Video Presentation Board : A Semantic Visualization of Video Sequence

Tao Chen[1], Ai-Dong Lu[2], Shi-Min Hu[1]

[1]Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology,
Tsinghua University, Beijing 100084, China
[2]Department of Computer Science, University of North Carolina at Charlotte, USA

Abstract

This paper presents a video summarization method to visualizing video sequences in a static image for the purposes of efficient representation and quick overview. It can assist viewers to understand important video contents by revealing essential information of video story units and their relations. We have designed a new method called Video Presentation Board to extract, organize, and synthesize important video information in a succinct and meaningful visual format, which also preserves the elegance of original videos. Specifically, we propose a new video shot clustering method using both visual and audio data for analyzing video contents and collecting important shot information. To represent these important video information properly, we design a multi-level video summarization method to represent the contents and relations of video shots through abstracting both locations and interested objects and characters. Suitable amount of selected video information, with the assistants of special visual languages, is then organized and synthesized according to the relations between video events and temporal structure of the video sequence. We have designed and performed a preliminary user study to evaluate our approach and collected very encouraging results. We believe that this approach seamlessly integrates video analysis and visualization methods to provide visually appealing results. Our approach can be used to assist viewers grasp video contents efficiently, and are especially suitable for documentations and reports.

Categories and Subject Descriptors (according to ACM CCS): Video Summarization, Information Visualization

1. Introduction

In recent years, both the quality and quantity of digital videos have been increasing impressively with the development of visual media technology. A vast amount of movies, TV programs, and home videos are being produced each year for various entertainment or education purposes. New techniques have made it pretty easy for people to browse, share videos from all over the world, however it is still very challenging to find desired videos just by using traditional keyword searching methods. To understand a video clip, viewers have to go through the entire video, that is not an effective way to search. Therefore, we need an effective video summarization method to assist the understand of main video contents, which gives users a quick way to choose their interested videos before watching them.

To assist the understanding of video contents, researchers have developed several video summarization and representation methods that use still images to visualize an entire video [HLz05, CGL04a, YY97, MZ05, WMH*07]. Since an image is generally much smaller and easier for viewers to understand, this is an effective approach to give viewers a taste of a video without actually going through the entire sequence. Currently, many existing video summarization methods mainly focus on news programs or home videos, which usually contain simple spatiotemporal structures and straightforward storylines. They are difficult to handle professional movies and TV programs, where directors tent to use more sophisticated screen techniques. For example, some movies may have two or more storylines which are alternately depicted in an irregular sequence. Technically, many existing methods use collections of key frames or regions of interest (ROIs) to summarize video sequence. These methods do not consider some high-level information such as location and occurrence. We believe that these information should be carefully embeded in the video summarization to represent video contents effectively.

This paper presents a new video summarization method to visualizing video sequences in one static image. We have designed a *Video Presentation Board* to assist viewers to understand important video contents by revealing essential information of video story units and their relations. Our approach can produce a concise and visually pleasing representation of video sequences, which highlights important video contents and preserves the balance coverage of original sequences. Accompanying the original text description of videos, these results assist viewers to understand video topics and select their desired ones without watching all of them. We have designed and performed a preliminary user study to evaluate our approach and collected very encouraging results.

Our approach seamlessly integrates video analysis and visualization methods to provide visually appealing results. Specifically, we propose a new video shot clustering method using both visual and audio data for analyzing video contents. We also design a multi-level video summarization method to represent the contents and relations of video shots through abstracting locations, interested objects and characters. Suitable amount of selected video information, with the assistants of special visual languages, is then organized and synthesized according to the relations between video events and temporal structure of the video sequence. Our main contributions are designing a new segmentation algorithm of video events that can better describe video contents and event relations than previous shot division approaches; developing several automatic visual analysis and representation tools to highlight important video contents and semantic storylines; and presenting a new multi-level image and information synthesis approach for producing visually pleasing results.

The remainder of this paper is organized as follows. We first summarize video summarization and related video analysis and representation approaches in Section 2. Section 3 presents our automatic approach for analyzing video structures and relations and collecting video information. Section 4 describes our multi-level video representation method that uses results from Section 3 to generate seamlessly synthesized video summarization. We will provide experimental results and user study to evaluate our approach in Section 5. Finally, Section 6 concludes the paper and lists important future works.

2. Related Work

Video summarization has been an important topic in the fields of Computer Vision, Multimedia, and Graphics. In this paper we concentrate on approaches which producing static visual representations. The video booklet system [HLz05] proposed by Hua *et al.* selected a set of thumbnails from original video and printed them out on a predefined set of templates. Although this approach achieved a variety of forms, the layout of predefined booklet templates were usually not compact. Stained-glass visualization [CGL04a] was another kind of highly condensed video summary technique, in which selected key-frames with an interesting area were packed and visualized using irregular shapes like a stained-glass. Different from this approach, this paper synthesizes images and information collected from video sequences to produce smooth transitions between images and visual forms. Yeung *et al.* presented a pictorial summary of video content [YY97] by arranging video posters, which summarized the dramatic incident in each story unit, in a timeline to tell an underlying story. Ma and Zhang presented a video snapshot approach [MZ05] that not only analyzed the video structure for representative images, but also used visualization techniques to provide an efficient pictorial summary of video. However, in the above two approaches, the key frame based representative image was not compact enough and representation results were insufficient to recover important relations between story units. Among all forms of video representations, Video Collage [WMH*07] was the first to give a seamlessly integrated result. Different from their approach, we try to reveal the information of locations and relations between interested objects and preserve important storylines.

