THE DYNAMIC VIDEOBOOK: A HIERARCHICAL SUMMARIZATION FOR SURVEILLANCE VIDEO

Lei SUN, Haizhou AI
Computer Science and Technology Department
Tsinghua University
Beijing, 100084, China
ahz@mail.tsinghua.edu.cn

Shihong LAO
Development Center
OMRON Social Solutions Co., LTD
Kyoto 619-0283, Japan
lao@ari.ncl.omron.co.jp

ABSTRACT
This paper proposes a novel hierarchical summarization system for surveillance video in which synopsis video and original video are at the lower layer and dynamic collage at the higher layer. A user can efficiently browse the video content through dynamic zooming into collage images and further skimming synopsis video or original video for details, enabling an interactive interface for video navigation. We call this system Dynamic VideoBook because its structure is akin to a book, that is, dynamic collage images resemble the cover and the contents while synopsis video gives an abridged edition of the video. Unlike previous collage techniques for films or entertainment videos, collage for surveillance video is challenging since such video imposes restrictions on spatiotemporal relationship and focuses on multiple foreground objects. So we put forward a new collage method which better conveys the spatiotemporal motion and structurally extracts the visual contents from the video simultaneously. And we also enable dynamic zooming in to reveal more information in real-time. In addition, by mapping the upper layer of sketchy visualization to the lower layer of visual data, this system can efficiently explore and browse the video content, highlight one motion or interactively search one specific fragment.

Index Terms— Video summarization, surveillance video, dynamic video collage

1. INTRODUCTION
Video summaries provide compact, coherent and succinct representations of a video’s content such as video collage [1], video synopsis [2], video narratives [3], tapestry [4] and storylines [5]. Since explosive growth of surveillance video data presents formidable challenges to its browsing, retrieval and storage, video summarization is of significance to relieve human from laborious manual browsing and examining. Of existed work, video synopsis [2] has proved to be an effective method to summarize the video in a compact and coherent manner. Unlike video synopsis, video collage [1] and related similar work [3–5], commonly applied in movies or entertainment videos with one or several characters and varying background, is not quite well suited for surveillance video where multiple targets are solely or collectively moving. In such scenarios, collage method aids to compactness, but conveys neither the flow of time nor the varying of motion. The most similar applicable scene to our work is [3], which dynamically revealed the motion of one or several actors. Unfortunately, it cannot work well in the case of occluding motion which occurs often in surveillance videos.

Fig. 1: The Dynamic VideoBook: Organize and summarize a visual book

We observe that considerable efforts have been invested in the summarization technique while very little attention is paid to organizing the video content that is quite useful in video summarization and navigation, e.g., the character-story organization of movie helps the summary to eliminate dispensable plots. Another observation is that hierarchical collage incorporated with synopsis provides a better user interface. Based on these we present a novel hierarchical summarization system called Dynamic VideoBook. It assumes that the surveillance video can be organized as a visual book with one visual cover, visual contents and visual pages and consequently the video would be hierarchically structured and efficiently indexed by the contents. Under this assumption, the system consists of two layers illustrated in Figure 1, of which the upper one is collage layer to extract visual contents and the lower is synopsis layer to summarize visual data. We adopt our previous work [6] to do online video synopsis to summarize general motions, during which foreground object track-
s are extracted, recorded and selectively arranged in a more condensed way. Meanwhile, video collage can offer an explicit comprehensible understanding of the video. However, the common collage technique selecting region of interests (ROIs) and then blending them together impacts the motion integrity and ignores the spatiotemporal relation of surveillance videos. So we present a well motion conveying collage method to structurally extract visual contents. And, we also enable dynamic browsing by zooming into the collage in real-time. Moreover, in order to efficiently search or index the visual book, we build the mappings from collage layer to synopsis layer. As showed in Figure 1, the upper image in the collage layer is the overall collage image treated as the visual cover composed of three images boxed with blue regarded as the visual chapters. Users can review more information of each chapter by zooming in and then obtain the sub chapters as depicted in the lower image of collage layer. By continuously browsing the entire visual contents would be gradually formed. And for more details in the yellow rectangle, users can efficiently index to synopsis layer for corresponding synopsis sequences and original video sequences. With this hierarchical design, our system is not only effective in video navigation by dynamic video collage but also efficient in video searching and indexing by mapping to synopsis or original video.

2. SYSTEM OVERVIEW

Given a surveillance video, we aim at establishing compact yet comprehensible video collage enabling zoom-in and zoom-out, synopsis video and mappings from collage layer to synopsis layer. Our synopsis [6] process summarizes the video content in real-time. Then the collage process would begin subsequently based on the object tracks extracted during synopsis.

![Fig. 2: System overview (red rectangle: pROI, green rectangle: detailed patches)](image)

In order to make descriptive collage and concerning that simply blending small ROIs would generally ruin the ensemble motions and spatiotemporal relations, we adopt the mechanism that assembles larger and representative ROIs (pROIs) into several images appropriately that best describe the video plot called plot images (red rectangle in Figure 1) and selectively stitch smaller ROIs called detailed patches (yellow rectangle in Figure 1) around in temporal order to generate a collage image. We believe such mechanism facilitates the representation of general motions by plot images and detailed description by detailed patches.

