IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 1, Issue 5, Oct-Nov, 2013
ISSN: 2320 - 8791
www.ijreat.org

Transition Clues Based Video Summarization By Using Scale Invariant Feature Transform

S.Bhuvaneswari(Asst.Prof)[1], **Mohanavalli**(Asst.Prof)[2], **K.Sudha**(P.G Scholar)[3]

DEPARTMENT OF ELECTRONIC AND COMMUNICATION SYSTEMS
PRATHYUSHA INSTITUTE OF TECHNOLOGY AND MANAGEMENT, Chennai.

ABSTRACT— A large amount of short, single-shot videos are created by personal camcorder every day, such as the small video clips in family albums, and thus a solution for presenting and managing these video clips is highly desired. From the perspective of professionalism and artistry, long-take/shot video, also termed oneshot video, is able to present events, persons or scenic spots in an informative manner. This paper presents a novel video composition system "Video Puzzle" which generates aesthetically enhanced long-shot videos from short video clips. Our task here is to automatically composite several related single shots into a virtual long-take video with spatial and temporal consistenc

We propose a novel framework to compose descriptive long-take video with content-consistent shots retrieved from a video pool. For each video, frame-by-frame search is performed over the entire pool to find start-end content correspondences through a coarse-to-fine partial matching process. The content correspondence here is general and can refer to the matched regions or objects, such as human body and face. The content consistency of these correspondences enables us to design several shot transition schemes to seamlessly stitch one shot to another in a spatially andtemporally consistent manner. The entire long-take video thus comprises several single shots with consistent contents and fluent transitions. Meanwhile, with the generated matching graph of videos, the proposed system can also provide an efficient video browsing mode. Experiments are conducted on multiple video albums and the results demonstrate the effectiveness and the usefulness of the proposed scheme.

keyword—Image retrieval, one-shot video, video authoring,video transition.

I.INTRODUCTION

These digital videos have several characteristics: (1) comparedwith former videos recorded by non-digital camcorder, nowadays videos are usually captured more casually due to the less constraint of storage, and thus the number of clips is often quite large; (2) many videos may only contain a single shot and are very short; and (3) their contents are diverse yet related with few major subjects or events.

Users often need to maintain their own video clip collections captured at different locations and time. These unedited and unorganized videos bring difficulties to their management and manipulation. For example, when users want to share their story with others over video sharing websites and social networks, such as YouTube.com and Facebook.com, they will need to put more efforts in finding, organizing and uploading the small video clips. This could be an extremely difficult **"Puzzle"** for users.Previous effortstowards efficient browsing such large amount of videos mainly focus on video summarization. These methods aim to capture the main idea of the video collection in a broad way, which, however, are not sufficiently applicable for video browsing and presentation.

In this paper, we introduce a scheme, "Video Puzzle", which can automatically generate a virtual one-shot presentation from multiple video clips. Given a messy collection of video clips, Video Puzzle can select a clip subset with consistent major topic (similar with finding the clues and solving the Puzzle Games among the images). The topic can refer to a person, object, or a scene here. It can be specified by users or found with an automatic discovery method. The start-end frame correspondences of these clips are then established with an efficient coarse-to-fine method, and we compose them into a long clip in a seamless manner accordingly, i.e., a one-shot presentation. Therefore, Video Puzzle provides a novel presentation of video content that enables users to have a deeper impression of the story within the video collection.

Fig. 1 shows the working process of Video Puzzle via two examples. The system can automatically discover video clips with "similar/continuous topics" in a video album and naturally stitch them into a single virtual long-take video, which can yield a cohesive presentation and convey a consistent underlying story. It is challenging as 1) it is generally hard to find shots which can be naturally combined among a large amount of candidate videos, and 2) generating

seamless transition between video shots is difficult usually.

Fig. 1. An illustration of the Video Puzzle presentation scheme, which generates one-shot videos by selecting and composing short video clips.

