SnapVideo: Personalized Video Generation for a Sightseeing Trip

Luming Zhang, *Member, IEEE*, Peiguang Jing, *Member, IEEE*, Yuting Su, Chao Zhang, and Ling Shao, *Senior Member, IEEE*

***Abstract*—Leisure tourism is an indispensable activity in urban people's life. Due to the popularity of intelligent mobile devices, a large number of photos and videos are recorded during a trip. Therefore, the ability to vividly and interestingly display these media data is a useful technique. In this paper, we propose SnapVideo, a new method that intelligently converts a personal album describing of a trip into a comprehensive, aesthetically pleasing, and coherent video clip. The proposed framework contains three main components. The scenic spot identification model first personalizes the video clips based on multiple prespecified audience classes. We then search for some auxiliary related videos from YouTube[1] according to the selected photos. To comprehensively describe a scenery, the view generation module clusters the crawled video frames into a number of views. Finally, a probabilistic model is developed to fit the frames from multiple views into an aesthetically pleasing and coherent video clip, which optimally captures the semantics of a sightseeing trip. Extensive user studies demonstrated the competitiveness of our method from an aesthetic point of view. Moreover, quantitative analysis reflects that semantically important spots are well preserved in the final video clip.**

***Index Terms*—Aesthetic, comprehensive views, scenic spots, SnapVideo, video clips.**

I. Introduction

IN MODERN society, tourism has become a regular and indispensable leisure activity for many people. Owing to the popularity of mobile devices such as iPhone, users nowadays can conveniently record their sightseeing trips by photos. For example, the famous Flickr App offers users a simple and effective platform to upload, display, and share their travel shots. However, displaying a trip using photos may be less intuitive and user-friendly, due to the semantic gap between temporal neighboring photos and the incomplete views of some spots. For example, a user may only take photos of a scenic spot from the front, but there are also other Web photos or videos taken from the aerial. When the user or other audiences experience this trip, they may expect to view this spot from multiple views. In addition, compared with the discontinuous travel shots, integrating multiple views into a video clip will undoubtedly enhances the audiences' browsing experiences. Motivated by this, we utilize users' travel shots as the prior queries to generate a semantically smooth and aesthetically pleasing video clip, by leveraging a rich variety of Web videos. We name this technique SnapVideo, which is a useful technique that can facilitate a number of applications. For example, we can embed SnapVideo as a plugin into the Google photo to automatically convert the photos of each album into a video clip. This can greatly enhance the viewing experience of Google users. Besides, capturing long-time high definition videos might be too expensive for mobile photos, both in storage and the processing ability for video compression. Our SnapVideo technique can off-line convert a set of photos captured during the trip into a video clip. It is a useful technique for common travellers equipped with mobile phones with mediate/low storage and processing power.

We utilize users' travel shots as the prior queries to generate a semantically smooth and aesthetically pleasing video clip, by leveraging a rich variety of Web videos. We name this technique SnapVideo, which can be deemed to be an reverse operation of video summarization. Different from video summarization, however, SnapVideo is a more challenging task due to the following factors.

1) Travellers typically take photos from a few views, each of which reflects a specific viewing angle and depth-of-field. Describing a spot in incomplete views is less informative, since each view has a unique glamor. Besides, optimally determining an appropriate number of views is difficult.
2) Toward an aesthetically pleasing video clip, we have to quantify the aesthetics, representativeness, and informativeness of the source video frames. Moreover, it is difficult to sort the selected frames to ensure the overall smoothness of the resulting video clip.
3) Video quality evaluation is a subjective task. Audiences from different groups might have different opinions on

Manuscript received November 20, 2015; revised March 13, 2016; accepted June 25, 2016. Date of publication July 19, 2016; date of current version October 13, 2017. This work was supported in part by the Project from National Nature Science Foundation of China under Grant 61572169, and in part by the Fundamental Research Funds for the Central Universities. This paper was recommended by Associate Editor W. Hu. *(Luming Zhang and Peiguang Jing contributed equally to this work.)*

L. Zhang is with the Department of CSIE, Hefei University of Technology, Hefei 230009, China (e-mail: zglumg@zju.edu.cn).

P. Jing and Y. Su are with the School of Electronics and Information Engineering, Tianjin University, Tianjing 300072, China.

C. Zhang is with the Department of Computer Science, University of Illinois at Urbana–Champaign, Champaign, IL 61801, USA.

L. Shao is with the Department of Computer Science and Digital Technologies, Northumbria University, Newcastle upon Tyne, U.K.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2016.2585764

[1]https://www.youtube.com/

beautiful shots. As shown in the user study,[2] family members prefer videos with shots including the users themselves; strangers appreciate sceneries during the trip; while friends expect videos containing both the users and sceneries. That is to say, our SnapVideo technique should satisfy the taste of different types of audiences.

To handle the aforementioned challenges, a tri-stage framework is proposed for SnapVideo. First, a scenic spot identification model is designed to select a few referred photos as the prior queries, which are employed to collect external videos on Web. The selection process is audience-aware[3] where the audience types are prespecified. It theoretically guarantees that the taste of different audiences are maximally satisfied. Afterward, a comprehensive views generation model clusters the video keyframes into multiple views, where the optimal number of views is calculated automatically. Finally, a probabilistic model is proposed to optimally integrate the keyframes into an aesthetically pleasing, representative, and semantically coherent video clip. This module can be divided into two parts: 1) keyframes selection and 2) ranking. The former selects the aesthetically pleasing and representative keyframes, while the latter adjusts keyframes order to preserve the intraview and interview smoothness simultaneously.

The remainder of this paper is organized as follows. Section II briefly reviews the related work. Section III introduces the proposed VideoSnap, including scenic spot identification, comprehensive views generation, and probabilistic model for SnapVideo. Comprehensive experiments are conducted in Section IV. Section V concludes this paper.

II. Related Work

Our SnapVideo is closely related to the previous video summarization algorithms, which condense one or multiple video clips into a representative and succinct set of keyframes or skims. In the past few years, the summarization of a single video has been extensively studied in multimedia [64], and fairly satisfactory performance has been achieved. As a more complicated technique, multivideo summarization tries to fuse multiple videos taken from nearby places (e.g., prominent scenery spots in Singapore) into a condensed one. A detailed review of the previous of single video summarization techniques is beyond the scope of this paper. Interested readers can refer to Truong and Venkatesh [48] and Liu *et al.* [4]'s article to get an overview. Reference [48] provides a systematic classification of these works. For each approach, the authors identified and detailed the underlying components and how they are addressed in the specific work. Tompkin *et al.* [65] developed a system that analyzes unstructured but correlated videos to generate a Videoscape: a data structure that enables interactive exploration of video collections by visually navigating spatially and/or temporally between video clips. Wang and Merialdo [66] proposed an approach for multivideo summarization by exploring the redundancy between different videos. Zhu *et al.* [67] presented a hierarchical video content description and summarization strategy based on a joint semantic and visual similarity optimization. Cong *et al.* [68] formulated video summarization as a dictionary selection problem using sparsity consistency. Noticeably, different from the video summarization, our SnapVideo fits multiple travel photos into an aesthetically-pleasing, smooth and representative video clip. It can be roughly considered as a reverse operation of video summarization. Fu *et al.* [13] and Wei *et al.* [61] constructed a spatial-temporal shot graph and formulated the summarization problem as a graph labeling task. The spatial-temporal shot graph is derived from a hypergraph and encodes the correlations with different attributes among multivideo video shots. Then, the shot graph is partitioned into multiple clusters of event-centered shots with similar contents. The final summarization result is generated by solving a multiobjective optimization. Noticeably, Fu *et al.*'s [13] work is significantly different from this paper in the following aspects.