The system framework is showed in Figure 2. We first segment the video into several volumes, extract pROIs of each volume subsequently and then assemble pROIs to plot images. The final collage is obtained by blending plot images and selected detailed patches in a storyboard layout. Last but not least, we build the mappings. And if users request to thoroughly explore into the collage, we iteratively process the above procedure by treating the corresponding depicted volume as a new video, enabling dynamic video navigation.

3. DYNAMIC VIDEO COLLAGE

Given a video sequence \(V\) containing \(M\) frames \(\{I_i\}_{i=1}^M\) and \(N\) foreground objects \(\{O_j\}_{j=1}^N\) obtained from synopsis process, each \(O_j\) contains a set of variables \(O_j = (f_j^i, f_j, T_j, W_j)\), where \(f_j^i\) and \(f_j\) are the start and end frame number of this object, \(T_j\) is a set of tracks and \(W_j\) is a set of the corresponding weight of each track. We segment \(V\) to \(\{V_k\}_{k=1}^K\), select pROIs \(\{R\}\) and detailed patches \(\{P\}\) and compose them to a collage \(C\). Let \(\lambda\) be one feasible solution \(\lambda = \{\{V_k\}_{k=1}^K, \{R\}, \{P\}\}\), we formalize the video collage as minimize \(E(\lambda)\) process:

\[
E(\lambda) = \omega_1 E_{\text{seg}}(\lambda) + \omega_2 E_{\text{rep}}(\lambda) + \omega_3 E_{\text{comp}}(\lambda)
\]

where \(E_{\text{seg}}(\lambda)\) denotes the segmentation cost, \(E_{\text{rep}}(\lambda)\) denotes representation cost caused by pROIs and detailed patches selection and \(E_{\text{comp}}(\lambda)\) denotes the composition cost in assembling pROIs and detailed patches, \(\omega_1, \omega_2, \omega_3\) are pre-defined to control the relative strength of each energy term. We solve this minimization problem in a step-wise manner.

3.1. Segmentation cost \(E_{\text{seg}}(\lambda)\)

A fine segmentation should enlarge the within-volume similarity and decrease the between-volume similarity. We measure the similarity \(S(V_i, V_j) = \sum_{i=1}^{N} \min(w_i^j, w_i)\), where \(w_i^j\) is the weight sum of tracks of object \(O_i\) appeared in \(V_i\). In this way, we define \(E_{\text{seg}}(\lambda)\) as following:

\[
E_{\text{seg}}(\lambda) = \sum_{k=1}^{K} \min(S(V_k, V_j), S(V_j, V_k))
\]

where \(V_j\) is the neighbor volume of \(V_k\).

3.2. Representation cost \(E_{\text{rep}}(\lambda)\)

The representation cost is associated with to what degree the selected pROIs and patches represent the volume. Taking the preservation of spatiotemporal relation, objects and motions into account, we define the cost as a combination of each term.
where $A(V_k)$, $B(V_k)$ and $D(V_k)$ denote the spatiotemporal cost, object cost and motion cost, respectively, and $\alpha$, $\beta$ and $\gamma$ measure their relative importance.

As the spatiotemporal relation and foreground objects are strongly emphasized, we construct candidate pROIs such that the region in small volume has foreground pixels and in the meantime it is distant from other regions, ensuring small motions and clustered objects preserved in pROIs while undesired background eliminated. In practical implementation, for each frame we cluster the foreground regions of neighbor small volume (e.g., 30 frames) into pROIs and regions of smaller volume (e.g., 10 frames) into detailed patches. Given candidate pROIs and detailed patches, $A(V_k)$ can be measured as the weighted sum of clustering moving objects missed in the selected pROIs, $B(V_k)$ is calculated by the amount of lost objects while $D(V_k)$ computes the weighted sum of lost tracks of each object. In each volume, we select a predefined number of pROIs and detailed patches. We test our approach on sequence S2LI from VS-PETS 2009 data set.\(^1\)

The three selected pROIs and detailed patches of the first segmented volume are depicted in Figure 3 (a) and (b), respectively. It can be easily perceived that there exist 7 pedestrians moving of which two coupled walked, three discussed for a while and one in blue cloth walked through. In addition, given the pROIs the detailed patches reveal more information.

