Video Adaptation for Small Display Based on Content Recomposition

Wen-Huang Cheng, *Student Member, IEEE*, Chia-Wei Wang, and Ja-Ling Wu, *Senior Member, IEEE*

***Abstract*—The browsing of quality videos on small hand-held devices is a common scenario in pervasive media environments. In this paper, we propose a novel framework for video adaptation based on content recomposition. Our objective is to provide effective small size videos which emphasize the important aspects of a scene while faithfully retaining the background context. That is achieved by explicitly separating the manipulation of different video objects. A generic video attention model is developed to extract user-interest objects, in which a high-level combination strategy is proposed for fusing the adopted three types of visual attention features: intensity, color, and motion. Based on the knowledge of media aesthetics, a set of aesthetic criteria is presented. Accordingly, these objects are well reintegrated with the direct-resized background to optimally match the specific screen sizes. Experimental results demonstrate the efficiency and effectiveness of our approach.**

***Index Terms*—Content recomposition, media aesthetics, region of interest, video adaptation, visual attention model.**

I. Introduction

ON THE INTERNET, multimedia content has been widely used for sharing information among users. Their transparent access from almost everywhere at anytime through all kinds of devices is desired and often required. To enable such universal multimedia access (UMA), one key technology is *video adaptation* [1]–[4]. In general, it is defined as the mechanism of transforming a video stream with one or more operations to meet specific application needs, such as device capabilities, network characteristics, and user preferences. At the user's end, hand-held devices including cellular phones, Smartphones, PDAs, and Pocket PCs are now in widespread use for their mobility and portability. In order to compete with desktop computers for practical computing tasks, they are not only developed for more powerful functionality but also equipped with more storage capacity. However, one exception is the display. For the portable requirement of hand-held devices, the screen size is kept permanently unchanged and even as small as possible. With the rapid growth of quality video sources (e.g., mobile TV, VCD/DVD on demand), the physical limitation would seriously disrupt user's viewing experience [2], [5], [6]. Thus, it is crucial to develop an efficient tool for facilitating video presentation on devices with limited display.

The conventional schemes that have been proposed for adapting videos on a small display can be divided into two categories: *spatial transcoding* and *frame cropping* [7], [8]. The former subsamples each frame to preserve intact video contexts and the latter discards partial surroundings to highlight specific user interests. Due to the bias of their design purposes, an adaptation engine has to make visual tradeoffs between the subject readability and content completeness [9]–[11]. However, sacrificing either aspect is usually intolerable because they are both important in our viewing experience. For example, when watching sports programs, player recognition and full-court variation are both important visual concerns [9], [11]. The difficulties with the conventional schemes arise because they both passively attempt to adapt the plain frame but not the actual content it contains. Consequently, the adapting process is forced to specify a desired area of the source frame (maximally itself) and uniformly stuff it onto the target screen. Until we move away from that paradigm, the obtained performance will fall short of our expectations [10], [11].

In this work, we propose a novel framework for video adaptation based on content recomposition. Our objective is to provide *effective* small size videos that emphasize important aspects of the scene while retaining the background context for adaptive delivery. We focus on nonuniform processing of different video regions by giving more display resource (i.e., space) to the important ones and less to the other parts. Specifically, we use visual attention analysis to extract user-interest objects (UIOs) of a scene. With regard to the background, these objects are downsized at a light level and with constant aspect ratio (AR). Then, according to the principles of media aesthetics [12]–[14], they are well reintegrated with the direct-resized background to optimally match various screen sizes of client devices (cf. Fig. 1). Note that in this paper the term *video objects* will be used interchangeably to indicate the collection of UIOs and the background. The recomposing-based framework provides a number of advantages over the conventional schemes. First, it improves the visibility of user's interests as well as retains faithful context information, e.g., the viewer can see not only who but also where a person is in the video. Second, it allows multiple key objects to be emphasized at the same time and we can easily control the visual importance by adjusting their relative sizes. Third, it is robust to the shape distortion of objects caused by changes in video aspect ratio, which gives consistent content experience to different viewers.

Manuscript received March 21, 2006; revised September 22, 2006. This work was supported in part by the CIETNTU (MOE) and by the National Science Council of R.O.C. under Contracts NSC94-2622-E-002-024, NSC94-2752-E-002-006-PAE, and NSC94-2213-E-002-078. This paper was recommended by Associate Editor P. L. Correia.

W.-H. Cheng and J.-L. Wu are with the Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei 10617, Taiwan, R.O.C. (e-mail: wisley@cmlab.csie.ntu.edu.tw; wjl@cmlab.csie.ntu.edu.tw).

C.-W. Wang is with the Department of Computer Science and Information Engineering, National Taiwan University, Taipei 10617, Taiwan, R.O.C. (e-mail: nacci@cmlab.csie.ntu.edu.tw).

Color versions of Figs. 1–11 and 14 are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2006.885717

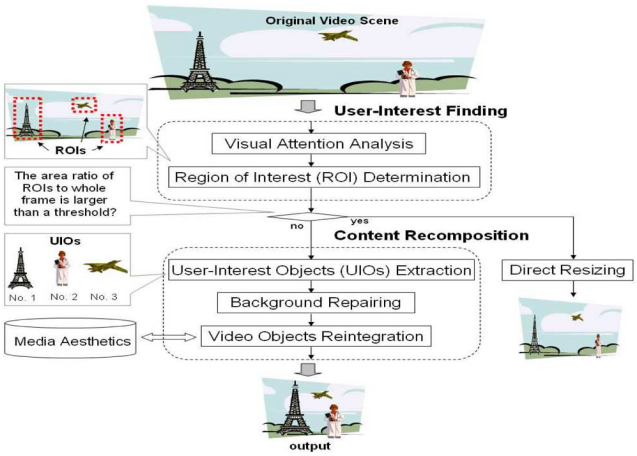

Fig. 1. Flowchart of the proposed framework for conducting video adaptation.

The main contributions of our work are twofold. First, a generic visual attention model is developed for video user-interest finding. The model is universal for its adequate utilization of inherent video characteristics, such as object and camera motions. Specifically, a high-level feature combination strategy based on the camera motion information is proposed. In addition, the motion feature model is integrated with confidence measures to improve its robustness and reliability. Second, based on the knowledge of media aesthetics, a set of aesthetic criteria is presented for guiding relevant decisions-making during video objects reintegration, such as the background positions to place UIOs. Without requiring user intervention, video content is automatically recomposed with satisfactory resultant visual rationality. We have conducted many experiments on various kinds of video data to demonstrate the efficiency and effectiveness of our approach.

The rest of this paper is organized as follows. After a discussion of related work, Section III presents a video attention model and associated algorithms for user-interest analysis and determination. The media aesthetics based content recomposition are described in Section IV. Section V shows experimental results, and Section VI presents our concluding remarks and the directions of our future work.

II. Related Work

In this section, we review previous studies on visual content adaptation. According to the design purposes, they are classified into three major categories, including *transcoding*, *cropping*, and *hybrid*. Meanwhile, their advantages and disadvantages to small displays will be briefly described as well.

In earlier works [4], [7], [8], the techniques of video transcoding have been extensively explored. The basic transcoding process is to convert a coded video signal from its original format into another one. An output format is determined entirely based on network and device constraints. Well-known transcoding methods include spatial resolution adjustment, temporal resolution adjustment, bit-rate adjustment, and coding syntax conversion [7], [8]. Scalable video coding is considered a special kind of transcoding techniques [15], [16]. The scalability is accomplished by providing multiple versions of a video stream so that the same contents of lower qualities are obtainable in different clients, e.g., Tung [16] developed a unified MPEG-4 video codec that supports universal scalability. For clients with small displays, spatial transcoding is always required but it causes excessive spatial resolution reduction or visual quality degradation. Once a visual content is scaled down more than its *minimal perceptible size*, the quality of service (QoS) or quality of experience (QoE) is usually far from acceptable [6], [9]. For example, some important details, such as the gesture of a drama actor or the ball location of a sport game, are not easy or even impossible to be recognized. Another difficulty with the spatial transcoding occurs when the aspect ratio of a target screen is inconsistent with the source video. If we linearly reduce both dimensions of the video to fit into the screen, it leaves black borders (sometimes known as the letterbox) and wastes valuable display resource. On the other hand, if the video is nonlinearly resized to occupy the whole screen, the resulting shape distortion of objects will annoy the viewer [13].

Much attention is then put on cropping-based approaches [17], [18]. First, Mohan *et al.* [2] proposed a general framework for adapting multimedia web documents, in which each media item (e.g., a video clip) is described with a multimodal and multiresolution representation hierarchy called the InfoPyramid. An importance value is subjectively assigned to each of the item combinations as the transcoding hint for content servers to dynamically select the best output. Similar ideas are also applied in Lee's work [5]. Instead of treating one video frame as a whole, selective presentation (or frame cropping) is allowed to improve the visibility of user's regions of interest (ROIs). Following their work, Chen *et al.* [6] developed an image adaptation system based on visual attention model. Using a simulated cognitive mechanism of human visual system (HVS) [19]–[21], the most important region is automatically determined. Better perceptual results have also been reported in other ROI-based video applications [22]–[24]. However, from the viewpoint of content authors, it not only destroys the carefully worked-out compositions but also distorts the overall conveyed messages. For example, if a visual scene is composed of multiple key objects, some of them are necessarily thrown away and a single ROI would fail to show the overview of their interrelationship. Moreover, significant information loss leads to viewer's misinterpretation about the original meaning that the authors want to communicate.

To preserve a complete video context or to clarify the specific user interest is not an either-or problem. Some hybrid approaches that lie between the two opposite extremes have been proposed. Liu *et al.* [25] presented a novel solution for browsing large pictures on mobile devices. All of the important regions are serially displayed and an optimal browsing path is calculated according to predicted shifting of visual attention. *Pan and Scan* [13] addressed an analogous technique for high-resolution video sources, but its discontinuous nature severely annoys the audience [13], [26]. Besides, the requirement of additional temporal resolutions conflicts with the primary video structures. The FilkFX corporation developed an awarded commercial system for transferring wide-screen films to the 4:3 aspect ratio of TV screens [26]. The intention is to generate a visually approximate replacement without object distortions. There-

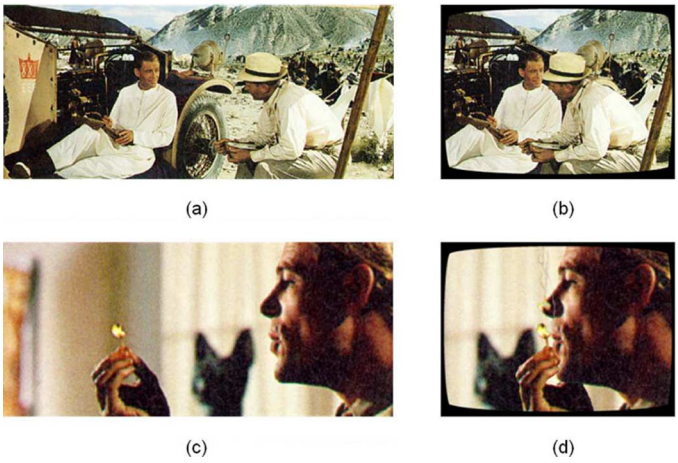

Fig. 2. Examples of semantic distortion in adapted videos. (a) and (c) Two original frames from the classical film "*Lawrence of Arabia*." (b) and (d) Corresponding adapted results using [26], respectively. With partial coverage, the two men of (a) no longer look at each other's eyes when they are chatting in (b), and the man in (d) seems more like to burn himself with the burning match rather than just hold it in (c). (Courtesy of FlikFX Ltd.).

fore, each of the film frames is condensed by eliminating the background portions of little significance and the main actors are artificially brought together to concentrate viewer's attention. However, without considering the original spatial interrelationship of video objects, semantic distortions are often generated, cf. Fig. 2. Recent work introduces nonuniform manipulations of the background and foreground information. Setlut *et al.* [10] decomposed an image into separate objects and unequally shrank them according to their relative visual importance. A side effect is that the relative size of different objects may be altered. Liu *et al.* [11] exploited a nonlinear warping transformation to emphasize the attractive foreground image regions but severe visual distortions are inevitable. Overall, the maintenance issue of user-perceived visual rationality in adapted results is not well addressed in the adaptation literature. Furthermore, although experiments show that nonuniform processing is more flexible to achieve superior performance, most discussions are confined to still images. These observations motivate our approach for motion pictures.