There are also related work focusing on video scene structure analysis, visual attention detection, and visualization. Rui *et al.* [RHM98] and Yeung *et al.* [YYL96] both presented methods to group video shots and used finite state machine to incorporate audio cues for scene change detection. Since these approaches are either bottom-up or top-down, they are difficult to achieve the global optimization result. Ngo *et al.* [NMZ05] solved this problem by adopting normalized cut on a graph model of video shots. Our work improves their method by counting on audio similarity between shots. Zhai and Shah [ZS06] provided a method for visual attention detection using both spatial and temporal cues. Daniel and Chen [DC03] visualized video sequences with volume visualization techniques. Goldman *et al.* [GCSS06] presented a schematic storyboard for visualizing a short video sequence and provided a variety of visual languages to describe motions in the video shot. Although this method was not suitable for exploring relations of scenes in a long video sequence, their definition of visual languages inspires our work.

3. Video Contents Analysis and Extraction

Our video summarization approach is composed of two main components: video contents analysis and extraction and representative image visualization. We use the first component to analyze video sequences and collect video information (this Section). Then, we select and organize important video information to generate meaningful video representations, which will be described in Section 4.

In the first component, we present a new video shot clustering algorithm through integrating several video features to segment a video into multiple meaningful events. We also automatically calculate event relations and select suitable background images and foreground ROIs. All these procedures help us analyze video contents and collect important information for video summarization.

3.1. Video Shot Clustering and Relation Calculation

To generate a successful video summarization, it is crucial to segment a video into a suitable amount of video shot sets. A shot is a continuous strip of motion picture film that runs for an uninterrupted period of time. Since shots are generally filmed with a single camera, a long video sequence may contain a large number of short video shots. These video shots can assist us to understand video contents; however, they do not reflect the semantic segmentation of original videos well, which may affect the efficiency of video summarization. Therefore, we propose a video shot clustering approach to better divide a video into multiple related meaningful video segments.

We cluster video shots through integrating both visual and audio features of a video sequence. Previously, Rui *et al.* [RHM98] and Yeung *et al.* [YYL96] presented methods to group video shots by using thresholds to decide whether a shot should belong to an existing group. Since a single threshold is usually not robust enough for a whole sequence, these approaches may lead to overfull segmentation. We find that in many movie sequences, we can see that several characters talk alternatively under similar scenes or images may change greatly while a character is giving a speech. We believe that combining both visual and audio features of a video sequence can improve the results of shot clustering, leading to more meaningful segmentations for video summarization. Also, we adopt a graph modeling approach to avoid selection of single threshold values. Figure 1 illustrates our video shot clustering algorithm, where we integrates several important video features to cluster video shots and calculate their relations.

Specifically, our shot clustering algorithm integrates the following visual and audio features: shot color similarity, shot audio similarity, and temporal attraction between shots. Our shots are segmented using the approach proposed in [Lie98], which can handle complex scene transitions, such as hard cut, fade and dissolve. We will describe the definition of these three video features and our clustering procedure to calculate the similarity/relation between shots.

- **Shot color similarity**
  For a video shot sequence $S = \{Shot_1, Shot_2, \ldots, Shot_n\}$, The color similarity between $Shot_x$ and $Shot_y$ is defined as:

$$SimC_{x,y} = 1 - Diff_{x,y}$$

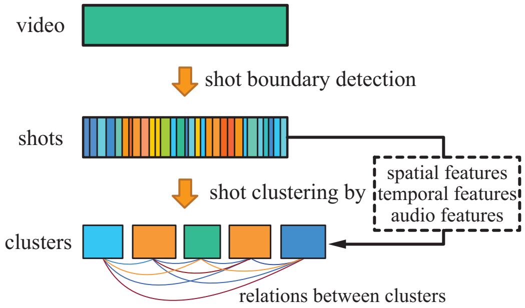


Figure 1: *Our video shot clustering algorithm combines both visual and audio features to generate suitable video segmentations and calculate shot relations for meaningful video summarization.*

where

$$Diff_{x,y} = \begin{cases} Dist(Hist(e_x), Hist(b_y)) & \text{if } x < y \\ Dist(Hist(b_x), Hist(e_y)) & \text{otherwise} \end{cases}$$

Dist is a normalized Chi-square Distance between two histograms. $Hist(b_i)$ and $Hist(e_i)$ are the color histogram of first and last five frames of $Shot_i$ respectively.

- **Shot audio similarity**
  The audio MFCC feature similarity between $Shot_x$ and $Shot_y$ is defined as:

$$SimA_{x,y} = 1 - Dist(MFCC(x), MFCC(y))$$

where $MFCC(i)$ is the average MFCC feature vector [RS78] of audio from $Shot_i$.