1) The hypergraph in Fu *et al.*'s [13] work describes both the spatial and temporal information, while the hypergraph in our method only clusters a large number of online crawled video frames into multiple views.
2) Fu *et al.*'s [13] method use a discriminative model to summarize multiple consumer videos, whereas our approach designs a probabilistic model to fit the selected representative keyframes into an aesthetically-pleasing, coherent, and smooth video clip. One potential benefit of our probabilistic model is its high extensibility, other factors such as video stability and exposure can be easily encoded into the model.
3) Fu *et al.*'s [13] work specifically focuses on a few scenic spots, such as office and campus. Comparatively, our proposed SnapVideos can handle travel photos with a rich variety of semantic types, due to the proposed robust probabilistic video fitting model.
4) Fu *et al.*'s [13] work condenses multiple videos into a set of video skims, whereas our approach fits a set of travel photos into an optimal video clip. The two operations are reverse practically.

This paper is also closely related to the aesthetic models for media quality evaluation. Yeh *et al.* [54] proposed a video quality assessment model consisting of two modules: aesthetic features construction and temporal integration. Seshadrinathan *et al.* [42] presented a subjective study of video quality using a collection of videos distorted by different application-relevant settings. Yeh *et al.* [54] proposed a personalized photo ranking system. It extracts the low-level visual features from top-ranked photos, where their weights are calculated based on ListNet [2]. Afterward, the photos are ranked based on the learned weights. Moorthy *et al.* [35] proposed a model which describes the aesthetic appeal of consumer videos and then classifies them into high/low aesthetic ones. They evaluated low-level features which are hierarchically integrated to model the aesthetics of consumer videos. Saini *et al.* [40] proposed an online video mashup algorithm, which allows for smooth shot transitions covering the performance from

[2]According to the experiences of ordinary life, family members prefer videos with shots including the users themselves; strangers appreciate sceneries during the trip; while friends expect videos containing both the users and sceneries.

[3]Audiences denote the friends, strangers and relatives to the two users whom collect the trip videos for our experiments.

multiple views. Herranz and Martínez [17] proposed to conveniently adapt the summary to a suitable length. The inherent analytic step adopts an iterative ranking algorithm where each summary is incrementally extended from the previous one. Cong *et al.* [7] formulated video summarization as a dictionary learning algorithm under a sparsity constraint. The objective is to train a dictionary where the original video can be reconstructed with the minimum error. Pongnumkul *et al.* [39] developed a system to enhance the experience of sharing and viewing tour videos, where frame-level salience features are extracted. With the popularity of mobile phones, activity cameras and Google glass, video capturing devices have become omni-present. Gygli *et al.* [62] presented a novel method for summarizing raw and casually-captured videos. The objective of this method is to create a short summary that conveys the story. The proposed method can generate both interesting and representative summary of the input video. Zhu *et al.* [63] proposed a novel multicamera joint video synopsis algorithm which is able to jointly produce multiple synopsis videos by preserving objects' chronological orders among camera views. The objective of this framework is primarily achieved by optimizing a chronological disorder cost function based on key time stamps of objects' appearing, merging, splitting and disappearing. Recently, Chang *et al.* [71] proposed a novel video-based event detection algorithm based on semantic saliency. It prioritizes the video shots using a notion of semantic saliency. Base on an isotonic regularizer, the authors designed a so-called "informed" nearly-isotonic SVM classifier that is able to exploit the carefully constructed ordering information. Ma *et al.* [58] proposed a multimedia event detection framework by leveraging few exemplars.

III. Proposed SnapVideo

This section elaborates the proposed SnapVideo framework. We first introduce the proposed scenic spots identification. Comprehensive views generation is described subsequently, including a preliminary of hypergraph, multiview clustering, optimization and adaptive inference of cluster number. The probabilistic model for SnapVideo is formulated in the last part.

A. Scenic Spot Identification

Generally, a user's travel album in Flickr covers photos of scenery and people, which were taken from a variety of scenic spots. Since the popularity of scenic spots is various, a scenic spot with high popularity is more likely to be attended. However, restricted by some factors, the numbers of times that observers directly watch a photo might be different. From the perspective of audiences, they prefer to watch a video clip with high views, attractive scenic spots, and satisfying their tastes. By encoding a photo with three attributes (correlation, popularity, and views), we proposed scenic spot identification to select referred photos corresponding to a few scenic spots as the prior queries. Afterward, we adopt the names of scenic spots into YouTube to crawl videos.

Venkatanathan *et al.* [49] categorized users into four groups (i.e., family, friends, colleagues, and strangers) to analyze their complex relationships. Inspired by this, when users sharing their travel photos, we divide the potential audiences into three social groups (i.e., family-group, friend-group, and stranger-group) according to their relationships with the traveller. Audiences from the same group are assumed to have a common preference. Given an album taken by a user, we conduct a survey with 30 potential audiences. Out of these 30 audiences, 10 are family members, 10 are friends, and 10 are strangers. It is observed that family members, such as father and mother, are likely to be attracted by users themselves. They tend to view photos containing portraits of the user instead of the pure sceneries. The audiences from the friend-group are interested in not only the user but also the beautiful sceneries. While audiences belonging to the stranger-group are likely to attend only those attractive scenery photos.

Based on the above observations, we characterize audiences' browsing preference by encoding a user's photo with three attributes: 1) correlation; 2) popularity; and 3) views. Correlation reflects whether the tourist is embedded in the given photo. It is a binary variable with 1 indicating embedded while 0 nonembedded. Popularity measures the frequency of a scenic spot involved the given photo, which is calculated by the number of Web pages indexed by Google. A higher popularity means that the photo is more likely to be added into the final video. Views reflect the times observers directly watch the user's photo. A photo with higher views tends to be more attractive. Therefore, given a collection of travel photos, we design a heuristic scheme to automatically select audience-preferred photos as queries by considering correlation, popularity, and views. We denote cor_x, p_x, and v_x as the values of correlation, popularity, and views of photo x, $v_{\max}$, and $p_{\max}$ as the maximum of popularity and views of photos. Then, the selection criteria can be described as follows.

1) For the family-group, we select photos with positive correlation, and rank them based on the popularity level and the number of views
$$f(x) = \text{cor}_x\left(\frac{1}{\phi_v}e^{\frac{v_x}{v_{\max}}} + \frac{1}{\phi_p}e^{\frac{p_x}{p_{\max}}}\right) \tag{1}$$
where ϕ_v and ϕ_p are normalized parameters. ϕ_v is determined by the largest value of $e^{vx/vmax}$, and similarly, ϕ_p is decided by the largest value of $e^{px/pmax}$.
2) For the stranger-group, we select photos with negative correlation, and rank them based on the popularity level and the number of views
$$f(x) = (1 - \text{cor}_x)\left(\frac{1}{\phi_v}e^{\frac{v_x}{v_{\max}}} + \frac{1}{\phi_p}e^{\frac{p_x}{p_{\max}}}\right). \tag{2}$$
3) For the friend-group, we rank photos by considering correlation, popularity, and views jointly
$$f(x) = \varphi\text{cor}_x + \frac{1}{\phi_v}e^{\frac{v_x}{v_{\max}}} + \frac{1}{\phi_p}e^{\frac{p_x}{p_{\max}}} \tag{3}$$
where $\varphi \in [0, 1]$ is a balance parameter. In our implementation, we tune the value of ϕ from 0 to 1 with a step of 0.01, and 0.33 is determined by cross-validation. Noticeably, the exponential functions in (1)–(3) have a smoothing effect, which weaken the discrepancy of the numbers of users viewed different photos.