The contribution of our work can be summarized as follows:
(1) We propose a video puzzle scheme. It is able to extract video contents about a specific topic and compose them into a virtual one-shot presentation. The scheme is flexible and several components can be customized and applied to different applications.
(2) We propose an efficient method to find the content correspondences of multiple videos and then compose them into a clip with an optimized approach.
(3) We introduce two applications based on the video puzzle scheme, one about home video presentation and the other about landmark video generation.

Specifically, the two specific applications introduced are:
(1) Personal video presentation.With a large set of personal video contents, we can generate a video matching graph which explicitly shows the content-consecutive relation of videos. The storyline of the video album found by Video Puzzle will automatically pop up. Besides, user only needs to appoint a specific person or scene and then we can generate a one-shot presentation to describe the corresponding person or scene by mining the video graph.
(2) Comprehensive landmark video generation. With multiple web videos that describe the same landmark, we are able to generate a one-shot visual description of the landmark, which contains more comprehensive visual description of the landmark, such as the visual content captured from different view.

II. RELATED WORK

A. Video Summarization

Many previous works focus on producing effective video summarization with visual friendliness and in a compact form. Existing methods can be classified into two categories, i.e., dynamic representation and static representation. Dynamic Representation generates a video sequence that is composed of a series of sub-clips extracted from one or multiple video sequences or generated from a collection of photos [1,2,3] . Whereas static representation generally generates one or multiple images from video key-frames to facilitate not only their viewing but also transmission and storage[4,5,6]. Although video summarization can reduce the cost for video browsing, there is a risk of missing details and the possibly inaccurate summarization also may cause inconvenience .

B. Video Editing/Composition

Our work is also related to video editing and composition. In comparison with still image editing, content-based video editing faces the additional challenges of maintaining the spatial-temporal consistency with respect to geometry. This brings up difficulties of seamlessly modifying video contents, such as inserting or removing an object. Zhang *et al.* [7] provide a solution based on an unsupervised inference of view-dependent depth maps for all video frames. Yan *et al.* [8] transfer desired features from a source video to the target video such as colorizing videos, reducing video blurs, and video rhythm adjustment.

The overall scheme "Video Puzzle" aims to discover content- consistent video shots and composes them into a virtual long-take video. To this end,we propose a novel graph-based visualization and path finding approach. The graph is constructed based on geometry matching (homograph mapping) and object matching (human, face). Based on the multi-cue content matching, the transition of video shots becomes meaningful and seamless.

III. AN OVERVIEW OF THE SCHEME

Our task is to automatically compose several related video shots into a virtual long-take video with spatial and temporal consistency, and it is different from the traditional works that try to either find a group of similar video clips or fit the composed video with extra information such as music or metadata ,. For a given video collection that contains video clips ,3 the system mainly contains three key components, as illustrated in Fig. 2. Firstly, we implement a coarse-to-fine partial matching scheme to generate a matching graph of the video collection. The matching scheme serves as a three-level matching, i.e., video pair selection, sequence-

sequence correspondence finding, and frame-level exact matching. The video pair selection acts as an evidence for ensuring the *non-redundant* and *complete* quality of the generated one-shot video. It uses a hashing-based method to quickly obtain the video similarity measurement. We then find sequence correspondence of the selected video pairs through local keypoints matching. The final frame-level matching aims to find different matched objects to provide variant and rich clues for video transition generation. We implement three object matching methods in this part, i.e., salient object matching using local visual pattern discovery [9], and human andfacelocation.