III. User-Interest Finding

UIOs are the semantic objects that catch part of the viewer's attention in videos, such as a walking person, a flower, an automobile, etc. Accordingly, an ROI is defined as the rectangular frame portion that contains some UIOs. Since the actual UIO shapes can be arbitrary, the ROIs serve as the tight bounding boxes of them. Their correct identification is the first key step for successful content recomposition. In intelligent image applications, a powerful mechanism for identifying the ROIs is visual attention modeling [19], [20]. Without truly understanding an image's content, several attentive features are extracted and combined into a single saliency map for representing local conspicuity. The computational attention methodology simplifies the problem of complex semantic analysis into a series of low-cost heuristic decisions. Several researches extend its capability for motion pictures by utilizing high-level video characteristics, such as speech, video genre, and lexical information [27], [28]. Unfortunately, most designs are too domain-specific to be applied to general purpose applications [21]. In the rest of this section, we will explain how to model generic visual attention in video clips. Rather than blindly adding semantic features (e.g., human face and text), a novel strategy for feature combination is presented to take the author's intentions into account. In addition, methods for dynamically determining the number of ROIs and their attributes (i.e., position and size) are also introduced.

A. Visual Attention Modeling

Visual attention refers to the ability of a viewer concentrating his attention on some visual objects or regions. Previous research showed that this physiological process could be modeled by a saliency-based attention model [19], [28], i.e., a saliency map computes an attractive value for each pixel or image block. Based on our previous studies [21], three types of video-oriented visual features (intensity, color, and motion) are adopted to model the visual attraction by using the same idea.

1) Contrast-Based Intensity and Color Feature Model: One of the most important ingredients of a visual attention model is the contrast [29]. In psychology, perceptual experiments have shown that the intensity and some color pairs possess high spatial and chromatic opposition. Accordingly, we include three contrast based feature models: intensity, red-green color contrast, and blue-yellow color contrast, into our visual attention analysis module. The contrast maps are, respectively, defined as follows:

$$\mathcal{M}_{\mathcal{I}}(p) = \max_{p' \in w_p} |\mathcal{I}(p) - \mathcal{I}(p')| \quad (1)$$

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

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

where $p = [x, y]^T$ is a position vector, w_p is a 3×3 window centered at p, and $\mathcal{I}, \mathcal{R}, \mathcal{G}, \mathcal{B}, \mathcal{Y}$ denote the intensity, red, green, blue, and yellow component value functions, respectively. That is, for each frame, the intensity and color feature values of a pixel p are computed from its local region w_p. For example, the intensity value of p is set to the maximum of the absolute intensity differences with its neighboring pixels p'.

2) Motion Feature Model: Object motion plays an essential role to direct an audience's attention across the scene space of a video [12]. Two feature models: x-motion and y-motion are, respectively, used to represent the horizontal and the vertical motion information in a scene. To find the motion activity of a specific direction, the two-dimensional (2-D) [30], [31] structure tensor (ST) is evaluated for each frame pixel. Compared with other motion descriptors, the 2-D ST is adopted for its confidence measure can also be estimated. The 2-D ST, J_x, for computing x-motion features is expressed as

$$J_x = \begin{bmatrix} J_{xx} & J_{xt} \\ J_{xt} & J_{tt} \end{bmatrix} = \begin{bmatrix} \sum_w H_x^2 & \sum_w H_x H_t \\ \sum_w H_x H_t & \sum_w H_t^2 \end{bmatrix} \quad (4)$$

where w is the 3×3 support window. H_x and H_t are, respectively, the partial derivatives of a horizontal slice along the

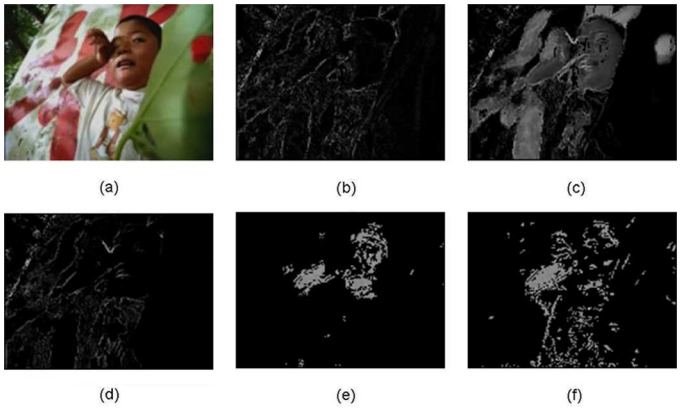

Fig. 3. Example of feature maps. (a) Original video frame. (b) Intensity. (c) Red-green color. (d) Blue-yellow color. (e) x-motion. (f) y-motion feature maps.

spatial and the temporal dimensions as defined in [30]. Consequently, the local motion angle θ_x and its corresponding confidence measure (cm_x) are computed as

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

and

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

The corresponding y-motion features, θ_y and cm_y, can be obtained in the same way. Finally, the x-motion and the y-motion maps are individually calculated as

$$\mathcal{M}_{\mathcal{X}}(p) = \theta_x \times cm_x \tag{7}$$
$$\mathcal{M}_{\mathcal{Y}}(p) = \theta_y \times cm_y \tag{8}$$

where $p = [x, y]^T$ is, again, a position vector. That is, for each frame, the motion feature values of a pixel p are its local motion angle θ_x (θ_y) multiplied by the corresponding confidence measure cm_x (cm_y). The confidence measure is used to suppress uncertain motions and to improve the reliability of obtained motion feature maps.

3) Camera Motion Based Saliency Map Generation: For each video frame, the distributions of individual visual features are calculated and constructed as five feature maps, as shown in Fig. 3. Then, according to the theory of visual attention model [19], [32], one saliency map is generated by their linear combinations, as described in the rest of this subsection. In the combining process, the selection of relative feature weights is important, which influences the accuracy of obtained saliency maps [32]. In the literature, the typical solution is assigning a single set of fixed feature weights for a whole video, e.g., the equal-weights [19], [27] and the video-genre-based schemes [28]. However, they are inflexible to adapt themselves to content variations of a video. Later, some dynamic fusion schemes were proposed, e.g., [32] and [33]. Although better results are obtained, the blindness to content semantics limits their extension for high-level applications.

TABLE I
WEIGHTS FOR THE FEATURE MAPS UNDER DIFFERENT CAMERA MOTION TYPES

	$\mathcal{M}_{\mathcal{I}}$	$\mathcal{M}_{\mathcal{RG}}$	$\mathcal{M}_{\mathcal{BY}}$	$\mathcal{M}_{\mathcal{X}}$	$\mathcal{M}_{\mathcal{Y}}$
zoom	0.2	0.2	0.2	0.2	0.2
pan	0.05	0.05	0.05	0.75	0.1
tilt	0.05	0.05	0.05	0.1	0.75
static	0.15	0.075	0.075	0.35	0.35
motion	0.05	0.05	0.05	0.425	0.425

From the viewpoint of media aesthetics [12], [13], different camera motions have different impacts on the audience's reception. They influence the relative importance of each visual feature and reveal what and where the author wants viewers to see. For example, a pan camera usually implies a tracking intention of some fast moving objects, e.g., a car in racing [12]. At this time, the horizontal motion feature would be more attractive to viewers and should have a larger weight than the other visual features. The study of task-oriented gaze control confirms the phenomenon. Under the same camera motion type, users' perceptual responses to the visual stimuli are generally consistent regardless of the video genres [21]. In videography, the techniques have been widely used in video productions especially in expert-produced videos [13]. Accordingly, the fact that directors purposely move their camera to control the audience's fixations appropriately serves as a high-level hint for integrating visual features [12], [13]. Therefore, we propose a camera motion based feature combination strategy for saliency map generation. Conceptually, this strategy can be viewed as one kind of the dynamic fusion schemes as prescribed. However, the fusion weights are determined by available high-level information (i.e., camera motion type) rather than the to-be-fused data itself.

In our work, five camera motion types are labeled: zoom, horizontal-pan, vertical-tilt, static-with-no-motion, and static-with-object-motion. Using an algorithm based on structure tensor histograms [30], one camera motion type is registered for every video frame. The saliency map S is generated according to the following equation:

$$S = \alpha_{c,1} \times FM_1 + \cdots + \alpha_{c,n} \times FM_n \tag{9}$$

where FM_i is the ith feature map of that frame, and $\alpha_{c,i}$ is the weight of corresponding FM_i under a given camera motion type c. Table I lists the feature weights for the adopted camera motion types. Instead of manual assignment, these parameters are defined via a supervised training process for feature weights selection (please refer to [21, Sec. 3.4.4] for the details). Note that the physical camera arguments (e.g., focal length) are not involved in the training process. The camera motion types are only used to classify the training data, and the same training process is separately conducted. The training data includes fifty video segments for each camera motion type, all of them are carefully chosen from various expert-produced films and TV programs. Each segment is 0.5 s long (roughly 15 frames) and contains a single camera motion type that is determined using the algorithm of [30].

The proposed feature combination strategy offers a number of advantages over the conventional ones [27], [28], [32]. First, it provides the capability of instant reaction to content variations. That is, the feature weights are dynamically selected according

to the high-level hint of registered camera motion types. Next, it is generic because the camera operations are always available in videos and their classifications have been well-defined [12], [13], [30]. Finally, it adds only moderate computational overhead since the required information (i.e., ST values) had been collected in the motion feature model.

B. Video ROIs Determination

In our work, an ROI is defined with two attributes: centroid position and region size (as described in the following). In this subsection, we describe how to compute the attributes for each ROI from a saliency map. Since there may be multiple key objects in a video frame, a method for dynamically determining the number of ROIs is also presented.

1) ROI Attributes Calculation: Saliency weighted regular moments [34] are effective to calculate the center coordinate of a set of weighted data points. They are adopted in our work to determine the centroid of each ROI. Let

$$m_{pq} = \sum_{x=1}^{M}\sum_{y=1}^{N} x^p y^q s(x,y), \quad p,q = 0,1,2,\ldots \tag{10}$$

where M, N are the dimensions of a saliency map and $s(x,y)$ is the saliency value function corresponding to the pixel (x,y). In the saliency map of the kth frame, the centroid position of an ROI is given by $(x_k, y_k) = (m_{10}/m_{00}, m_{01}/m_{00})$. Further, based on our observations, the region size of each ROI is proportional to the spatial distribution (area) of saliency values on a saliency map. A saliency weighted invariant [34] is defined to measure the variation of a computed centroid as follows:

$$\eta_{pq} = \frac{\sum_{x=1}^{M}\sum_{y=1}^{N}(x-\overline{x})^p(y-\overline{y})^q s(x,y)}{m_{00}}. \tag{11}$$

Consequently, the region size is set as $(e\sqrt{\eta_{20}}) \times (e\sqrt{\eta_{02}})$, where $e = 2$ is the expansion factor.