We compute the total score per scenic spot and select the top three as the audience-aware queries. The corresponding photos of each scenic spot are treated as the audience-aware query photos.

To achieve the above goals, we acquire a collection of photos from a user's albums. Viola–Jones algorithm [50] is employed to detect human faces using Haar-like features [23], which is a scalar product between images and the Haar-like templates. Afterward, we remove faces with regions smaller than 5% of the photo, since there might be nonrelevant users appearing in photos and occupying a small part of the space. Based on the detected faces, we construct an undirected graph to recognize the user. The vertices correspond to faces, the edges represent the neighboring relationships, and the edge-weights measure the relationship between vertices. To find the underlying graph patterns, a dense subgraph detection algorithm [53] is employed to partition the graph into homogeneous subgraphs. The vertices in the same subgraph have high connections while those in different subgraphs have low connections. Generally, different subgraphs correspond to different persons and their sizes reflect the frequency of a person. Besides, the density of the subgraph denotes the closeness of the faces within the same group. This observation is based on that a person occurring in albums with high frequency indicates a closer connection with the user. Obviously, the largest subgraph should be adopted to represent the user. The person corresponding to the largest subgraph may not contain the user himself/herself, but is very likely that the person is one of his/her closest person. Based on this, given a set of photos of a user, we generate queries for different types of audiences. We use the scenery names corresponding to queries into YouTube to crawl a large number of videos for each scenic spot. These videos are employed to reconstruct a sightseeing trip.

B. Comprehensive Views Generation

This section describes the generation of comprehensive views for scenic spots using the retrieved videos. Here, a clustering scheme [59] is needed. A multiview clustering algorithm is proposed to categorize the crawled video frames, where each cluster represents a specific view of the scenic spot. We first overview the hypergraph theory. The proposed multiview clustering algorithm is elaborated subsequently. We then present an alternating algorithm to optimize the clustering objective function. Finally, we illustrate how to calculate the optimal number of clusters.

1) Preliminaries of Hypergraph: A hypergraph $G(V, E, \mathbf{W})$ consists of the vertex set V, the hyperedge set E, and the hyperedge weight $\mathbf{W}$. E is a family of hyperedges e connecting arbitrary subsets of V such that $\cup_{e\epsilon E} = V$, and each hyperedge e is assigned a positive weight $w(e)$. A hypergraph G can be represented by a $|V| \times |E|$ incidence matrix $\mathbf{H}$ whose entities are defined as

$$h(i,j) = \begin{cases} A(i,j) & \text{if } v_i \epsilon e_j \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where $A(i,j)$ denotes the probability of vertex v_i belonging to hyperedge e_j.

Based on $\mathbf{H}$, the vertex degree of $v_i \in V$ is calculated as

$$d(v_i) = \sum_{e_j \epsilon E} w(e_j) h(i,j). \tag{5}$$

The edge degree of hyperedge $e_j \in E$ is defined as

$$\delta(e_j) = \sum_{v_i \epsilon e_j} h(i,j). \tag{6}$$

We use $\mathbf{D}_v$, $\mathbf{D}_e$, and $\mathbf{W}$ to denote the diagonal matrices of vertex degrees, the hyperedge degrees, and the hyperedge weights, respectively.

Hypergraph is useful in a variety of machine learning tasks, e.g., classification [55], clustering [15], ranking [19], and embedding [60]. Typically, the regularizer $\Omega(\mathbf{f})$ on hypergraph is

$$\frac{1}{2} \sum_{e\epsilon E} \sum_{u,v\epsilon e} \frac{w(e)h(u,e)h(v,e)}{\delta(e)} \left(\frac{f(u)}{\sqrt{d(u)}} - \frac{f(v)}{\sqrt{d(v)}} \right)^2. \tag{7}$$

Let $\Theta = \mathbf{D}_v^{-(1/2)} \mathbf{H} \mathbf{W} \mathbf{D}_e^{-(1/2)} \mathbf{H}^T \mathbf{D}_v^{-(1/2)}$, we can further reorganize (7) into

$$\Omega(\mathbf{f}) = \mathbf{f}^T (\mathbf{I} - \Theta) \mathbf{f} = \mathbf{f}^T \mathbf{L} \mathbf{f} \tag{8}$$

where $\mathbf{I}$ is an identity matrix; $\mathbf{L} = \mathbf{I} - \Theta$ is a positive semi-definite matrix, i.e., the hypergraph Laplacian matrix.

In a hypergraph, each vertex is treated as a centroid and forms a hyperedge by circling around its k-nearest neighbors. For each type of feature, a hypergraph is constructed whose vertices represent the keyframes. It is noticeable that the hypergraph-based multiview modeling is more effective than the conventional binary graph. We can use a hyperedge to connect all the views related to an object, which is natural and intuitive. Comparatively, for conventional binary graph, this relationship will be described by a set of edges, which may be somewhat misleading.

2) Multiview Clustering: For each scenic spot, we first crawl a few top videos by issuing its name as query to YouTube. Following that, we extract keyframes and abandon the irrelevant ones from the crawled videos. These keyframes are then represented by H types of low-level features both globally and locally. Denoting $\{\mathbf{X}^k\}_{h=1}^H$ as a set of low-level visual features, where $\mathbf{X}^h = [\mathbf{x}_1^h, \ldots, \mathbf{x}_N^h]^T \in \mathbb{R}^{N \times d_h}$ is a feature matrix for the kth feature, and N is the total number of keyframes. Based on the hypergraph theory, the hypergraph Laplacian matrix for the hth type of feature is represented as $\mathbf{L}_h = \mathbf{I} - \Theta_h$, where $\Theta_h = \mathbf{D}_{vh}^{-(1/2)} \mathbf{H}_h \mathbf{W}_h \mathbf{D}_{eh}^{-(1/2)} \mathbf{H}_h^T \mathbf{D}_{vh}^{-(1/2)}$.

Fu *et al.* [13] studied the multiview video summarization [12], [13], and proposed a multiview metric learning framework that combines the advantage of maximum margin clustering with the disagreement minimization criterion. We follow this framework to discover a latent feature space $\mathbf{X}$ by minimizing the following objective function:

$$R(\mathbf{X}) = R_{\text{stru}}(\mathbf{X}) + \gamma R_{\text{diff}}(\mathbf{X}) \tag{9}$$

where $R_{\text{stru}}(\mathbf{X})$ and $R_{\text{diff}}(\mathbf{X})$ are the structural and disagreement losses of $\mathbf{X}$, respectively; γ is a weight controlling the tradeoff between the structural loss and disagreement loss.

Spectral graph [6], [51] clarifies that a v-clustering is determined by the first v smallest eigenvalues of $\mathbf{L}$. It is the normalized Laplacian matrix of the similarity matrix of $\mathbf{X}$. In our implementation, we assume that $\mathbf{L}$ is the latent hypergraph Laplacian matrix. Thereby, the structural loss can be formulated as

$$R_{\text{stru}}(\mathbf{L}) = \frac{\sum_{i=1}^{v}\lambda_i}{\sum_{i=1}^{N}\lambda_i} = \frac{1}{tr(\mathbf{L})}\sum_{i=1}^{v}\lambda_i \tag{10}$$

where $\lambda_1 \leq \lambda_2 \cdots \leq \lambda_N$ are the eigenvalues of $\mathbf{L}$ and v is the expected number of clusters.

