Video^{M}: Multi-Video Synopsis

Teng Li[†*], Tao Mei[‡], In-So Kweon[†], Xian-Sheng Hua[‡]
[†] Korea Advanced Institute of Science and Technology, Daejeon 305-701, KOREA
tengli@rcv.kaist.ac.kr, iskweon@ee.kaist.ac.kr
[‡] Microsoft Research Asia, Beijing 100190, P. R. China
tmei@microsoft.com, xshua@microsoft.com

Abstract

Conventional video representation methods focus predominantly on a single video, aiming at reducing the space-time redundancy as much as possible, while this paper describes a novel approach to simultaneously presenting dynamics of multiple videos, aiming at a less intrusive viewing experience. Given a main video and multiple supplementary videos, the proposed approach automatically constructs a synthesized multi-video synopsis, called Video^{M}, by integrating the supplementary videos into the most suitable space-time portions within the main video. We formulate the problem of Video^{M} as a Maximum a Posterior (MAP) problem which maximizes the desired properties related to less intrusive viewing experience, i.e., informativeness, consistency, visual naturalness, and stability. This problem is solved by the Viterbi beam search algorithm to optimally find the suitable integration between the main video and supplementary videos.

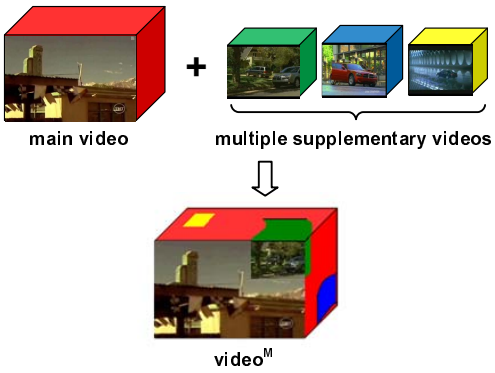

Figure 1. Idea of multi-video synopsis. Video^{M} temporally integrates multiple supplementary videos at the most suitable space-time holes within a main video (the yellow, green, and blue portions). For better viewing, please see the color pdf file.

1. Introduction

The popularity of video capture devices and internet has caused an exponential increase in the amount of available video data and in the number of users. The technology of video presentation becomes more and more important, which can be used for summarizing videos for efficient browsing, automatic new videos generating for games or other applications.

Video^{M} is a compact temporal representation of multiple videos which integrates multiple supplementary videos into the most *suitable* space-time holes within a main video. The *suitableness* is characterized by the least intrusive viewing experience. Fig. 1 shows the basic idea of Video^{M} in which three supplementary videos are simultaneously integrated into a main video. Given these input videos, the detection of space-time holes within the main video and the matching between these holes and supplementary videos have taken informativeness, consistency, and naturalness into consideration. Such manipulation provides a region-level representation of multiple videos.

There are many situations that multiple supplementary videos are expected to be integrated into a space-time portion within a main video. For example, it is a complement to existing video browsing and summarization if we can display the dynamics of multiple videos at the same time. Another applications include region or object level video advertising (or product placement) and gaming, in which advertisements are inserted into a source to replace some unimportant regions or objects.

Research on video representation has proceeded along two dimensions in terms of the input: (1) single video representation which compactly represents the information of a single video, aiming at reducing the redundancy as much

*This work was performed when the first author visited Microsoft Research Asia as a research intern.

978-0-7695-3503-6/08 $25.00 © 2008 IEEE
DOI 10.1109/ICDMW.2008.84

as possible [1, 2, 9, 11, 12, 14], and (2) multi-video representation which simultaneously displays the dynamics of multiple videos, aiming at delivering information as much as possible in a limited space or sequence [5, 7]. The first dimension has attracted much attention of research, while the second still needs to be investigated. The innovative multi-video synopsis falls into the second dimension. However, we argue that existing work in this dimension has not considered the non-intrusive viewing experience, as well as not reached region-level integration of multiple videos. As video has much space-time redundancy, it is reasonable to detect the redundant portions and achieve region or even object level integration of multiple videos.

The proposed Video$^{\mathrm{M}}$ is a novel approach to multi-video representation. Given a main video which is usually long enough and a set of supplementary videos which are shorter than the main video, we aim to synthesize a new video by inserting those supplementary videos into the most suitable space-time portions within the main video. Such manipulation is expected to achieve the least intrusive viewing experience which is regarded to be related to the following visual properties.