Meanwhile, we propose using a tracking technique of the discrete Kalman filter [35], [36] to estimate and correct the computed ROI attributes. Generally, an ROI centroid with its corresponding positions in the previous frames constitute a smoothly continuous trajectory on the screen. Therefore, a predicted centroid $(\bar{x_k}, \bar{y_k})$ of the ROI can be obtained with the prior information as follows [36]:

$$\begin{bmatrix} \bar{x_k} \\ \bar{y_k} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_{k-1} \\ y_{k-1} \\ \Delta_{x_{k-1,k-2}} \\ \Delta_{y_{k-1,k-2}} \end{bmatrix} + \begin{bmatrix} w_{x_{k-1}} \\ w_{y_{k-1}} \end{bmatrix} \tag{12}$$

where $(\Delta_{x_{k-1,k-2}}, \Delta_{y_{k-1,y-2}})$ is the centroid difference of the ROI between the $(k-1)$th and the $(k-2)$th frames, and $w_{x_{k-1}}$ and $w_{y_{k-1}}$ are two independent white noises with normal probability distribution $\mathbf{N}(0, 0.5)$. If the Euclidean distance of the computed (x_k, y_k) and the predicted $(\bar{x_k}, \bar{y_k})$ positions is larger than a dynamic threshold τ_k, the computed centroid (x_k, y_k) will be treated as unreliable. For example, a flashlight event often alters the spatial distribution of a saliency map and sharply shifts away the computed ROI centroid from where it should be. In this case, we would "propagate" the ROI from the previous frame instead. That is, the ROI centroid is set to the predicted position $(\bar{x_k}, \bar{y_k})$, and the region size would be the same as that in the $(k-1)$th frame. Finally, the dynamic threshold τ_k of the ROI in the kth frame is defined as

$$\tau_k = \gamma \| (\Delta_{x_{k-1,k-2}}, \Delta_{y_{k-1,y-2}}) \|_2 \tag{13}$$

where $\|\cdot\|_2$ denotes the two-norm operation of vectors and $\gamma = 5$ is the tolerance factor.

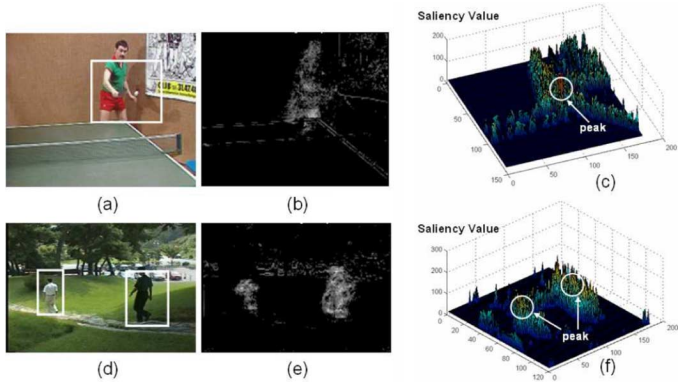

Fig. 4. Examples of a video frame with (a) one and (d) two ROIs (indicated by the white squares). (b) and (e) Corresponding saliency maps. (c) and (f) The 3-D profiles of the saliency maps of (a) and (d), respectively.

2) Dynamic Determination of ROIs: Sometimes, there are more than one ROI in a video frame. For example, in one view of a tennis game, two players may form two different ROIs. This scenario has to be explicitly addressed. In a saliency map, each ROI usually consists of a set of saliency values peaked at the center of its 3-D profiles. For example, if a video frame has two ROIs [e.g., there are two separate moving persons in Fig. 4(d)], its saliency map usually has two separate peaked sets, as shown in Fig. 4(f). We assume that the saliency value ranges from 0 to R (in this work, $R = 255$). If a pixel's saliency value is greater than a predefined threshold, it is added to the peak set (PS). All pixels in the PS are further grouped via an unsupervised clustering algorithm called the adaptive sample set construction [37]. Euclidean distance is chosen as the similarity measure because it works well when a data set has compact or isolated clusters [38]. Then, the peak set is divided into several disjoint subsets. That is

$$PS = \bigcup_{i=1}^{n} PS_i, \text{ where } PS_i \cap PS_j = \phi \text{ if } i \neq j. \tag{14}$$

In this way, a saliency map is partitioned into n regions, and each region corresponds to a peak subset PS_i. One ROI is declared for each region. With this scheme, the number of ROIs can be automatically and dynamically determined for each video frame.

Note that in practice the number of ROIs is purposely kept no more than three [12], [13]. Too many attractive objects in a frame will distract the attention of viewers. In such a case, like the scene of a busy street, the global view is preferred over individual objects. Therefore, we terminate the on-going clustering process and determine a single ROI based on the whole saliency map.

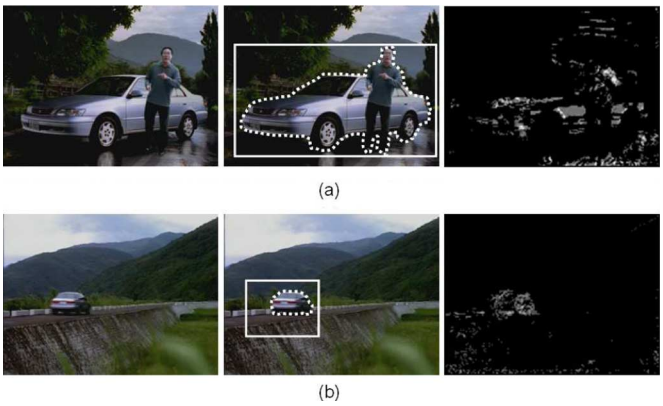

Fig. 5. Comparisons of the ROI and the UIO representations for user-interests. They are, respectively, indicated by the solid and dotted lines. In (a) and (b), the number of contained semantic objects (man together with a car versus one single car) is different.

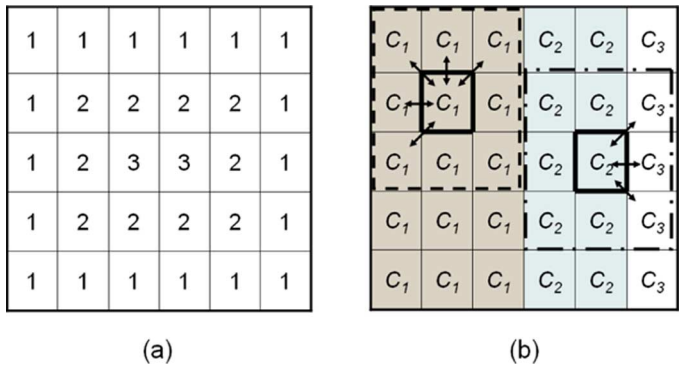

Fig. 6. Example of flooding operations with a 6×5 ROI. In (a), the number of each pixel indicates which border it belongs to. In (b), the left and the right pixels of a thick solid line are marked as the background and UIO, respectively. Their valid neighbors are connected with the arrows. (Let C_i be a color in RGB space and $d_\theta(C_2, C_3) > T_d$).

IV. Content Recomposition

In this section, we explain in detail the process of recomposing video content to fit within a target screen. After obtaining the ROI information, exact UIOs are separated from the background. Since the removal of UIOs leaves some scene holes on the background, an inpainting algorithm is applied to refill them. To emphasize the UIOs, we increase their relative scales with regard to the scene; meanwhile, to retain the video context, we resize the background to match the screen dimensions. The adapted result is then obtained by reintegrating all of the modified video objects. To ensure the resultant visual rationality, a set of criteria based on media aesthetics are developed for guiding the recomposition.

Before we proceed, a natural question may be raised here is whether all video scenes need to be recomposed. Obviously, if the UIOs have been appropriately emphasized by the author (i.e., large enough) as shown in Fig. 5(a), other lightweight options such as direct resizing seem to be sufficient. Therefore, we compute an area ratio of ROIs to the whole frame as a simple condition for thresholding (cf. Fig. 1). The threshold is empirically set to 0.65. For example, if the total area ratio of all ROIs in a frame is greater than the threshold, the direct resizing is applied instead.

A. UIOs Extraction

For simplicity, we assume that each ROI contains one single UIO. Aforementioned, the only difference between ROI and UIO is in the inclusion of partial background or not, cf. Fig. 5. In this definition, a UIO can possess one to several semantic objects. For example, in Fig. 5(b) the car itself constitutes a UIO, and in Fig. 5(a) the car together with a man form another one. Since the ROI information has to be available for all video frames, the extraction task is transformed to explicitly segment the UIO mask from its corresponding regions. For video presentation, the appearance of each frame is very short to the viewer so that the segmented results need not to be perfect, and the efficiency (i.e., the processing speed) seems to play a more important role. Therefore, we apply a real-time flooding procedure to mark the redundant background parts of an ROI [39], in this work.

Conceptually, an ROI is composed of a set of nonoverlapped rectangular borders with one-pixel width. For explanation, these borders are successively numbered as 1 to n from the most exterior one, e.g., the case of $n = 3$ is shown in Fig. 6(a). Initially, all pixels of the first border are marked as the background. Next, every pixel of the second border is compared with its neighbors that belong to the previous adjacent border. The valid neighbors are those not 2 pixels away from the target pixel, i.e., within a 3×3 support window [cf. Fig. 6(b)]. If the difference from any of its neighbor pixels is less than a fixed threshold T_d, it is marked as the background, otherwise the UIO. Fig. 6(b) gives an example. The same process continues through out the following borders. Meanwhile, two stop conditions are set and either one will end the flooding. The first condition is when it reaches the nth border, i.e., all pixels of the ROI have been scanned. The second one is when all pixels of the checked border at hand are marked as the UIO, i.e., it has got into the UIO interior and no more background pixels are left. Finally, the desired mask is obtained and used for extracting the UIO. Some examples of UIO extraction are shown in Fig. 7.

In the above, the difference of any neighboring pair (p_1, p_2) is defined by the *color vector angle* [40], whose value is computed as

$$d_\theta(C_1, C_2) = \left(1 - \frac{\left(C_1^T C_2\right)^2}{C_1^T C_1 C_2^T C_2}\right)^{1/2} \tag{15}$$

where C_1 and C_2 are the RGB color vectors of p_1 and p_2, respectively. In addition, the threshold T_d is set to 0.09 as suggested in [40]. The similarity measure d_θ is adopted for its insensitive to variations in intensity, yet sensitive to differences in hue and saturation. This property is useful for identifying meaningful object edges.

