Multimedia Systems (2012) 18:201–214
DOI 10.1007/s00530-011-0234-0

REGULAR PAPER

@ICT: attention-based virtual content insertion

Huiying Liu · Qingming Huang · Changsheng Xu ·
Shuqiang Jiang

Received: 18 August 2010 / Accepted: 1 April 2011 / Published online: 28 April 2011

Abstract In this paper, we propose an attention-based virtual content insertion solution, called @ICT. Virtual content insertion (VCI) is an emerging application of video analysis and has been used in video augmentation and advertisement insertion. An ideal VCI solution should make the inserted virtual content being noticed by audiences and at the same time should not interfere with audiences' viewing experience on the original content. To balance these two conflicting issues, meaning high attention and low intrusiveness, we choose higher attentive shots as insertion time while determine insertion place and content interdependently by considering lower attention together with visual consistency. We also propose a measurement of intrusiveness from the viewpoint of visual attention. Furthermore, @ICT includes an in-scene insertion module, which embeds the virtual content into the videos with higher vividness and lower intrusiveness. @ICT is able to obtain an optimal balance between the noticing of the virtual content by audiences and disruption of viewing experience to the original content. It needs little prior knowledge and is applied to general videos. Extensive quantitative and qualitative evaluations on the VCI result have verified the effectiveness of the solution.

Keywords Video content analysis · Virtual content insertion · Visual attention

1 Introduction

Virtual content insertion (VCI) is an emerging application of video analysis and has been applied in video augmentation, to improve the audiences' viewing experience to the original content [33], and advertisement insertion to provide more advertising opportunities to the advertisers [1, 11, 20, 21, 32]. Manual insertion is a time-consuming and labor-intensive work for the huge amount of video data. To tackle this problem, automatic VCI approaches and systems have been studied in the past years.

The challenge of VCI is to balance its two conflicting tasks, which are to make the inserted content more probable to be noticed by the audiences and meanwhile not to interfere with the audiences' viewing experience on the original content. A conventional method is to insert the virtual content (VC), usually advertisement, at the beginning or the end of a video. But it is believed that the advertisement should be inserted at appropriate positions within video streams. Regarding this, VideoSense inserts advertisements at the time of higher discontinuity and lower attractiveness to avoid disturbing the audience from watching the video [21]. It chooses advertisement clips under the principle of textual and visual-aural consistency to improve the advertising effect.

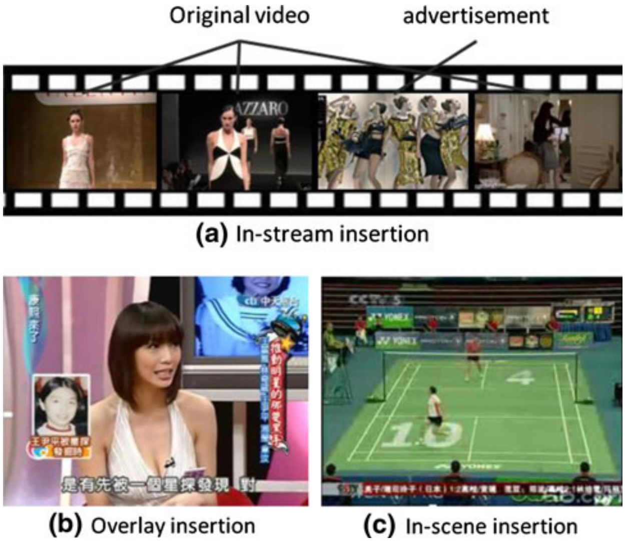
In the aforementioned methods, the advertisement clips are inserted into the video stream, which is referred to as in-stream insertion. Figure 1a illustrates an example of

Communicated by Thomas Haenselmann.

H. Liu (✉) · S. Jiang
Key Lab of Intelligent Information Processing,
Institute of Computing Technology, Chinese Academy
of Sciences, 100190 Beijing, China
e-mail: hyliu@jdl.ac.cn

Q. Huang
Graduate University of Chinese Academy of Sciences,
100049 Beijing, China

C. Xu
National Lab of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, 100190 Beijing, China



(a) In-stream insertion

(b) Overlay insertion

(c) In-scene insertion

Fig. 1 Examples of virtual content insertion. Overlay insertion and in-scene insertion belong to in-video insertion

in-stream insertion. Another choice is to insert the virtual content into the video frames, referred to as in-video insertion. In-video insertion can be overlay or in-scene insertion. In overlay insertion, the virtual content flows over the original content, as shown in Fig. 1b. In-scene insertion embeds the VC into the video scene, as shown in Fig. 1c. In-video insertion has two advantages over in-stream insertion. First, in-video insertion does not increase the video length while in-stream insertion does. Second, under in-video insertion, the inserted contents cannot be avoided without loss of the original video and so are more probable to be noticed by audiences. In spite of the two advantages, in-video insertion is more challenging for the risk of annoying the audiences. This risk is usually reduced through insertion time, insertion place, insertion content, and insertion method. Existing approaches of in-video insertion mainly focus on sports video for the reason that there is plenty of domain knowledge available which can be used in VCI to determine the insertion time [30] and place [32], and to calibrate the camera [3, 27, 33]. These methods rely on domain knowledge of sports video, such as the structure of the play field. They lack generality and are difficult to be extended to other video types. Visual attention analysis is a rational way to generalize virtual content insertion. AdOn [22] is an instance of visual attention-based generic contextual in-video advertising system. It chooses attractive shots as insertion time, less attentive region as insertion place, and chooses advertisement according to textual relevance, user preference, and visual content consistency.

In this paper, we concentrate on in-video insertion and aim to construct a generic virtual content insertion solution that is applied to general videos without specific domain knowledge. We balance the two conflicting tasks of VCI by taking human visual characteristics into account and propose an *a*ttention based virtual content insertion solution, called @ICT. To increase the chance of the inserted content to be attended, @ICT chooses Higher Attentive Shots (HAS) as insertion time. It reduces the intrusiveness caused by insertion through insertion place and insertion content choosing. However, there is not a measurement to evaluate intrusiveness. In this paper, we propose a measurement of intrusiveness from the viewpoint of visual attention. The measurement covers two aspects, ROI interference and distraction. ROI interference is caused by occluding the region of interest (ROI). Distraction usually happens when the inserted content outstands of the original content. To decrease the intrusiveness caused by insertion, @ICT determines insertion place and insertion content interdependently. To avoid ROI interference, it first detects lower attentive regions (LAR). Then for a particular VC, within the LAR, it chooses the position which is most consistent with the VC to decrease distraction.

@ICT includes both overlay and in-scene insertion. To perform in-scene insertion in general videos, we propose a new method by using affine rectification and camera tracking. This method needs only two pairs of parallel lines, which are relatively easy to be obtained in most videos of artificial locations, e.g., indoor or urban location. For in-scene insertion it is necessary to find the homography matrix between frames, which usually fails under fierce camera motion. We tradeoff between insertion effect and audience attention and choose the shots of little camera motion as insertion time. For in-scene insertion, we detect dynamic LAR as insertion place.

Part of our work has been published in [15]. Compared with our previous work, three major improvements have been made in this paper:

1. Besides insertion time, place, and method, insertion content choosing is also investigated. The general consideration is that the inserted content should change audience attention as little as possible. So we choose the insertion content according to visual consistency.
2. We propose a measurement of intrusiveness, which covers both ROI interference and distraction.
3. The in-scene insertion method is enhanced by using a more reliable camera motion method. In experiment, we found that the error of global motion estimation (GME) accumulates with time and results in virtual content displacement. In this paper, we detect SIFT points and calculate the homography matrices using the matched points to generate better results.