- **Temporal attraction between shots**
  Similar to the approach in [RHM98], we define a temporal attraction between shots:

$$Attr_{x,y} = \max(0, 1 - \frac{FrmDist(x,y)}{10 * avgShotLength})$$

where

$$FrmDist(x,y) = \begin{cases} bFrmIdx_y - eFrmIdx_x & \text{if } x < y \\ bFrmIdx_x - eFrmIdx_y & \text{otherwise} \end{cases}$$

Here, *avgShotLength* is the average shot length of the whole video stream; $bFrmIdx_i$ is the index of the first frame of f$Shot_i$; $eFrmIdx_i$ is the index of the last frame of f$Shot_i$.

Thus, we define the overall similarity between two shots as:

$$ShtSim_{x,y} = Attr_{x,y} \times (W_C * SimC_{x,y} + W_A * SimA_{x,y})$$

where W_C and W_A are the weights for color and audio measures. Since we have the observation that larger similarity is more reliable, we define the weights as follows:

$$W_C = \frac{\omega_c}{\omega_c + \omega_a}, \; W_A = \frac{\omega_a}{\omega_c + \omega_a},$$

where

$$\omega_c(x,y) = \begin{cases} \beta_c e^{\lambda_c(x,y)} & \text{if } SimC_{x,y} > \sigma_c \min(SimC) \\ & \quad +(1-\sigma_c)\max(SimC) \\ \beta_c e^{-1} & \text{otherwise} \end{cases},$$

$$\omega_a(x,y) = \begin{cases} \beta_a e^{\lambda_a(x,y)} & \text{if } SimA_{x,y} > \sigma_a \min(SimA) \\ & \quad +(1-\sigma_a)\max(SimA) \\ \beta_a e^{-1} & \text{otherwise} \end{cases},$$

$$\lambda_c(x,y) = -\frac{(\max(SimC) - SimC_{x,y})^2}{\sigma_c^2(\max(SimC) - \min(SimC))^2},$$

$$\lambda_a(x,y) = -\frac{(\max(SimA) - SimA_{x,y})^2}{\sigma_a^2(\max(SimA) - \min(SimA))^2}.$$

We set $\beta_c = \beta_a = 1,\ \sigma_c = \sigma_a = 0.2$.

After calculating pairwise similarities, we build weighted undirected graph and adopt normalized cut technique in [NMZ05] to cluster the shots. Our incorporation of an audio feature significantly improves the clustering result. Ideally, each cluster represents a video event (or sub-story), and we denote clusters as $E = \{Event_1, Event_2, \ldots, Event_m\}$. Those video events are usually not independent to each other, especially in movies. Some video events may strongly related while others may not. For example, some movies often contain more than one storyline and different events occurred at different locations synchronously. To demonstrate this, filmmakers may cut two stories to multiple sub-stories and depict them alternately. To capture this important information in video summarization, we calculate the relations between two video events and define it as follows:

$$\begin{aligned} EvtRlt_{i,j} &= W_C * \max_{x\in E_i, y\in E_j} SimC_{x,y} \\ &+ W_A * \max_{x\in E_i, y\in E_j} SimA_{x,y} \end{aligned}$$

For robustness, we usually use the average value of top 10 similarities as the relation result. The video events with larger similarity values are viewed as being more related. We will integrate the relation information during the process of video summarization in Section 4.

3.2. Background Image Selection

This step aims to find a frame which can best describe the location (or background) of an video event. Typically, it should be an image with the largest scene during this event. Although detecting the scale of an image is still an unsolved problem in the areas of computer vision and machine learning, we can simplify this problem under assumptions summarized from our observations:

1. Shots containing scenes of larger scales usually have smoother temporal and spatial optical flow fields. If the optical flow fields indicate a zooming-in or zooming-out transition, the first or the last frame should be selected.
2. We can remove the frames with good respondence to face detection to avoid the violation of characters' feature shots.
3. Very often, a shot containing this kind of frames appears at the beginning of the video event which is called establishing shot.

Therefore, we can detect the image with the largest scale semi-automatically using additional information collected from a video sequence. We run a dense optical flow calculation [BA96] and face detection algorithms [LM02] through the video event and discard shots with stable face detection respondence. The remaining shots are sorted in the ascending order of *adjusted optical flow discontinuity* defined as follow.

Adjusted optical flow discontinuity for $Shot_i$ from a video event (i is shot index in the video event):

$$Discont(i) = \frac{Ws(i)}{numFrm_i} * \sum_{j=1}^{numFrm_i - 1} (DscS_j + DscT_j)$$

where

$$Ws(i) = \begin{cases} \frac{1}{\lambda+1-i} & \text{if } i \le \lambda \\ 1 & \text{if } i > \lambda \end{cases}$$

Here, $numFrm_i$ is the frame number of $Shot_i$, $DscS_j$ is spatial of frame j, and $DscT_j$ is temporal optical flow discontinuity between frame j and j+1 [BA96]. We set $\lambda = 3$ for all the results in this paper.

After sorting, a proper frame from the first event will be selected (due to zooming order) as the background of video summarization. The background images from first ten events will be stored as replacements to be used with further user interaction. We will describe our user interface that allows user to reselect their desired background image in Section4.3.

