Journal of Zhejiang University-SCIENCE C (Computers & Electronics)
ISSN 1869-1951 (Print); ISSN 1869-196X (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Salient object extraction for user-targeted video content association*

Jia LI[1,2], Han-nan YU[3], Yong-hong TIAN[†‡3], Tie-jun HUANG[3], Wen GAO[1,3]

(*[1]Key Lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China*)
(*[2]Graduate University of Chinese Academy of Sciences, Beijing 100049, China*)
(*[3]National Engineering Lab for Video Technology (NELVT), School of EE & CS, Peking University, Beijing 100871, China*)
[†]E-mail: yhtian@pku.edu.cn

Received Sept. 14, 2010; Revision accepted Oct. 8, 2010; Crosschecked Sept. 14, 2010

Abstract: The increasing amount of videos on the Internet and digital libraries highlights the necessity and importance of interactive video services such as automatically associating additional materials (e.g., advertising logos and relevant selling information) with the video content so as to enrich the viewing experience. Toward this end, this paper presents a novel approach for user-targeted video content association (VCA). In this approach, the salient objects are extracted automatically from the video stream using complementary saliency maps. According to these salient objects, the VCA system can push the related logo images to the users. Since the salient objects often correspond to important video content, the associated images can be considered as content-related. Our VCA system also allows users to associate images to the preferred video content through simple interactions by the mouse and an infrared pen. Moreover, by learning the preference of each user through collecting feedbacks on the pulled or pushed images, the VCA system can provide user-targeted services. Experimental results show that our approach can effectively and efficiently extract the salient objects. Moreover, subjective evaluations show that our system can provide content-related and user-targeted VCA services in a less intrusive way.

Key words: Salient object extraction, User-targeted video content association, Complementary saliency maps
doi:10.1631/jzus.C1001004 **Document code:** A **CLC number:** TP391.7

1 Introduction

In recent years, the amount of videos on the Internet has increased tremendously. For example, the most popular video-sharing site to date, YouTube, possessed up to 78.3 million videos by March 17, 2008. Meanwhile, this number still increases by 150 000 to over 200 000 per day (Gao *et al.*, 2010). Thus, it would be encouraging to provide incentive services such as personalized video recommendation and online video content association (VCA). In particular, VCA refers to the service that associates additional materials (e.g., texts, images, and video clips) with the video content to enrich the viewing experience. A promising application of VCA is user-targeted video advertising. This new model of advertising is less intrusive, displaying only advertising information when the user makes the choice by clicking on an object in a video. By learning the user's preference, the hot-spots that correspond to brands can be further highlighted so as to extract more interest of the users. Therefore, VOA presents a significant opportunity for marketers to extend the reach of their campaigns with compelling content (Gao *et al.*, 2010). VOA can also be easily applied in digital libraries, where the related materials can be automatically associated with the video content being viewed so as to provide supplementary information or enrich the viewing experience.

[‡] Corresponding author

[*] Project supported by the CADAL Project and the National Natural Science Foundation of China (Nos. 60973055 and 90820003)

There have been many approaches to associate additional materials (e.g., advertising logos and video clips) with video content. In these approaches, different guidelines have been proposed in different VCA services. For example, some works (Li *et al.*, 2005; Chang *et al.*, 2008; Guo *et al.*, 2009) point out that a good VCA service should not interrupt the viewing experience, while some others argue that the associated materials should be relevant to the video content (Mei *et al.*, 2007; Wang *et al.*, 2008) or user preference (Lekakos *et al.*, 2001; Thawani *et al.*, 2004) to enrich the viewing experience. To sum up, a good VCA service should be: (1) Non-intrusive. The associated materials should not interrupt, clutter, or delay the viewing experience. (2) Content-related. The associated materials should be relevant to the important video content. (3) User-targeted. The associated materials should match individual preferences of different users.

Existing studies on VCA focused mainly on the first two requirements. For example, Li *et al.* (2005) presented an approach to overlay less intrusive images into a baseball video, while Wang *et al.* (2008) proposed an approach to detect predefined semantic concepts in video. Ad images were then associated with these concepts through concept-to-ad and ad-to-concept relevancies. However, these approaches may have difficulties in meeting the requirement of user-targeted VCA. That is, they had not taken the individual preferences of different users into account, which is important for the satisfactory VCA services. In fact, the user preference is an important factor that should be considered. From the perspective of intrusiveness, people often show high tolerance to the associated materials with preferred contents. Moreover, individual preference plays an essential role in determining which video content is important and interesting. Therefore, it necessitates the development of content-related and user-targeted VCA services.

In this paper, we propose a novel approach for user-targeted video content association. In this approach, images are associated with video content using two modules, including the ‘pull’ and ‘push’ modules (Fig. 1). The ‘pull’ module refers to an interactive approach that allows users to retrieve the images related to the interesting video content through simple interactions by the mouse or by an infrared pen. In contrast, the ‘push’ module can automatically extract the salient objects from the video stream by using complementary saliency maps. Given these salient objects, our VCA system can push the content-related images to the users. Moreover, the user feedbacks on the pulled or pushed images are also collected to mine individual preferences, which can be further used for user-targeted VCA. Experimental results show that our approach can effectively and efficiently extract the salient objects. In addition, subjective evaluations show that our system can provide content-related and user-targeted VCA service in a less intrusive way.