- **Informativeness**. The created Video$^{\mathrm{M}}$ should lose minimal space-time information of the main video. The selected 3D holes from the main video for integration should be the least informative.
- **Consistency**. The composed frames of Video$^{\mathrm{M}}$ should be consistent in appearance of the main and the supplementary frames. In other words, the frames from supplementary videos should be visually similar to those from the main video where these frames are inserted.
- **Visual naturalness**. The connecting boundary areas across different videos are visually natural or smooth.
- **Stability**. Frames of the supplementary videos should be presented continuously and orderly, and their spatial positions should be relatively stable in Video$^{\mathrm{M}}$.

Basically this is a difficult problem since some properties are defined on the frame level and they are hard to be combined together. Furthermore, the searching space is huge for finding the optimal 3D holes in the main video, which is also a 3D space. In the proposed approach, the properties are formulated to the probabilistic form, accordingly a posterior model is defined to measure the desired properties of the inserting positions for supplementary video frames. Optimal inserting positions are obtained by the Viterbi Beam search algorithm, which maximizes the posterior probability while keeping the video stability to obtain the best visual effect. The supplementary video frames are then inserted to the corresponding optimal positions and an effective information guided probabilistic seamless blending is adopted to naturalize the connecting boundary.

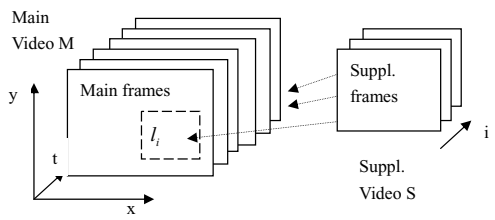

Figure 2. Illustration of problem formulation.

We will formulate the problem of Video$^{\mathrm{M}}$ in Section 2, and then present the solution to the problem, i.e, searching the optimal inserting positions for supplementary videos in Section 3. Section 4 presents the informative blending method for naturally integrating video frames. Section 5 gives the experimental results, followed by the conclusion in Section 6.

2. Problem Formulation

There are two main steps in Video$^{\mathrm{M}}$: finding the optimal inserting holes for arranging the supplementary video frames, and seamless integration. Optimal inserting holes detection is the key for the system. Assuming that the inserting region for each frame is a rectangle of a proper size, detection of inserting holes is to decide the spatial-temporal positions of the region centers in the main video and the corresponding supplementary frame to be inserted simultaneously. If we directly search each possible inserting position for each supplementary frame, obviously the searching space is too huge to handle, also the resulted inserting positions most probably will break the order and stability of the supplementary video.

In the proposed method, the visual properties on the frame level considered in Video$^{\mathrm{M}}$ are measured in a probabilistic form and the detection of inserting holes is formulated as a Maximum a Posterior (MAP) problem. The solution is to decide a sequence of inserting positions for frames of each supplementary video. In the next we will show the probabilistic formulation of Video$^{\mathrm{M}}$.

Fig. 2 illustrate the problem of Video$^{\mathrm{M}}$ using a main video $\{M_t\}_{t=1}^{T}$ containing T frames, let (x, y) denote the spatial coordinate in the frame, the supplementary video frames are arranged in the 3D space of (x, y, t). Without lose of generality, we first consider the case of one supplementary video and then extend to multi-video. The supplementary video is denoted by $\{S_i\}_{i=1}^{T'}$, T' is the number of frames of this supplementary video. The objective is to decide the optimal inserting positions $\mathbf{l} = \{l_i\}_{i=1}^{T'}$ for frames $\{S_i\}$, where $l_i = (x_i, y_i, t_i)$ is the position in (x, y, t) space. Using z to represent the desired properties of Video$^{\mathrm{M}}$ introduced in section 1, and the probability $P(z)$ as the mea-

(a) (b)

Figure 3. An example of saliency map calculation. (b) is the saliency map of (a). Higher intensity indicates higher saliency.

surement, optimal l_i is obtained by maximizing a posterior probability:

$$\mathbf{l}^* = \arg\max_{\mathbf{l}} \prod_i P(z|l_i, S_i)P(l_i) \quad (1)$$

i.e.,

$$\mathbf{l}^* = \arg\max_{\mathbf{l}} \sum_i \{\log(P(z|l_i, S_i)) + \log(P(l_i))\} \quad (2)$$