3.3. Foreground ROIs Selection

There are three kinds of objects which are very likely to become foreground ROIs and draw visual attentions:

1. Character faces. Characters often play a main role in many movies, with more than half of the frames containing human characters.
2. Objects with different motion from the background often draw temporal attentions.
3. Objects with high contrast to the background often draw spatial attentions.

Therefore, we propose a method that integrates the detection algorithms of human faces and spatiotemporal attentions. We reuse the per frame face detection result from Section 3.2 and only preserve those stably detected in temporal space(detected in continuous 5 frames). Then, we define a face-aware spatiotemporal saliency map for each frame as:

$$Sal(I) = \kappa_T \times SalT(I) + \kappa_S \times SalS(I) + \kappa_F \times SalF(I),$$

Here, the spatiotemporal terms are exactly the same as in [ZS06]. We add the face detection result to the saliency map with the last factor. Specifically, for pixels falling in the detected face regions, we set its saliency value $SalF(I)$ as 1, or zero otherwise. κ_F is the weight for $SalF(I)$.

Next, we carefully select ROIs for each video event. To prevent duplicate object selection, we restrict that only one frame can be used for ROI selection in each video event. This frame is the one with the largest saliency value in the event. Then for a new selected ROI, we check the difference between its local histogram and those of existed ROIs. If it is smaller than a threshold, only the one with the larger saliency value will be preserved. Those ROIs will be sorted by their saliency value per pixel.

4. Representative Image Visualization

After we automatically collect video information, we select and arrange them to generate video summarization which represent both important video contents and shot relations. We propose a multi-level video summarization approach though arranging and synthesizing both image and video information onto one still image. Our approach also integrates several image and information synthesis tools to produce both semantic and visual appealing results.

In previous work, visualization is usually done by finding a keyframe from the sequence [YY97, HLz05, MZ05] or further more by finding a ROI (region of interest) from keyframe [CGL04b, WMH*07]. But consider a video event with basic components of time, location, characters and occurrence, one single keyframe or ROI is insufficient for representing all those information. Besides, "stack" all the information together will make user losing focus, previous work like "VideoCollage" is plagued with this problem: Although selected ROIs represent most important information in the video, after putting together, due to lack of relations and emphasis, user can not tell the story line or the importance of different characters.

For that reason, we will not follow the collage-like visualization scheme. Instead, our multi-level video summarization approach comes from an observation: although the most basic events are only involved of several key factors, such as characters and locations; complex events are usually combinations of multiple relatively simpler events. For videos containing more than one basic event, the procedures to describe these videos will not be significantly different according to their complexity levels because of the following reasons.

- First, since one still image only provides a limited space to represent information, we need to control the total information amount, so that they can be presented at a suitable scale for viewers to observe. Also, the time that viewers spend on understanding an image is generally exponential to the information amount contained in this image. Therefore, we prefer to visualize a proper amount of information from one video summarization for the best effect.
- Second, it is more important to describe the events that are closer to the whole video sequence than those that are only relevant to portions. Thus, the procedures to represent the several top levels, including the contents from relatively simpler events and their relations, are the same despite the different event levels a video may contain.

Therefore, this multi-level approach allows us to represent general videos with different complexity levels.

We have developed several tools to synthesize image and information collected from video sequences. The following will describe our basic events presenter, events layout, and assisted visual language and user interface.

4.1. Basic Event Presenter

The basic video sequence we aim to represent is a single video event. Our approach is inspired by those popular commercial movie posters. They usually have a large stylized background and featured character portraits, along with multiple (relative smaller) most representative film shots. This layered representation not only induces the user to focus on the most important information, but also provides state-of-the-art visual appearance.

Our basic event presenter contains at least four layers. The bottom layer is the background image frame extracted in section 3.2. The layer next to bottom contains ROIs with no face detected, while other layers are composed of other ROIs extracted in section 3.3. The higher layer contains ROIs with higher order, i.e. higher saliency values. We use a greedy algorithm to calculate the layout, as illustrated in Figure 2.

We start from the bottom layer, i.e. the background image. We initialize the global saliency map with the saliency map of background image. Then we add each layer overlapping on the presenter from the lowest layer to the top layer. For each layer, we add ROIs from the one with the highest saliency value to the lowest. For each ROI, we first resize it by its saliency, then search for a position that minimizes the global saliency value of the presenter covered by the ROI. After adding a new ROI, global saliency map should be updated by replacing covered region's saliency with last added ROI's.

In this progress, We use a threshold $\varphi = 50\%$, which we called level of detail controller, to control the amount of presented ROIs. That means, when adding a new ROI, every objects in the presenter (including background image) must preserve at least φ portion of its original saliency value in the global saliency map (detected face region has the exception that it should never be covered, to prevent half face). When this is violated, the ROI with least saliency will be removed from the presenter, and recalculate the layout. With this LOD control, when the video sequence we represented