In this paper, we propose a generic virtual content insertion solution based on visual attention analysis, called @ICT. It needs little domain knowledge and is applied to general videos, such as TV play series and movies, etc. The basic idea of our overlay insertion method is similar to the one of AdOn [22]. Compared with AdOn, our work has the following characteristics:

1. We propose a measurement of intrusiveness, which covers both ROI interference and distraction. AdOn considers only ROI interference and calculates intrusiveness as the average saliency value of the insertion place.
2. We determine insertion place and insertion content interdependently. Given an image/video shot, different VCs should be inserted at different positions. We choose the position by considering both lower saliency and visual consistency.
3. @ICT includes an in-scene insertion module, which embeds the virtual contents into general videos by using affine transformation and camera tracking.

The rest of the paper is structured as follows: In Sect. 2, we review the related works and present the overview of @ICT. In Sects. 3 and 4 we detail the overlay insertion and in-scene insertion modules, respectively. We report the evaluation result in Sect. 5 and conclude this paper with future work in Sect. 6.

2 Related work

The two conflicting tasks of VCI are usually balanced through insertion time, insertion place, insertion content, and insertion method. In this section, we review the existing works from these four aspects and present the overview of @ICT.

2.1 Insertion time

For time choosing, an important factor to consider is to make the inserted content noticeable to the audiences. Therefore, the virtual content is usually inserted into video highlights as they are usually paid more attention to by the audiences [30]. Besides highlights, the consecutive frames with little camera motion are also selected as candidates to hold the inserted content for a period of time [32]. Highlight extraction usually needs domain knowledge while the frames with little camera motion cannot ensure the frames to be attended. AdOn chooses attractive shots, through motion intensity and shot length, as insertion time. In this paper, @ICT performs temporal attention analysis and chooses higher attentive shots as insertion time for overlay insertion to increase the opportunity of the inserted contents to be noticed. For in-scene insertion, it detects the shots of little camera motion as insertion time to ensure the insertion effect.

2.2 Insertion place

Several approaches have been proposed to detect proper places for VCI to avoid damaging the original content of the image/video. For sports video, domain knowledge can be exploited to determine the insertion place. For instance, static region, goalmouth, central circle, and boundary line in soccer video are detected to identify suitable locations for insertion [32]. More generic approaches include visual relevance measure [30] and lower informative region [11], which do not need any domain knowledge and can be extended to all types of videos. Other methods tackle this problem from the viewpoint of visual attention. ImageSense [20], AdOn [22], and GameSence [12] insert the advertisements at the non-salient corner or side regions to avoid occluding the informative content of the image. In our previous work [16], a notation of lower attentive region is proposed and defined, from the cognitive point of view, as a region of the video frame which attracts less audience attention. For @ICT, to reduce intrusiveness, we detect lower attentive region as candidate insertion place, then choose the final position by taking into account the visual consistency with the VC to be inserted.

2.3 Insertion content

Virtual content itself also plays a role in the effect of VCI. It should be coherent with its spatial and temporal context to maintain the visual effect and to reduce the intrusiveness. Under this principle, ImageSense chooses the advertisements according to local textual relevance, global textual relevance, and local content relevance [20]. Besides VC choosing, another way is to adjust the appearance of the insertion content. For example, ViSA re-colors the VC by using computational esthetics to ensure visual harmony [1]. In our approach, we maintain the appearance of the virtual content and choose virtual content according to visual consistency to avoid annoying the audiences.

2.4 Insertion method

Insertion method can be overlay or in-scene insertion. Overlay insertion is relatively simple, while in-scene insertion utilizes the camera parameters to embed the virtual content into the real circumstance. VC embedding can be performed by using predetermined landmarks [26]. If the model of the scene is available, it can be used to estimate camera parameters [33] and to distort the inserted VC [10, 27]. In this paper, @ICT performs in-scene insertion through affine rectification and camera motion estimation. This method needs weak condition and is applied to most videos of indoor or outdoor urban scenes.

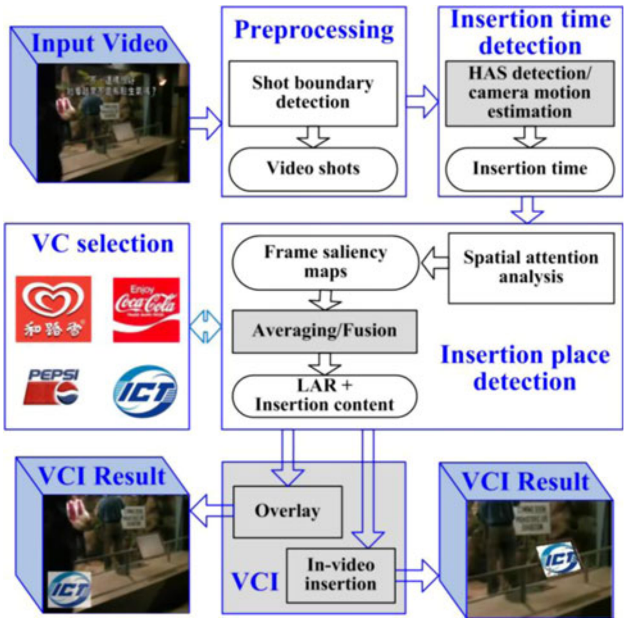


Fig. 2 The overview of @ICT. It has two modules, overlay insertion and in-scene insertion. The two modules share five parts but are different at the *shaded* contents

2.4.1 *Our approach*

Figure 2 illustrates the overview of @ICT. It has an overlay module and an in-scene insertion module. The two modules share five parts, including video preprocessing, insertion time detection, insertion place detection, virtual content selection, and virtual content insertion. But they are slightly different at the shaded contents in the figure, i.e., insertion time detection, insertion place detection, and virtual content insertion. In the video preprocessing part, the input video is segmented into shots by using the method of [13]. Within a shot, the content is of spatially and temporally continuum. Thus taking shot as basic unit of VCI makes the insertion result visually influent. The insertion time module detects HAS for overlay VCI and static shots for in-scene insertion. The insertion place module chooses LAR and then the VC selection module determines the virtual content according to visual consistency. Finally, the VCI module inserts the chosen content into the chosen shot, at the chosen place, through overlay or in-scene insertion. The details of the two modules will be presented in the following two sections.

3 Overlay VCI

For overlay VCI, the insertion method is relatively simple. Therefore, we focus on insertion time choosing, insertion place, and insertion content determination.

3.1 Insertion time choosing

While watching a video, audiences pay different amount of attention to the video content at different time. The virtual content inserted at the time when the video attracts more audience attention is more probable to be noticed and remembered by audiences. So we detect the shots more attentive as insertion time.

Generally speaking, the shots different from the preceding ones attract more attention. Here we adopt the notation of novelty to evaluate the attention of each shot. In our work, a shot's novelty is evaluated through its difference to its preceding ones. Let S_t be a shot of the video, its novelty is

$$\mathrm{Nov}(S_t) = \sum_{k=t-N}^{t-1} \mathrm{diff}(S_k, S_t) w(k) \tag{1}$$

where the length of context window N is the number of shots included, which is set as 5 in our work. $w(k)$ is the weight of shot k for the consideration that the nearer shots have more influence on the current one. In our work, we adopt linear weight $w(k) = (t-k)/(1+2+\cdots+N)$, which is the relative distance between the two shots. diff (S_k, S_t) is the dissimilarity between the two shots. The feature used here is the normalized color histogram of the shot, calculated by averaging the normalized frame histograms. For each shot, an 8 × 8 × 8 RGB histogram is calculated and the dissimilarity is calculated using histogram intersection:

$$\mathrm{diff}(S_k, S_t) = 1 - H_t \cap H_k \tag{2}$$

Besides novelty, shot length also determines a shot's attention value. The longer a shot is, the more probable it is to be attended. In this regard, the attention value of a shot is calculated as $L_t \times \mathrm{Nov}(S_t)$, where L_t is its length.