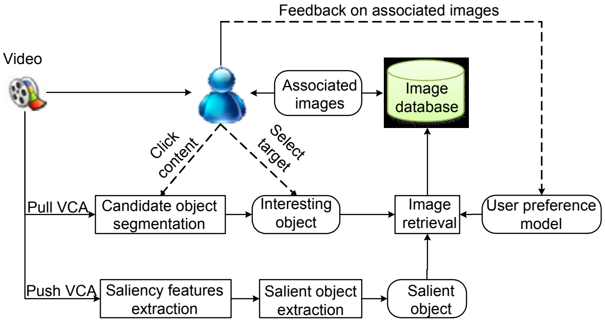

Fig. 1 System framework of our proposed approach for user-targeted video content association, including two modules ‘pull’ and ‘push’

Compared with existing approaches, our main contributions are summarized as follows:

1. We propose a novel approach to automatically extract salient objects. In this approach, two complementary saliency maps (i.e., sketch- and envelope-like maps) are used to locate the accurate borders of the salient objects. Experimental results show that our approach outperforms four state-of-the-art approaches remarkably.

2. We present an intelligent system to provide content-related and user-targeted VCA services in a less-intrusive way. Subjective evaluations show that the effect of VCA service is promising and acceptable.

2 Related work

In existing approaches, there are mainly two ways to associate other materials with video content, including in-banner and in-stream VCA. Generally speaking, in-banner VCA places associated materials on banners outside of the video content, while in-stream VCA refers to inserting or overlaying multimedia materials into video streams; i.e., the main difference between in-banner and in-stream VCA approaches is the placement of associated materials. We focus on in-stream VCA in this survey.

2.1 In-stream video content association

In general, there are two ways for in-stream VCA, i.e., linear and non-linear VCA. In linear VCA, associated materials are usually inserted before, in the middle of, or after the related video segment. The key characteristic is that users are forced to watch these materials as they take over the full view of the video. This approach is popular in existing video-sharing sites. Existing research focused mainly on inserting content-related video clips in a less intrusive way. For example, Mei *et al.* (2007) proposed an approach to insert content-related clips into video. The insertion point was selected as the time point with high content discontinuity and low content attractiveness. In Srinivasan *et al.* (2007), scene transitions were detected for inserting content-related video clips. Typically, linear VCA is prevalent and can achieve promising results. However, the original video is inevitably prolonged, which may interrupt the viewing experience.

Another means for in-stream VCA is non-linear overlay, which provides associated materials in parallel to the video content. The simplest means for non-linear VCA is to detect attractive clues in videos and associated images/videos with them. For example, Lee *et al.* (2009) extracted keywords from video scripts for further association. In contrast, Liao *et al.* (2008) first detected the popular objects in video through image matching. These objects are then used for scheduling the associated images. Similarly, Wang *et al.* (2008) proposed an approach based on attention-relevancy computing. They detected predefined semantic concepts in video and associated them with ad images through concept-to-ad and ad-to-concept relevancies. In general, these methods can give encouraging results, but their performance relies heavily on other techniques such as semantic concept detection and trademark extraction.

Beyond these approaches, some research focused on embedding logos or clips into videos of specific genres. For example, Li *et al.* (2005) proposed an approach to overlay images onto the simple background in baseball video. Similarly, Chang *et al.* (2008) selected the insertion positions in tennis video by attention calculation, and then inserted harmonically re-colored images to these positions. In contrast to these approaches, Liu *et al.* (2008) proposed a generic approach for inserting virtual contents into the less attractive regions within higher attentive shots. Guo *et al.* (2009) presented an approach to detect a set of spatio-temporal non-intrusive positions for overlaying contextually relevant images. Overall, these approaches are effective but intrusive since some video content is replaced by other irrelevant materials. Moreover, they have not taken the individual preferences of different users into account, which absolutely play an important role in achieving satisfactory VCA.

To avoid interrupting the natural viewing experience and to guarantee the integrity of video content, we propose a novel approach to associating images to interesting or salient objects in a user-targeted way. Therefore, we also give a brief review of interesting/salient object extraction.

2.2 Interesting/Salient objects extraction

Typically, interesting objects are visually salient (Elazary and Itti, 2008). As a consequence, the two terms, ‘interesting object’ and ‘salient object’, are often used interchangeably, particularly for the scenes with unique foreground objects. As a straightforward approach, Hou and Zhang (2007) extracted salient objects using simple thresholds on saliency maps, while Achanta *et al.* (2009) binarized the saliency map with adaptive thresholds for salient object extraction. Moreover, Park and Moon (2007) presented an approach to extract salient objects in images using feature maps. Candidate regions were selected using the convex hull algorithm for salient object extraction. Kwak *et al.* (2005) first localized the salient regions in contrast maps and then extracted the borders of salient objects using salient points. In Liu *et al.* (2007), the condition random field (CRF) was used for salient object extraction. Often, these approaches can well extract the interesting or salient objects in simple scenes, but may have difficulty in detecting interesting objects from complex scenes.