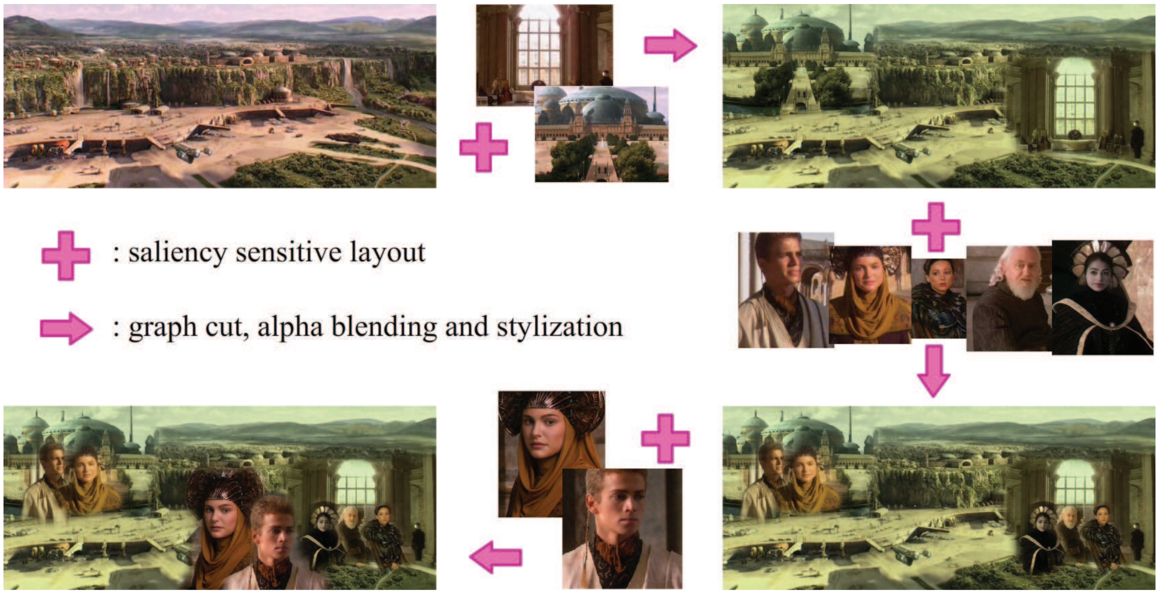

: saliency sensitive layout
: graph cut, alpha blending and stylization

Figure 2: *Synthesis progress of the basic event presenter*

becomes more complicated, we can ensure each presented part still provides sufficient information.

After adding each layer, we use graph cut to solve labeling problem followed by α-poisson image blending [RBHB06]. To emphasize the importance of foreground objects, we stylize each layer as shown in Figure 2. We compute the average hue value of background image, use this value to tint each layer, and lower layers will be tinted by larger proportions. Figure 3 shows six basic event presenters synthesized by our approach. They successfully represented most important information of the video event such as locations, characters, and also preserves the original video style.

4.2. Events Layout

When we represent a longer video sequence with more basic video events, the only additional step is an events layout. With the relations and basic event presenters we calculated in above section, one can make a layout by combining those information use proper visualization approaches. Here we present a layout that can utilize all our collected information.

Given n basic event presenters $\{R_1, R_2, \ldots, R_n\}$ for n video events and their relations, and a canvas of size $l \times m$, we first resize all the basic event presenters:

$$size(R_i) = \max(0.25, \frac{L(R_i)}{L_{max}}) \times \frac{l * m}{1.5n},$$

where $L(R_i)$ is the length (in frame) of the $Event_i$, L_{max} is the maximal length among the video events. Let (x_i, y_i) denotes the shift vector of the basic event presenters R_i on canvas, then we minimize the following energy function:

$$E = E_{ovl} + w_{sal} * E_{sal} + w_{rela} * E_{rela} + w_{time} * E_{time},$$

overlay term $E_{ovl} = -A_{ovl}$ is the negative of the overlay area of all the basic event presenters on the canvas; Saliency cost E_{sal} is negative saliency value of composed saliency map; Relation term is defined as:

$$E_{rela} = \sum_{i=0}^{n} \sum_{j=i+1}^{i+3} (Dist(i,j) - \frac{\sqrt{lm}(EvtRlt_{max} - EvtRlt_{i,j})}{EvtRlt_{max} - EvtRlt_{min}})^2,$$

where $EvtRlt_{i,j}, EvtRlt_{max}$ and $EvtRlt_{min}$ are relation between $Event_i$ and $Event_j$, $Dist(i,j)$ is the distance between the centers of two basic event presenters. maximal relation and minimal relation respectively. This term attempts to position basic event presenters with larger relation closer to each other in x coordinate; Temporal order term is defined as:

$$E_{time} = \sum_{i=0}^{n-1} \delta_i$$

where

$$\delta_i = \begin{cases} 0 & \text{if } y_i + \varepsilon < y_{i+1} < y_i + h_i - \varepsilon \\ 1 & \text{otherwise} \end{cases}$$

h_i is the height of resized R_i, and $\varepsilon = 30$. This term attempts to position basic event presenters with respect to temporal order in y coordinate while preserve some overlapping.

