5,808,695, 1998.

[10] Kreitman, et al., “Method and System for Perspectively Distorting an Image and Implanting Same into a Video Stream”, US Patent 5,731,846, 1998.

[11] Y-F. Ma, X-S. Hua, L. Lu, H-J. Zhang. “A Generic Framework of User Attention Model and Its Application in Video Summarization”, *IEEE Trans on Multimedia*, Vol. 7, No. 5, pp: 907- 919, 2005.

[12] Y. Zhai, M. Shah. “Visual Attention Detection in Video Sequences Using Spatiatemporal Cues”. *Proceedings of the 14th annual ACM international conference on Multimedia*, pp: 815-824, 2006.

[13] O L. Meur, P L. Callet, D. Barba, D. Thoreau, “A Coherent Computational Approach to Model Bottom-up Visual Attention”, *IEEE Trans on Pattern Analysis and Machine Intelligence,* Vol. 28, No. 5, pp: 802-816, May 2006.

[14] L. Itti, C. Koch, E. Niebur. “A Model of Saliency-Based Visual Attention for Rapid Scene Analysis”, *IEEE Trans on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 11. pp: 1254-1259, 1998.

[15] H. Liu, S. Jiang, Q. Huang, C. Xu, and W. Gao. “Region-Based Visual Attention Analysis with Its Application in Image Browsing on Small Displays”. *15th ACM International Conference on Multimedia*, pp: 305-308, 2007.

[16] Y. Hu, D. Rajan and L-T Chia. “Robust subspace analysis for detecting visual attention regions in images”, *Proceedings of the 13th annual ACM international conference on Multimedia*, pp: 716-724, 2005.

[17] Y. Li, Y-F. Ma, H-J. Zhang. “Salient Region Detection and Tracking in Video”, *International Conference on Multimedia and Expo,* Vol. 2, pp: 269-272, 2003.

[18] Y. Sun and R. Fisher. “Object-based Visual Attention for Computer Vision”, *Artificial Intelligence,* Vol. 146, No. 1, pp: 77-123, 2003.

[19] T.N. Topper, “Selection Mechanisms in Human and Machine Vision”, Ph.D. Thesis, University of Waterloo, 1991.

[20] N. D. B. Bruce, “Features That Draw Visual Attention: an Information Theoretic Perspective”, *Neurocomputing*, Vol. 65-66, pp: 125-133, 2005.

[21] M.T. Lopez, A. Fernandez-Caballero, M. A. Fernandez, J. Mira, A. E. Delgado, “Visual Surveillance by Dynamic Visual Attention Method”, *Pattern Recognition*, Vol. 39, pp: 2194-2211, 2006.

[22] L. Itti, P. Baldi, “A Principled Approach to Detecting Surprising Events in Video”, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* Vol. 1, pp: 631-637, 2005.

[23] L. Itti. “Automatic Foveation for Video Compression Using a Neurobiological Model of Visual Attention”, *IEEE Transactions on Image Processing,* Vol. 13, No. 10, pp: 1304-1318, 2004.

[24] G. Qiu, X. Gu, Z. Chen, Q. Chen and C. Wang, “An Information Theoretic Model of Spatiotemporal Visual Saliency”, *IEEE International Conference on Multimedia & Expo*, pp: 1806-1809, 2007.

[25] J. You, G. Liu, L. Sun, H. Li, “A Multiple Visual Models Based Perceptive Analysis Framework for Multilevel Video Summarization”, *IEEE Trans. Circuits and Systems for Video Technology*, Vol. 17, No. 3, pp: 273-285, 2007.

[26] C. Liu, H. Liu, S. Jiang, Q. Huang, Y. Zheng and W. Zhang: “JDL at Trecvid 2006 Shot Boundary Detection”, TRECVID 2006 Workshop (2006).

[27] R. E. Soodak, “Two-dimensional modeling of visual receptive fields using Gaussian subunits”, *Proc. Antl. Acad, Sci. USA*, Vol. 83, pp: 9259-9263, 1986.

[28] E. T. Rolls, N. C. Aggelopoulos, and F. Zheng, “Effective size of receptive fields of inferior temporal visual cortex neurons in natural scenes”, *The Journal of Neuroscience*, Vol. 23, No. 1, pp:339-348, 2003.

[29] L-Y. Duan, M. Xu, Q. Tian, C-S. Xu, J. S. Jin, “A Unified Framework for Semantic Shot Classification in Sports Video”, *IEEE Trans on Multimedia*, Vol. 7, No. 6, pp: 1066-1083, 2005.

[30] F. Dufaux and J. Konrad, “Efficient, robust, and fast global motion estimation for video coding,” *IEEE Trans. Image Processing*, vol. 9, no. 3, pp. 497-501, 2000.

[31] X. Chen, J. Yang, J. Zhang, and A. Waibel. “Automatic Detection and Recognition of Signs from Natural Scenes”, *IEEE Transactions on Image Processing*, Vol. 13, No.1, pp: 87-99, 2004.