To process the complex scenes, Ko and Nam (2006) presented a novel approach for extracting salient objects in images. In their approach, images were segmented into regions and a support vector machine (SVM) was trained to select and merge the salient regions to construct multiple salient objects.

Hua *et al.* (2006) proposed an energy minimization approach for interesting objects extraction. Interesting objects were extracted by iteratively estimating the object/background models and optimizing the boundaries. As a learning-based approach, Allili and Ziou (2007) first learned the relevance of the features characterizing the interesting objects. These feature relevancies were then used to optimize the object contours. Furthermore, Pinneli and Chandler (2008) presented a Bayesian formulation to predict the perceived interest of objects. Based on the subjective data obtained from psychological experiments, the algorithm then estimated the interest of each object with respect to their visual, spatial, and semantic attributes. Although these approaches may work well in some cases, they may fail to generalize to all scenes since the learned mappings between objects and interest/saliency may not always hold. Moreover, these approaches may have difficulties in extracting the accurate contours of the salient objects.

To sum up, interesting/salient object extraction requires automatically determining the location and extent of the most attractive targets in a scene. Toward this end, we propose an approach for salient object extraction using complementary saliency maps. In this approach, a sketch-like saliency map is used to roughly locate the most salient targets. After that, the sketch- and envelope-like maps are both used to extract the accurate contours. Details of our approach will be discussed in the following sections.

3 System framework

Our VCA system (Fig. 1) can associate images with videos through two VCA modules, 'pull' and 'push'. The 'pull' module allows a user to interact with the video content to pull the images that are related to the preferred content. Meanwhile, the 'push' module can extract the salient object from video and push related images to the user. The user feedbacks on associated images are also collected to mine individual preferences. In retrieving related images, these preferences can be used to obtain user-targeted VCA services.

The flowchart of the 'pull' module is illustrated in Fig. 2. Users are allowed to directly interact with the video content by the mouse or by an infrared pen. Note that here the interaction points made by the infrared pen are located using a Wiimote and the approach proposed by Lee (2008). Since a complex interaction (e.g., drawing a rectangle) would confuse non-experienced users, we only request the user to click a preferred object either in the viewing process or during a pause. Since it is often difficult to accurately segment the object with a single click, we segment multiple candidate objects by selecting different parameters. These candidate objects are then displayed and the user is asked to click the one that best meets his/her intention (Fig. 2). After that, the selected target is used to retrieve the most relevant images. The associated images and other information are displayed in two additional windows attached to the video window to avoid interrupting the video viewing process (Fig. 2).

The flowchart of the 'push' module is shown in Fig. 3. In this module, salient objects are first extracted from a video stream using the complementary saliency maps. For each salient object, a feature vector is extracted to retrieve the most relevant images on the image server. Finally, the server will push the associated images and other information to the client player. Often, the salient objects correspond to the important content of the video. Moreover, the user preference, which can be simply mined from the historical user feedbacks on associated images, will

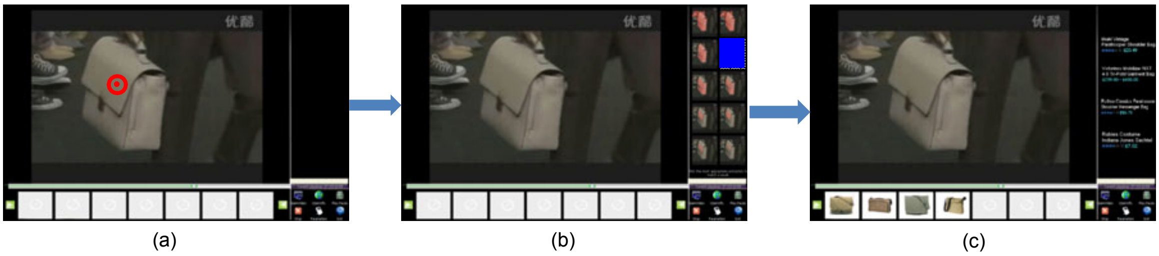

Fig. 2 Flowchart of the 'pull' module
(a) Click the preferred content; (b) Select a desired object; (c) Retrieve related images

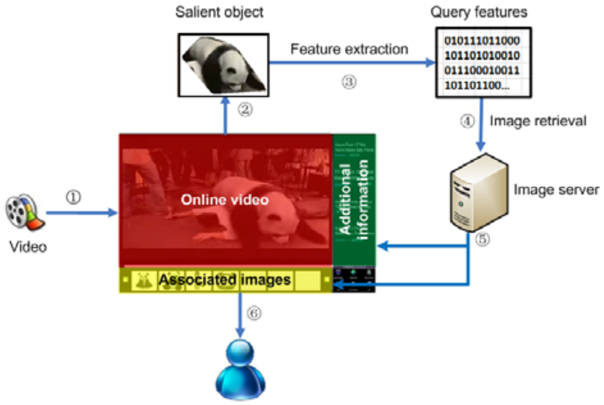

Fig. 3 Flowchart of the 'push' module

be considered in the retrieval processes. In this way, the VCA service is expected to be content-related and user-targeted.