Finally, it should be noted that too frequent insertion will annoy the audiences. So the insertion time is chosen under the restriction of minimum time interval.

3.2 Insertion place and insertion content determination

The objective of insertion place and insertion content choosing is to decrease the intrusiveness caused by insertion. In this section, we first present a quantitative measure of intrusiveness from the viewpoint of visual attention. Then we present how we determine the insertion place and the insertion content interdependently to reduce intrusiveness.

3.3 Measurement of intrusiveness

Intrusiveness is defined as a perception or psychological consequence that occurs when an audience's cognitive processes are interrupted [17]. According to this definition, we have proposed a measurement of intrusiveness from the perspective of visual attention [14]. This measurement covers two aspects. First, if the inserted advertisement covers the main content of the image/video, it is definitely intrusive. This type of intrusiveness is referred to as ROI interference in this paper. Second, if the advertisement visually outstands of the image, it will distract audience attention from the original attending point. To measure distraction, we employ the difference between audience attention distributions, which are represented through attention maps, before and after content insertion. The attention maps are normalized to $\sum_{(x,y)} \mathrm{AM}(x,y) = 1$; thus they can be looked as probability density function for convenient comparison. Several comparison methods are available, such as correlation coefficient, Kullback–Leibler divergence, and intersection etc. In our work, we calculate the consistency between attention maps as their intersection because it best discriminates distraction from non-intrusiveness in experiment. Intersection is calculated as follows:

$$C = \sum_{(x,y)} \min(\mathrm{AM}_{\mathrm{before}}(x,y), \mathrm{AM}_{\mathrm{after}}(x,y)) \qquad (3)$$

Then we calculate the distance as $1 - C$. Finally, we calculate intrusiveness, taking ROI interference into account, as follows:

$$\mathrm{Intr} = \begin{cases} 1 & \text{if the brand covers the ROI} \\ 1 - C & \text{others} \end{cases} \qquad (4)$$

According to (4), the intrusiveness locates between 0 and 1. When the virtual content covers the ROI, the intrusiveness reaches its maximum. This measurement provides a straightforward objective for in-video VCI. Given a shot, we first detect its LAR to avoid ROI interference. Then for a given VC, we traverse all possible positions of the LAR to find the insertion position of minimal distraction. Figure 3 illustrates an example of insertion place choosing, by taking an image as instance.

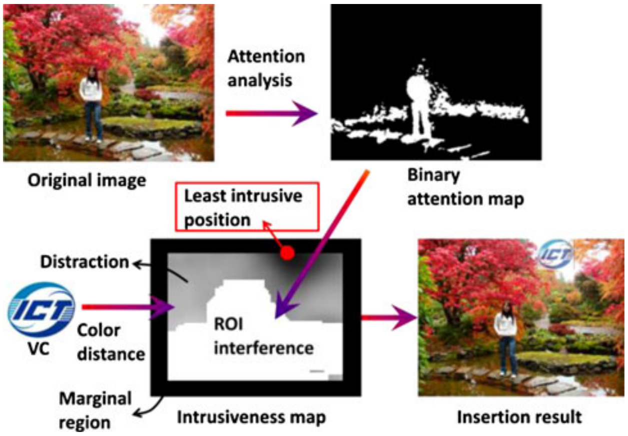


Fig. 3 Illustration of insertion place and insertion content determination

The details of LAR detection and virtual content determination will be presented in the rest of this section.

3.4 LAR detection

There are already many works about visual attention analysis, such as [7, 9, 19, 23, 29]. In this paper, we employ our previous method proposed in [15]. This method adopts region as perceptive unit, calculate the attention value of a region through color contrast, color rarity, motion contrast, and color novelty, as shown in Fig. 4. Here we present this method briefly. For details, the readers are referred to [15].

3.4.1 Image segmentation

The perceptive unit can be chosen as pixel [23], image block [9, 19], region [7], or object [29]. A pixel/block contains little perceptive information. Comparatively, an object contains much perceptive information but is difficult to be obtained for its complexity. 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, for which there are already many methods available. Therefore in our work we adopt region as perceptive unit. This choice also enables the proposed method to analyze visual attention at multiple 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. An issue for K-Means is to determine the cluster number. To avoid over segmentation, we set the maximum cluster number as 8 empirically. To choose a suitable cluster number for an image according to image content, we calculate its $8 \times 8 \times 8$ RGB histogram and use the minimum number of the highest bins which in total cover over 95% of the pixels as the cluster number. If the number is bigger than 8, 8 is adopted. After color quantization, the neighboring pixels of the same color are regarded as a region.

3.4.2 Color contrast

It has been verified that human visual system are sensitive to contrast for the center-surround structure of the receptive field.

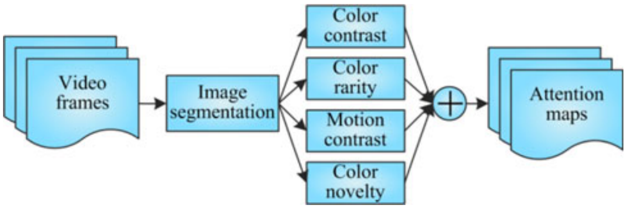


Fig. 4 Overview of the attention analysis method

Receptive field is proved to be an ellipse with its main axis 20° to the horizon and is modeled with Difference of Gaussian (DoG) [28]. For simplicity we adopt an isotropic model:

$$\mathrm{DoG}(x,y) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) - \frac{1}{2\pi\lambda^2\sigma^2}\exp\left(-\frac{x^2+y^2}{2\lambda^2\sigma^2}\right) \tag{5}$$

where $\sigma^2 E$ and $\lambda^2\sigma^2 E\ (\lambda > 1, E\,\text{is identity matrix})$ are the covariance matrices of the two Gaussians. We experimentally set $\lambda \approx 2.0984$ and $\sigma = 0.5104R$, with R be the region's radius. Then a region's contrast is calculated as

$$\mathrm{Con(k)} = \sum_{i=1}^{K} d(f_k, f_i) \times G_k(i,k) \times S_i \tag{6}$$

where $d(f_k, f_i)$ is the distance between two features, G_k is the *DoG* function of region k, and S_i is the area of region i.

3.4.3 Color rarity

While watching the scene, our purpose is to pursue information. So the informative content usually attracts our attention. According to Shannon's information theory, the rarer an event is, the more informative it is. The rarity of each region is calculated, by using the color quantization result, as follows:

$$\mathrm{Rar}(k) = \log p(f_k) \tag{7}$$

where f_k is the feature of region k and $p(f_k)$ is its probability calculated from the color quantization result.

3.4.4 Motion contrast

Motion vector can be obtained by several methods such as optical flow [6]. 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) is adopted to alleviate the negative impact caused by camera motion [4]. This method presents the MVS through HSV color space as follows:

$$\begin{aligned} \text{Angle} &\rightarrow H \\ \text{Magnitude} &\rightarrow S \\ \text{Texture} &\rightarrow V \end{aligned} \tag{8}$$

where the motion magnitude and the texture are normalized to [0, 255]. The choosing 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 the existence of camera motion, the V component can still provide a good presentation of the frame. Finally, by using the HSV presentation, motion contrast is calculated through (6).