The disagreement loss $R_{\text{diff}}(\mathbf{X})$ measures the difference between $\mathbf{X}$ and all $\mathbf{X}_h$, and we define it as follows:

$$R_{\text{diff}}(\mathbf{X}) = \sum_{i=1}^{h} S(\mathbf{L}, \mathbf{L}_h) = \sum_{i=1}^{h} \mathbf{O}\mathbf{O}_h\mathbf{O}_h^T\mathbf{O} \tag{11}$$

where $\mathbf{O}, \mathbf{O}_h \in \mathbb{R}^{N\times p}$ is the orthogonal matrix of $\mathbf{L}$ and $\mathbf{L}_h$, respectively.

We define the orthogonal subspace of $\mathbf{L}_h$ as $\mathbf{O}_h \in R^{N\times p}$, which is a point on the Grassmannian manifold $G(N, p)$ [14], [16]. The Grassmannian manifold $G(N, p)$ is a set of N-D linear subspaces of $\mathbb{R}^p$. A typical way to handle orthogonal subspaces is to formulate them on the Grassmannian manifold. The orthogonal subspace $\mathbf{O}_h$ can be easily obtained by calculating the eigen-decomposition of a positive semi-definite matrix $\mathbf{L}_h$.

The similarity between points on the Grassmannian manifold is computed through the projection kernel. Let $\mathbf{Y}_1$ and $\mathbf{Y}_2$ be two orthonormal matrices of size $D\times m$, then the projection kernel can be defined as

$$K_P(\mathbf{Y}_1, \mathbf{Y}_2) = \left\|\mathbf{Y}_1^T\mathbf{Y}_2\right\|_F^2. \tag{12}$$

The consistency between the latent Laplacian subspace and the hth type of Laplacian subspace is calculated based on the kernel function [70].

The disagreement loss $R_{\text{diff}}(\mathbf{X})$ measures the difference between $\mathbf{X}$ and $\mathbf{X}^h$. By combining the above definitions, we reorganize the objective function (9) as

$$\min_{\mathbf{L},\mathbf{O}} \frac{1}{tr(\mathbf{L})}\sum_{i=1}^{v}\lambda_i - \gamma\sum_{h=1}^{H} tr\left(\mathbf{O}^T\mathbf{O}_h\mathbf{O}_h^T\mathbf{O}\right). \tag{13}$$

3) Optimization: We propose an alternative algorithm to solve (13). Once $\mathbf{L}$ is calculated, we can derive the first v eigenvectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_v$ of $\mathbf{L}$, then K-means is employed to cluster all points into v clusters. In other word, we just have to compute $\mathbf{L}$. For a series of hypergraph Laplacian matrix $\mathbf{L}^h$, denoting $\tilde{\mathbf{\Theta}} = (1/H)\sum_{h=1}^{H}\Theta_h$ and $\tilde{\mathbf{L}} = (1/H)\sum_{h=1}^{H}\mathbf{L}_h$, it can be derived that $\mathbf{\Theta} = \mathbf{I} - (1/H)\sum_{h=1}^{H}\Theta_{\mathbf{h}} = \mathbf{I} - \tilde{\mathbf{\Theta}}$. A matrix $\mathbf{W} \in \mathbb{R}^{N\times N}$ is defined to guarantee the symmetric positive semi-definite of $\mathbf{L}$. Then, matrix $\mathbf{L}$ can be written as $\mathbf{L} = \mathbf{W}^T\tilde{\mathbf{L}}\mathbf{W}$.

Since $\mathbf{O}$ is the orthogonal matrix solved using the eigen-decomposition of $\mathbf{L}$, we can obtain a set of eigenvalues $\{\lambda_i\}_{i=1}^{v}$ accordingly.

Toward an efficient solution, we impose an auxiliary constraint on the columns of $\mathbf{W}$ and $\mathbf{O}$, i.e., $\mathbf{W}^T\mathbf{W} = \mathbf{I}$ and $\mathbf{O}^T\mathbf{O} = \mathbf{I}$. The optimization then becomes

$$\begin{aligned}\min_{\mathbf{W},\mathbf{O}}\ & tr\left(\mathbf{O}^T\mathbf{W}^T\tilde{\mathbf{L}}\mathbf{W}\mathbf{O}\right) \\ & - \gamma\, tr\left(\mathbf{W}^T\tilde{\mathbf{L}}\mathbf{W}\right)\sum_{h=1}^{H} tr\left(\mathbf{O}^T\mathbf{O}_h\mathbf{O}_h^T\mathbf{O}\right) \\ \text{s.t.}\ & \mathbf{W}^T\mathbf{W} = \mathbf{I}, \mathbf{O}^T\mathbf{O} = \mathbf{I}.\end{aligned} \tag{14}$$

Minimizing the above objective function is based on the Lagrangian multipliers and eigen-decomposition. We introduce an alternating descent algorithm. First, we fix $\mathbf{W}$ and then the partial derivative of the objective function with respect to $\mathbf{O}$ is calculated as

$$\left(\mathbf{W}^T\tilde{\mathbf{L}}\mathbf{W} - \gamma tr\left(\mathbf{W}^T\tilde{\mathbf{L}}\mathbf{W}\right)\sum_{h=1}^{H}\mathbf{O}_h\mathbf{O}_h^T\right)\mathbf{o}_i = \beta_i\mathbf{o}_i \tag{15}$$

where $\mathbf{o}_i$ is the generalized eigenvector of matrix $\mathbf{W}^{-1}\tilde{\mathbf{L}}\mathbf{W} - \gamma(\mathbf{W}^T\mathbf{W})^{-1}\sum_{h=1}^{H}\mathbf{O}_h\mathbf{O}_h^T$ and β_i is the corresponding eigenvalue. $\mathbf{o}_i$ and β_i can be solved via the eigen-decomposition of $\mathbf{T}$.

We then fix matrix $\mathbf{O}$ and obtain the partial derivative of the objective function with respect to $\mathbf{W}$

$$\left(\tilde{\mathbf{L}} - \gamma\sum_{h=1}^{H} tr\left(\mathbf{O}^T\mathbf{O}_h\mathbf{O}_h^T\mathbf{O}\right)\tilde{\mathbf{L}}\right)\mathbf{W}\mathbf{O} = \mathbf{W}\mathbf{O}\Phi \tag{16}$$

where Φ is the diagonal matrix constituted by the eigenvalue of matrix $\tilde{\mathbf{L}} - \gamma\sum_{h=1}^{H} tr(\mathbf{O}^T\mathbf{O}_h\mathbf{O}_h^T\mathbf{O})\tilde{\mathbf{L}}$.

By defining $\acute{\mathbf{L}} = \tilde{\mathbf{L}} - \gamma\sum_{h=1}^{H} tr(\mathbf{O}^T\mathbf{O}_h\mathbf{O}_h^T\mathbf{O})\tilde{\mathbf{L}}$, we perform the eigen-decomposition of matrix $\acute{\mathbf{L}}$, and obtain $\acute{\mathbf{L}}\mathbf{U} = \mathbf{U}\Lambda$. Λ is the diagonal matrix formed by the eigenvalues of matrix $\tilde{\mathbf{L}}$, and $\mathbf{W}\mathbf{O}$ equals to the matrix $\mathbf{U}$ whose columns correspond to the eigenvectors of $\tilde{\mathbf{L}}$. Since $\mathbf{U}$ and $\mathbf{O}$ are orthogonal, performing QR decomposition will produce a common upper triangular matrix $\mathbf{R}$, unitary matrix $\tilde{\mathbf{Q}}$, and $\tilde{\mathbf{O}}$

$$\mathbf{U} = \tilde{\mathbf{Q}}R,\ \ \mathbf{O} = \tilde{\mathbf{O}}R. \tag{17}$$

Afterward, matrix $\mathbf{W} = \tilde{\mathbf{Q}}^{-1}\tilde{\mathbf{O}}$ is obtained. The pipeline of the above procedure is elaborated in Algorithm 1.