Here measurement for a single supplementary frame S_i being inserted to the position l_i is formulated as $P(z|l_i, S_i) \cdot P(l_i)$. $P(l_i)$ is the prior model that defines how much a given area in the main video is expected to be inserted, i.e., the informativeness measurement. $P(z|l_i, S_i)$ is the fitness measurement of inserting frame S_i to the area centering at l_i, which is used for evaluating the properties of consistency and visual naturalness, given by

$$\log(P(z|l_i, S_i)) = \log P^{con}(M_{t_i}, S_i) + \lambda \log P^{nat}(l_i, S_i) \quad (3)$$

where M_{t_i} is the corresponding main video frame into which S_i is integrated, λ is a weighting parameter. $P^{con}(M_{t_i}, S_i)$ is the measurement of the consistence between the two frames to be combined. $P^{nat}(l_i, S_i)$ measures the connecting naturalness of putting S_i in the area defined by l_i. In the next these measurements defined for these properties on the frame level will be detailed.

2.1. Informativeness measurement

The informativeness is a prior knowledge for inserting positions l_i defined in terms of the saliency measurement $P^{sal}(l_i)$ and the smoothness measurement $P^{smo}(l_i)$.

$$P(l_i) = P^{sal}(l_i) \cdot P^{smo}(l_i) \quad (4)$$

To minimize the informational loss of the main video and the intrusiveness of inserting, highly smooth areas should have more prior while areas containing salient parts are less expected to be inserted.

To measure the saliency and smoothness of the inserting area defined by a position, firstly a saliency map is calculated for each main frame using the visual attention model of [3] which combines static and temporal saliency maps. The static contrast-based saliency map investigating the effects of contrast in human perception while the temporal saliency map integrates the motion inductors. Fig. 3 gives an example of the saliency map of a frame.

The saliency measurement of l_i is calculated over the inserting area centering at it. Since high saliency of even a small part of this area causes information loss, we measure this informativeness using the highest J saliency value in the area. J is defined as $\frac{1}{8}$ of the area size according to the viewing experience, and $\{I_j\}_{j=1,\ldots,J}$ are the highest J saliency values, I_j ranges over $(0\ 255)$.

$$P^{sal}(l_i) = 1 - \sum_{j=1}^{J} \frac{I_j}{255 \times J} \quad (5)$$

High saliency prior indicates there is almost no salient part in this area.

Since the saliency map also contains the region segmentation information, the smoothness measurement for l_i can be defined using all the saliency values in the inserting area $\boldsymbol{I} = \{I_j\}_{j=1,\ldots,8J}$:

$$P^{smo}(l_i) = \exp\Big\{ - \frac{\sqrt{var(\boldsymbol{I})}}{255} \Big\} \quad (6)$$

where $var(\boldsymbol{I})$ is the variance of vector $\boldsymbol{I}$. High P^{smo} means the inserting area defined by l_i is quite smooth and less informative.

2.2. Consistency measurement

By consistency measurement, we expect that the inserted supplementary frames and the corresponding main frames appear to be visually similar in Video$^{\mathrm{M}}$, which makes the result look natural. Inspired by the recent success of modeling scenes using texture and colore distribution [4, 13], the combination of color similarity $P^{ccon}(M_{t_i}, S_i)$ and texture similarity $P^{tcon}(M_{t_i}, S_i)$ is used to measure the consistency between main frame M_{t_i} and supplementary frame S_i:

$$P^{con}(M_{t_i}, S_i) = P^{ccon}(M_{t_i}, S_i) \cdot P^{ccon}(M_{t_i}, S_i) \quad (7)$$

The color similarity is calculated using the color histogram correlation. For each pixel of color (R, G, B), we calculate its chromaticity color $[g, b] = [\frac{16 \times G}{R+G+B}, \frac{16 \times B}{R+G+B}]$. Based on which a 16×16 color histogram H_M for a main frame and H_S for a supplementary frame can be obtained by accumulating the points. The color relevance likelihood is obtained by:

$$P^{ccon}(M_{t_i}, S_i) = \frac{\sum_{i=1}^{16}\sum_{j=1}^{16} H_M(i,j)H_S(i,j)}{\sqrt{\sum_{i=1}^{16}\sum_{j=1}^{16} H_M(i,j)^2}\sqrt{\sum_{i=1}^{16}\sum_{j=1}^{16} H_S(i,j)^2}}$$