Minimize the energy function above by heuristic approach will maximize the overlay area of all basic event presenters which visualize temporal order in y coordinate and visualize

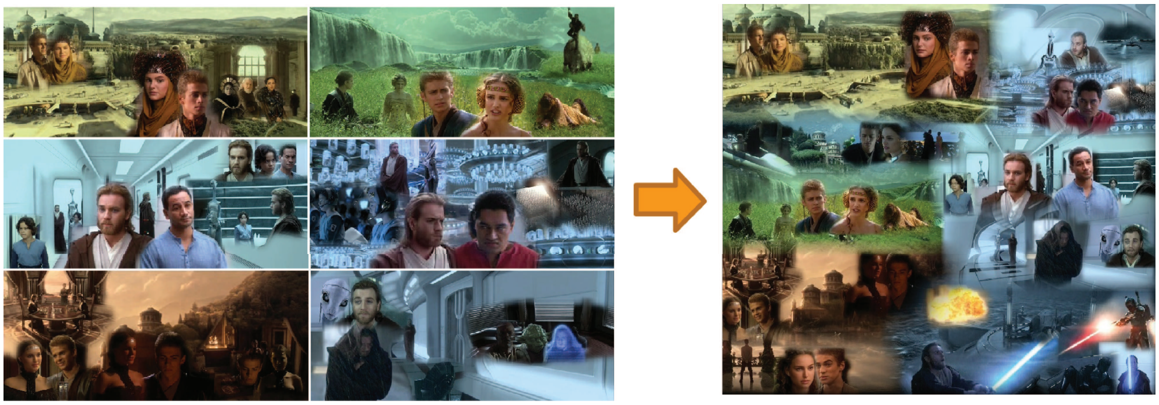


Figure 3: *Events Layout. The right part is a synthesized representative image for a video sequence of 30 min, which is clustered to 10 video events. For limited spaces, the left part of the figure only show six basic event presenters.*

relations in x coordinate. As this method can not ensure that all pixels are covered, we can choose those obsoleted ROIs from adjacent basic event presenters to fill the hole. Overlapped region will be labeled by graph cut and α-poisson image blending. Since overlapping may cause the violation of LOD control, it is necessary to recalculate the layout for basic event presenters. Figure 3 shows the events layout and the LOD control effect. It shows when the represented video sequence becomes complicated, our results will not be in a clutter as other methods while still provide essential video information.

4.3. Assisted Visual Language and User Interface

We also embed abstract visual forms in the video summarization, so that we can represent more complex relations among multiple events. To generate seamless synthesized results, we synthesize visual forms with their relevant events presenters and the final background. If space is available, we also allow information to be embed inside the visual forms to represent more detailed relations.

We have mainly designed the following two types of visual forms for describing storylines and event types. First, we build a storyline graph and use a sequence of arrow shapes to represent key storylines. The storyline graph uses video events as nodes, for two adjacent video events in the representative image. If the relation between them is larger than a threshold, we add an edge in between. After traversing all the nodes, circles will be cut off at the edge between two nodes with largest temporal distance. Then each branch in this acyclic graph represent a story line. We add an arrow around the intersection location between any two connected basic event presenters with the restrict that no ROI will be covered. To produce a smooth storyline, we calculate the arrow directions according to a B-spline generated by connecting all the arrow centers and saliency-weighted centers of involved basic event presenters on this storyline. The arrow bottom is reduced to disappear among the previous event to emphasize the direction of storylines. Different storylines will be distinguished by the color of arrows.

Second, we select several visual shapes to represent types of video shots. For example, we can use a pink heard shape for loves, a red knife shape for actions, and a blue question mark for suspenses. The user interface is shown in Figure 4. Users can add their selected visual shapes on a basic event presenter to represent special information, the system will automatically find a position for the shape to minimize saliency cost. Although our objects selection and layout can be done fully automatically, users are not always satisfied with the result. Besides, in special cases, background, face, saliency detector may fail. Thus, we present very flexible functions for user adjustment. User can choose any objects presented in the summarization, including backgrounds, ROIs, arrows and shape, then the selection will be able to move, rotate, scale, change colors and remove. When backgrounds or ROIs are selected, precalculated candidates will be pop-up in a dialog for replacement. The system even supports importing other resources for synthesis, for example, in Figure 5(a), user adds title of the movie into summarization. After user adjustment, a new synthesized result will be calculated.

5. Experiments and Evaluations

5.1. Experimental Results

Figure 5 shows our final representation results. Figure 5(a) represented a sequence of 30 minutes long from a commercial movie, Figure 5(b) represented a sequence of 20 minutes long from a TV program. The pre-computing times for video structure analysis and information extraction are 2 hours and 1 hour 20 min respectively. Synthesizing times are much less. After each user interaction, resynthesis takes less than 1 min. Figure 5(a) takes a veteran user 10 min to adjust (2

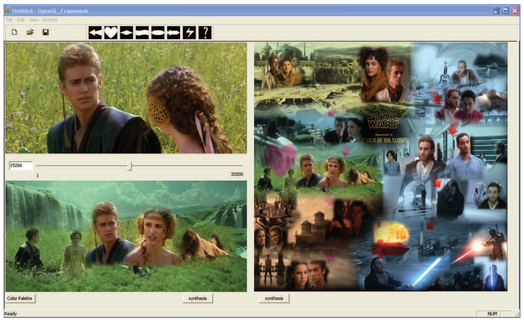


Figure 4: *User Interface of Video Presentation Board*

backgrounds and 5 ROIs are reselected and adjusted), while Figure 5(b) takes 7 min (1 backgrounds and 5 ROIs are reselected and adjusted).