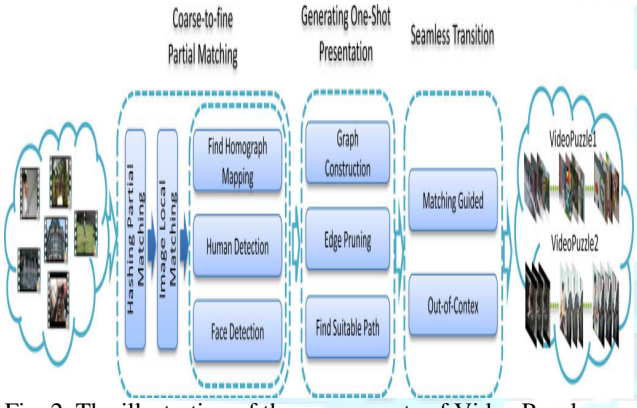


Fig. 2. The illustration of the components of Video Puzzle

Secondly, we design a flexible scheme to select the optimal video compositions from a constructed video matching graph. The video selection task turns out to find the longest path in the graph by constructing a video matching graph following three criteria, i.e., continuity, completeness and diversity. This selection scheme can either work fully automatically by creating one-shot videos with globally optimal content consistency or work interactively with users by generating one-shot videos with optional topics (such as the specified key objects or persons). We will introduce the details in Section V

. Finally, we compose the video correspondence pair one by one. We propose a space-temporal morphing-based transition through matched local patterns, i.e.,matched local common pattern, matched human or face. The produced transition is more natural than the traditional transitions such as fade-in, fadeout, wipes, and dissolve.4 Since both image-level and sequencelevel matching for video pairs are available, we can accomplish a content-based continuous transition. The proposed content- based transition produces virtually consistent link for the final composition. The Video Puzzle system, which can automatically generate a virtual one-shot presentation frommultiple video clips, provides a novel presentation of video contents and enables users to have a deeper impression of the story from the video collection. We will provide two applications in detail.

IV. Coarse-to-Fine Partial Matching

In this section, we introduce the first component of the system, namely, coarse-to-fine partial matching. The target of this part is to (a) produce video similarity measurement acting as evidence for ensuring the *non-redundant* and *complete* quality of the generated video; (b) fast and accurately locate the sequences in video pairs with start-end content correspondence; and (c) find the keyframe pairs with transition clues in the correspondence sequences. We first use a hashing-based method to quickly obtain the video similarity measurement. Then we try to match two video sub-sequences in order to generate continuous transition. Finally, specific transition clues are obtained for video composition through local common pattern discovery, human appearance modeling and face appearance modeling.

A. Hashing-Based Video Pair Selection

In this part, we adopt the recently proposed Partition Min-Hashing (PmH) [10] algorithm to rapidly calculate the frame partial similarity between every pair of videos and the computed frame similarity is accumulated to estimate the video similarity measurement. Then, the video pairs with high similarity are selected as candidate pairs to generate one-shot videos. A graph of video similarity is built based on the results of video pair selection.

B. Sequence Matching

In this subsection, we aim to accurately match two video subsequences within the selected video pairs in order to generate continuous transition. We propose a method that uses image local matching to get the correspondence of two sub-sequences.

1) Image Local Keypoints Matching: We use SIFT [11] + Color Moments with Difference of Gaussians (DOG) keypoint detector. Existing studies demonstrate that the SIFT descriptors and Color Moments are complementary to each other, one describing the local structure and the other providing higherorder information of local differences. We concatenate these two features to describe each local keypoint. To determine the local match, we use the method proposed by [11]. Given two frames (the source image, frame in video , and the target image, frame in video), the best candidate match for each keypoint of the source image is found by identifying its nearest neighbor among the keypoints from the target image. The nearest neighbor is defined as the keypoint with the minimum Euclidean distance. Since there will be many keypoints from the source image that do not have any correct match in the target image, such as those that arise from background clutter or are not detected in the target image, it is useful to discard them.

An effective measure is obtained by comparing the distance of the closest neighbor to that of the second-closest neighbor. We then get the keypoints matching set and their matching scores which is the similarity measurement of the keypoint matching pair.

2) Frame Similarity to Sequence Correspondence: To locate the sequence correspondence of two videos, we sample the video frames in a constant rate.

Since the ultimate video clip is expected to contain only one shot,we compose two videos only in their starting or ending part in order to keep the storyline within the clips. Thus the video sequence correspondence is also found within the starting and ending part.