Algorithm 1 Optimization of Multiview Clustering

1) Compute the mean matrix of $\mathbf{L}_h$: $\tilde{\mathbf{L}} = \frac{1}{H}\sum_{h=1}^{H}\mathbf{L}_h$;
2) Initialize a random matrix $\mathbf{W}$;
3) Iterate steps 4)$\sim$ 7) until convergence:
4) Find matrix $\mathbf{O}$ by applying eigen-decomposition to the matrix $\mathbf{T} = \mathbf{W}^T\tilde{\mathbf{L}}\mathbf{W} - \gamma\cdot tr\left(\mathbf{W}^T\tilde{\mathbf{L}}\mathbf{W}\right)\sum_{h=1}^{H}\mathbf{O}_h\mathbf{O}_h^T$;
5) Calculate matrix $\mathbf{WO}$ by conducting eigen-decomposition to $\tilde{\mathbf{L}}\mathbf{WO} = \mathbf{WO}\Lambda$;
6) Calculate matrix $\tilde{\mathbf{Q}}$ and $\tilde{\mathbf{O}}$ by applying QR decomposition to $\mathbf{U}$ and $\mathbf{O}$;
7) $\mathbf{W} = \tilde{\mathbf{Q}}^{-1}\tilde{\mathbf{O}}$ is obtained.
8) Output matrix $\mathbf{L}$

4) Adaptive Inference of Cluster Number: Choosing an optimal v is important in many clustering applications. A small v cannot represent scenic spots from diverse views, while a large v may result in an incoherent video clip.

In this paper, the structural loss measures the quality of a v-clustering by the ratio of the first v smallest eigenvalues to the sum of all eigenvalues. Besides, maximizing the ratio

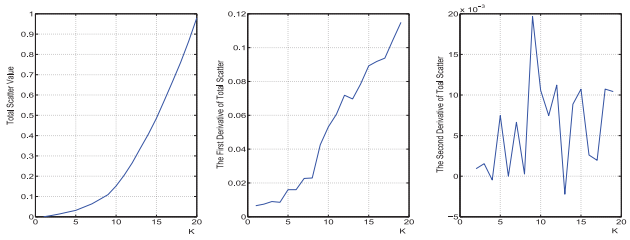


Fig. 1. Process of determining the number of clusters on the photos from the Merlion on Marina Bay scenic spot.

of intercluster scatter to that of intracluster scatter is another criterion to guide the clustering [36]. Assuming that we have K clusters, denoting $\mathbf{c}_k (1 \le k \le K)$, where $\mathbf{c}_k = \{c_k^1, c_k^2, \ldots, c_k^{n^k}\}$. We use $\mathbf{g}(c_k^i) \in R^d$ to represent the visual feature of the i samples in the kth cluster, then the intracluster scatter $S_{\text{intra}}(K)$ and the intercluster scatter $S_{\text{inter}}(K)$ is represented as

$$\begin{cases} S_{\text{intra}}(K) = \dfrac{1}{\sum_{k=1}^{K} n^k} \sum_{k=1}^{K} \sum_{i=1}^{n^k} \left\| \mathbf{g}(c_k^i) - \boldsymbol{\mu}_k \right\|^2 \\ S_{\text{inter}}(K) = \dfrac{1}{K(K-1)} \sum_{i=1}^{K} \sum_{j=1}^{K} \left\| \boldsymbol{\mu}_i - \boldsymbol{\mu}_j \right\|^2 \end{cases} \tag{18}$$

where $\boldsymbol{\mu}_i$ is the mean vector of ith cluster.

To measure the quality of the K clusters, we define the total scatter function $F(K)$ by linearly combining the structural loss and the ratio of intercluster scatter to intracluster scatter

$$F(K) = R_{\text{struct}}(K) + \beta \frac{S_{\text{intra}}(K)}{S_{\text{inter}}(K)} \tag{19}$$

where $R_{\text{struct}}(K)$ is the value of structural loss for the K clusters. $0 \le \beta \le 1$ is a weight controlling the tradeoff between two scatter measurement.

In general, the second order derivative reflects the changing rate of first order derivative. For a curve, the second derivative corresponds to the curvature or concavity. The curve of a function with a positive second order derivative curve upward, while that with a negative second order derivative curve downward. We choose the point with the largest second order derivative as the optimal cluster number. Fig. 1 elaborates the calculation of the cluster number on the photos from the "Merlion on Marina Bay" scenic spot, where $v = 9$.

C. Probabilistic Model for SnapVideo

Our proposed SnapVideo is a semantically-aware framework. In the scenic spot identification and comprehensive view generation steps, users can flexibly encode various video semantics, such as assigning audience types and clustering frames into multiple prespecified views. After encoding semantics, the smoothness of frames is achieved by the interview and intraview smoothness as elaborated in the keyframes ranking step in the following.

1) Keyframes Selection: Inspired by Zhang *et al.* [56], the aesthetics of keyframes can be quantified by how much aesthetic cues from beautiful photos $\{p_1, p_2, \ldots, p_M\}$ can be transferred into a keyframe. In this paper, the beautiful photos come from Zhejiang University dataset [43], which contains 6000 aesthetically pleasing photos crawled from Photosig.[4]

[4]http://www.photosig.com

For the ith cluster $\mathbf{c}_i = \{c_i^1, c_i^2, \ldots, c_i^{n^i}\}$, we suppose that c is a random sample belonging to $\mathbf{c}_i$. The aesthetics of keyframe c can then be calculated as

$$Q_{\text{aes}}(c) = P(c|p_1, p_2, \ldots, p_M) = \prod_{m=1}^{M} p(c|p_m) \tag{20}$$

where $p(c|p_m) \propto \exp(-(\|\mathbf{g}(c) - \mathbf{g}(p_m)\|/\sigma_{ae})^2)$ is defined as a Gaussian kernel and $\mathbf{g}(c)$ and $\mathbf{g}(p_m)$ denote the low-level visual feature vector of c and p_m, respectively.

We also take the representativeness of keyframes into account, which measures the centroid of a keyframe in each cluster. It can be computed by the kernel density estimation [44]

$$Q_{\text{rep}}(c) = P(c|\mathbf{c}_i) = \frac{1}{n^i} \sum_{j=1}^{n^i} F(\mathbf{g}(c) - \mathbf{g}(c_i)) \tag{21}$$

where $F(\mathbf{x})$ is a kernel function satisfying $F(\mathbf{x}) > 0$ and $\int F(\mathbf{x}) d\mathbf{x} = 1$. In this paper, we adopt the exponential kernel, i.e., $F(\mathbf{x}) \propto \exp(-\|\mathbf{x}\|^2/\sigma_{\text{rep}})$. The above equation can be explained as: c_i and each of its cluster member can be considered as a family and its members, respectively. Then, the closeness of a sample to the family is estimated by averaging the soft voting of all the family members.