From the discussions above, we can see that there are two key issues both in the two modules, including interesting/salient object extraction and associated image retrieval. We will discuss these two issues in the following sections.

4 Interesting/Salient object extraction

In the 'pull' module, the extraction of an interesting object involves only two simple clicks. After the first click, multiple scaled pairs of rectangles are generated according to the size of the frame. For each pair, pixels outside the large rectangle are labeled as background seeds while pixels inside the small rectangle are labeled as object seeds. After that, the other pixels are classified based on their similarities to the seeds. Here we adopt the approach proposed by Friedland *et al.* (2005).

Step 1: Seeds clustering. Establish two KD-trees by clustering the background and object seeds, respectively. Each node in a tree is a cluster of pixels.

Step 2: Pixel assignment. For the other pixels, find the tree nodes with similar colors and assign them to these nodes.

Step 3: Post-processing. Connect isolated components or smooth the boundary.

After these three steps, an interesting object can be extracted for each rectangle pair. Here we generate multiple candidate objects with different rectangle pairs, and users can select the desired one with the second click.

While in the 'push' module, salient objects should be extracted in a fully automatic way and without any human interaction. Compared with the interactive approach, automatic extraction of salient objects is more difficult. Typically, visual saliency serves as the selection criterion of the important video content. Therefore, extracting the salient object can facilitate the content-related VCA. To accurately extract the entire object, we first introduce the concepts of complementary saliency maps.

Let $\mathcal{O}$ be a salient object in an image $\mathcal{I}$. The fragment $\mathcal{E}$ can be called the 'envelope' of the object if $\mathcal{O}\subseteq\mathcal{E}$, and the fragment $\mathcal{S}$ can be called the 'skeleton' of the object if $\mathcal{S}\subseteq\mathcal{O}$. In other words, the envelope is a fragment with a high recall while the skeleton is a fragment with a high precision. Correspondingly, the saliency maps that emphasize only the envelope and skeleton are called the envelope-like and sketch-like maps, respectively. Here we denote these two maps as complementary maps since they capture two different but 'complementary' aspects for salient objects in an image.

To construct the object and background models, we need to obtain the envelope and skeleton of an object, as in the interactive object extraction. Intuitively, we expect that:

1. The envelope covers the object as much as possible, containing just a few redundant parts of background.

2. The skeleton includes the most representative parts of the objects with little background.

To extract these two fragments, we propose a novel approach for salient object extraction using complementary maps. The framework of this approach is illustrated in Fig. 4. First, the envelope and skeleton are generated from two complementary

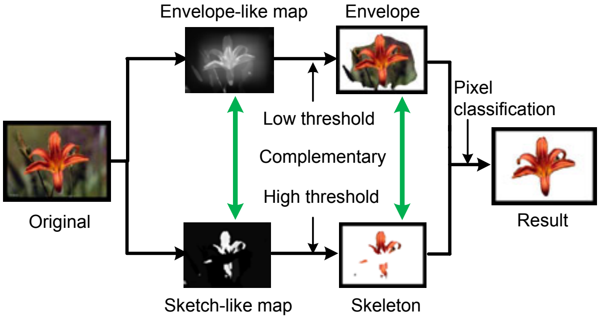

Fig. 4 The framework of our approach for extracting a salient object by using complementary saliency maps

saliency maps. Second, the seeds for object and background are generated from these two fragments. Third, the remaining pixels are classified as object or background pixels according to the seeds.

From these discussions, we can see that how to select the complementary saliency maps is the most important issue. Typically, the envelope-like map should highlight a large area covering the objects while the sketch-like map should emphasize only the most representative parts in the objects. Considering these properties, we can generate the envelope-like map by integrating two simple feature maps. The first map is the frequency-tuned map (FTM) proposed by Achanta *et al.* (2009) and the second map is the center-surround contrast map (CCM) proposed by Liu *et al.* (2007). The reason of choosing these two feature maps is that FSM together with CCM can pop-out almost every salient location from different perspectives. Therefore, the linear combination of these two maps can yield a blurred saliency map covering nearly all the salient objects. For a pixel *p*, the envelope-like map can be integrated as

$$s_{\text{env}}(p)=\lambda\cdot s_{\text{FSM}}(p)+(1-\lambda)\cdot s_{\text{CCM}}(p), \tag{1}$$

where $\lambda\in[0, 1]$ can be determined by cross-validation. In this study, we empirically set λ=0.55.

For the sketch-like map, we exploit an existing feature map known as the 'color distribution map'. The main idea is that the wider a color distributes, the more likely it appears on the background (Liu *et al.*, 2007). Therefore, the representative colors of the objects can be determined by measuring the spatial color distribution. If the spatial derivation of one class of colors is large, this class is likely to serve as the background color. Here we use the color distribution map to approximate the sketch-like map.