For detail, the video is first partitioned into two parts with equivalent duration. The sequence correspondence represents a sequence in the second part of the video is matched with a sequence in the first part of the video . Then, the sequence similarity is scaled by a preference factor which is set to 1 when the two sequences are close to the start or end of the videos and gradually turns to 0 when one of the sequences is far from the video border. For each video pair and , we obtain the sequence correspondence by finding the sequence pair with the largest sequence similarity. The video similarity in the graph is also replaced by this sequence similarity.

C. Transition Clues for Video Composition

Given two matched sequences, we select the frame pairs as the transition key frames according to several transition clues.

(1)*Cross-Frame Common Pattern Discovery:* The first transition clue we use is based on image matching. Since image matching often contains a large amount of outliers, we need a robust fitting method to find the common pattern. Specifically,common pattern denotes those matching pairs that share the same or similar homogeneous transformation parameters. RANSAC method is utilized to find the matching transformation parameters. However, practically RANSAC may perform poorly when the ratio of inliers falls below 50%. It means that a large overlap between the pair of images is required for the matching, which is rare case for location representation in a video. Therefore, we need an extra method to determine the true matches within the matched pairs. It is worth noting that [12] incorporated the local matching within a large image collection scenario with RANSAC. For this part, we adopt the "Graph Shift" method . The main idea is to introduce spatial constraint for the matched pair to find a dense common pattern. This algorithm has three advantages over RANSAC: 1) it is robust to outliers; 2) it is able to discover all common visual patterns, no matter the mappings among the common patterns are one-to-one, one-to-many, or many-to-many; and 3) it is computationally efficient.

2) Human Appearance Matching: The frames from two videos are also matched according to the appearance of human contained in the video. Firstly, automatic human body detection is accomplished. We implement the part-based model in learnt with the annotated human images from the PASCAL Visual Object Classes (VOC) Challenge 2010 dataset [14] for human detection. Some examples of the training samples are shown in Fig. 3. The part-based detection model contains two parts, one describing full view (denoted as root model) and the other describing part views (denoted as part models

3) Face Appearance Matching: We also implement the state-of-the-art multi-view face detector [15] and active shape model for fac alignment. For each frame, we perform the near-frontal face detector to localize the face area as well as several facial parts, such as eyes, mouth, nose and face contour.

A frame with face is assumed to be matched with another frame with face according to the following criteria:

1) Both face areas should be large enough. Small face areas are much less important since video matching and transition on small area frequently lead to unnatural effects. In our implementation, we set the threshold to 3,600 pixels.

2) The faces should belong to the same person. We first perform the face alignment procedure to align the faces and then calculate the Euclidean distance for the feature vectors extracted from each face pair. A threshold is empirically set to remove most mismatched candidates.

3) The two face poses should not vary much. The output of the face detector [18] includes the pose view information.

V. GENERATION OF ONE-SHOT PRESENTATION

A. Edge Pruning

We prune the cycle paths in the graph to avoid the repeated clips in the composed video. An important and also the most straightforward criterion is time constraint. People are used to watching videos in the order of time, especially for home video browsing. We use the timestamp metadata of the video clips to ensure that the shots maintain the temporal relationships in the composition process. However, we also notice that many video clips lack of such metadata. Therefore, for those video clips, we need to design extra content-based edge pruning method to reduce the cycle graph.

B. Path Finding

1) Automatic Path Finding: Themaximal paths can be found automatically. After finding the longest path over the graph , all the edge weight linking to those nodes in the path should be scaled by a factor (in our implementation) to reduce the possibility for these nodes to be selected again. We then find the longest path again in the updated graph. This procedure can be iterated until reaching the criterion that the sum of weights in the final path is less than a threshold.

2) Interactive Path Finding: For personal usage, the one shot technique can help to find and composite consecutive video clips with human interaction. A user may expect a one-shot video that contains a specified key video clip or focuses on a specific object or scene.

VI. SEAMLESS VIDEO COMPOSITION

Here we introduce how to compose the selected video clips into one-shot video and the key problem is to smooth the visual discontinuities at the transitions. For each two best matched frames, all the matches are local, such as common patterns, human bodies, and faces. Since directly stitching the two videos based on these two frames may lead to abrupt change, we need to consider adding natural transition, which act as the link between the two consecutive videos, in the final virtual long-take video.