3.4.5 Color novelty

Besides motion saliency, novelty, an event's standing out of its temporal background, also affects audience attention. Itti [8] 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. The novelty of each pixel at each time is calculated as the distance between the prior and the posterior distributions. 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:

$$\mathrm{Nol}(t) = \mathrm{KL}(M_{t-1}, M_t) = \int_{\mathrm{X}} M_t(x)\log\frac{M_t(x)}{M_{t-1}(x)}\mathrm{d}x \tag{9}$$

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

$$M_t \sim N\left(\mu_t, \sigma_t^2\right) \tag{10}$$

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

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.

3.4.6 LAR detection

After attention analysis, we obtain four maps, including color contrast map (M_c), color rarity map (M_r), motion contrast map (M_m), and color novelty map (M_n). Suitable fusion of the maps produces the final attention map. 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 ROI or LAR from the saliency/novelty maps, we model this progress as binary classification. We use the maximum inter-class variance to determine the fusion coefficients for the reason that the higher a map's inter-class variance is, the more powerful the map's discriminability is [24]. The maximum inter-class variance of map M_c is

$$\mathrm{Var}_c = \max_k \left(n_1(k)(\mu_1(k)-\mu)^2 + n_2(k)(\mu_2(k)-\mu)^2\right)\Big/n \tag{11}$$

where $n_1(k)$, $n_2(k)$, $\mu_1(k)$, and $\mu_2(k)$ are the number of samples and the means of the two classes when using threshold k, n is the total number of samples, μ is the mean of all the samples. Let Var_r, Var_m, and Var_n be the maximum inter-class variance of M_r, M_m, and M_n. Then the fusion weight for color contrast map is

$$w_c = \mathrm{Var}_c / (\mathrm{Var}_c + \mathrm{Var}_r + \mathrm{Var}_m + \mathrm{Var}_n) \tag{12}$$

The weights for color rarity map, motion contrast map, and color novelty map are similarly calculated. The final attention map is

$$\mathrm{AM} = w_c M_c + w_r M_r + w_m M_m + w_n M_n \tag{13}$$

Then the attention map is binarized to obtain the ROI and the LAR. In the binary attention map in Fig. 3, the white region is ROI and the black region is LAR.

3.5 Insertion place detection

After LAR detection, for a given VC, we first exclude the marginal regions and the positions which may lead to ROI interference, as shown in the intrusiveness map in Fig. 3. Then we traverse all the possible positions to find the optimal one for a given VC. Instead of repeating attention analysis on the insertion results, which is straightforward but time consuming, visual consistency is utilized as the criterion for position choosing. Visual consistency is calculated through color histogram. At position (x, y), the visual consistency between the virtual content and the shot is

$$C_{x,y}(S, \mathrm{VC}) = \frac{1}{L} \sum_{t=1}^{L} \left(H_{x,y,t} \cap H_{\mathrm{VC}} \right) \tag{14}$$

where $H_{x,y,t}$ and H_{VC} are the normalized color histograms of the insertion position of frame t and the virtual content, L is the shot length. Then we get the optimal insertion position and the visual consistency between the VC and the shot:

$$\begin{aligned} (x*, y*) &= \arg \max_{(x,y) \in LAR} C_{x,y}(S, \mathrm{VC}) \\ C(S, \mathrm{VC}) &= C_{x*,y*}(S, \mathrm{VC}) \end{aligned} \tag{15}$$

Here (x*, y*) is the optimal insertion position and C(S, VC) is the consistency between the VC and the shot.

3.6 Virtual content determination

For a given shot and a set of candidate VCs, we choose the VC which has the lowest intrusiveness as insertion content. For a VCI task includes multiple VCs to be inserted into the video, without loss of generality, we insert into each shot no more than one VC, and each VC will not be chosen for more than one time. Then we search for a best match between the shots and the VCs according to visual consistency. Let m and n be the numbers of shots and VCs; then the objective function is defined as

$$\max_{\{(i_k, j_k)\}} \sum_{k=1}^{\min(m,n)} C\left(S_{i_k}, VC_{j_k}\right) \tag{16}$$

Function (16) has a huge solution space. For the task of 70 video shots and 75 brands in our experiment, it has $C_{75}^{70} \times 70!$ solutions. We employ genetic algorithm to search for the optimal solution [31]. Genetic algorithm mimics the process of natural evolution. Canonical genetic algorithm produces the next generation from the parents through crossover and/or mutation. Then a portion of the existing solutions are chosen as new parents. For our matching problem, we adopt single-parent genetic algorithm, which produce the next generation from just one parent. We first generate a random solution as the parent. Then we randomly choose two points in the solution and inverse the part between the two points to produce the next generation. The more optimal one is chosen for the next iteration. This progress is repeated until convergence.

Through (16), the virtual content and the insertion place are determined interdependently.

4 In-scene VCI

For in-scene VCI, the insertion content choosing method is similar to the one of overlay VCI. It is different with the overlay insertion module in insertion time choosing, insertion place detection, and the insertion method. So we focus on these three aspects.

4.1 Insertion time choosing

In-scene insertion embeds the virtual content into the videos according to camera parameters. The current camera motion estimation methods usually fail under fierce camera motion. Furthermore, under fierce camera motion, e.g., under fast pan, a region may be visible for just a very short while. To obtain better results, we tradeoff between attention and insertion effect and perform in-scene insertion on the shots of little camera motion. Scale Invariant Feature Transformation (SIFT) [18] is performed and the matched SIFT points are used to obtain the homography matrix through the RANSAC algorithm. Let $H_{t,t+1}$ be the homography matrix between frames t and t + 1:

$$H = \begin{pmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{pmatrix} \tag{17}$$

Then for a point P_t in frame t, its corresponding point in frame $t+1$ is P_{t+1}:

$$\begin{aligned} P_{t+1} &= P'_{t+1}/w' \\ P'_{t+1} &= H_{t,t+1}P_t \end{aligned} \tag{18}$$

w' is to transform the point to homogeneous coordinate. Then for the point $(0 \quad 0 \quad 1)^{\mathrm{T}}$ in frame t, its corresponding point in frame $t+1$ is $(h_{13}/h_{33} \quad h_{23}/h_{33} \quad 1)^{\mathrm{T}}$. The displacement of the point, $\sqrt{h_{13}^2 + h_{23}^2}\Big/h_{33}$, is used to describe the camera motion. The shots of little camera motion are detected as candidates. For a chosen shot, if suitable insertion place is detected, it will be chosen for in-scene VCI.

4.2 Insertion place detection

For the chosen candidate shots, we detect dynamic LAR as insertion place. Different with a static LAR, which is a region of the same position on each frame, a dynamic LAR is a region of the same position in the real circumstance. It should be kept on the corresponding position on each frame. For the existence of camera motion, mosaic image stabilization is adopted to present the appearance of a locked down camera. Through the homography matrices, for a point P_t in frame t, we obtain its corresponding points in each frame. Then for a point P_M in the mosaic attention map, its attention value is calculated as

$$\mathrm{AV}(P_M) = \sum_{t=1}^{L} \mathrm{AV}(P_t) \tag{19}$$

where P_t is the corresponding point of P_M in frame t and L is the total number of frames of the shot.

Finally, the proposed in-scene insertion method (will be detailed in the following section) utilizes two pairs of parallel lines to embed the VC into the video. So we detect straight lines, through Hough transform, of the shot. A pair of parallel lines in real world may be not parallel in the image/video. So we detect nearly parallel lines. If a region is of lower attention value and is surrounded by two pairs of parallel lines, it is detected as insertion place.

4.3 In-scene insertion

Compared with overlay insertion, in-scene insertion needs more techniques. The existing methods perform in-scene insertion in sports video because the structure of the playfield can be used to calibrate the camera [33]. However, in general videos there is too little prior knowledge to perform camera calibration. In this paper we propose to insert VC with as little as circumstance information by using affine transformation. To insert the virtual content into the videos with reality, affine rectification is first performed to obtain the front view of the circumstance. 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, we obtain the affine matrix by using two pairs of parallel lines. The method is briefly presented here. For more details, the readers are referred to [2].