For texture, following the texton histogram representation in [13], a set of filter-banks are applied to each frame

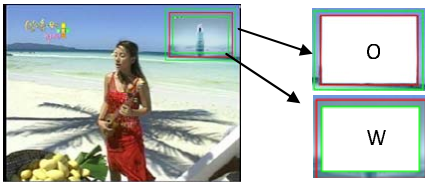

Figure 4. Connecting boundary areas. In the left image, a supplementary frame is inserted into the red rectangle. The right column shows the outside and inside areas of the connecting boundary.

using the intensity value of each pixel. The filter-banks include three Gaussians (with the scale parameter $\sigma = 1, 2, 4$), four Laplacian of Gaussians (with $\sigma = 1, 2, 4, 8$) and four order-1 derivatives of Gaussians (with $\sigma = 2, 4$ and x, y directions). Therefore, each pixel is associated with an 11-dimensional texture feature vector. We randomly select some training pixel features from the main frames and cluster to a vocabulary of texton by k-means. By mapping each pixel to one texton in the vocabulary and accumulating, a texton histogram can be gotten for each frame. The texture relevance $P^{tcon}(M_{t_i}, S_i)$ is then calculated using the same histogram correlation method as used for color relevance.

2.3. Visual naturalness measurement

Inconsistent appearance of the two sides of connecting boundaries in Video$^{\mathrm{M}}$ causes visual unnaturalness. The connecting boundary area between different videos in Video$^{\mathrm{M}}$ should be natural and has no much contrast between two sides. The naturalness can be evaluated by judging the consistency in appearance between the two side areas. The left column of Fig. 4 shows an example of the connecting boundary areas between an inserted supplementary frame and a main frame. The red rectangle indicates the inserting region. In the right column, O and W areas between the green and red rectangles are the neighboring areas outside and inside the boundary, respectively. The visual naturalness is defined as the consistence of O and W.

To consider the naturalness in different directions of the inserting area, boundary areas O_i and W_i are divided evenly to eight sub areas $\{O_i^j\}_{j=1}^8$ and $\{W_i^j\}_{j=1}^8$ according to the direction, as shown in Fig. 5. Color and texture feature are extracted and the consistency is measured between the corresponding sub inside area W_i and sub outside area O_i respectively. The same feature extraction and relevance calculation methods as used for the previous frames consistency property are adopted. The naturalness measurement $P^{nat}(l_i, S_i)$ given supplementary frame S_i and inserting position l_i is obtained by summing the consistency measurements of the sub areas

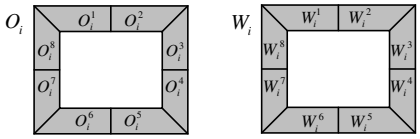

Figure 5. Illustration of boundary area feature extraction.

$$P^{nat}(S_i, l_i) = \sum_{j=1}^{8} P^{con}(O_i^j, W_i^j)/8 \tag{8}$$

where the method for calculating P^{con} is defined in equation (7).

If part of the inserting region boundary is the boundary of the main frame, the consistence measurement of sub areas in this direction is set to a relative high value since the overlapped boundary does not bring any visual intrusiveness

3. Searching for Optimal Insertion Positions

The optimal inserting positions must maximize the likelihood defined in equation (1) while satisfying some constraints such as the video frames should be inserted continuously and orderly. Therefore the problem is to to search an optimal inserting path, i.e., a series of inserting positions and their corresponding frames of a supplementary video in the 3D main video space. We adopted the Viterbi Beam search algorithm which is designed for find an optimal sequence of states for the likelihood measurement. In this process, the stability property is considered by limiting the searching paths' extension of spatial position. Most previous works utilize the energy minimization algorithms to find a local maximal or minimal value of the measurement such as [2]. The computation is high and increases significantly when the number or the size of videos increases.