B. Background Repairing

To repair unfilled scene holes left by UIOs extraction, we develop an exemplar-based inpainting algorithm based on the work [41], in which the visible parts of a frame serve as a source set of exemplars (i.e., image patches) to infer the target region. Unlike other kinds of inpainting schemes, the missing data is replicated rather than synthesized from available information. It is superior in reducing blurring artifacts. More importantly, it

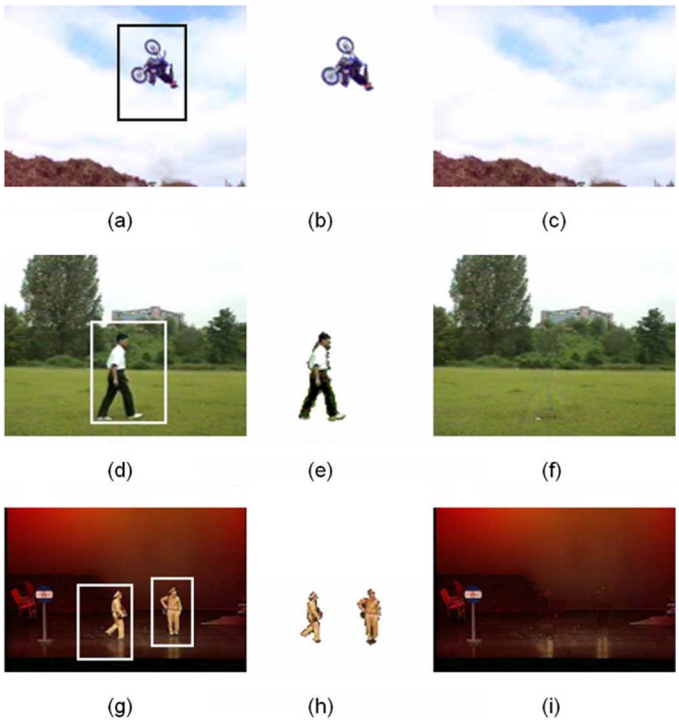

Fig. 7. Examples of video objects separation. The columns from left to right are successively the original frames with ROIs, extracted UIOs, and repaired backgrounds.

has better speed efficiency [41]. Some examples are shown in Fig. 7. The detailed algorithm and more corresponding results are reported in our previous work [42].

Based on the experiments, it is worthy to notice that most parts of a scene hole will be covered with the re-pasted UIOs. Few inpainted pixels, especially those along the scene hole boundary, are visible to viewers. Therefore, it is generally safe to moderate the computational overhand by merely repairing a partial region, e.g., we empirically suggest about 50% of the whole. For the same reason, we adopt image-based rather than video-based inpainting algorithm. The later often requires complex spatial-temporal analysis of videos, such as exact camera registration [43], [44]. For our applications, there is not much gain in doing so.

C. Media Aesthetics Based Video Objects Reintegration

As the final step, we reintegrate all of the separated video objects for content recomposition. Since the background is directly resized to match the target screen size, the reintegration task becomes making a proper arrangement of enlarged UIOs on the resized background. The relevant issues of a UIO include the following two points: one is where to paste it and the other is how large it should be. The discussion is difficult for its subjective nature [4]. Careless handling often incurs negative effects on the visual rationality. That is, the visual structure of a content would be altered to distort the conveyed message, e.g., Fig. 2. Fortunately, media aesthetics is an efficient process of examining media elements for identifying their roles in manipulating human perceptual reactions and synthesizing effective media productions [12]–[14]. It provides us a reliable basis for the automatic decisions-making in an objective and rational way.

1) Determination of UIO Positions: From the viewpoint of media aesthetics, the idea of increasing the relative scales of

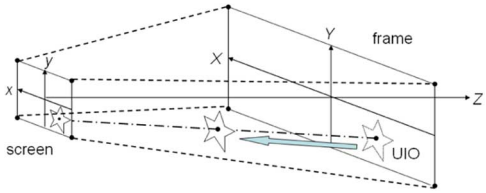

Fig. 8. Virtual 3-D scene model. All video objects of a frame are reprojected onto a target screen. An object is perceived larger (i.e., the star-shaped UIO) while it comes closer to the screen.

UIOs acts as enhancing the *depth cue* of a video scene [13]. If an object is larger in the scene, it seems closer to the viewer. The relation can be described with a virtual 3-D scene model [15] (cf. Fig. 8). It is obtained by using the perspective projection camera model. The major advantage of this model is its ability to enable transformations between different aspect ratios, as illustrated in Fig. 8. In this way, the corresponding coordinates of each pixel in a frame can uniquely be determined on the screen. Accordingly, we take the centroid of a UIO, say (X_c, Y_c), as its control point to obtain the target position (x_c, y_c)

$$(x_c, y_c) = \left(\frac{\text{Width}_S}{\text{Width}_F} X_c, \frac{\text{Height}_S}{\text{Height}_F} Y_c\right) \tag{16}$$

where Width_F, Width_S, Height_F, and Height_S are widths and heights of the source frame and the target screen, respectively.

2) Determination of UIO Size: For emphasizing the perceptual feeling, we would like to make the perceived UIOs as large as possible. However, to ensure the recomposed visual rationality, we introduce some aesthetic criteria to upper-bound the limits. For explanation, video objects of the fth frame are represented as follows:

$$VO_f = \{BG_f, \text{UIO}_{f,1}, \ldots, \text{UIO}_{f,k}\}, \quad 1 \le k \le 3 \tag{17}$$

where BG_f and $\text{UIO}_{f,i}$ are the repaired background and the ith UIO, respectively. Let BG_f^R be the resized BG_f to match the screen size, and $\text{UIO}_{f,i}^R$ be the enlarged $\text{UIO}_{f,1}$ by a scaling factor $r_{f,i}$. Conceptually, they are the "projected" images of the video objects on the screen surface (cf. Fig. 8). Since the aspect ratio of each UIO is kept fixed to avoid shape distortion, the same scaling factor is applied to both of its dimensions. In the following, we describe in detail and formulate each of the adopted aesthetic principles.

1) *Principle of Object Closure*: When showing only part of an object on the screen, we must frame the objects so that the viewer can easily fill in the missing parts and perceive the whole. That is, enough parts of the enlarged UIOs should be visible on the screen S to be faithfully recognized. This principle can be formulated as

$$\frac{\left|\text{UIO}_{f,i}^R \cap S\right|}{\left|\text{UIO}_{f,i}^R\right|} > \delta \tag{18}$$

where δ is empirically set to 0.9 and $|A|$ denotes the size of a video object A.

2) *Principle of Overlapping Planes*: When an object is partially covered by another, we perceive that the one that is doing the covering must be in front of the one that is partially covered. That is, the enlarged UIOs should not

be overlapped since they are originally nonoccluded. This principle can be formulated as

$$\mathrm{UIO}_{f,i}^{R} \cap \mathrm{UIO}_{f,j}^{R} = \phi, \quad i \neq j. \tag{19}$$

3) *Principle of Relative Size*: When the relative size of an object is smaller than another, we perceive the smaller one as being farther away and the larger one as being closer. That is, the relative size of the enlarged UIOs should be consistent with that of the originals to keep their inter-relationship. This principle can be formulated as '

$$\frac{|\mathrm{UIO}_{f,i}|}{|\mathrm{UIO}_{f,j}|} = \frac{|\mathrm{UIO}_{f,i}^{R}|}{|\mathrm{UIO}_{f,j}^{R}|}, \quad i \neq j. \tag{20}$$

It implicitly implies that $r_{f,i} = r_{f,j}$.

4) *Principle of On-Screen Continuity*: When the visual setting of a scene has been established, we must keep its consistency in the following frames to maintain the viewer's mental map. That is, the size changing pattern of the enlarged UIOs should be consistent with that of the originals. This principle can be formulated as

$$\frac{|\mathrm{UIO}_{f+1,i}|}{|\mathrm{UIO}_{f,i}|} = \frac{|\mathrm{UIO}_{f+1,i}^{R}|}{|\mathrm{UIO}_{f,i}^{R}|} \tag{21}$$

where $\mathrm{UIO}_{f+1,i}$ corresponds to the same object of $\mathrm{UIO}_{f,i}$ in its next frame. It implicitly implies that $r_{f,i} = r_{f+1,i}$.

From the principles 3 and 4, we know that values of the scaling factor are identical for all UIOs within the same shot, which effectively reduces the solution space for exploration. To avoid biasing, we compute a valid value range $[r_{\min}, r_{\max}]$ for each frame (as described in the following) and take the maximum from their intersections as our final result r^*. In this way, all enlarged UIOs are promised to satisfy the adopted aesthetic criteria. Obviously, the smaller scaling factor of the background would be the natural lower bound, i.e.,

$$r_{\min} = \min\left(\frac{\mathrm{Width}_S}{\mathrm{Width}_F}, \frac{\mathrm{Height}_S}{\mathrm{Height}_F}\right). \tag{22}$$

If the obtained scaling factor r^* is roughly equal to $r_{\min}$, the whole frame seems to be directly resized while the aspect ratio of UIOs is kept constant. Generally, we search the possible valid maxima $r_{\max}$ for each frame by linearly increasing the value of $r_{\min}$. Specifically, the value $r_{\max}$ of a frame is the maximum of possible scale factors that satisfying the adopted aesthetic principles. The search precision (sp) depends on the speed efficiency requirement of the application. In this work, it is empirically set to 0.1. Let a and b be the indexes of the first and the last frames of a shot, respectively. The process to obtain r^* of the shot can be summarized as follows.

• Step 1 (Initialization): $[r_{\min}, r_{\max}] \Leftarrow [\min(\mathrm{Width}_S/\mathrm{Width}_F, \mathrm{Height}_S/\mathrm{Height}_F), \infty]$

• Step 2:

for $f = a$ to b **do**

$r_{\mathrm{temp}} \Leftarrow r_{\min}$

while $r_{\mathrm{temp}} < r_{\max}$ **and** r_{temp} satisfies ALL the adopted aesthetic criteria of the fth frame **do**

$r_{\mathrm{temp}} \Leftarrow r_{\mathrm{temp}} + sp$

end while

$r_{\max} \Leftarrow r_{\mathrm{temp}}$

end for

• Step 3: $r^* = r_{\max}$.

It should be noted that if the additive incremental mechanism take a long time to find the solution, i.e., the selected search precision is very high, other fast algorithms like the variant binary search [45] can be applied instead.

TABLE II
SCREEN SIZES USED IN THE EXPERIMENTS

Type	Size (pixel^2)	Aspect Ratio (AR)
1	240×180	4:3
2	208×156	4:3
3	168×126	4:3
4	120×90	4:3

TABLE III
SOURCE CLIPS USED IN THE EXPERIMENTS

Clip	Resolution	AR	Duration	Content description
A	320×240	4:3	232 sec.	A motorcycle with a man flying into the sky and droping down on a hill.
B	320×240	4:3	245 sec.	An automobile gradually approaches from a distant place.
C	320×240	4:3	196 sec.	Chinese comic dialogue: two actors interact with plentiful facial expressions and body languages.
D	320×240	4:3	277 sec.	Chinese opera: an actress performs with fine gesture on the stage.
E	320×240	4:3	168 sec.	A man walks leisurely in the park from the left to the right sides.
F	320×240	4:3	217 sec.	A man enters the bathroom and looks around.
G	640×360	16:9	258 sec.	Including one scene of the film *Team America - World Police*: a police fights against a terrorist.
H	640×360	16:9	323 sec.	Including three scenes of the film *Homerun*: a boy runs, two student succor teams negotiate, and the boy and his sister happily walked together.

V. EXPERIMENTAL RESULTS

In this section, we conduct several experiments and compare our results with those of the conventional approaches [4], [8]. Then, we carry out user studies to verify the effectiveness of the proposed framework. Finally, the time efficiency of our approach is analyzed. Here, the technology of spatial resizing is chosen as the conventional approaches for the following two reasons. First, it is currently the most popular and dominant solution for adaptive video delivery [7]. Second, to our best knowledge, although some improved solutions other than the spatial resizing are proposed, there is no relevant work that focuses on video-based applications. Most efforts are put on still images as described in Section II.