Given these two maps, the envelope $\mathcal{E}$ and the skeleton $\mathcal{S}$ can be generated as

$$\begin{cases}\mathcal{E}=\{p\,|\,s_{\text{env}}(p)\geq\alpha_{\text{env}}\cdot s_{\text{env}}^{\mathcal{I}}(\varDelta), p\in\mathcal{I}\},\\ \mathcal{S}=\{p\,|\,s_{\text{ske}}(p)\geq\alpha_{\text{ske}}\cdot s_{\text{ske}}^{\mathcal{E}}(\varDelta), p\in\mathcal{I}\},\end{cases} \tag{2}$$

where $s_{\text{env}}^{\mathcal{I}}(\varDelta)$ and $s_{\text{ske}}^{\mathcal{E}}(\varDelta)$ are the average saliency values over $\mathcal{I}$ and $\mathcal{E}$, respectively. The parameter α_{env} is set to a small value to obtain a high recall fragment, while α_{ske} is set to a large value to reach a high precision fragment. The envelope and skeleton are then used to derive the seeds. Intuitively speaking, pixels outside the envelope are highly likely to be part of the background and thus can be labeled as background seeds, while pixels inside the skeleton have a high probability of belonging to the salient objects and can be labeled as object seeds. In this way, the classification method used to extract the interesting objects can be applied here to automatically extract the salient objects from the video stream. For efficiency, only the salient objects in the key frames are extracted as candidates for further video content association.

5 Associated image retrieval

An important issue in VCA is to retrieve the images that are similar to the interesting/salient objects. In the retrieval process, the individual preference should be considered to obtain user-targeted results. In this study, for simplicity, such preference refers to the interest on specific image categories (e.g., shoes, animals, cars). Without loss of generality, we suppose that there exist *N* categories and the preference of the *k*th user can be represented by an *N*-dimensional vector $\boldsymbol{v}_k$. In this vector, the *n*th component v_{kn} represents the user interest on the *n*th image category.

To learn the user interest on each image category, we have to collect their feedbacks on the associated images. For the pulled/pushed images in the *n*th category, the *k*th user has labeled them as 'preferred' or 'disliked' for T_{kn}^{+} and T_{kn}^{-} times, respectively. Thus, we can empirically calculate the interest on this category as

$$v_{kn}=\frac{1}{1+\exp(T_{kn}^{-}-T_{kn}^{+})}. \tag{3}$$

We can see that once the images in a category are frequently labeled as 'preferred', v_{kn} will approximate 1. Otherwise, the 'disliked' category will correspond to zero interest. For other image categories, the interest value will be ranged in [0, 1].

Given the user preference, we select the following features to compute four kinds of similarities between interesting/salient objects and images:

1. Point-based feature. We extract SIFT (scale-invariant feature transform) points from 2000 images and group them into 256 clusters. Then we can extract the 256-bin SIFT histogram and represent it with 256 bytes.

2. Edge-based feature. We filter images or objects with 12 Gabor filters in three scales and four directions. For each output map, we calculate a 10-bin energy histogram. Then a Gabor histogram of 120 bins is formed by combining these histograms. Similarly, the Gabor histogram is represented with 120 bytes.

3. Contour-based feature. The image signature proposed by Brasnett and Bober (2007) involves tracing and analysis of lines in images with trace transform. The generated signature forms a vector of 64 bytes.

4. Color-based feature. Here we extract a 128-bin color histogram for each image in the HSV space and represent the histogram with a vector of 128 bytes.

In our application, these features are combined to form a packet and are transferred to the image server using the TCP/IP protocol. In the retrieval, the score of an image $\mathcal{I}$ in the nth category is computed as follows:

$$\mathrm{Score}(\mathcal{I}) = v_{kn} \cdot (\mathrm{Sim}_\mathrm{P} + \mathrm{Sim}_\mathrm{E} + \mathrm{Sim}_\mathrm{T} + \mathrm{Sim}_\mathrm{C}), \quad (4)$$

where Sim_P, Sim_E, Sim_T, and Sim_C are the point-, edge-, contour-, and color-based similarities, respectively. Note that here all these four similarities are normalized into [0, 1]. Finally, only the top four images are delivered to the client player to provide content-related and user-targeted VCA.

6 Experiments

Two experiments were performed to test the effectiveness and efficiency of our VCA system. In the first experiment, we mainly tested the effectiveness of our approach in salient object extraction. In the second experiment, subjective evaluations were adopted to test the efficiency of our VCA system.

To evaluate the performance of our salient object extraction approach, two public image datasets were adopted. Dataset $\mathcal{A}$ is as introduced in Achanta *et al.* (2009) and contains the accurate masks for all salient objects in 1000 images. Dataset $\mathcal{B}$ was proposed by Martin *et al.* (2001); it contains 300 complex images with object contours manually labeled by seven subjects in Movahedi and Elder (2010). For dataset $\mathcal{B}$, we manually integrated these contours to obtain the object masks for each image.

We compared our approach with four state-of-the-art approaches (Itti *et al.*, 1998; Walther and Koch, 2006; Hou and Zhang, 2007; Achanta *et al.*, 2009). Our evaluation metrics are most frequently used ones: precision (P), recall (R), and F-score (F). F-score can be calculated as follows:

$$F = \frac{2PR}{P+R}. \quad (5)$$

The performances of these approaches are illustrated in Table 1. Some representative results are shown in Figs. 5 and 6.