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:

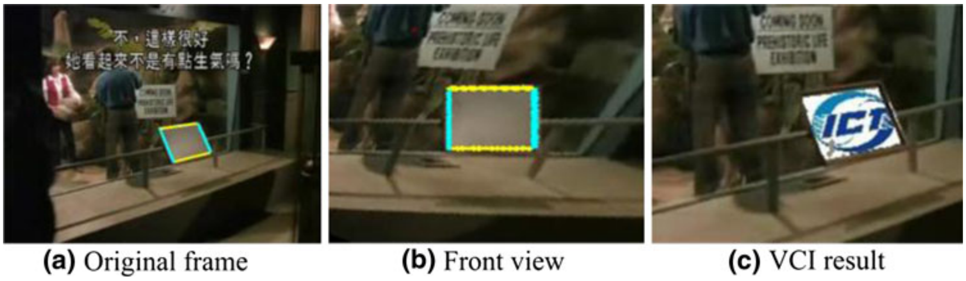
$$l_i : a_i x + b_i y + c_i = 0, i = 1,2 \tag{20}$$

If l_1 and l_2 are not parallel in the image plane, we obtain the vanishing point v_1, with aligned coordinate $V_1 = (x_1 \quad y_1 \quad f)^{\mathrm{T}}$. Here f is the focal length of the camera. It can be calculated through camera calibration or be set as 1 for simplification. If l_1 and l_2 are parallel in the image plane, the vanishing point is infinite and its aligned coordinate is written as $V_1 = (a_1 \quad -b_1 \quad 0)^{\mathrm{T}}$. The normalized spatial direction of the lines L_1 and L_2 is$r_1 = V_1/|V_1|$. From the other pair of parallel lines we obtain $r_2 = V_2/|V_2|$. Then the norm of the spatial plane is $r_3 = r_1 \times r_2$. The three vectors compose the affine matrix$A = (r_1 \quad r_2 \quad r_3)$. 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 $P_A(3)$ equal to the focal length f. P_0 is a point keeps unchanged during affine transformation. It is set as the center of the LAR to keep the LAR within the scope of the frame after affine rectification.

After the virtual content is inserted into the first frame, the homography matrices are utilized to embed it into the following frames. To improve the precision of region tracking, we re-estimate the homography by using the SIFT points of the insertion region. For a point P_A on the front view, its position on frame t is calculated as follows:

$$P_t = \prod_{k=0}^{t-1} H_{k,k+1} \times (AP_0 + P_0 - AP_A) \tag{21}$$

where $H_{k,k+1}$ is the homography matrix to track the insertion region. An example of in-scene insertion is illustrated in Fig. 5. In the example, the parallelogram in the frame (Fig. 5a) is detected as insertion place. It is a rectangle in real circumstance, as shown in the front view (Fig. 5b). Then the virtual content is inserted into the frame through affine transform, as shown in the VCI result (Fig. 5c).



(a) Original frame (b) Front view (c) VCI result

Fig. 5 Example of in-scene insertion. In the original frame (**a**) two pairs of parallel lines are detected. Then the front view (**b**) of the frame is obtained through affine rectification. Finally, the VCI result (**c**) is obtained by using the affine matrix

5 Performance evaluation

In this section, the performance of the proposed VCI solution will be evaluated. The VCI solution is evaluated on the videos in Fig. 6. There are three major types of location, urban, rural, and indoor. As shown in the figure, the test videos include all the three types of locations. 75 famous brands including car brand such as Benz, noshery brand such as Mcdonald’s, and other brands which frequently appear in life are chosen as virtual content. We adopt these brands for two considerations. One is that advertising is an important application of VCI. The other one is that it should be easier for the users to select the ones he/she has seen in the video. That is to decrease the noticing difference caused by the brands themselves.

5.1 Performance of overlay insertion

We first test the HAS detection method and then the insertion place and content determination method.

5.2 HAS detection

The HAS detection method is tested on two of the testing videos, “Adventure to the west” and “Children at home”. These two videos are of different types. “Adventure to the west” is of rural scene and contains many fighting shots. “Children at home” is sitcom and is of indoor scene. From the beginning of each of the two videos, excluding the prolog, two consecutive video clips, each of which includes 50 shots, are chosen for test. The clips are test in experiment independently. To study to what degree each shot is

Title: Adventure to the west
Type: teleplay, fantasy
Length: 0:52:10
Location: rural

Title: Friends
Type: sitcom
Length: 22:56
Location: indoor

Title: Snow White
Type: movie, cartoon
Length: 1:23:22
Location: rural, indoor

Title: Full house
Type: movie
Length: 1:10:27
Location: urban, indoor

Title: White snake
Type: teleplay, fantasy
Length: 0:46:10
Location: rural, indoor

Title: Children at home
Type: sitcom
Length: 22:47
Location: indoor

Title: ROB-B-HOOD
Type: movie
Length: 2:05:36
Location: indoor, urban

Fig. 6 The representative frames and details of the test videos

attended by audiences, each shot is inserted with a unique VC randomly chosen from the VC database. 16 users, aged between 22 and 30 years, are invited to watch the result videos for one time. After watching, the users are shown to a set of brands, which include the ones inserted into the videos. The users are required to choose the brands he/she has seen in the videos. The users' feedback for each clip is then averaged to obtain the noticing rate of each brand, i.e., each shot. The attention curve and the noticing curve are compared by using consistency and correlation coefficient (cc). Consistency is calculated similarly with formulation (3). The attention curve and the noticing rate are first normalized to have sum 1; then the intersection of them is calculated as consistency. The higher the consistency is, the more precise the result is. cc between two variables, x_1 and x_2, is calculated as $\mathrm{cc}(x_1, x_2) = \mathrm{cov}(x_1, x_2)/\sigma_1\sigma_2$. Here $\mathrm{cov}(x_1, x_2)$ are the two variables' co-variance; σ_1 and σ_2 are the standard deviations of the two variables. The value of cc locates between -1 and 1. The larger the absolute value of cc is, the stronger the linear relationship between the two variables is. A value of $1/{-1}$ indicates a perfect positive/negative linear relationship. A value of 0 means totally irrelative.

The results are illustrated in Fig. 7. The first row shows the attention curves and the noticing rates of the four test video clips, together with the consistency. The consistencies are about 0.8. The second row plots the noticing rate versus attention value, with the correlation coefficient. The cc values are all higher than 0, but regretfully, less than 0.5. Generally speaking, the consistency and correlation coefficient indicate that the proposed HAS detection method is effective and inserting virtual content into higher attentive shots does increase the opportunity of the contents to be noticed. However, low-level features, without semantic content, can describe just a part of human cognition. This limitation can also be seen in the experimental result. The results of "Adventure to the west" are better than the ones of "Children at home". This is because the latter one is indoor sitcom, so the visual scene changes just slightly and audience attention is more affected by semantic content, which is out of the scale of the low-level features adopted in our experiment.