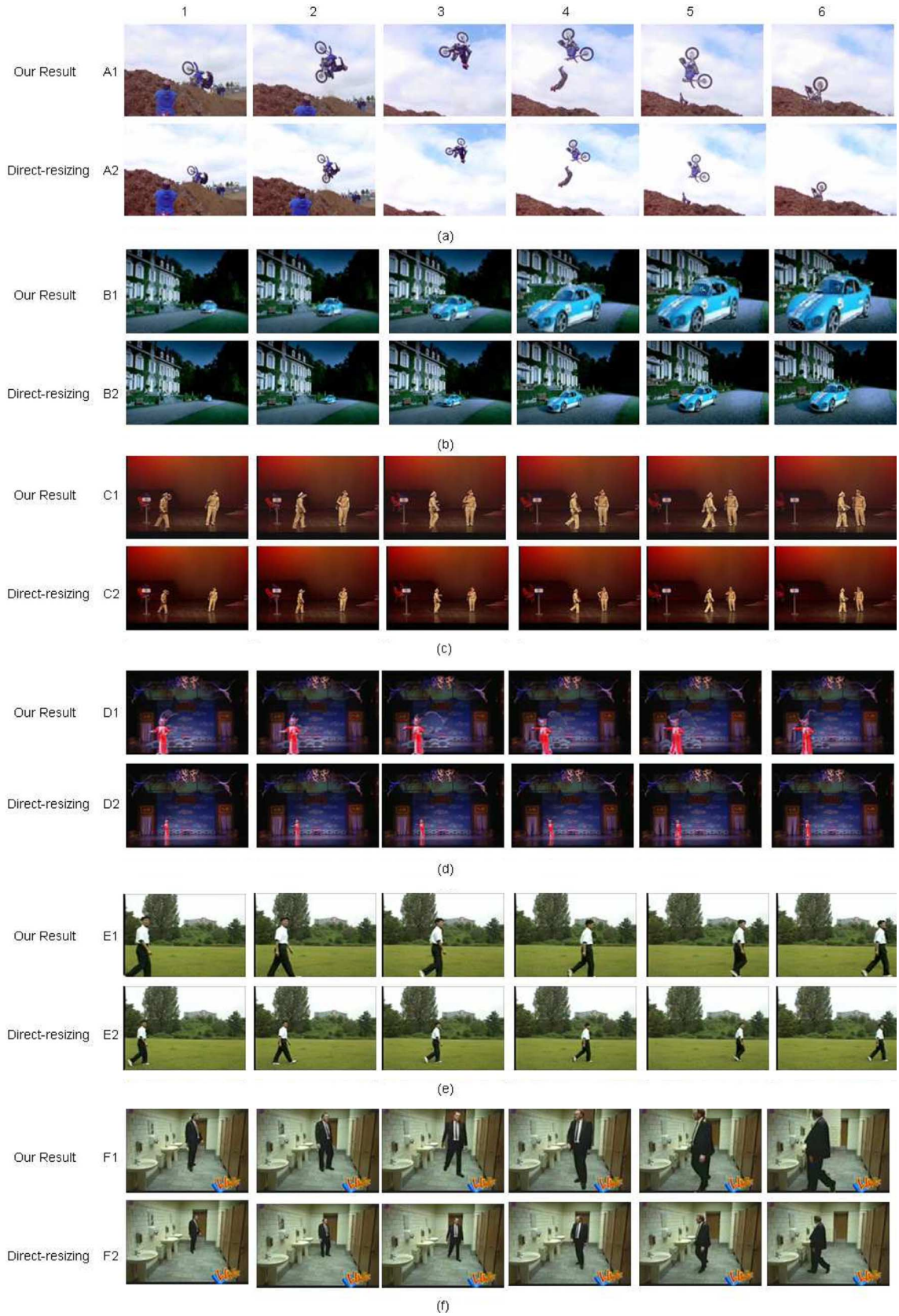

Fig. 9. Comparison of our approach with the conventional approach (direct-resizing) for the clips of subgroup 1.

As listed in Table II, four typical screen sizes of hand-held devices are adopted in our work [9]. All of them have a 4:3 AR. On the other hand, we have eight source clips as described in Table III. They are all expert-produced and each of them is about three to five minutes long. Some sample frames of each clip are illustrated in Figs. 9 and 10. Note that we use short

Fig. 10. Comparison of our approach with the conventional approaches (direct-resizing and linear-resizing) for the clips of subgroup 2.

clips rather than long sequences in the experiments, because observers' viewing fatigue has been reported to have a severe interference with user study [9]. In this way, it would be affordable to cover more kinds of video data. Further, taking into account the effect of video changes in ARs, the source clips are divided into two subgroups according to their ARs. Each clip of the subgroup 1 (i.e., clips A–F) is direct-resized into four testing clips with different resolutions as shown in Table II. Also, four testing clips are automatically generated by our approach. In addition to the direct-resizing and recomposition, linear-resizing is another adapting choice for the clips of subgroup 2 (i.e., clips G–H) since they have a different 16:9 AR with the adopted screen sizes. By definition, linear-resizing keeps the original video AR intact but direct-resizing changes it to match the adopted screen AR. Therefore, each clip of subgroup 2 will have two kinds of spatial resized testing clips for one specific resolution. In the rest of this section, the term conventional approaches will be used interchangeably to indicate both the direct-resizing and the linear-resizing approaches.

A. Recomposition Results

Figs. 9 and 10 illustrate partial results of our approach and the conventional approaches for clips of the subgroups 1 and 2, respectively. The frames are taken at the resolution format of screen type 3 (i.e., 168×126). For explanation, each frame is depicted by the adopted clip number followed by its virtual temporal index, e.g., C2–3. Generally, our approach outperforms the conventional ones based on the following observations.

1) Although the background contexts are faithfully preserved in all results, the key subjects are more effectively emphasized in our approach. In terms of visibility, our approach provides more useful information to the viewers.
2) It can be found that some of our results have lots of gain in the subject visibility and others have moderate improvement. For example, the UIO of Fig. 9(b) (i.e., the automobile) is emphasized more clearly than that of Fig. 9(a) (i.e., the motorcycle). The difference is due to the consideration of visual rationality. For example, in frame A1–4, the two UIOs (i.e., the man and the motorcycle) are originally located at a close distance to each other. According to the aesthetic principle of overlapping planes (i.e., (19)), mild enlargement is made to keep their spatial interrelationship. Similar phenomenon is observed in the results of Fig. 9(c) and (d).
3) The UIOs in clips C and D (i.e., the main actors and the actress) have fine gestures, facial expressions, and plentiful body language. They are important visual cues in dramatic

TABLE IV
TEST CONDITIONS OF THE USER STUDIES. (SEE SECTION V-B FOR DETAILS)

	TRIAL-I	TRIAL-II
Methodology	pair-comparison	score-rating
Display Device	17" LCD	17" LCD & 3.6" Smartphone
Viewing Distance	40 cm	40 cm & 30 cm
Video Resolution	screen type 3	all screen types (see Table II)
Testee Number	20	20

performance but most of them are almost unrecognizable with the conventional approach, cf. Fig. 9(c) and (d). Furthermore, some small things that carry specific meaning are also invisible, e.g., in Fig. 9(d), the white long feather on top of the actress's head. In contrast, the visibility of these details are improved with our approach. Even if the results are not perfect, they are at least "visible" to the viewers.

4) Our approach is flexible to deal with the case of video AR changes. We fully utilize the valuable space resource of a target screen and keep the UIO AR to avoid degradation of the viewer's visual comfort. Both advantages of the linear- and direct-resizing are effectively integrated in our approach. Fig. 10 demonstrates two real examples.
5) Sometimes our approach generates new visual artifacts in the recomposed video, e.g., the actress's incomplete feather ornament in the frame D1–2 and the defective boundary of the girl's head in the frame H1–6. Based on the experiments, these artifacts come from the imperfect UIO segmentation. Currently, as described later, we find that most viewers are not well aware of those artifacts. If visual artifact becomes a major concern in the application, other more accurate segmentation algorithms can be applied to resolve this shortage.

B. User Studies

To evaluate our approach, two user studies (TRIAL-I, TRIAL-II) are separately carried out. The objective of TRIAL-I is to determine if the accompanied content changes of our approach (e.g., enlarged UIOs) are visually acceptable to users, and to determine which of our approach and the conventional approaches is visually preferred. In TRIAL-II, we aim to investigate the effectiveness of our approach in practical usage. The viewer's viewing experience of our approach is compared with that of the conventional approaches on hand-held devices. For reference, the test conditions are listed in Table IV. The detailed methodology of each experiment is explained in the following:

1) Trial-I: For our purpose, two aspects of the results need to be investigated. One is whether UIOs are effectively emphasized, and another is whether recomposed videos look reasonable. Furthermore, the usefulness of our approach depends on whether it is actually preferred by viewers. Therefore, we apply a pair-comparison technique [6], [11] to study viewers' visual acceptability and preference. That is, at each time, an observer will be shown two different adapted results of the same video, and asked to subjectively decide which one would be better based on some predefined questions. In this way, it allows us to know the relative advantage and disadvantage of our approach.

In this study, the used testing clips are at the resolution format of screen type 3 as prescribed. Twenty participants are randomly invited in our campus. They are in the ages of 20 to 27, all with Chinese as their native language. Before joining the study, they have no ideas about our research work. Since the study will be conducted via the web, every participant is assigned a 17-in LCD at the viewing distance of 40 cm.

Fig. 11. Example of the displayed web page for TRIAL-I. For reality, both the pair of testing clips are presented on a virtual cellular phone. (See Section V-B1).

Initially, the testing goal, process, and relevant details are explained to the participants, such as descriptions of the predefined questions. For fair comparison, they are not told any details about our video adaptation algorithm, e.g., UIOs are extracted and reintegrated with the background. In addition, they are required to conceal personal interests in different video clips since the study focuses on viewer's visual experience rather than his/her emotional perception. After making sure that all participants understood the instructions clearly, we begin the experiment. At each time, a pair of testing clips (one is generated from our approach and another is from the conventional approaches) is displayed side by side on a web page, cf. Fig. 11. To be fair, the source videos are not presented to the participants in advance, and names of the corresponding approaches will not be prompted. Both the order of presentation and which clip to be appeared on the left or right side are independently randomized for each participant. After browsing the pair of clips, the participants are asked to answer the following eight predefined questions (Q1–Q8).

- Q1: In terms of UIOs, which clip is more visible to be recognized?
- Q2: In terms of UIOs, which clip appears with better motion and shape continuity on the screen?
- Q3: In terms of UIOs, which clip would you visually prefer?
- Q4: According to the relative size of video objects, which clip looks more reasonable?
- Q5: According to the interactive behavior of video objects, which clip looks more natural?
- Q6: According to the scene composition, which clip would you visually pleasant?
- Q7: Generally, for content comprehension, which clip would be more informative?
- Q8: Generally, for browsing on your hand-held device, which clip would you prefer to receive?