**Fig. 3**: (a) The extracted pROIs, (b) extracted detailed patches, (c) the plot image obtained from (a) and (b)

### 3.3. Composition cost $E_{comp}(\lambda)$

The composition cost penalizes for the weak composition quality of pROIs $\{R\}$ and detailed patches $\{P\}$. To make the collage comprehensible, it is critical to make sense of the composed plot image. Particularly for surveillance video, we expect general motions kept intact and groups well preserved. However, there are often occluding situations from objects of different frames which results in unpleasant visual experience in collage by stitching objects from each pROI with respect to their spatial locations. In contrast, simply blending pROIs together avoids the occlude occasion but suffers from the loose structure and implicit motions. Concerning the above, we adopt the mechanism that if two pROIs share a certain number of objects we blend them else we simply concatenate them together. If small regions intersect we resize plot image (foreground object areas from pROIs are retained) using seamless carving [7] and selectively stitch objects with slight deviations subsequently, or else if large regions occlude we concatenate occluding area using graph cuts. In this way, we give the composition cost of $\{R\}$ is computed by $C_r(\{R\})=L(\{R\})+\frac{\text{Area}(I_p)}{\text{Area}(R)}$, where $L(\{R\})$ denotes the weighted sum of lost objects and the deviations from original positions, $\text{Area}(\cdot)$ denotes the area size and $I_p$ denotes the plot image, respectively. Arranging $\{P\}$ is quite simple: resizing and blending in temporal order. As a consequence, the composition cost of $\{P\}$ is computed by $C_p(\{P\})=\sum_{k=1}^{K}(1-\frac{\text{Area}(\hat{P})}{\text{Area}(P)})$, where $\hat{P}$ denotes the resized image of $P$. So the overall composition cost can be described as the following function:

$$E_{comp}(\lambda) = \sum_{k=1}^{K}(C_r(\{R\}) + C_p(\{P\}))$$

Since it is expected to show the temporal relation that helps understanding the flow of time and conveying the motions, there is a falloff for each object that makes older objects more transparent and highlights the newer ones in the plot image, as illustrated in Figure 3 (c).

### 3.4. Dynamic browse

To better explore the video content, we enable dynamic browsing by clicking the plot images, which would iteratively minimize the energy and generate a new detailed collage image containing its own plot images and detailed patches subsequently. However, the collage must be showed and zoomed in or out in real-time, which imposes certain cost limitations in the video processing. Given the energy definition, the first and second term can be pre-computed, so the segmentation results as well as the selected pROIs and detailed patches can hence be stored. When users request for zooming in, the corresponding pROIs and detailed patches are composed appropriately with respect to the third term of energy function. In our implementation, it costs averagely three seconds for generating a new collage.

### 4. Dynamic VideoBook

Video collage extracts succinct and explicit visual contents for visualization while video synopsis condenses informative video content for quick browsing. However, neither of them is workable independently to support users to efficiently search or index. So we establish a hybrid system with dynamic collage as the upper layer and original video as well as synopsis video as the lower layer. Building the mappings from the collage layer to synopsis layer is just like linking the contents to book texts, which obviously improves the searching efficiency. In detail, we map each plot image and detailed patches

---

\(^1\)http://www.cvg.rdg.ac.uk/PETS2009
to one certain time span of synopsis video as well as original video, and we also map each synopsis frame to one certain time span of the original video. In this way, if a user sees an interested object showed in the collage image, it would be a lot easier and quicker to check its sketchy motion from the synopsis video. For more details the correlated original video sequence can be quickly indexed and browsed by mapping either through the synopsis or the collage.

With this design, the upper layer of the Dynamic Video-Book takes charge of generally dynamic interactive browsing, the lower layer summaries the content while the hierarchical design enables efficient search and indexing. Users can also highlight or search specific objects at the collage panel. As showed in Figure 1, the patches boxed with yellow can be quickly indexed to synopsis and original sequences.

5. EXPERIMENTS

We test our system on a public surveillance video, and we also compare our collage method with the Wang et. al.’s method [1]. As illustrated in Figure 4, (a) gives the collage result using Wang et. al.’s method, which selects ROIs and then blends them together. (b) shows three frames of generated synopsis video in which the objects are coherently and densely arranged. In (c), the upper image is the overall collage image using our collage method and the lower image is the dynamic collage depicting the zoomed-in area of the upper collage, in which the motions are more explicitly depicted. It is quite clear that the collage image using our method is more comprehensible than the traditional method [1], which we believe benefits from the plot image with motion well preserved and spatiotemporal relation revealed. We can easily perceive that people are walking in some certain area from our collage image. In contrast, the traditional collage images confuse users for ignoring the spatiotemporal information. In addition, through this hierarchical system we can highlight objects like the two boxed in yellow in the lower collage in (c), the corresponding pedestrians have been boxed in red through the collage and can be indexed to relative synopsis sequences.

6. CONCLUSIONS

In this paper, we present a hierarchical video summarization system particularly for surveillance video: The Dynamic VideoBook, which incorporates video collage structurally extracting general visual contents and video synopsis summarizing informative motions. Since the common used video collage methods do not work well in the surveillance video, we put forward a new dynamic video collage method, which makes the result more comprehensible. In addition, the collage can be dynamically zoomed in to get details. And having the contents of the visual book and visual data including the synopsis video and original video, we hierarchically build mappings from the contents to visual data for efficient interactive searching or indexing.

7. ACKNOWLEDGEMENT

This work is supported by National Science Foundation of China under grant No.61075026, and it is also supported by a grant from Omron Corporation.

8. REFERENCES