3.1. The Viterbi beam search

The Viterbi algorithm is a dynamic programming algorithm that discovers the most likely explanation of a hidden states sequence for an observation. It is an efficient, recursive algorithm that performs an optimal exhaustive search along the time line. The computing of the most likely sequence up to a certain time point t depends only on the observation at point t, and the most likely sequence at point $(t-1)$ [8]. The Viterbi method searches every possible pathes and yields the global best results. But the number of possible pathes increases exponentially as the length increases, which brings the heavy load of computation and storage for long sequential data.

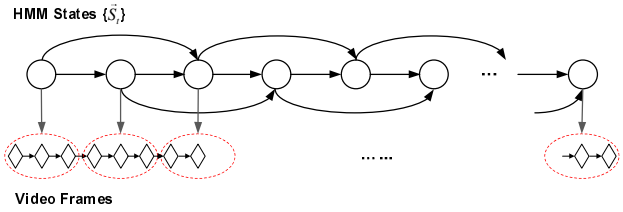

Figure 6. The HMM representation of the supplementary video.

To improve the efficiency, the beam search method [6] is usually applied to reduce the effective size of the search space. The beam search defines a pruning beam width Δ relative to the most probable path likelihood $P_{max}(t)$ at frame t. Hypotheses outside the beam, i.e., with likelihoods less than $P_{max}(t) - \Delta$, are pruned from the search. It is a heuristic search method and may miss the global best result when pruning. However, it shows to be effective in practice such as speech recognition.

3.2. HMM representation

A supplementary video is modeled by a discrete left to right Hidden Markov Model (HMM), and each state corresponds to a number of frames which is fixed to three in our implementation. The graphical view is shown in Fig. 6. Each state of the supplementary video corresponds to three inserting positions with the same spatial coordinates in three continuous main video frames. Because of the temporal continuous property of video, assigning three frames to one state does almost no influence on the fitness of inserting paths, but can improve the efficiency and video stability.

According to the HMM structure, if the current state is $\overrightarrow{S_t}$, the next candidate states to be inserted can be $\overrightarrow{S}_{t+1}$ or $\overrightarrow{S}_{t+2}$. The transition probability is

$$P(\overrightarrow{S_t} \rightarrow \overrightarrow{S}_{t+1}) = 1.0;\ P(\overrightarrow{S_t} \rightarrow \overrightarrow{S}_{t+2}) = 0.9 \qquad (9)$$

Thus some states could be jumped over, by this way some parts of the inserted video can be accelerated. By this dynamical adaptation the better visual effect may be achieved. It also does not cause the information loss of the supplementary video for its temporal continuous property.

To find the inserting positions for the HMM states sequence, the possible paths are extended along the time line of the main video. Each path X contains a historic list of the inserting positions $\{l_t\}$ and a historic list of states $\{\overrightarrow{S}_t\}$. The probability of the path X_t up to state t whose state $\overrightarrow{S}_t$ is assigned to the position l_t is:

$$\begin{aligned} & P(X_t = (l_t, \overrightarrow{S}_t)) \qquad (10) \\ = & \max_{X_{t-1}} \left\{ P(X_{t-1} \rightarrow (l_t, \overrightarrow{S}_t)) P(z|l_t, S_i) P(l_t) + P(X_{t-1}) \right\} \end{aligned}$$

where X_{t-1} is the path at state $(t-1)$ whose likelihood is $P(X_{t-1})$ and $P(X_{t-1} \rightarrow (l_i, \overrightarrow{S}_t))$ is the transition probability of extending to the next state to $(l_t, \overrightarrow{S}_t)$. S_i denote the current corresponding frames of the supplementary video.

At each possible inserting position, we initialize the paths from the beginning state of the supplementary video. Paths that reach the end of the supplementary video are saved to the output and deleted from the current path list. Finally the path with the highest likelihood in the output list is obtained as the optimal solution.

To guarantees the spatial stability property of Video$^{\mathrm{M}}$, we limit the HMM state extension of $\{l_t\}$ in a spatial neighboring area of the current position. Lower transition probability is set to farther positions in the neighborhood and vice versa while probability for the same position is maximum.

3.3. Implementation details

Considering each pixel in the main video as a possible inserted point is computationally heavy, and slightly moving the inserting positions does not yield much difference in view. Therefore we densely sample some points evenly in the search space as the candidates. All these designs are under the consideration that they do not influence the viewing measurement much.