For each of the questions, three given comments are allowed for participants to choose as their answer: "the left one is better," "no difference," and "the right one is better." Note that the answering time is unrestricted and the pair of clips are allowed to

TABLE V
USER STUDY OF THE RELATIVE PREFERENCE (RP) OF OUR APPROACH WITH REGARD TO THE CONVENTIONAL APPROACHES

	Better	**No Diff.**	**Worse**	μ_{RP}	σ_{RP}
(A) OUR APPROACH VERSUS THE DIRECT-RESIZING for the clips of subgroup 1					
Q1	**95.00%**	3.33%	1.67%	+0.9333	0.0972
Q2	10.00%	**76.67%**	13.33%	-0.0333	0.2362
Q3	**53.33%**	30.00%	16.67%	+0.3667	0.5751
Q4	6.67%	**63.33%**	30.00%	-0.2300	0.3180
Q5	3.33%	**63.33%**	33.34%	-0.3000	0.2814
Q6	**51.67%**	38.33%	10.00%	+0.4167	0.4506
Q7	40.00%	**56.67%**	3.33%	+0.3667	0.3040
Q8	**83.33%**	10.00%	6.67%	+0.7667	0.3175
(B) OUR APPROACH VERSUS THE DIRECT-RESIZING for the clips of subgroup 2					
Q1	**90.00%**	7.50%	2.50%	+0.8750	0.1635
Q2	7.50%	**85.00%**	7.50%	+0.0000	0.1538
Q3	**77.50%**	20.00%	2.50%	+0.7500	0.2436
Q4	27.50%	**70.00%**	2.50%	+0.2500	0.2440
Q5	25.00%	**72.25%**	2.50%	+0.2250	0.2301
Q6	**70.00%**	25.00%	5.00%	+0.6500	0.3359
Q7	**62.50%**	17.50%	20.00%	+0.4250	0.6609
Q8	**92.50%**	5.00%	2.50%	+0.9000	0.1436
(C) OUR APPROACH VERSUS THE LINEAR-RESIZING for the clips of subgroup 2					
Q1	**97.50%**	2.50%	0.00%	+0.9750	0.0250
Q2	5.00%	**82.50%**	12.50%	-0.0750	0.1737
Q3	**72.50%**	7.50%	20.00%	+0.5250	0.6660
Q4	20.00%	27.50%	**52.50%**	-0.3250	0.6350
Q5	22.50%	**57.50%**	20.00%	+0.0250	0.4353
Q6	**62.50%**	22.50%	15.00%	+0.4750	0.5635
Q7	**70.00%**	20.00%	10.00%	+0.6000	0.4513
Q8	**87.50%**	10.00%	2.50%	+0.8500	0.1821

be repeated. The same process continues until all combinations of possible clips are tested for each participant.

For our testing purpose, Q1–Q3 concentrate on the UIO itself. Specifically, Q1, Q2, and Q3 examine whether our UIOs are visually emphasized, acceptable, and preferred to viewers, respectively. Here, the acceptance refers to user-perceived motion smoothness and shape consistency of UIOs, which is affected by adopted underlying algorithms, such as the UIO segmentation. Therefore, Q2 in some sense serves as a performance index of our system. Next, Q4–Q6 relate to the user-perceived visual rationality of the whole recomposition. The static and dynamic visual perceptions are individually explored in Q4 and Q5. Further, Q7 explores the assistance in content comprehension and helps us to know the functional role of our approach. Finally, Q8 investigates whether viewers would like to receive recomposed videos in practical applications, which demonstrates the usefulness of our approach.

Table V shows the statistical results of our approach. According to the categories of clip subgroups and competitive approaches, the results are further divided into three subtables [Table V(a)–(c)]. Note that the fourth "worse" column denotes the percentage that the competitive approaches are chosen as better by viewers. For reference, we compute the weighted value μ_{RP} as an index of the user's relative preference (RP) to our approach, that is

$$\mu_{\mathrm{RP}} = \frac{((+1)\cdot f_b + 0\cdot f_n + (-1)\cdot f_w)}{100} \tag{23}$$

where f_b, f_n, and f_w are the "better," "no difference," and "worse" percentages for a specific question in a subtable, respectively. Clearly, μ_{RP} is in the range of $[-1, 1]$. If the value is positive, our approach would be more preferred by users, otherwise the conventional approaches. The RP strength is measured by its absolute magnitude, i.e., $|\mu_{\mathrm{RP}}|$. Meanwhile, a corresponding RP variance σ_{RP} is estimated.

According to Q1's statistics in Table V, our approach is really helpful to improve the visibility of UIOs for viewers. As shown in Q3's statistics, most of the viewers also prefer such an improvement, but there is a 10%-40% decrease in the "better" percentage and the RP variance is high. Based on our observations, it is mainly caused by two reasons: First, the motion and shape continuity of emphasized UIOs is not perfect in our approach, e.g., shape inconsistency of the actress's feather tail in clip D1 as prescribed. As shown in Q2's statistics, some viewers are displeased to this kind of artifacts [e.g., there is a 13.33% "worse" in Table V(a)] and would rather visually prefer the conventional approaches with smaller UIOs. Second, the effectiveness of UIO emphasis is content dependent. For example, in Fig. 10, it is useful to emphasize the boy of clip H for showing his important details to viewers, such as the facial expression. However, in Fig. 9, it becomes less meaningful for the car of clip B since viewers can easily recognize its appearance even in a smaller form. In this case, our approach is not specially preferred by viewers. That is also the reason why we have a lower "better" percentage (53.33%) and a higher "no difference" percentage (30.00%) of Q3 in Table V(a) than those in Table V(b) and (c).

Further, according to statistics of Q4 and Q5 in Table V, the visual rationality of our approach is generally acceptable to viewers. We find an exception is in Q4's statistics of Table V(c). One reason is due to the artificial essence of our approach. Since we recompose videos with software-based techniques rather than real video reshooting, the visual rationality of our approach could not be so realistic as that in the original. Another reason is that the object distortion caused by AR change is undesirable to viewers, which makes the perceived relative size of video objects visually unreasonable. It is found that the shape distortion seems more intolerable to viewers. For example, compared with the Q4 statistics in Table V(a), the "worse" percentage decreases to 2.50% in Table V(b), but increases to 52.50% in Table V(c). The more the video objects are distorted in AR, the more the relative size of them looks unreasonable. An interesting phenomenon is that even if the viewers are aware of those visual imperfection, in average, they still prefer the scene composition of our approach, cf. Q6's statistics in Table V. Notice that the terms "better" and "worse" in Table V do not mean absolute success or failure but the relative performance of the proposed or the conventional approaches. Therefore, we can say that although the visual rationality of recomposed videos is not perfect, it does not fall far short either when compared with the original one. The recomposed visual quality seems good enough to be accepted by most viewers.

Finally, in Table V, Q8's statistics show that most of the participants are willing to receive our results for practical usage. Furthermore, as shown in Q7's statistics, our approach improves the viewer's comprehension of video contents; however, the corresponding RP variance (σ_{RP}) is high, as shown in Table V(b).

From the perspective of information delivery, this makes it plausible that viewers would prefer our approach for its informative benefits that attributed to the enhanced visibility of important details. Overall, while the results of our preliminary experiments may be inconclusive, we find it encouraging. The proposed approach seems really helpful to improve the video experience for mobile users. Also, the mobile users would prefer our approach to obtain the improvements.

2) Trial-II: To further evaluate the effectiveness of our approach in practical use, we carefully design the experiment to assess viewers' subjective satisfaction with the viewing experience on hand-held devices. Specifically, the satisfaction refers to the level that a viewer is satisfied with his/her viewing experience of an adapted video on a hand-held device when compared with that of the original video on a standard display. For our purpose, overall, the viewing experience is defined as the user-perceived detail visibility, visual rationality, and browsing ease of video content. Specifically, the detail visibility indicates the user-perceived clarity of small objects or things in a scene; the visual rationality relates to the user-perceived relative size and interactive behavior of video objects, cf. TRIAL-I; the browsing ease refers to if the user can comfortably view a whole video. In this way, we are able to measure how effective our approach would be in improving the viewing experience for mobile users.

In this study, the used testing clips are at all resolution formats as described in Table II. To ensure the validity, another twenty participants different from TRIAL-I are randomly invited. A 17-in LCD and a Dopod 900 Smartphone (with a 3.6-in LCD) are assigned to each participant as the testing platforms. They are set at the viewing distance of 40 and 30 cm, and treated as the standard display and the hand-held device, respectively.

Initially, the testing purpose, process, and relevant details are explained to the participants, e.g., definitions of the viewing experience and subjective satisfaction. For fair comparison, they are not told any details about our video adaptation algorithm and required to conceal personal interests in different video clips, as like in TRIAL-I. Then, one of the source clips at its original format (cf. Table III) is shown to participants through the standard display. To avoid viewers' misconception, the participants are asked to read the corresponding content descriptions in Table III. The playing is repeated until all of the participants have well understood the video content. Next, all the corresponding testing clips of that source clip, one at a time, are presented on the hand-held device. For each of the testing clips, when the playing is finished the participants have one minute to give a subjective score in the range of 0–1 with two decimal places at most, e.g., 0.75. The score value is proportional to the viewer's relative satisfaction with that clip as prescribed. For example, if the perceived viewing experience for a viewer is about the same as that on the standard display, the viewer will give a large score value, otherwise a small one instead. Note that the replay is inhibited and the answering time is restricted since we believe the viewer's first impression without reconsideration reveals his/her true satisfaction about the viewing experience. To avoid biasing, testing clips of the same resolution format are displayed in series and at a random order. In the following, the same process is conducted for all of the other source clips.

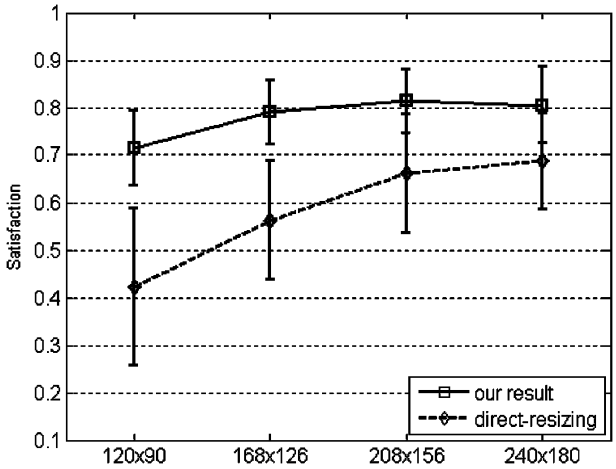

Fig. 12. Comparison of the user study between our approach and the conventional approach (direct-resizing) for the clips of subgroup 1 at different resolution formats.

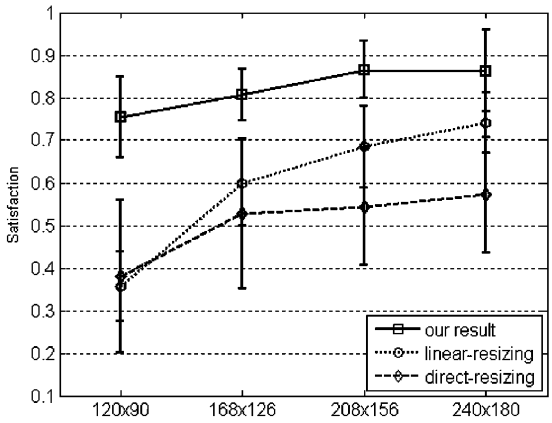

Fig. 13. Comparison of the user study between our approach and the conventional approaches (direct-resizing and linear-resizing) for the clips of subgroup 2 at different resolution formats.