Table 1 Performance of salient object extraction approaches

Dataset	Approach	Precision	Recall	F-score
$\mathcal{A}$	Itti98	0.78	0.49	0.65
	Achanta09	0.82	0.75	0.80
	Hou07	0.67	0.57	0.63
	Walther06	0.45	0.38	0.41
	Ours	**0.88**	**0.89**	**0.88**
$\mathcal{B}$	Itti98	0.42	0.37	0.39
	Achanta09	0.57	0.45	0.50
	Hou07	0.43	0.48	0.45
	Walther06	0.52	0.34	0.41
	Ours	**0.73**	**0.70**	**0.71**

The highlight indicates the best result

Table 1 shows that our approach outperforms the other approaches remarkably on both datasets with the highest F-scores 0.71 and 0.88. Figs. 5 and 6 show that our approach can work well in most cases. That is, the accurate boundaries of the salient object can be well located, even when the contrast between the salient objects and the background is low (e.g., the fourth column in Fig. 6), or the object has a complex structure (e.g., the second column in Fig. 5 and the first column in Fig. 6). However, our approach still has some drawbacks. As shown in the second column of Fig. 6, some parts of the salient object are missed due to the inaccurate envelope.

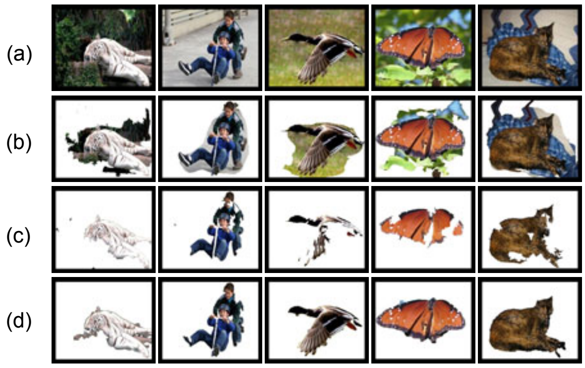

Fig. 5 Representative results of our salient object extraction approach in dataset $\mathcal{A}$

(a) Original images; (b) Envelope of the object; (c) Skeleton of the object; (d) Final result

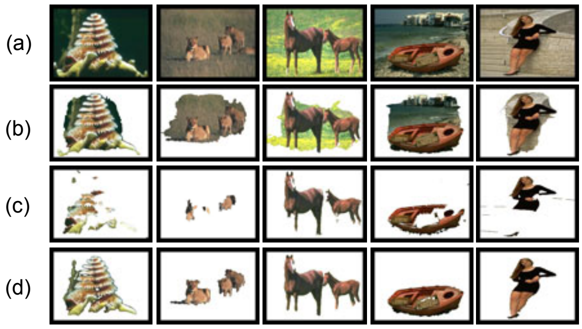

Fig. 6 Representative results of our salient object extraction approach in dataset $\mathcal{B}$

(a) Original images; (b) Envelope of the object; (c) Skeleton of the object; (d) Final result

To sum up, our approach can extract salient objects with precise boundaries. However, generating the envelope-like and the sketch-like saliency maps with predefined strategies may lead to unsatisfactory results. To solve this problem, constructing scene-adaptive strategies (e.g., learning the optimal strategies in different situations) to generate these two maps could be a feasible solution.

In the second experiment, we tested the performance of our VCA system by subjective evaluation. We selected 350 short videos (6.5 h) in five genres: documentary, ad, cartoon, news, and movie. We also grabbed 189432 images in 28 categories from Amazon to construct the image database. We run the client player on a PC with a Quad-core 2.66 GHz CPU (one thread for video play and one thread for video analysis). The associated image retrieval was performed on a server with two Quad-core 2.33 GHz CPUs (four threads were used for the retrieval process).

First, we tested the efficiency of the VCA system. We tested the average processing time used in each major step of the system:

1. Attractive object segmentation. When the user clicked the video content, it took about 3.36 s to segment the candidate objects using 10 different parameters.

2. Query feature extraction. Given an interesting or salient object, it took about 0.31 s to extract all the four features for querying the related images.

3. Associated image retrieval. It took about 0.15 s to retrieve the most relevant four images on the server.

From these data, we can see that it took little time to associate images with video content. This indicates that VCA can be performed efficiently.

Beyond the efficiency test, subjective evaluations were also adopted for the effectiveness test. In the experiment, we requested eight subjects to use our VCA system and give scores (1 for unsatisfactory, 2 for acceptable, 3 for satisfactory) on intrusiveness, content-relevance, and preference. Moreover, these subjects were also asked to report the representative results. Some representative examples of successful VCA are given in Fig. 7.

Fig. 7 shows that both the 'pull' and the 'push' modules can achieve satisfactory results. Investigation of these subjective evaluations showed that our VCA system can achieve promising performance in 'content-relevance' (2.13) and 'preference' (2.25). Moreover, the 'intrusiveness' was also acceptable (2.38). We also find that the subjective scores in 'intrusiveness' were often proportional to scores in 'content-relevance' and 'preference'. In fact, when the associated images are tightly correlated to the video content and individual preferences, subjects will often feel less intruded upon.

Survey of the subjective evaluations also showed that there are mainly five reasons for the unsatisfactory association effect:

1. Semantic gap. Often, visual similarity does not always guarantee semantic similarity. There still exists a great gap between finding a visually similar image and finding a semantically related image.