5.3 Overlay insertion evaluation

Then we evaluate the intrusiveness caused by overlay insertion. In the proposed solution, the lower attentive region is chosen as insertion place to avoid covering the ROI. What we need to do is to evaluate the intrusiveness caused by distraction. We choose ten attentive shots, under the minimum time interval of 2 min, from each of the testing videos in Table 1 to construct a testing set of 70 shots. Then the task is formulated as inserting the 75 logos into the 70 shots. In experiment, the area of the insertion region is set as 1% of the video frame. The following three sets of results are compared:

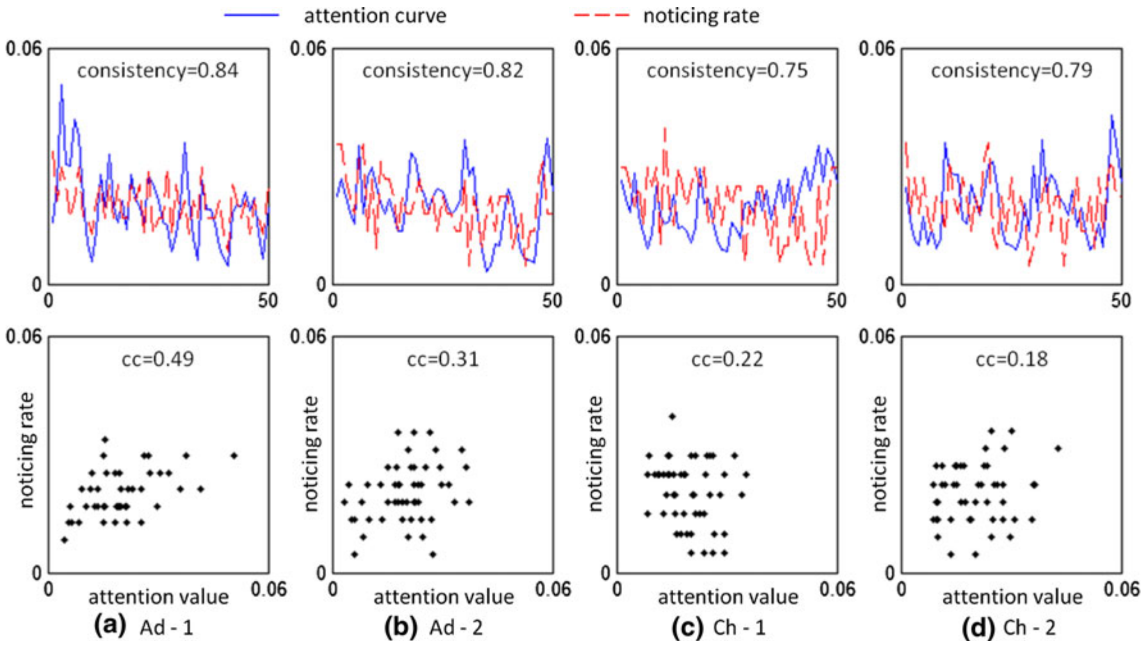


Fig. 7 Results of HAS detection. The first row shows the attention curves and the noticing rates of the four test video clips, together with the consistency. The second row plots the noticing rate versus attention value, with the correlation coefficient. In the figure, *Ad-1* and *Ad-2* are the two clips from "Adventure to the west". *Ch-1* and *Ch-2* are the two clips from "Children at home"

Table 1 Insertion shot detection result

Video	Shot	Insertion shot	Detected
Friends	326	8	7
Children at home	253	9	8
ROB-B-HOOD	3228	29	25
Full house	693	27	20

1. This set is like Mei's method and is adopted as baseline to evaluate our approach. Mei chooses the advertisement according to textual relevance, user preference, and global visual consistency. The insertion place is chosen among the upper fifth and the bottom fifth of the frame. To make the comparison as fair as possible, we choose insertion content according to only global visual consistency and drop the textual relevance and user preference. Moreover, we choose the insertion place among the upper tenth and the bottom tenth of the frame to make the insertion region 1% of the video frame, the same with our setting.
2. The logos are randomly matched with the shots and are inserted into the videos at the position determined by lower attention and visual consistency.
3. The logo and insertion place are determined by using the proposed solution.

Figure 8 displays the consistencies of the above three methods. To keep the figure neat, the shots are sorted in descending order according to the consistencies of the baseline. The second set outperforms the baseline on 54 (77%) shots. This result indicates that our insertion place determination method is more rational than the one of the baseline. The third set outperforms the second one on 54 (77%) shots, meaning that choosing the logos through visual consistency further reduces the intrusiveness. The third set outperforms the baseline on 59 (84%) shots. The average consistencies of the three methods are 0.93, 0.95, and 0.96. Figure 9 shows two examples of overlay insertion, including the shot saliency maps, the baseline method, random VC at LAR, and the final result. Figure 9a shows a shot on which our approach failed. Figure 9b shows a result that virtual content choosing reduces the intrusiveness. In this result, global visual consistency and local visual consistency give the similar results. The above comparison result verifies the effectiveness of @ICT.

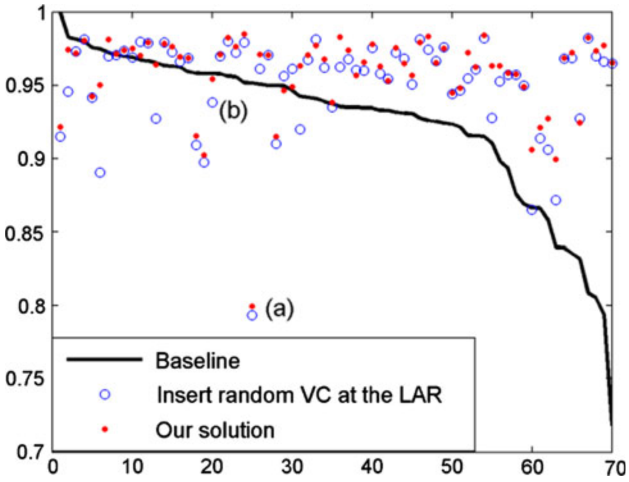


Fig. 8 Consistency of the three sets of results of overlay insertion. The insertion results of **a** and **b** are shown in Fig. 9

5.4 In-scene insertion evaluation

In this section, we test the proposed in-scene insertion method both quantitatively and qualitatively. In-scene insertion should be performed on videos of indoor or outdoor of urban scenes. Therefore, "Friends", "Children at home", "ROB-B-HOOD", and "Full house" are chosen for in-scene insertion. The insertion time and insertion place detection result is shown in Table 1, which shows the total shot number, the number of manually chosen insertion shots, and the number of the detected insertion shots. Among the four videos, "ROB-B-HOOD" contains smaller ratio of insertion shots because it is an action movie and many shots are of fierce camera motion. The result verifies that our insertion time and place detection methods for in-scene insertion are effective.

Among the detected insertion shots, 10 ones are chosen for evaluation. For the detected insertion place, we manually label the corresponding region in each frame as ground truth. The tracking precision is evaluated through the distance between the centers of the tracked region and the ground truth region:

$$\|(x,y)-(x_{\mathrm{GT}},y_{\mathrm{GT}})\|_2 \Big/ \sqrt{W^2+H^2} \qquad (22)$$

where (x, y) and $(x_{\mathrm{GT}}, y_{\mathrm{GT}})$ are the centers of the tracked region and the ground truth region, respectively. W and H are the width and height of the video. Considering that the tracking method may lose the region, we calculate the average tracking precision of each shot on the tracked frame and we also calculate the losing ratio for each shot. The method of this paper is compared with the GME based method [5], which is adopted in our previous paper [15]. The result is illustrated in Table 2. It can be seen that the SIFT method performs better than the GME method. The SIFT method obtains much more precise region tracking result and does not lose any region.