Figs. 12 and 13 illustrate the statistical comparisons of the user studies between our approach and the conventional approaches for the clips of subgroup 1 and subgroup 2, respectively. In the figures, each of the points is obtained by averaging the participants' satisfaction of the adapted results of an approach at a fixed resolution format. The symmetric error bar indicates two standard deviation units in length. In addition, the techniques of hypothesis testing are applied to obtain the statistical significance (P-value) of our approach [46]. Since the claim is that the viewer's satisfaction with our approach is higher than that of the conventional approaches, the P-value provides the probability that the difference (i.e., improvement) in the experiment happened by chance.[1] For each resolution format in Figs. 12 and 13, we compute a P-value between our approach and one conventional approach using the upper-tailed t-test with $n-2$ degrees of freedom [46], where n is the number of observed viewers' satisfaction. Specifically, for each resolution format, we obtain one P-value between our approach and the direct-resizing in Fig. 12. Similarly, we obtain two P-values (one between our approach and the linear-resizing, another between our approach and the direct-resizing)

[1][Online]. Available: http://teachmefinance.com/Scientific_Terms/p-value.html

TABLE VI
TIME EFFICIENCY ANALYSIS OF THE PROPOSED FRAMEWORK FOR RECOMPOSING A 320 × 240 VIDEO FRAME

Components	Time (sec.)	Percentage (%)
Attention analysis	0.7500	3.45
ROI determination	0.0780	0.36
UIO extraction	0.1418	0.65
Background repairing	20.6661	95.02
VOs reintegration	0.1135	0.52
Total	21.7497	100

in Fig. 13. The P-value results show that except for the cases at resolution 240×180 with the linear-resizing and at 168×126 with the direct-resizing in Fig. 13 (i.e., 0.006 and 0.001, respectively), the other P-values are far less than 0.001.

Generally, according to the average satisfaction in Figs. 12 and 13, our approach outperforms the conventional approaches in all cases. It is found that the satisfaction of our approach remains high (above 0.7) throughout all resolution formats, but that of the conventional approaches drop rapidly down to an unacceptable level as the screen size decreases. This phenomenon indicates that important visual details (e.g., UIOs) have dominant effects on the viewing experience, which confirmed the statements given in [9]. Fig. 13 exhibits another fact that viewers prefer linear-resizing to direct-resizing when there is an AR mismatch between the source video and the target screen. As also indicated in TRIAL-I, it is interesting to find that although the linear-resizing wastes a large amount of screen space, the UIO distortion seems more intolerable to viewers. In summary, our approach is helpful to maintain acceptable video property and generates comfortable viewing results for viewers.

C. Time Efficiency Analysis

We analyze the time efficiency of our approach by logging the computational time costs. Without loss of generality, we only include the time costs for clips of the subgroup 1. The proposed framework is programmed using Matlab 6.5. Our test bed is Acer VT7600 PC with Intel P4 3.0 GHz CPU, 1.0 GB memory, and MS Windows XP system. The average processing time for recomposing a 320×240 video frame is currently about 21 s. The time cost of each underlying component is shown in Table VI. Obviously, the inpainting algorithm is the most time-consuming one. It takes more than 95% of the total time. However, as prescribed, we can reduce the time cost by merely repairing less than half parts of a scene hole. Moreover, with the help of some advanced techniques, such as the program porting to compiled languages like C/C++, code optimization, and system on chip (SoC) design, our approach could achieve real-time performance with confidence. For example, the technique of field-programmable gate array (FPGA) has been recently adopted by some researchers as a fast and low-cost way for creating real-time software applications [47], [48]. In other words, the proposed framework is general and practical enough to be employed on various kinds of adaptive content delivery systems.

VI. DISCUSSION AND CONCLUSION

This paper presents a novel framework for video adaptation based on content recomposition. Our approach is superior to existing schemes in that it emphasizes the important aspects of a scene while faithfully retaining the background context. It also considers the visual rationality of recomposed content and is robust to video changes in aspect ratio. Therefore, the proposed framework can provide more effective and informative video experience to viewers, in an automatic way.

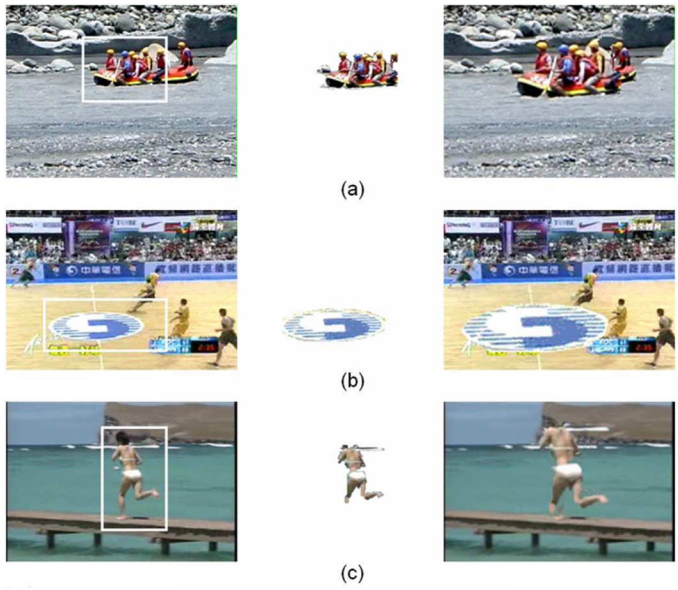

Fig. 14. Failure examples of our approach. The columns from left to right are successively the original frames with ROIs, extracted UIOs, and recomposed frames.

Many aspects of our approach can be improved. For example, currently, we have a fixed expansion factor in the ROI determination module. A risk is that actual semantic objects may not be completely contained in a determined ROI, e.g., Fig. 14(a). The phenomenon partially comes from the fact that the visually salient regions are not exactly corresponding to semantic objects. It is one essential limitation of the visual attention models [19], [20]. Therefore, a promising direction for future research is to integrate the proposed framework with other semantic-level techniques of video understanding and computer vision, such as the "task-relevance map" [49] which tells where human eye's attentions are voluntarily focused on one or more objects that are predefined or meaningful goals to the viewers. For example, in Fig. 14(b), the detected ROI (i.e., the ground logo) does not match the viewer's semantic attention. Another example is the UIOs extraction. Since the automatic and precise object segmentation from normal videos is extremely difficult and still an open problem [50], [51], we use a simpler algorithm to trade segmentation accuracy for processing time. However, as showed in the user studies, the segmentation accuracy does have impacts on the viewer's perceptual satisfaction. The development of a robust segmentation algorithm will be one of our future research directions. Another failure example is shown in Fig. 14(c). Since the woman's head color is more similar to that of the cliff surface rather than that of her body skin, it is segmented as a part of the background and erroneously separated from the body. Besides, speed efficiency is still a big issue in our approach. The underlying components should be effectively optimized and efficiently coupled together. Finally, it is obvious that the recomposed results will be better if we can "discuss" (i.e., interact) with the content authors in some ways. Therefore, the proposed

framework should be integrated with the standardized description schemes for content authors to specify some usage rules. For example, the fifth part of MPEG-21 [3], [52] specifies a machine-readable rights expression language (REL) for declaring rights and permissions, which provides mechanisms to protect digital contents and honors the rights of content authors. In addition, the tenth part digital item processing (DIP) specifies digital item methods (DIMs) as a way for content authors to provide manipulation suggestions of a digital content.

More extensive and complete evaluation of our approach is of importance. One task is to assess the viewer's perceptual response to the recomposed content. We believe that UIOs should be highly emphasized to provide more important information, but we have no idea whether it is appropriate to all kinds of video data and where is the threshold limit value (TLV) of viewers' perceptual comfort. Specifically, the TLV represents the subjective limit that viewers would like to accept such an emphasis. Its investigation assists in clarifying the application scope of our approach. The influence of accompanying audio in the viewer's visual experience is another issue. The study of human visual and aural perceptual interaction would be very helpful. Besides, a fundamental problem is the lack of standardized testing video database. In the experiments, we have attempted to describe and illustrate all of our testing clips as clearly as possible. However, if the number is largely increased, it will be tedious to do this and hard to reproduce the experiments.

A limitation of our approach is that the modified spatial cues (e.g., scene depth and object size) of videos may not be acceptable to some applications, e.g., sports programs, medical teaching clips, astronomical observing videos, etc. Specifically, our approach is unsuitable for those accuracy-sensitive or distortion-intolerant applications. Another limitation is that the adaptation is performed only in the spatial domain. Although it is already highly useful in most existing application scenarios, more flexible and economic methods should be studied further. For example, spatio-temporal based recomposing techniques are effective to reduce the computational overhead. Besides, corresponding methods in the compressed domain are always required for practical demands. The synchronization between the adapted video and its original audio tracks is also an untouched issue. In the future, we will continue our investigation in these directions.

ACKNOWLEDGMENT

The authors would like to thank the members of the Communication and Multimedia Laboratory (CMLab) at the National Taiwan University, Taiwan, R.O.C., for their assistances in conducting the experiments, and the anonymous reviewers for their valuable comments.

REFERENCES

[1] F. Pereira and I. Burnett, "Universal multimedia experience for tomorrow," *IEEE Signal Process. Mag.*, vol. 20, no. 2, pp. 63–73, Mar. 2003.

[2] R. Mohan, J. R. Smith, and C.-S Li, "Adapting multimedia internet content for universal access," *IEEE Trans. Multimedia*, vol. 1, no. 1, pp. 104–114, Mar. 1999.

[3] J. Bormans, J. Gelissen, and A. Perkis, "MPEG-21: The 21st century multimedia framework," *IEEE Signal Process. Mag.*, vol. 20, no. 2, pp. 53–62, Mar. 2003.

[4] S.-F. Chang and A. Vetro, "Video adaptation: Concepts, technologies, and open issues," *Proc. IEEE*, vol. 93, no. 1, pp. 148–158, Jan. 2005.

[5] K. Lee, H.-S. Chang, S.-S. Chun, H. Choi, and S. Sull, "Perception-based image transcoding for universal multimedia access," in *Proc. 8th Int. Conf. Image Process. (ICIP'01)*, 2001, vol. 2, pp. 475–478.

[6] L.-Q. Chen, X. Xie, X. Fan, W.-Y. Ma, H.-J. Zhang, and H.-Q. Zhou, "A visual attention model for adapting images on small displays," *Multimedia Syst.*, vol. 9, no. 4, pp. 353–364, Oct. 2003.

[7] I. Ahmad, X. Wei, Y. Sun, and Y.-Q. Zhang, "Video transcoding: an overview of various techniques and research issues," *IEEE Trans. Multimedia*, vol. 7, no. 5, pp. 793–804, Oct. 2005.

[8] J. Xin, C.-W. Lin, and M.-T. Sun, "Digital video transcoding," *Proc. IEEE*, vol. 93, no. 1, pp. 84–97, Jan. 2005.

[9] H. Knoche, J. D. McCarthy, and M. A. Sasse, "Can small be beautiful? Assessing image resolution requirements for mobile TV," in *Proc. 13th ACM Int. Conf. Multimedia (MM'05)*, 2005, pp. 829–838.

[10] V. Setlur, S. Takagi, M. Gleicher, R. Ramesh, and B. Gooch, Automatic Image Retargeting Comp. Sci. Dept., Northwestern Univ., Evanston, IL, 2004, Tech. Rep. NWU-CS-04-41.