In the search process only path list of the current frame needs to be kept in memory. For an T'-frame supplementary video, which has N_s states, and an T-frame main video which has L possible inserting positions in each frame, the maximal paths number we need to keep is $(N_s \cdot L)$. If the number of extension candidates for a state is limited to E, the size of paths searching space is at most $(N_s \cdot L \cdot E)$. The proposed method is computationally efficient and practical for long videos and online videos processing.

3.4. For multiple supplementary videos

It is straightforward to apply the approach in previous sections to the cases of multiple supplementary videos. Each supplementary video keeps an individual path list when searching along the time line. Finally an optimal solution with no overlap between the inserting paths of different supplementary videos and the relative highest overall likelihood is outputted. It is also convenient to output a list of candidates to combine with the user interaction.

4. Information Guided Probabilistic Blending

From previous steps, we can get the optimal positions that are well suitable for inserting corresponding frames naturally. But still there are clear seams that impair the visual effect of the output. Therefore, a seamless blending

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

Figure 7. Effect of the informative blending. (a) the frame to be inserted, (b) saliency map of (a), (c) no blending, (d) prob. blending, (e) informative prob. blending.

algorithm is necessary to create smooth transitions between different videos. The idea of a probabilistic alpha matting approach designed for videos is adopted [10].

The blending algorithm is applied on an extended area which covers both the inserting and its neighboring area. For each pixel e, a vector $\{P_e(M), P_e(S)\}$ representing the probabilities it belongs to main frame and supplementary frame is assigned. The output value of this pixel is obtained by:

$$e = P_e(M) * e_M + P_e(S) * e_S, P_e(M) + P_e(S) = 1, \quad (11)$$

where e_M and e_S represent the corresponding pixel value in original main frame and supplementary frame respectively.

Since the overlapped cases of different supplementary frames is not considered, in one blending step only a main frame and a supplementary frame are considered. Frames from different supplementary videos can be inserted one by one by using the previous integrated frame as the main frame. Then, an iterative process is taken to distribute each pixel's probabilities equally to its four-connected neighboring pixels, to drive neighboring pixels to have similar probability vector. The resulted probabilities are used as alpha values for alpha matting between the inserted frame and the extended inserting area of the main frame.

The probability that a pixel belongs to the supplementary frame $P_e(S)$ is associated according to its information, i.e., saliency value in the supplementary video. The saliency value is obtained using the method in section 2.1. The probability $P_e(S)$ is set to 1 if its saliency value I_e is above a threshold Th, or $\frac{I_e}{Th}$ if not. Through which the informative parts of the inserted frame are kept while the integration is highly natural. Fig. 7 compares the visual effects of an example integrated frame with no blending, probabilistic blending of [10], and the proposed informative probabilistic blending. Clearly we can see blending is crucial for visual naturalness while the proposed informative probabilistic blending shows good effect.

5. Experimental Results

We collected 16 videos as the main videos which consist of four home videos, four TV programs, four movies, and four sports videos. The durations of these videos vary from less than $1,000$ frames to more than $6,000$ frames. For each main video a set of related supplementary videos were collected from internet. In our experiments, for the convenience of processing, all the main frames were resized to 300 in width. The supplementary frames are resized to $1/4$ to $1/3$ of the size of the main frames in width. The size of extended patches for calculating connecting naturalness of section 2.3 is set to $5/4$ size of the inserted frames.

5.1. Impact of different properties

To compare the influence of different properties to the result, in this experiment we detect the optimal inserting positions for one supplementary frame in a main video of movie, using different properties settings in the MAP formulation, i.e., informativeness, naturalness, consistence and their combinations. The result is shown in Fig. 8. We can observe that considering the informativeness property alone, the inserting region is smooth and non-informative but there is obvious contrast between this region and its neighborhood which brings visual unnaturalness. The result yielded by the naturalness property has less visual contrast but this area is not smooth and contains much information of the main frame. The combination of the informativeness and naturalness properties yields a better result. However, it looks awkward to integrate these two frames together. Further combination with the consistency property gives result the most reasonable shown in Fig. 8 (d). This result validates the reasonableness and complementary characteristics of the properties we considered in Video$^{\mathrm{M}}$.