2. Limited dataset. For some interesting/salient objects, it is difficult to find related images on the limited image database.

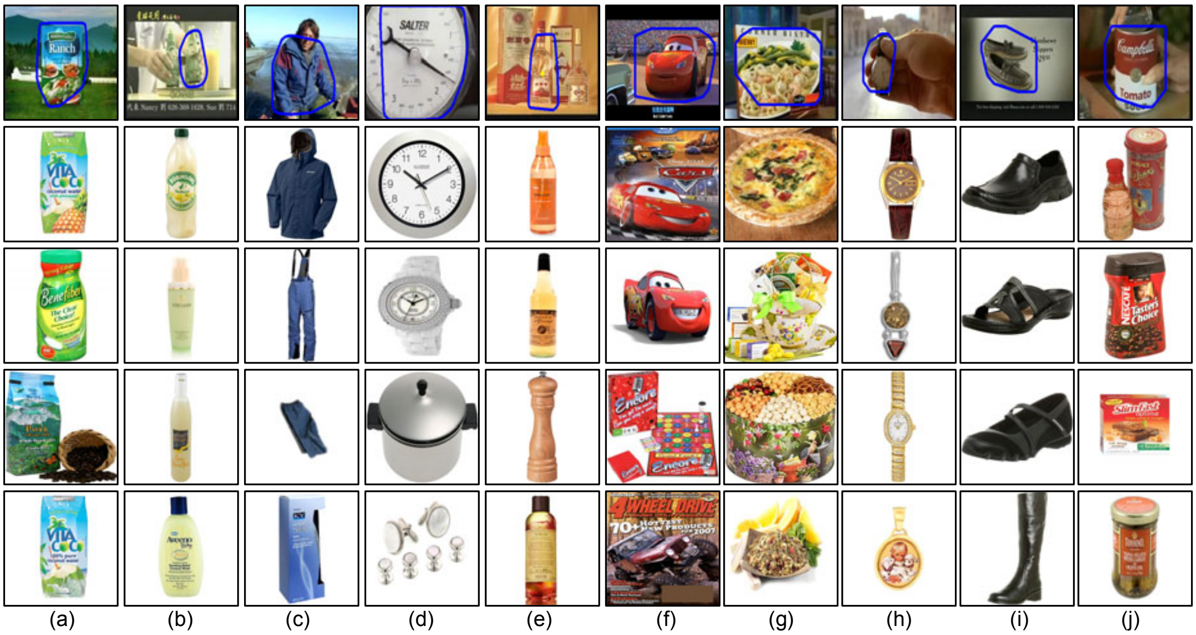

Fig. 7 Representative examples of successful video content association (VCA)
The first row shows the frames used for interesting or salient object extraction and the convex hulls indicate the interesting/salient objects. The second to the fifth rows are the top four associated images. (a)–(e) Examples for 'pull' VCA; (f)–(j) Examples for 'push' VCA

3. Non-experienced user. Usually, most users of the VCA system are non-experienced. They have no idea on which kinds of interactions are suitable for image retrieval (e.g., they may select small or fuzzy objects for further association in the pull module).

4. Moving object or poor video quality. The interesting/salient objects in video are usually moving and are encoded with poor quality. Such motion and poor video quality will make the object edges illegible, leading to inaccurate edge-based similarity.

5. Incomplete object. The interesting/salient objects in videos are usually incomplete. Often, some parts of these objects could be concealed by other objects or cut out by poor object segmentation algorithms. For these incomplete objects, existing features cannot well describe their visual properties, leading to inaccurate image retrieval.

From the discussions above, VCA is still a challenging task that requires the advances in many other related research areas such as semantic-based image retrieval and image segmentation. Among these five reasons, the most important issue is to obtain semantically similar images other than the visually similar ones. Toward this end, using image tags may be a feasible solution to improve the retrieval performance.

7 Conclusions and future work

In this paper, we propose a novel approach for user-targeted video content association (VCA). We have shown that it is necessary to provide content-related and user-targeted VCA services in a less intrusive way. Inspired by this requirement, we have developed a VCA system that can associate user-targeted images to the interesting or salient objects. From the subjective evaluations, it seems that our system can provide promising VCA services.

In the future, we will focus on learning the optimal strategies for complementary map generation by selecting and integrating various kinds of feature maps. Moreover, we will incorporate the tags of images to narrow the gap between visual and semantic similarities for better association effect.

References

Achanta, R., Hemami, S., Estrada, F., Susstrunk, S., 2009. Frequency-Tuned Salient Region Detection. IEEE Conf. on Computer Vision and Pattern Recognition, p.1597-1604. [doi:10.1109/CVPR.2009.5206596]

Allili, M.S., Ziou, D., 2007. Object of Interest Segmentation and Tracking by Using Feature Selection and Active Contours. IEEE Conf. on Computer Vision and Pattern

Recognition, p.1-8.

Brasnett, P., Bober, M., 2007. Proposed Improvements to Image Signature XM 31.0. MPEG Doc No. M14983.

Chang, C.H., Hsieh, K.Y., Chung, M.C., Wu, J.L., 2008. Visa: Virtual Spotlighted Advertising. Proc. ACM Int. Conf. on Multimedia, p.837-840.