[11] F. Liu and M. Gleicher, "Automatic image retargeting with fisheye-view warping," in *Proc. 18th ACM Symp. User Interface Technol. (UIST'05)*, 2005, pp. 153–162.

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

[13] H. Zettl, *Sight, Sound, Motion: Applied Media Aesthetics*, 3rd ed. Belmont, CA: Wadsworth, 1998.

[14] C. Dorai and S. Venkatesh, "Computational media aesthetics: Finding meaning beautiful," *IEEE Multimedia*, vol. 8, no. 4, pp. 10–12, Oct./Dec. 2001.

[15] Y. Wang, J. Ostermann, and Y.-Q. Zhang, *Video Processing and Communications*. Upper Saddle River, NJ: Prentice Hall, 2001.

[16] Y.-S. Tung, "The design and implementation of an MPEG-4 based universal scalable video codec in layered path-tree structure," Ph.D. dissertation, Dept. Comp. Sci. Inf. Eng., National Taiwan University, Taipei, Taiwan, R.O.C., 2002.

[17] J. Nam, Y.-M. Ro, Y. Huh, and M. Kim, "Visual content adaptation according to user perception characterstics," *IEEE Trans. Multimedia*, vol. 7, no. 3, pp. 435–445, Jun. 2005.

[18] A. Cavallaro, O. Steiger, and T. Ebrahimi, "Semantic video analysis for adaptive content delivery and automatic description," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 10, pp. 1200–1209, Oct. 2005.

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

[20] C. M. Privitera and L. W. Stark, "Algorithms for defining visual regions-of-interest: Comparison with eye fixations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 9, pp. 970–982, Sep. 2000.

[21] W.-H. Cheng, W.-T. Chu, and J.-L. Wu, "A visual attention based region-of-interest determination framework for video sequences," *IEICE Trans. Info Syst.*, vol. E-88D, no. 7, pp. 1578–1586, Jul. 2005.

[22] C.-W. Lin, Y.-C. Chen, and M.-T. Sun, "Dynamic region of interest transcoding for multipoint video conferencing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 10, pp. 982–992, Oct. 2003.

[23] C.-C. Ho, J.-L. Wu, and W.-H. Cheng, "A practical foveation-based rate-shaping mechanism for MPEG videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 11, pp. 1365–1372, Nov. 2005.

[24] M. M. Hannuksela, Y.-K. Wang, and M. Gabbouj, "Isolated regions in video coding," *IEEE Trans. Multimedia*, vol. 6, no. 2, pp. 259–267, Apr. 2004.

[25] H. Liu, X. Xie, W.-Y. Ma, and H.-J. Zhang, "Automatic browsing of large pictures on mobile devices," in *Proc. 11th ACM Int. Conf. Multimedia (MM'03)*, 2003, pp. 148–155.

[26] Digital Recomposition System, FlikFX Pty GmbH Ltd, (1999) [Online]. Available: http://www.widescreenmuseum.com/flikfx/

[27] Y.-F. Ma, L. Lu, H.-J. Zhang, and M.-J. Li, "A user attention model for video summarization," in *Proc. 10th ACM Int. Conf. Multimedia (MM'02)*, 2002, pp. 533–542.

[28] C.-C. Ho, W.-H. Cheng, T.-J. Pan, and J.-L. Wu, "A user-attention based focus detection framework and its applications," in *Proc. 4th Pacific-Rim Conf. Multimedia (PCM'03)*, 2003, pp. 1341–1345.

[29] S. Engel, X. Zhang, and B. Wandell, "Colour tuning in human visual cortex measured with functional magnetic resonance imaging," *Nature*, vol. 388, no. 6637, pp. 68–71, 1997.

[30] C.-W. Ngo, T.-C. Pong, and H.-J. Zhang, "Motion analysis and segmentation through spatio-temporal slices processing," *IEEE Trans. Image Process.*, vol. 12, no. 3, pp. 341–355, Mar. 2003.

[31] B. Jahne, *Spatio-Temporal Image Processing: Theory and Scientific Applications*. New York: Springer-Verlag, 1991.
[32] L. Itti and C. Koch, "A comparison of feature combination strategies for saliency-based visual attention systems," in *Proc. SPIE Human Vision and Electronic Imaging IV (HVEI'99)*, 1999, pp. 473–482.
[33] X.-S. Hua and H.-J. Zhang, "An attention-based decision fusion scheme for multimedia information retrieval," in *Proc. 5th Pacific-Rim Conf. Multimedia (PCM'04)*, 2004, pp. 1001–1010.
[34] R. Paramesan, P. Ramaswamy, and S. Omatu, "Regular moments for symmetric images," *Electron. Lett.*, vol. 34, no. 15, pp. 1481–1482, Jul. 1998.
[35] P. S. Maybeck, *Stochastic Models, Estimation, and Control*. New York: Academic, 1979, vol. 1.
[36] G. Welch and G. Bishop, An introduction to the Kalman filter Dept. Comp. Sci., Univ. North Carolina at Chapel Hill, 2004, Tech. Rep. 95-041.
[37] S. T. Bow, *Pattern Recognition and Image Preprocessing*. New York: Marcel Dekker, 2002.
[38] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Comput. Surveys*, vol. 31, no. 3, pp. 264–323, Sep. 1999.
[39] C.-H. Chang, C.-H. Wu, J.-C. Chen, J.-H. Kuo, and J.-L. Wu, "A real-time semi-automatic video segmentation system based on mathematical morphology," in *Proc. Vis. Comm. Image Process. (VCIP'05)*, 2005, pp. 2152–2162.
[40] H.-Y. Lee, H.-K. Lee, and Y.-H. Ha, "Spatial color descriptor for image retrieval and video segmentation," *IEEE Trans. Multimedia*, vol. 5, no. 3, pp. 358–367, Sep. 2003.
[41] A. Criminisi, P. Perez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE Trans. Image Process.*, vol. 13, no. 9, pp. 1200–1212, Sep. 2004.
[42] W.-H. Cheng, C.-W. Hsieh, S.-K. Lin, C.-W. Wang, and J.-L. Wu, "Robust algorithm for exemplar-based image inpainting," in *Proc. Int. Conf. Comput. Graphics, Imaging Vis. (CGIV'05)*, 2005, pp. 64–69.
[43] K. A. Patwardhan, G. Sapiro, and M. Bertalmio, "Video inpainting of occluding and occluded objects," in *Proc. 12th Int. Conf. Image Process. (ICIP'05)*, 2005, vol. 2, pp. 69–72.
[44] V. Cheung, B. J. Frey, and N. Jojic, "Video epitomes," in *Proc. CVPR'05*, 2005, vol. 1, pp. 42–49.
[45] T. G. Cormen, C. E. Leiserson, and R. L. Rivest, *Introduction to Algorithms*, 2nd ed. Cambridge, MA: MIT Press, 2001.
[46] J. L. Devore, *Probability and Statistics for Engineering and the Sciences*, 4th ed. Belmont, CA: Wadsworth, 1995.
[47] A. K. Gupta, S. Nooshabadi, D. Taubman, and M. Dyer, "Realizing low-cost highthroughput general-purpose block encoder for JPEG2000," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 7, pp. 843–858, Jul. 2006.
[48] J. Diaz, E. Ros, F. Pelayo, E. M. Ortigosa, and S. Mota, "FPGA-based real-time optical-flow system," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 2, pp. 274–279, Feb. 2006.
[49] V. Navalpakkam and L. Itti, "A goal oriented attention guidance model," *Lecture Notes Comput. Sci.*, vol. 2525, pp. 453–461, Nov. 2002.
[50] R. R. Sarukkai, "Video search: Opportunities and challenges," in *Keynote Speech at ACM MIR Workshop, 2005 ACM Multimedia Conf. (MM'05)* [Online]. Available: http://www.cs.utsa.edu/~mir/MIR2005_files/keynote/MIR-Keynote-2005-Sarukkai-Final.ppt
[51] H. Zhong, L. Wenyin, and S. Li, "Interactive tracker—A semi-automatic video object tracking and segmentation system," in *Proc. 2001 IEEE Int. Conf. Multimedia Expo (ICME'01)*, 2001, pp. 1167–1170.
[52] I. Burnett, R. V. Walle, K. Hill, J. Bormans, and F. Pereira, "MPEG-21: Goals and achievements," *IEEE Multimedia*, vol. 10, no. 4, pp. 60–70, Oct./Dec. 2003.

Wen-Huang Cheng (S'05) received the B.S. and M.S. degrees in computer science and information engineering from National Taiwan University, Taipei, Taiwan, R.O.C., in 2002 and 2004, respectively, where he is currently pursuing the Ph.D. degree in the Graduate Institute of Networking and Multimedia.

His research interest includes multimedia data management and analysis.

Chia-Wei Wang received the B.S. and M.S. degrees in computer science and information engineering from National Taiwan University, Taipei, Taiwan, R.O.C., in 2004 and 2006, respectively.

His research interest includes video processing and digital content analysis.

Ja-Ling Wu (SM' 98) received the B.S. degree in electronic engineering from Tamkang University, Tamshoei, Taiwan, R.O.C., in 1979, and the M.S. and Ph.D. degrees in electrical engineering from Tatung Institute of Technology, Taipei, Taiwan, in 1981 and 1986, respectively.

From 1986 to 1987, he was an Associate Professor with the Electrical Engineering Department, Tatung Institute of Technology, Taipei, Taiwan, R.O.C. In 1987, he transferred to the Department of Computer Science and Information Engineering, National Taiwan University (NTU), Taipei, Taiwan, R.O.C., where he is presently a Professor and the Director of the Communications and Multimedia Laboratory. From 1996 to 1998, he was the first Head of the Department of Information Engineering, National Chi Nan University, Puli, Taiwan, R.O.C. During his sabbatical leave (from 1998 to 1999), he was invited to be the Chief Technology Officer of the Cyberlink Corporation, Taipei. In this one-year term, he was involved with the developments of some well-known audio-video softwares, such as the PowerDVD. Since August 2004, he has been appointed the Head the newest research institute of NTU—the Graduate Institute of Networking and Multimedia. He has published more than 200 technique and conference papers. His research interests include digital signal processing, image and video compression, digital content analysis, multimedia systems, digital watermarking, and digital right management systems.

Prof. Wu was the recipient of the Outstanding Young Medal of the Republic of China in 1987 and the Outstanding Research Award three times of the National Science Council, R.O.C., in 1998, 2000, and 2004, respectively. He was the recipient of the Award for Distinguished Information People in 1993, the Special Long-Term Award for Collaboratory Research in 1994, the Best Long-Term paper Award in 1995, and the Long-Term Medal for Distinguished Researchers in 1996, all sponsored by the Acer Corporation. In 2001, his paper "Hidden Digital Watermark in Images" (coauthored with Prof. Chiou-Ting Hsu), published in IEEE TRANSACTIONS ON IMAGE PROCESSING, was selected to be one of the winners of the "Honoring Excellence in Taiwanese Research Award" offered by ISI Thomson Scientific. He started his Industrial-Academic Collaborative researches, which are sponsored by both the National Science Council of Taiwan and local industries, in 1992. During the last 12 years, he has conducted four three-year term Industrial-Academic Collaborative projects covering various kinds of multimedia related applications, such as multimedia office, multimedia classroom, multimedia home, etc. Due to his continuous contributions on promoting the cooperation between Industry and Academia, he was the recipient of the Special Award for Collaboratory Research, offered by of the Ministry of Education, Taiwan, R.O.C., in 1997.