We use informativeness to measure the popularity of keyframes. It reflects the number of near-duplicate keyframes in a cluster. We construct a graph by interlinking the keyframes with visual similarity. The graph is constructed where each vertex is linked to all other vertices except for those within the same video. After abandoning the edges below the detection threshold [18], [57], the graphs will be segmented into multiple keyframe groups. Assuming that $\{g_n, n = 1, 2, \ldots, G\}$ is a set of groups after graph segmentation, and keyframe c belongs to group g. The informativeness of a keyframe c is quantified as

$$Q_{\text{inf}}(c) = \frac{\log|b| - \min(\log|b_n|)}{\max(\log|b_n|) - \min(\log|b_n|)} \tag{22}$$

where $|b_n|$ denotes the number of near-duplicate keyframes within each group b_n.

Toward a set of aesthetically pleasing, representative and informative keyframes, we combine the above two measurements

$$\begin{aligned} c^* &= \arg\max_{c} Q(c) \\ &= \arg\max_{c} Q_{\text{aes}}(c) \cdot Q_{\text{rep}}(c) \cdot Q_{\text{inf}}(c). \end{aligned} \tag{23}$$

By optimizing the above objective function, we select the top t keyframes with the highest scores for each cluster.

2) Keyframes Ranking: Now, we quantify the smoothness of a video clip. Smoothness penalizes the sudden changes between neighboring keyframes and enhances the audiences' browsing experiences. Toward a smooth video clip, keyframes should be ranked appropriately by taking two factors into consideration: the intraview and interview smoothness. The former ensures smoothness between keyframes within the same view, while the latter transits the spot view progressively.

After the keyframes selection, we obtain a new sequence $\tilde{\mathbf{c}}_i = \{\tilde{c}_i^1, \tilde{c}_i^2, \ldots, \tilde{c}_i^t\}$ for the ith cluster. The intraview smoothness is defined by accumulating the differences of keyframes

$$Q_{\text{intra}}(\mathbf{c}) = \prod_{i=1}^{K} P_{\text{intra}}(\tilde{\mathbf{c}}_i) = \prod_{i=1}^{K}\prod_{j=1}^{t-1} \exp\left(-\frac{\left\|\mathbf{g}\left(\tilde{c}_i^{j+1}\right) - \mathbf{g}\left(\tilde{c}_i^{j}\right)\right\|^2}{\sigma_{\text{smo}}}\right). \quad (24)$$

The interview smoothness ensures a tolerable discrepancy between neighboring views

$$Q_{\text{inter}}(\tilde{\mathbf{c}}) = P\left(\theta(\tilde{\mathbf{c}}_1, \tilde{\mathbf{c}}_2), \theta(\tilde{\mathbf{c}}_2, \tilde{\mathbf{c}}_3), \ldots, \theta(\tilde{\mathbf{c}}_{K-1}, \tilde{\mathbf{c}}_K)\right) = \prod_{i=1}^{K-1} P_{\text{inter}}\left(\theta(\tilde{\mathbf{c}}_i, \tilde{\mathbf{c}}_{i+1})\right) = \prod_{i=1}^{K-1} \exp\left(-\frac{\left\|\mathbf{g}(\tilde{c}_i^t) - \mathbf{g}(\tilde{c}_{i+1}^1)\right\|^2}{\sigma_{\text{smo}}}\right) \quad (25)$$

where $p_{\text{inter}}(\theta(\tilde{\mathbf{c}}_i, \tilde{\mathbf{c}}_{i+1}))$ measures the smoothness between $\tilde{\mathbf{c}}_i$ and $\tilde{\mathbf{c}}_{i+1}$, which is defined as a Gaussian kernel; $\tilde{\mathbf{c}}_i^t$ is the last frame of a sequence $\tilde{\mathbf{c}}_i$ and $\tilde{\mathbf{c}}_{i+1}^1$ is the first frame of sequence $\tilde{\mathbf{c}}_{i+1}$. Noticeably, the parameters σ_{aes}, σ_{rep}, and σ_{smo} are all learned from the training videos through a Gaussian estimation.

We combine the above two smoothness measurements to obtain an objective function as

$$\tilde{\mathbf{c}}^* = \arg\max_{\tilde{\mathbf{c}}} Q(\tilde{\mathbf{c}}) = \arg\max_{\tilde{\mathbf{c}}} Q_{\text{inter}}(\tilde{\mathbf{c}}) \cdot Q_{\text{intra}}(\tilde{\mathbf{c}}). \quad (26)$$

The above objective function has no analytic solution, we employ a heuristic approach for acceleration. We first fix the sequence of scenic spots according to the photo taking time. Then, we adjust the view order followed by the order of keyframes within each view. Based on the ranking result, a video clip describing a scenic spot is generated[5] and obtain the final video clips. Therefore, the temporal and geographical information is automatically encoded in the video generation process. The pipeline of fitting frames from multiple views into a video clip is shown in Fig. 2.

IV. Experimental Results and Analysis

In this section, we first briefly introduce the compilation of our dataset. Then, we evaluate the effectiveness of proposed method using four experiments. We first show the feasibility of the audience-aware query generation model. The second experiment validates our proposed clustering algorithm by analyzing its convergence and comparative studies. We testify the necessity and effectiveness of the keyframes selection and ranking subsequently. Comparisons with other video summarization algorithms are conducted in the last experiment.

[5]Users' album in Flickr often contains data describing where and when shots were taken. According to the time stamp, we to obtain the tour routes of a user. We rank these generated video clips of scenic spots in accordance with user's tour route.

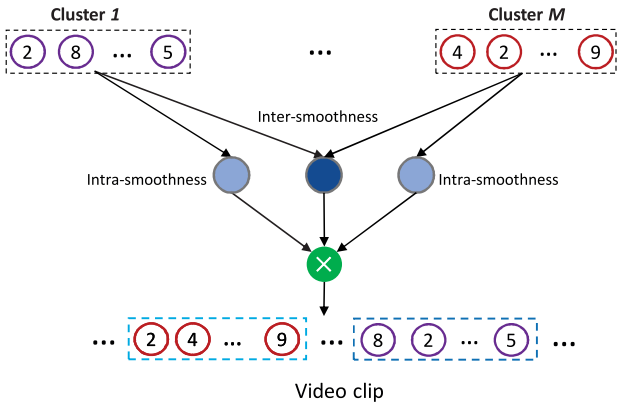


Fig. 2. Probabilistic model for SnapVideo (aes: aesthetic, rep: representativeness, inf: informativeness, inter: intersmoothness, and intra: intrasmoothness).

A. Experimental Settings

1) Data Collection and Feature Extraction: As far as we know, there is no publicly available dataset to evaluate our proposed SnapVideo. We manually collect a dataset to verify our approach. It contains two subdatasets: 1) the index-photo dataset and 2) the video dataset. We built the index-photo dataset based on two users' travel experiences, one male and one female. The male and female user took the photos in Singapore and Beijing, respectively. They created albums on Flickr to share their trip photos with others. The Beijing and Singapore trips contain 46 and 243 photos, respectively.

We downloaded their albums and constructed the index-photo dataset. The exact building or location name of the scenic spot is straightforward to acquire, as it is described by the photo title. Table I summarizes the scenic spots in users' albums. In total, our video dataset contains 11 scenic spots of Beijing and 16 scenic spots of Singapore. We first treat the name of each scenic spot as a query, which is then fed into the YouTube search engine. Afterward, the videos of the first three searching pages are collected. By eliminating those very long and duplicate videos, we obtain 468 videos for Beijing and 524 videos for Singapore, respectively. By abandoning those low-resolution and irrelevant videos, we finally obtain 182 videos with 61 515 keyframes for Beijing and 168 videos with 42 069 keyframes for Singapore.