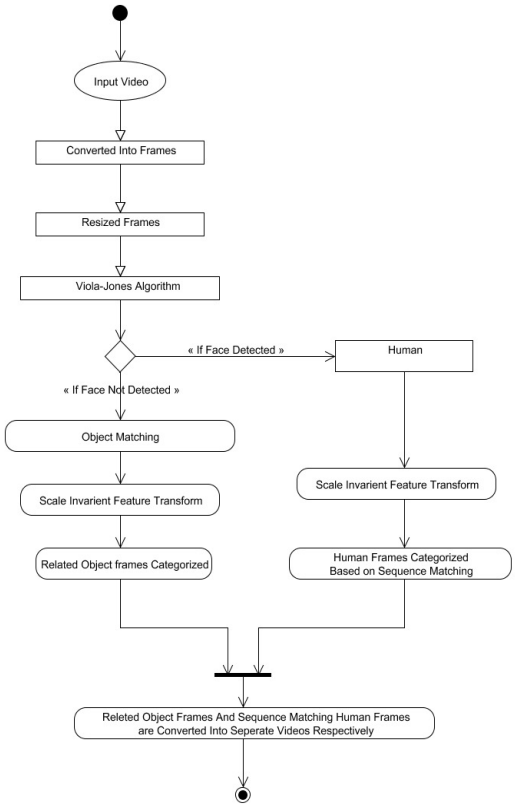
In video animation, transition is often accomplished by image morphing . The goal of morphing is to generate the in-between geometry which smoothly transforms the source shape into the target shape with interpolated texture smoothing. Morphing can produce appealing result for matched objects, but it may also cause the *ghost* phenomenon in transition for unmatched parts. This problem is even worse for our task since the video transition is based on partial matching.

A.IMPLEMENTATION DETAILS

For each video album, we first construct a bag-of-words(BoW) models using SIFT features for the PmH hashing[7]. The number of features per image ranges from 200 to 1000 , and we have quantized them using a visual word vocabulary with one million visual words. It take support 100ms per image to extract the features.

The min-hash contains 2sketches of size .each imageis divided into about 100 partitions with 50% overlap as recommended[7]. We uniformly sample 1/5 of the frames to accelerate the video similarity measurements.the frame similarities are accumulated to form video similarity. We then set a threshold to make the similarity graph sparse.for each connected video pairs, local matching-based sequence matching is performed.finally, extract matching frame pairs are located.

B.FLOW CHART



1)PRE PROCESSING

First Our Input short videos are converted into frames. . Then we eliminate some frames like information less frames (Mean of Input frame<15). we resize the each frames. Then all frames are merging into a single video for video categorization**.**

2) CATEGORIZATION BASED ON TRANSITION CLUE

Videos are categorized by using transition clues like human, object. Then we are taking human clue for first categorization by using Viola-Jones algorithm, if faces are not detected in frames that frames are separated into another process for object matching clue

.

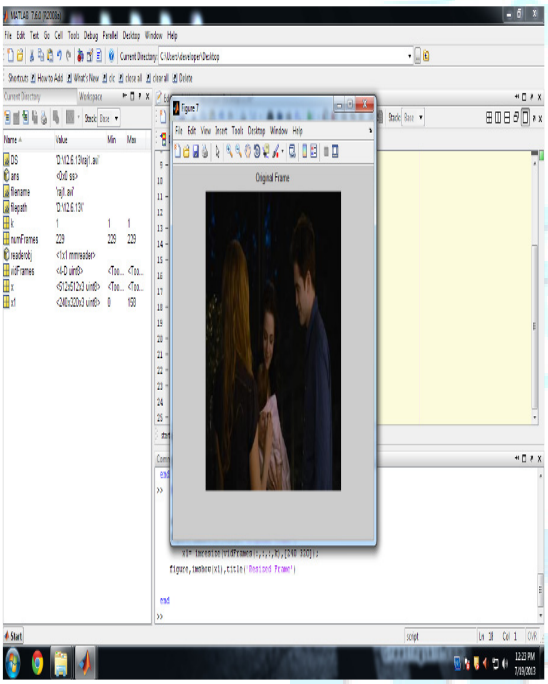
3) VIDEO COMPOSITION

Object & sequence matching process are done by using SIFT algorithm (**Scale-invariant feature transform**). Related Object frames and related sequence frames are categorized into separate folder respectively. Finally categorized frames are converted into Separate videos.