Then the insertion results are evaluated by users. In this experiment, our major purpose is to evaluate the in-scene insertion method, including deforming the VC through affine rectification and tracking the camera. Therefore, in the results, we did not match the VC and the shot through visual

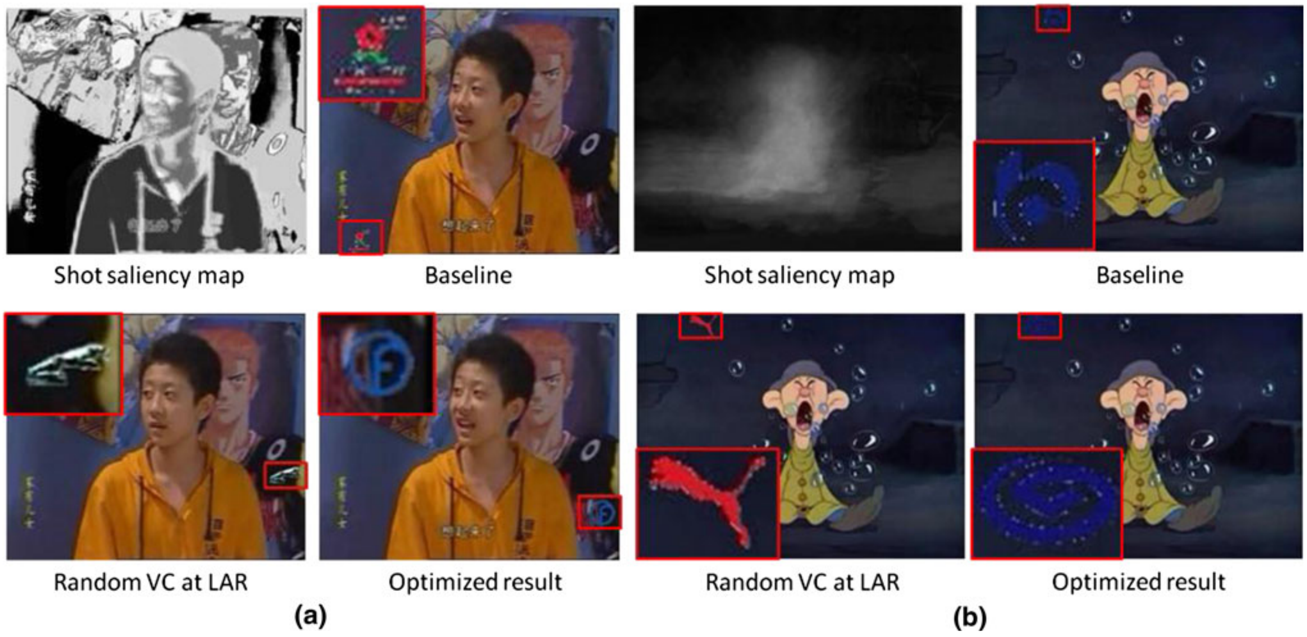


Fig. 9 Examples of overlay insertion. In each insertion result, the inserted content is enlarged and shown in the *red box*. The consistencies of the result can be found in Fig. 8

Table 2 Comparison of the GME based in-scene insertion method and our new method

No.	Center distance		Losing ratio	
	GME	SIFT	GME	SIFT
1	0.14	0.04	0.38	0
2	0.02	0.01	0	0
3	0.21	0.01	0.67	0
4	0.04	0.01	0.83	0
5	0.22	0.01	0	0
6	0.39	0.06	0.37	0
7	0.16	0.01	0	0
8	0.08	0.02	0	0
9	0.11	0.01	0.67	0
10	0.00	0.05	0.89	0
Average	0.14	0.02	0.38	0

Table 3 User study result of in-scene insertion

No.	Mean	Variance	No.	Mean	Variance
1	4.38	0.92	6	2.92	1.08
2	3.08	1.41	7	3.08	1.08
3	3.69	1.06	8	3.38	1.42
4	3.31	1.40	9	4.23	0.86
5	3.15	1.31	10	3.00	0.83
Average		Mean		Variance	
		3.42		1.14	

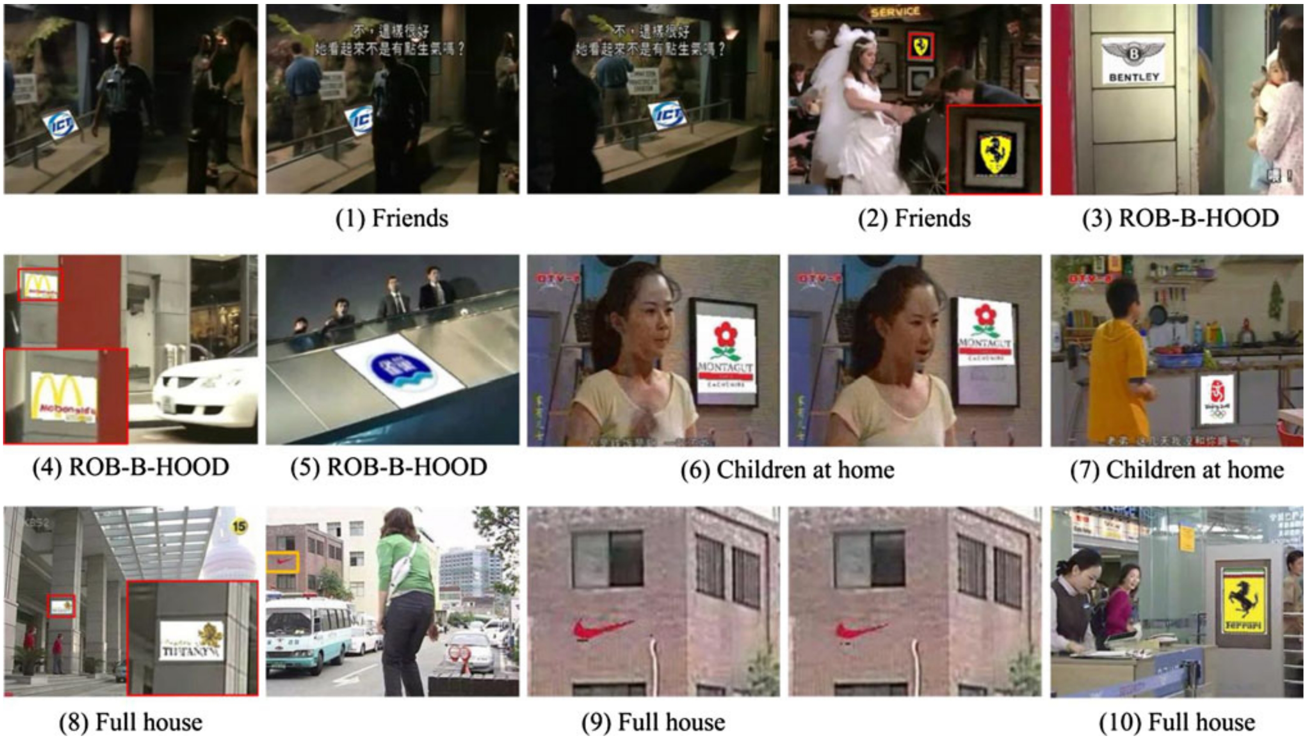
consistency. Instead, to make the result convenient for the subject to watch, we inserted into the shots with high-contrast VCs and kept the images' background. Thirteen users take part in the study. The users are requested to give each result an overall score under the following criteria:

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. The mean and variance of the user scoring result is illustrated in Table 3. The representative frames of the shots are shown in Fig. 10. Three of the results, including the 1st, 6th, and the 9th ones, have multiple frames shown in the figure. The 6th result gets the lowest score because the camera tracking method causes the displacement of the brand. The 1st and the 9th results are satisfying ones. In the 1st one, the proposed method tracks the camera motion well, even after the person walks in front of the insertion place. In the 9th shot, the logo of Nike is embedded into the circumstance with vividness. From the above evaluation result, it can be concluded that the overall performance of in-scene insertion is acceptable.