Five types of visual features are extracted to describe each video frame. They are: 1) 225-D block-wise color moments extracted over 5×5 sized grid partitions [46]; 2) 72-D HSV color histogram [24]; 3) 72-D edge direction histogram [21]; 4) 59-D local binary pattern [47]; and 5) 200-D bag-of-words [22] based on SIFT [8] descriptors. Afterward, we combine these features into a 628-D feature vector. To avoid the curse of dimensionality [20], PCA [45] is employed to convert the original feature vector to 120-D. The rest of parameters are: $\beta = 0.02$, $t = 5$, and $\gamma = 0.0025$.

Besides the above five features to represent each video frame, in the system implementation stage, we also employ several popular visual features to represent each video frame, i.e., RGB-domain spin image, color correlogram, wavelet texture, PCA mask, CTM descriptor, PHOG descriptor, region

TABLE I
LIST INCLUDING 11 SCENIC SPOTS FROM BEIJING AND 16 SCENIC SPOTS FROM SINGAPORE

Places	Scenic Spots
Beijing	Changling Tomb of the Ming Tombs, Drum Tower and Bell Tower, St Joseph's Church, Temple of Heaven, The Forbidden City, The Great Hall of the People, The Great Wall, The monument to the People's heroes, The Peninsula Palace Hotel, The Summer palace, Tiananmen
Singapore	Buddha Tooth Relic Temple, Cavenagh Bridge, Esplanade-Theatre on the Bay, Fort Siloso, Fountain of Wealth, St.Andrew's Cathedral, Raffles Hotel, Clarke Quay, Singapore Flyer, Singapore Parliament building, Raffles' Landing Sitbuhler2009spectrale, Tanjong Pagar Train Station, Sentosa Merlion, The Fullerton Hotel, Marina Bay Merlion, War Memorial

covariance, PCA V1plus, and Gist descriptor. All these features are briefed by Zhang *et al.* [69].We first use all these features to represent each video frame. Afterward, we remove those unimportant features one-by-one via cross-validation. Finally, the five most effective features are adopted.

2) Evaluation Setups: User study is a typical approach for subjective video quality evaluation. We invited 12 volunteers to our user study and reported the statistics of their votes. All the volunteers are master/Ph.D. students aged between 21 and 26, majoring in computer sciences. Half of them are female. Audience's subjective preference to video clips are quantified by a score ranging from 0 to 5 (5, 4, 3, 2, 1, and 0, respectively, stands for excellent, good, general, bad, worse, and worst). To facilitate evaluation, the length of video clips are uniformly reduced to less than 3 min. Each volunteer acts as a family, friend, or stranger of the users and rates the video clip. The final score of a video clip is an average of the 12 volunteers' rating scores.

In addition, *P*-value [11], [25] is adopted to evaluate the statistic significance against the null hypothesis, i.e., hypothesis without change. If the P-value is smaller than 0.05, it reflects that the observed data is inconsistent with null hypothesis. A smaller *P*-value corresponds to a greater significance against the null hypothesis. tenfold validations are conducted between each of the baseline and our method.

B. Experimental Results and Discussions

1) Evaluation of Multichannel Features of Each Video Frame: In this section, we testify the visual features adopted in our framework. We remove one of the five visual features and then report the corresponding two important metrics: 1) Rs: structural loss and 2) Sw/Sb: the ratio of within-class scatter and between-class scatter. The lower values of the above two metrics reflect a better multiview clustering performance. As shown in Fig. 3, both the structural loss and the Sw/Sb increase when abandoning one visual feature, which clearly demonstrates the inseparability of the five visual features. In addition, clustering time consumption increases slightly when using all the five features, reflecting that the high efficiency of the features extraction.

2) Audience-Aware Query Generation: In this section, we validate the audience-aware query generation module for Beijing and Singapore. It outputs different video clips for matching the three types of audiences' requirements. To demonstrate the necessity of this module, we replace it by randomly selecting three scenic spots of the travel photos as queries to generate the video clips. This method does not consider audience preferences. Thus, we ask all the volunteers to score the three videos for Beijing and Singapore. The average scores and the variation of volunteers' rating scores are reported in Table II. "Rand" denotes random sampling and "Beijing+Singapore" means the averaged results. From Table II, we observe that our method outperforms random sampling (RS) by a higher mean score and a smaller variance. It reflects that our personalized strategy can substantially enhance audiences watching experiences.

	Beijing		Singapore		Mean		
	Rs	Sw/sb	Rs	Sw/Sb	Rs	Sw/Sb	Times (s)
Non-HSV	0.0058	0.2388	0.0057	0.2012	0.0058	0.2162	10.26s
Non-LBP	0.0060	0.1933	0.0059	0.2255	0.0060	0.2126	9.67s
Non-EDH	**0.0056**	0.2688	0.0058	0.1449	0.0057	0.1940	9.62s
Non-BCM	0.0057	0.2154	0.0061	0.1566	0.0059	0.18014	9.69s
Non-BOW	0.0060	0.2764	0.0062	0.1369	0.0061	0.1927	9.73s
Our	**0.0056**	0.1603	**0.0057**	**0.1443**	**0.0056**	**0.1509**	10.5s

Fig. 3. Rs, Sw/Sb, and the clustering time consumption when abandoning one type of visual features.

TABLE II
AUDIENCE-BASED EVALUATION ON AUDIENCE-AWARE QUERY GENERATION

	Beijing	Singapore	Beijing+Singapore	P-value
Rand	2.90±0.48	2.92±1.11	2.91±0.75	< 0.05
Ours	**3.61±0.24**	**3.81±0.70**	**3.71±0.28**	**––**

3) Comprehensive Views Generation: In this section, we demonstrate the convergence and effectiveness of our proposed clustering. Based on the above analysis, the iterative criteria are essential to guarantee the convergency of objective function (14). We test the convergency of (14) based on our proposed alternating algorithm and randomly select four trials to report the results. Fig. 4 displays the variation of objective function value during the iteration. As can be seen, the value decreases rapidly by increasing iteration number. For the "Fountain of Wealth" and the "The Summer Palace," our algorithm converges after about 30 iterations. For the "St. Joseph's Church" and the "Temple of Heaven," the algorithm converges much faster.

To further evaluate the effectiveness of our proposed clustering, we replace it by five well-known clustering algorithms: RS, *k*-means clustering, spectral clustering (SP) [51], spectral

TABLE III
AUDIENCE-BASED EVALUATION ON COMPREHENSIVE VIEWS GENERATION. (FAM: FAMILY, FRI: FRIENDS, STR: STRANGERS, AVE: AVERAGE, AND VAR: VARIANCE)

	Beijing				Singapore				Beijing+Singapore				
	Fam	Fri	Str	Ave	Fam	Fri	Str	Ave	Fam	Fri	Str	Ave	P-value
RS	2.58	2.75	3.33	2.89±0.31	2.50	2.92	3.42	2.94±0.72	2.54	2.83	3.38	2.92±0.39	< 0.05
k-means	2.59	3.08	**4.00**	3.22±0.77	2.67	3.17	3.92	3.25±0.87	2.63	3.13	3.96	3.24±0.78	< 0.05
LSC	2.83	3.08	**4.00**	3.31±0.21	2.75	2.83	4.08	3.22±0.23	2.79	2.96	4.04	3.26±0.16	< 0.05
PSP	3.08	3.33	3.92	3.44±0.23	2.83	3.17	3.92	3.31±0.19	2.96	3.25	3.92	3.37±0.15	< 0.05
SP	2.92	3.25	3.92	3.36±0.09	3.00	3.33	3.75	3.36±0.21	2.96	3.29	3.83	3.36±0.07	< 0.05
Ours	**3.17**	**3.67**	**4.00**	**3.61±0.24**	**3.25**	**3.75**	**4.42**	**3.81±0.70**	**3.21**	**3.71**	**4.21**	**3.71±0.28**	– –