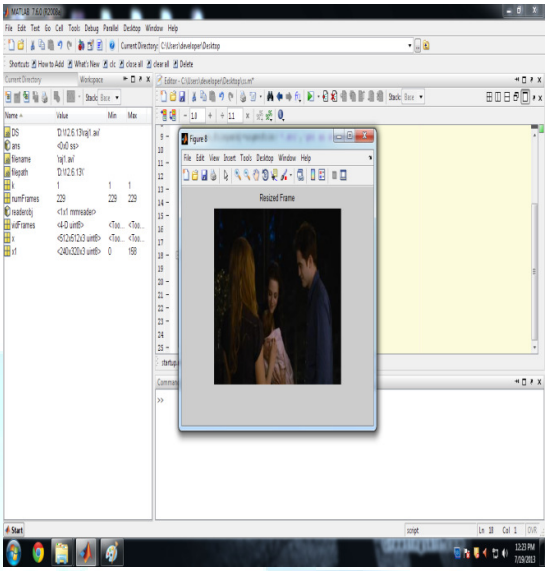
VII. CONCLUSIONS

In this paper, we proposed "Video Puzzle" using SIFT algorithm, an integrated system for both video summarization, browsing and presentation, based on large amount of personal and web video clips. This system automatically collects content-consistent video clips and generates an one-shot presentation using them. It can facilitate family album management and web video categorization. This system is used for improving matching accuracy and reduce the time cost.

VIII. RESULTS



In this figure, Our multiple Input short videos are converted into frames. . Then we eliminate some frames like information less frames.To get the single frames



In this figure, We resize the each frames. Then all frames are merging into a single video for video categorization.

REFERENCES

[1] C. Correa, "Dynamic video narratives," *ACM Trans. Graph.*, vol. 29, no. 4, Jul. 2010.

[2] J. Lee and J. Oh, "Scenario based dynamic video abstractions using graph matching," in *Proc. ACM MM*, 2005.

[3] J. Scharcanski and W. Gaviao, "Hierarchical summarization of diagnostic hysteroscopy videos," in *Proc. ICIP*, 2006

.

[4] E. Bennett, "Computational time-lapse video," *ACM Trans. Graph.*, vol. 26, no. 102, Jul. 2007.

[5] J. Calic, D. Gibson, and N. Campbell, "Efficient layout of comic-like video summaries," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 7, pp. 931–936, Jul. 2007

[6] P. Chiu, A. Girgensohn, and Q. Liu, "Stained-glass visualization for highly condensed video summaries," in *Proc. ICME*, 2004.

[7] G. Zhang, Z. Dong, J. Jia, and L.Wan, "Refilming with depth-inferred videos," *IEEE Trans. Visual. Comput. Graph.*, vol. 15, no. 5, pp.828–840,2009

[8] W.-Q. Yan, M. S. Kankanhalli, and J. Wang, "Analogies based video editing," *Multimedia Syst.*, vol. 11, no. 1, pp. 3–18, 2005.

[9] C. Barnes, D. Goldman, E. Shechtman, and A. Finkelstein, "Video tapestries with continuous temporal zoom," in *Proc. SIGGRAPH*, 2010.

[10] D. Lee and Q. Ke, "Partition min-hash for partial duplicate image discovery," in *Proc. ECCV*, 2010.

[11] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[12] I. Kemelmacher-Shlizerman, E. Shechtman, R.Garg, and S. Seitz, "Exploring photobios," in *Proc. SIGGRAPH*, 2011.

[13] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J.Comput. Vision*, vol. 88, pp. 303–338, 2010.