5.2. User Study

We invited twenty individuals for our user study. They include fourteen graduate students and six undergraduate students (majoring in computer science, architecture and art). We created four kinds of summaries for video sequences in Figure 5: Booklet, Pictorial, Video Collage and Video Presentation Board. After watching video sequences, users were asked to answer the following questions with 1 (definitely no) to 5 (definitely yes), as used in [RBHB06, WMH*07]. Here we list our questions and provide the average scores for each method after their names.

- Are you satisfied with this summary in general?
  Video Presentation Board(4.4), Video Collage(3.1), Pictorial(2.1), Booklet(2.5)
- Do you believe that this result can represent the whole video sequence?
  Video Presentation Board(4.2), Video Collage(3.3), Pictorial(2.8), Booklet(2.2)
- Do you believe this presentation is compact?
  Video Presentation Board(4.0), Video Collage(3.9), Pictorial(2.8), Booklet(2.6)
- Would you like use this presentation as a poster of the video?
  Video Presentation Board(4.7), Video Collage(3.8), Pictorial(1.4), Booklet(3.1)
- Do you believe that this presentation produces the correct storylines?
  Video Presentation Board(4.9), Video Collage(2.2), Pictorial(2.8), Booklet(1.5)

This results shows that Video Presentation Board achieves the highest scores in all the categories; therefore, it is the most representative and visual appealing summary among these four approaches. This also shows that Video Presentation Board is the only one that can extract and visualize video storylines. Although our multi-level representation may not be fully compact, it does help users quickly grasp the significant contents of a video while achieving artistic styles.

6. Conclusions and Discussions

This paper presents a video summarization method to generate meaningful and visually appealing results through designing and integrating the techniques of automatic video analysis and interactive image and information synthesis. We have proposed new methods to analyze videos by segmenting video events and selecting representative image segments. We present a multi-level video representation method to abstract and synthesize important video information into succinct still images. Our approach provides more meaningful information than previous approaches by preserving main storylines and highlighting important video contents. We have designed and performed a preliminary user study to evaluate our approach and collected very encouraging results. We think that video summarization results are an important addition to handle the enormous volume of digital videos, and it can save users a significant amount of time to grasp video contents quickly. With the efficiency provided by video summarization techniques, we believe that they can also be used to assist other video operations, such as browsing and documentation, especially for entertainment and educational purposes.

In the future, we plan to continue working on the following problems to extend the proposed approach. First, we will design a flexible interaction method that allows users to control more steps conveniently during the procedure of video summarization. Since selecting the preferred summarization styles is a subject issue, we believe that a small amount of interaction is worth to generate user-desired results. Second, we plan to explore the suitable range of information amount to be contained in one video summarization. This result will help us to improve our method and should be useful for other approaches with similar objectives. Finally, we plan to promote our approach to two real life applications that are currently using key frames. One is to producing more meaningful summarizations for series TV programs, which can take advantage of our contents/relations representation technique. The other is to help local newspapers to select representative images for reporting events with recorded videos. We believe that there are more real life applications that can benefit from effective video summarization approaches.

References

[BA96] Black M. J., Anandan P.: The robust estimation of multiple motions: parametric and piecewise-smooth flow fields. *Comput. Vis. Image Underst. 63*, 1 (1996), 75–104.

[CGL04a] Chiu P., Girgensohn A., Liu Q.: Stained-glass visualization for highly condensed video summaries. In *Proc. IEEE Intl. Conf. on Multimedia and Expo* (2004), pp. 2059–2062.

[CGL04b] CHIU P., GIRGENSOHN A., LIU Q.: Stained-glass visualization for highly condensed video summaries. *IEEE International Conference on Multimedia and Expo 3* (June 2004), 2059–2062 Vol.3.

[DC03] DANIEL G., CHEN M.: Video visualization. In *VIS '03: Proceedings of the 14th IEEE Visualization 2003 (VIS'03)* (Washington, DC, USA, 2003), IEEE Computer Society, p. 54.

[GCSS06] GOLDMAN D. B., CURLESS B., SEITZ S. M., SALESIN D.: Schematic storyboarding for video visualization and editing. *ACM Transactions on Graphics (Proc. SIGGRAPH) 25*, 3 (2006).

[HLz05] HUA X.-S., LI S., ZHANG H.-J.: Video booklet. *icme 0* (2005), 4 pp.

[Lie98] LIENHART R. W.: Comparison of automatic shot boundary detection algorithms. In *Proc. SPIE Vol. 3656, p. 290-301, Storage and Retrieval for Image and Video Databases VII, Minerva M. Yeung; Boon-Lock Yeo; Charles A. Bouman; Eds.* (Dec. 1998), Yeung M. M., Yeo B.-L., Bouman C. A., (Eds.), vol. 3656 of *Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference*, pp. 290–301.

[LM02] LIENHART R., MAYDT J.: An extended set of haar-like features for rapid object detection. In *IEEE ICIP 2002* (September 2002), vol. 1, pp. 900–903.

[MZ05] MA Y.-F., ZHANG H.-J.: Video snapshot: A bird view of video sequence. *Proceedings of the 11th International Multimedia Modelling Conference,* (Jan. 2005), 94–101.

[NMZ05] NGO C.-W., MA Y.-F., ZHANG H.-J.: Video summarization and scene detection by graph modeling. *IEEE Transactions on Circuits and Systems for Video Technology 15*, 2 (Feb. 2005), 296–305.