5.2. Single video synopsis

It is straightforward to apply the proposed algorithm to single video synopsis tasks. In this experiment we randomly extract four different parts from a home video and set the longest one as the main video and the other three as the supplementary videos. Fig. 9 shows some frames of the input videos and the resulted video, the red, green and blue lines indicating three inserted supplementary videos respectively are set below the inserting area. It can be seen that the detected inserting regions are reasonable since the inserted frames are well fusion with the neighboring snow background and they do not defect the main frame scene. This is an easy case since there are always some scenes similar or related to the supplementary videos in the main video as they come from a single video. While video synopsis [9] naturally packs many activities of instances in the video, only the information of activities matter can be kept, regardless of their temporal context and background information. Comparatively more information can be kept in Video$^{\mathrm{M}}$.

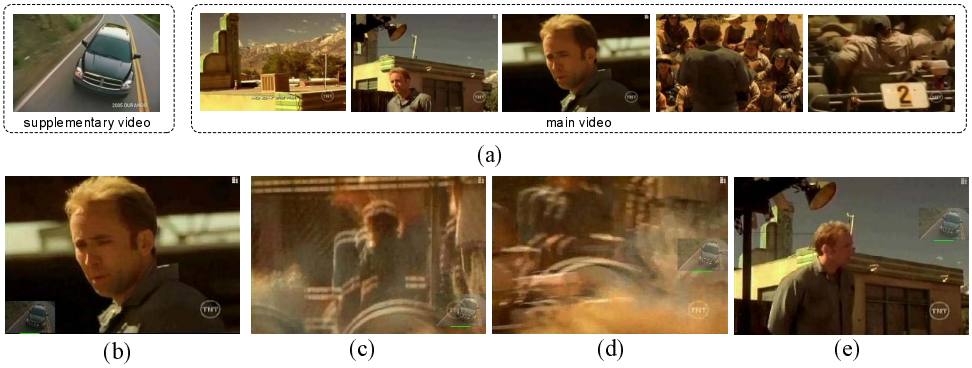

Figure 8. Detected inserting positions using different properties settings. (a) a supplementary frame (top left) and some frames of a movie. The detection results: (b) informativeness; (c) naturalness; (d) informativeness and naturalness; (e) informativeness, naturalness, and consistency. A green line is intentionally added at the bottom of the supplementary video for better viewing.

Figure 9. An example of single video synopsis using a home video. The red, green, and blue lines are intentionally added at the bottom of the supplementary videos for clarification.

5.3. Multi-video synopsis

Furthermore, the proposed synopsis approach is well designed for multiple videos of various scenes. Fig. 10 shows the result of integrating two different supplementary advertising videos to a TV program. Interestingly, there is the beach in both the main video and the supplementary video # 2. The proposed algorithm successfully detects the optimal inserting regions and naturally integrates the two beach scenes together. For another supplementary video, even though there is no similar scene in the main video, the optimal integration does not cause much intrusiveness while keeping the information.

Fig. 10 is a good example of applying Video$^{\mathrm{M}}$ to video advertising. In today's TV programs or videos on the Internet, hard insertion of advertisements often annoys the audience. The proposed Video$^{\mathrm{M}}$ can optimally insert advertising videos with much less intrusiveness. While the VideoSense system proposed in [5] has applied video synthesization to advertising, it does not achieve region-level advertisements insertion. The proposed Video$^{\mathrm{M}}$ is an efficient approach to region-based advertising.