Elazary, L., Itti, L., 2008. Interesting objects are visually salient. *J. Vis.*, **8**(3), Article No. 3. [doi:10.1167/8.3.3]

Friedland, G., Jantz, K., Rojas, R., 2005. Siox: Simple Interactive Object Extraction in Still Images. IEEE Int. Symp. on Multimedia, p.7-14.

Gao, W., Tian, Y.H., Huang, T.J., Yang, Q., 2010. Vlogging: a survey of video blogging technology on the web. *ACM Comput. Surv.*, **42**(4), Article No. 15. [doi:10.1145/1749603.1749606]

Guo, J.L., Mei, T., Liu, F.L., Hua, X.S., 2009. Adon: an Intelligent Overlay Video Advertising System. SIGIR, p.628-629.

Hou, X.D., Zhang, L.Q., 2007. Saliency Detection: a Spectral Residual Approach. IEEE Conf. on Computer Vision and Pattern Recognition, p.1-8. [doi:10.1109/CVPR.2007.383267]

Hua, G., Liu, Z.C., Zhang, Z.Y., Wu, Y., 2006. Iterative local-global energy minimization for automatic extraction of objects of interest. *IEEE Trans. Pattern Anal. Mach. Intell.*, **28**(10):1701-1706. [doi:10.1109/TPAMI.2006.209]

Itti, L., Koch, C., Niebur, E., 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, **20**(11):1254-1259.

Ko, B.C., Nam, J.Y., 2006. Automatic Object-of-Interest Segmentation from Natural Images. IEEE Int. Conf. on Pattern Recognition, p.45-48.

Kwak, S.Y., Ko, B.C., Byun, H., 2005. Automatic salient-object extraction using the contrast map and salient points. *LNCS*, **3332**:138-145.

Lee, J.C., 2008. Hacking the nintendo Wii remote. *IEEE Perv. Comput.*, **7**(3):39-45. [doi:10.1109/MPRV.2008.53]

Lee, J.T., Lee, H.D., Park, H.S., Song, Y.I., Rim, H.C., 2009. Finding Advertising Keywords on Video Scripts. SIGIR, p.686-687.

Lekakos, G., Papakiriakopoulos, D., Chorianopoulos, K., 2001. An Integrated Approach to Interactive and Personalized TV Advertising. Workshop on Personalization in Future TV.

Li, Y., Wan, K.W., Yan, X., Xu, C.S., 2005. Real Time Advertisement Insertion in Baseball Video Based on Advertisement Effect. Proc. ACM Int. Conf. on Multimedia, p.343-346.

Liao, W.S., Chen, K.T., Hsu, W.H., 2008. Adimage: Video Advertising by Image Matching and Ad Scheduling Optimization. SIGIR, p.767-768.

Liu, H.Y., Jiang, S.Q., Huang, Q.M., Xu, C.S., 2008. A Generic Virtual Content Insertion System Based on Visual Attention Analysis. Proc. ACM Int. Conf. on Multimedia, p.379-388.

Liu, T., Sun, J., Zheng, N.N., Tang, X.O., Shum, H.Y., 2007. Learning to Detect a Salient Object. IEEE Conf. on Computer Vision and Pattern Recognition, p.1-8.

Martin, D., Fowlkes, C., Tai, D., Malik, J., 2001. A Database of Human Segmented Natural Images and Its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics. IEEE ICCV, p.416-423.

Mei, T., Hua, X.S., Yang, L.J., Li, S.P., 2007. Videosense—Towards Effective Online Video Advertising. Proc. ACM Int. Conf. on Multimedia, p.1075-1084.

Movahedi, V., Elder, J.H., 2010. Design and Perceptual Validation of Performance Measures for Salient Object Segmentation. IEEE Computer Society Workshop on Perceptual Organization in Computer Vision, p.49-56.

Park, K.T., Moon, Y.S., 2007. Automatic Extraction of Salient Objects Using Feature Maps. Int. Conf. on Acoustics, Speech, and Signal Processing, p.617-620.

Pinneli, S., Chandler, D.M., 2008. A Bayesian Approach to Predicting the Perceived Interest of Objects. 15th IEEE Int. Conf. on Image Processing, p.2584-2587. [doi:10.1109/ICIP.2008.4712322]

Srinivasan, S.H., Sawant, N., Wadhwa, S., 2007. Vadeo-Video Advertising System. Proc. ACM Int. Conf. on Multimedia, p.455-456.

Thawani, A., Gopalan, S., Sridhar, V., 2004. Context Aware Personalized Ad Insertion in an Interactive TV Environment. Workshop on Personalization in Future TV.

Walther, D., Koch, C., 2006. Modeling attention to salient proto-objects. *Neur. Networks*, **19**(9):1395-1407. [doi:10.1016/j.neunet.2006.10.001]

Wang, J.Q., Fang, Y.K., Lu, H.Q., 2008. Online Video Advertising Based on User's Attention Relevancy Computing. IEEE Int. Conf. on Multimedia and Expo, p.1161-1164. [doi:10.1109/ICME.2008.4607646]