[RBHB06] ROTHER C., BORDEAUX L., HAMADI Y., BLAKE A.: Autocollage. In *SIGGRAPH '06: ACM SIGGRAPH 2006 Papers* (New York, NY, USA, 2006), ACM, pp. 847–852.

[RHM98] RUI Y., HUANG T. S., MEHROTRA S.: Exploring video structure beyond the shots. In *In Proc. of IEEE conf. Multimedia Computing and Systems* (1998), pp. 237–240.

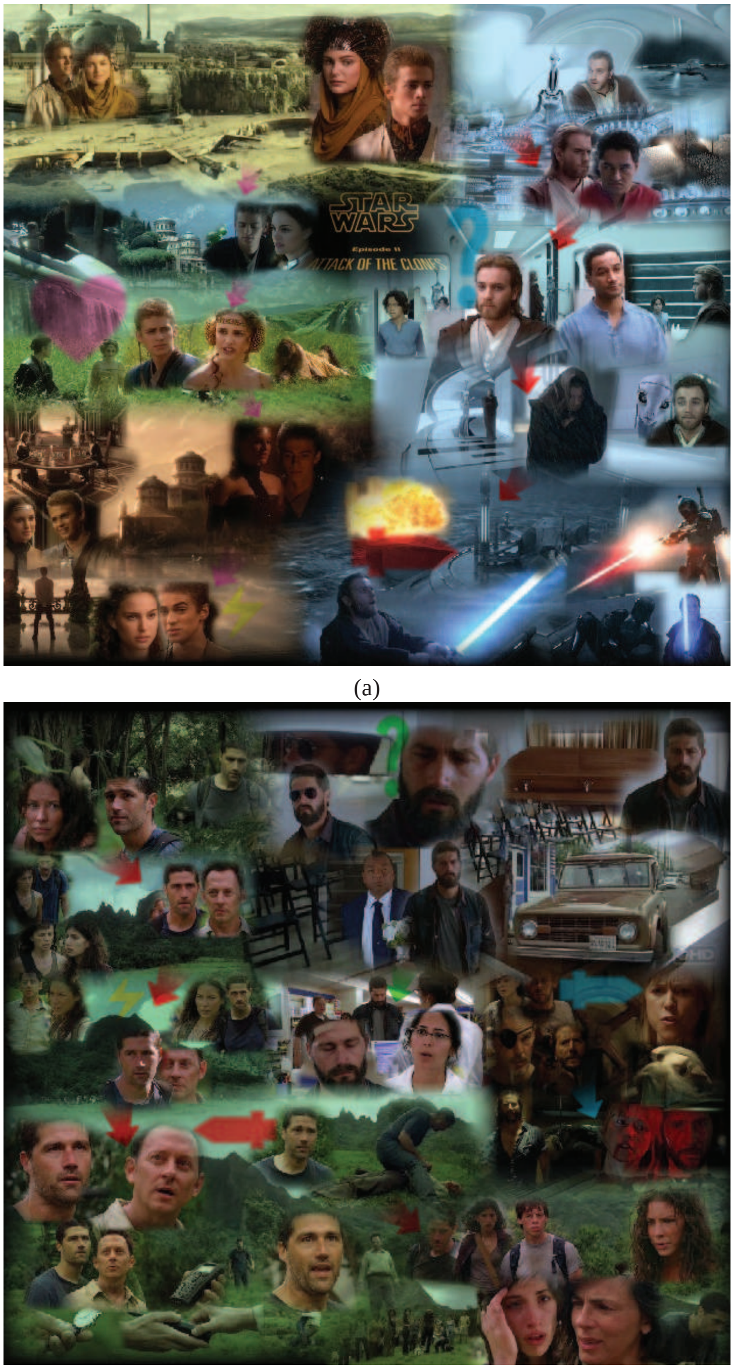
[RS78] RABINER L., SCHAFER R.: *Digital Processing of Speech Signals*. Englewood Cliffs: Prentice Hall, 1978.

[WMH*07] WANG T., MEI T., HUA X.-S., LIU X.-L., ZHOU H.-Q.: Video collage: A novel presentation of video sequence. *IEEE International Conference on Multimedia and Expo* (July 2007), 1479–1482.

[YY97] YEUNG M., YEO B.-L.: Video visualization for compact presentation and fast browsing of pictorial content. *IEEE Transactions on Circuits and Systems for Video Technology 7*, 5 (Oct 1997), 771–785.

[YYL96] YEUNG M., YEO B.-L., LIU B.: Extracting story units from long programs for video browsing and navigation. *Proceedings of the Third IEEE International Conference on Multimedia Computing and Systems* (Jun 1996), 296–305.

[ZS06] ZHAI Y., SHAH M.: Visual attention detection in video sequences using spatiotemporal cues. In *MULTIMEDIA '06: Proceedings of the 14th annual ACM international conference on Multimedia* (New York, NY, USA, 2006), ACM, pp. 815–824.



(a)

(b)

Figure 5: *Video Presentation Boards. a) represented a sequence of 30 minutes long from a commercial movie, b) represented a sequence of 20 minutes long from a TV program.*