6 Conclusion

In this paper, we propose a generic virtual content insertion solution, called @ICT. This solution trades off



(1) Friends (2) Friends (3) ROB-B-HOOD

(4) ROB-B-HOOD (5) ROB-B-HOOD (6) Children at home (7) Children at home

(8) Full house (9) Full house (10) Full house

Fig. 10 The in-scene insertion results. The numbers correspond to the ones in Table 3. The title of each sub figure is the video title. In the 2nd, 4th, and the 8th results, the enlarged logos are shown at the corner of the frames. For the 9th result, we mark the inserted Nike logo with a *yellow box* for better vision. In the two following images, the local parts of the frames are shown

between the two conflicting tasks of VCI by taking advantage of audience attention. It ensures the inserted content to be noticed by audiences through inserting the contents into the attentive shots. To decrease the intrusiveness caused by insertion, it detects the lower attentive region as insertion place and chooses virtual content according to visual consistency. Furthermore, it includes an in-scene insertion module. It embeds the virtual content into the video vividly through affine transformation and camera tracking.

Although the experiments have verified the effective of @ICT, there are several possible improvements and extensions for it. First, visual attention is a far more complex mechanism. It has two major progresses, bottom–up attention and top–down attention, corresponding to stimulus-driven and task-driven, respectively. In our work, we employed only bottom–up attention, which is incapable of semantic content of videos. This incapability is shown in the HAS detection result. Second, textual relevance and user preference, as used by AdOn, can be employed to improve the virtual content choosing method furthermore. Third, for application in broadcasting, the inserted virtual content can be encoded separately with the original video. For overlay insertion, it needs only the insertion time, place, and content. For in-scene insertion, besides insertion time, place, and content, it needs the affine matrix and the homography matrices. Separate encoding the virtual content can facilitate transferring different virtual content according to user preference.

References

1. Chang, C.-H., Hsieh, K.-Y., Chung, M.-C., Wu, J.-L.: ViSA: virtual spotlighted advertising. In: Proceedings of the 16th ACM International Conference on Multimedia, pp. 837–840 (2008)
2. Chen, X., Yang, J., Zhang, J. Waibel, A. Automatic detection and recognition of signs from natural scenes. IEEE Trans. Image Process. **13**(1), 87–99 (2004)
3. Deshpande, S., Naphade, P., Rao, C.V.K., Bhadada, K., Rangan, P.V.: Method and apparatus for including virtual ads in video presentations. US Patent, 7158666 (2007)
4. Duan, L.-Y., Xu, M., Tian, Q., Xu, C.-S., Jin, J.S.: A Unified framework for semantic shot classification in sports video. IEEE Trans. Multimedia **7**(6), 1066–1083 (2005)
5. Dufaux, F., Konrad, J.: Efficient, robust, and fast global motion estimation for video coding. IEEE Trans. Image Process. **9**(3), 497–501 (2000)
6. Horn, B.K.P., Schunck, B.G.: Determining optical flow. Artif. Intell. **17**, 185–203 (1981)
7. Hu, Y., Rajan, D., Chia, L.-T.: Robust subspace analysis for detecting visual attention regions in images. In: Proceedings of the 13th ACM International Conference on Multimedia, pp. 716–724 (2005)
8. Itti, L., Baldi, P.: A principled approach to detecting surprising events in video. In: Proceedings of the IEEE Computer Society

Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 631–637 (2005)

9. Itti, L., Koch, C., Niebur, E.: "A model of saliency-based visual attention for rapid scene analysis". IEEE Trans. Pattern Anal. Mach. Intell. **20**(11), 1254–1259 (1998)
10. Kreitman, H., Bar-El, D., Amir, Y., Tirosh, E.: Method and system for perspectively distorting an image and implanting same into a video stream. US Patent 5,731,846 (1998)
11. Li, Y., Wah Wan, K., Yan, X., Xu, C.: Real time advertisement insertion in baseball video based on advertisement effect. In: Proceedings of the 13th ACM International Conference on Multimedia, pp. 343–346 (2005)
12. Li, L., Mei, T., Hua, X.-S.: GameSense: game-like in-image advertising. Multimedia Tools Appl. **49**(1), 145–166 (2010)
13. Liu, C., Liu, H., Jiang, S., Huang, Q., Zheng Y., Zhang W.: JDL at Trecvid 2006 shot boundary detection. In: TRECVID 2006 Workshop
14. Liu, H., Qiu, X., Huang, Q., Jiang, S., Xu, C.: Advertise gently—in-image advertising with low intrusiveness. In: Proceedings of the IEEE International Conference on Image Processing (2009)
15. Liu, H., Jiang, S., Huang Q., Xu, C.: A generic virtual content insertion system based on visual attention analysis. In: Proceedings of the 16th ACM International Conference on Multimedia, pp. 379–388 (2008)
16. Liu, H., Jiang, S., Huang Q., Xu, C.: Lower attentive region detection for virtual content insertion. In: Proceedings of the IEEE International Conference on Multimedia & Expo (2008)
17. Li, H., Edwards, S.M., Lee, J.-H.: Measuring the intrusiveness of advertisements: scale development and validation. J. Advert. **31**(2), 37–47 (2002)
18. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
19. Ma, Y.-F., Hua, X.-S., Lu, L., Zhang, H.-J.: A generic framework of user attention model and its application in video summarization. IEEE Trans. Multimedia **7**(5), 907–919 (2005)
20. Mei, T., Hua, X.-S., Li, S.: Contextual in-image advertising. In: Proceedings of the 16th ACM International Conference on Multimedia, pp. 439–448 (2008)
21. Mei, T., Hua, X.-S., Yang, L., Li, S.: VideoSense-towards effective online video advertising. In: Proceedings of the 15th ACM International Conference on Multimedia, pp. 1075–1084 (2007)
22. Mei, T., Guo, J., Hua, X.-S., Liu, F.: "AdOn: toward contextual overlay in-video advertising". Multimedia Syst **16**, 335–344 (2010)
23. Meur, O.L., Callet, P.L., Barba, D., Thoreau, D.: A coherent computational approach to model bottom-up visual attention. IEEE Trans. Pattern Anal. Mach. Intell. **28**(5), 802–816 (2006)
24. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man Cybern. **9**(1), 62–66 (1979)
25. Rolls, E.T., Aggelopoulos, N.C., Zheng, F.: Effective size of receptive fields of inferior temporal visual cortex neurons in natural scenes. J. Neurosci. **23**(1), 339–348 (2003)
26. Rosser, R.J., Das, S., Tan, Y.: Method of tracking scene motion for live video insertion systems. US Patent 5,808,695 (1998)
27. Sharir, A., Tamir, M., Wilf, I.: Method and apparatus for automatic electronic replacement of billboards in a video image. US Patent 6,384,871 (2002)
28. Soodak, R.E.: "Two-dimensional modeling of visual receptive fields using Gaussian subunits". Proc. Natl. Acad. Sci. USA **83**, 9259–9263 (1986)
29. Sun, Y., Fisher, R.B., Wang, F., Gomes, H.M.: A computer vision model for visual-object-based attention and eye movements. Comput. Vis. Image Underst. **112**(2), 126–142 (2008)
30. Wan, K., Xu, C.: Automatic content placement in sports highlights. In: Proceedings of the IEEE International Conference on Multimedia & Expo, pp. 1893–1896 (2006)
31. Whitley, D.: A genetic algorithm tutorial. Stat. Comput. **4**, 65–85 (1994)
32. Xu, C., Wan, K.W., Bui, S.H., Tian, Q.: Implanting virtual advertisement into broadcast soccer video. In: Pacific-Rim Conference on Multimedia, pp. 264–271 (2004)
33. Yu, X., Yan, X., Chi, T.T.P., Cheong, L.F.: Inserting 3D projected virtual content into broadcast tennis video. In: Proceedings of the 14th ACM International Conference on Multimedia, pp. 619–622 (2006)