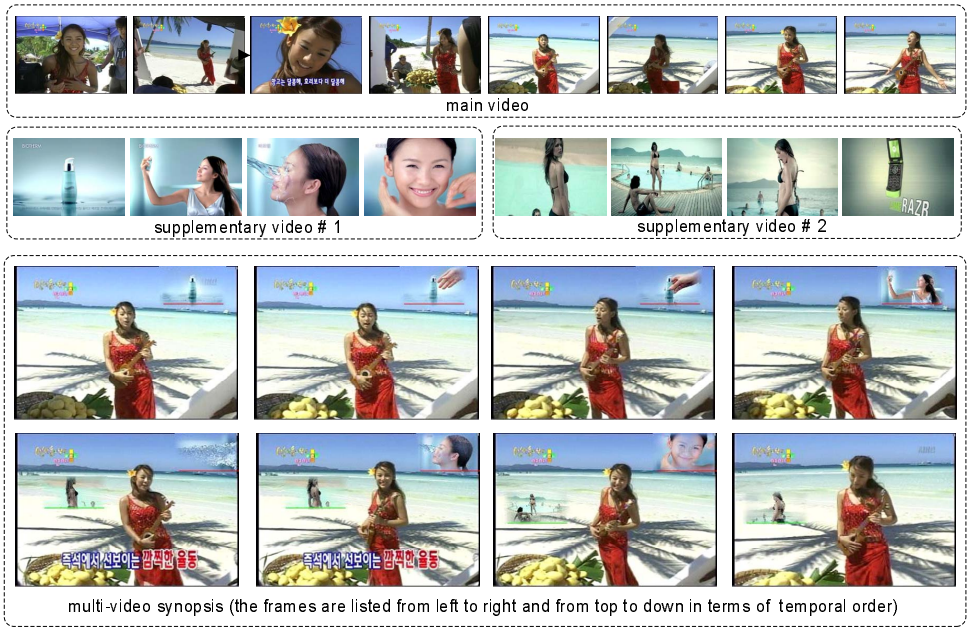

Figure 10. An example of multi-video synopsis using a TV program.

6. Conclusion and Future Work

In this paper, we propose a novel approach to multi-video synopsis which automatically creates a synthesized video from a main video and multiple supplementary videos. This synthesized video, called Video$^{\mathrm{M}}$, is a kind of video-to-video integration at region level while preserving the least intrusive viewing experience.

The future work and extensions to the current approach could include: 1) Automatically detecting the existence of the valid holes in the main video for a supplementary video can be integrated into Video$^{\mathrm{M}}$ as a pre-step. 2) The HMM representation does not consider the content summarization of supplementary videos. More complex sequential models for both the main and supplementary videos can be considered to improve the flexibility of Video$^{\mathrm{M}}$. Moreover, simple user interactions can be incorporated to improve the subjective visual effect.

References

[1] M. Irani, P. Anandan, and S. Hsu. Mosaic based representations of video sequences and their applications. *Proceedings of ICCV*, pages 605–611, 1995.

[2] H. Kang, Y. Matsushita, X. Tang, and X. Chen. Space-time video montage. *Proceedings of CVPR*, pages 1331–1338, 2006.

[3] Y.-F. Ma, X.-S. Hua, L. Lu, and H.-J. Zhang. A generic framework of user attention model and its application in video summarization. *IEEE Trans. on Multimedia*, 7(5):907–919, January 2005.

[4] T. Mei, X.-S. Hua, W. Lai, L. Yang, and et al. Msra-ustc-sjtu at trecvid 2007: High-level feature extraction and search. *TREC Video Retrieval Evaluation Online Proceedings (TRECVID)*, 2007.

[5] T. Mei, X.-S. Hua, L. Yang, and S. Li. Videosense: Towards effective online video advertising. *ACM Multimedia*, pages 1075–1084, September 2007.

[6] H. Ney, D. Mergel, A. Noll, and A. Paesler. Data-driven search organization for continuous speech recognition. *IEEE Trans. on Signal Processing*, 40:272–281, 1992.

[7] Online. Blinkx. *www.blinkx.com/wall*.

[8] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[9] A. Rav-Acha, Y. Pritch, and S. Peleg. Making a long video short: Dynamic video synopsis. *Proceedings of CVPR*, pages 435–441, 2006.

[10] C. Rother, L. Bordeaux, Y. Hamadi, and A. Blake. Autocollage. *Proceedings of ACM Siggraph*, 2006.

[11] A. M. Smith and T. Kanade. Video skimming and characterization through the combination of image and language understanding. *Proceedings of CAIVD*, pages 61–70, 1998.

[12] T. Wang, T. Mei, X.-S. Hua, X. Liu, and H.-Q. Zhou. Video collage: A novel presentation of video sequence. *IEEE International Conference on Multimedia & Expo*, pages 1479–1482, 2007.

[13] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2005.

[14] M. M. Yeung and B. L. Yeo. Video visualization for compact presentation and fast browsing of pictorial content. *IEEE Trans. on CSVT*, 7(5):771–785, 1997.