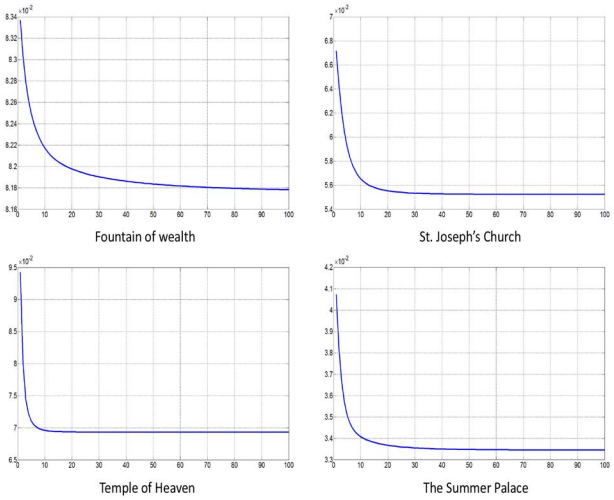


Fig. 4. Convergence curve on the four scenery spots. (Fountain of Wealth, St. Joseph's Church, Temple of Heaven, and The Summer Palace.)

clustering based on the graph p-Laplacian (PSP) [1], and the landmark-based spectral clustering (LSC) [5].

1) *RS:* We replace our proposed clustering by randomly selecting keyframes for the corresponding scenic spots.
2) *k-Means:* Partitioning the keyframes into k clusters, each of which belongs to the cluster with the nearest mean.
3) *SP:* A typical clustering based on eigen-decomposition of the Laplacian matrix corresponding to the sample graph.
4) *PSP:* A general version of spectral clustering based on the second eigenvector of graph p-Laplacian, which can be deemed to be a nonlinear generalization of the graph Laplacian.
5) *LSC:* Representative data points are selected as the landmarks, where the original data points are approximated by its linear sparse combination. The spectral embedding is computed using the landmark-based representation efficiently.

The above clustering algorithms cannot calculate the optimal cluster number. Thus, we uniformly set the cluster number to 9 and select five keyframes for each cluster using our probabilistic model. In total, we obtain 45 keyframes to describe a scenic spot. For RS, we select 45 keyframes for a scenic spot to generate the video clips.

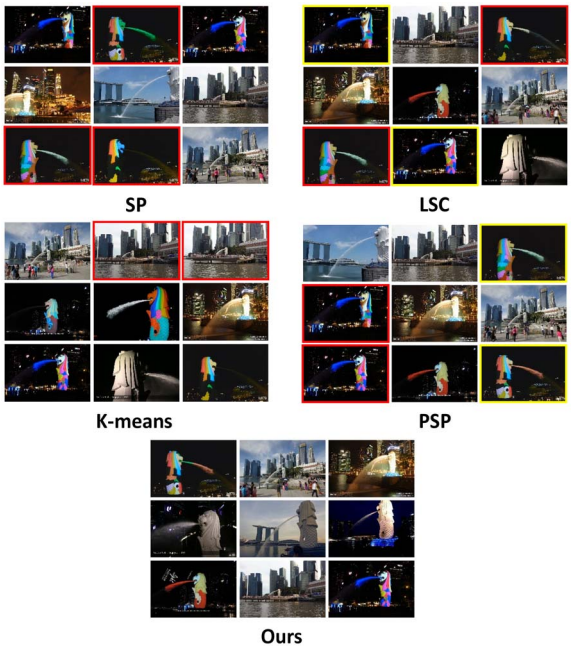


Fig. 5. Visual comparison of results for different clustering methods.

Table III presents the results of our user study. The following observations are made.

1) As expected, RS performs the worst, which indicates that clustering is contributive to enhance audience satisfaction.
2) Our method performs the best, which demonstrated that it is essential to generate comprehensive and appropriate descriptions for scenic spots.
3) The highest score is achieved by the stranger-group, the second is the friend-group, while the worst is the family-group. This is because photos containing the user's self-portrait occupies only a small fraction of the final video clips. This fails to meet the requirement of families and friends. To alleviate this problem, we can shorten video clips or increase the playtime of the photos.
4) All the P-values are less than 0.05, indicating that there is a strong support to our method.

Fig. 5 compares the five clustering algorithms on the Merlion on Marina Bay. In this example, the number of clusters is

TABLE IV
AUDIENCE-BASED EVALUATION USING THE PROBABILISTIC MODEL (RANDOM: RANDOM SAMPLING, NONAES: NONAESTHETICS, NONSMO: NONSMOOTHNESS)

	Beijing				Singapore				Beijing+Singapore				
	Fam	Fr	Str	Ave	Fam	Fri	Str	Ave	Fam	Fri	Str	Ave	P-value
Random	2.42	2.75	3.42	2.86±0.88	2.42	2.75	3.42	2.86±1.67	2.42	2.75	3.42	2.86±1.12	< 0.05
Non-aes	2.58	3.08	3.75	3.14±0.11	2.42	2.83	3.50	2.91±0.12	2.50	2.96	3.63	3.03±0.80	< 0.05
Non-smo	2.67	2.92	3.50	3.03±0.27	2.42	2.67	3.42	2.83±0.25	2.54	2.79	3.46	2.93±0.22	< 0.05
Ours	**3.17**	**3.67**	**4.00**	**3.61±0.24**	**3.25**	**3.75**	**4.42**	**3.81±0.70**	**3.21**	**3.71**	**4.21**	**3.71±0.28**	--

fixed to 9. For every clustering method, we display clustering results using a photo near the center of the cluster, which is considered as the sample most informative to the cluster. As shown in Fig. 5, most of the clustering methods perform satisfactorily to classify the day scenes. However, they are not effective to distinguish the night scenes. Keyframes that are highlighted by same-colored rectangles mean that they are taken at the same scenic spot. Obviously, our method extracts more comprehensive representations of the scenic spot without duplicated scenes.

4) Probabilistic Model for SnapVideo: Keyframes selection and ranking arrange frames from multiple views into an aesthetically pleasing and smooth video clip. Different from the traditional methods, the photo aesthetic measurement an important factor in keyframes selection. Ranking can recover the true order of keyframes in a video clip with the constraint of smoothness. Here, we validate our probabilistic model for SnapVideo using three baseline methods: 1) RS; 2) nonaesthetics; and 3) nonsmoothness. Nonaesthetic is a baseline that generates video clips by abandoning the aesthetic attribute. For nonsmoothness, we do not consider the keyframes ranking and sort the selected keyframes in an arbitrary order. Besides, RS is the simplest way to select keyframes from clustering and generate video clips without ranking.

Similar to the previous procedure, we produce three types of video clips for Beijing and Singapore, respectively, and then invite 12 volunteers to score them. Statistics of evaluation are shown in Table IV. We observe the following.

1) RS performs the worst because it fails to utilize the aesthetic and smooth attributes.
2) Nonaesthetics and nonsmoothness also perform unsatisfactorily. Both of them fail to jointly optimize the aesthetics and smoothness of keyframes.
3) Noticeably, the highest score is again obtained by the stranger-group, followed by the friend- and family-groups.
4) All the *P*-values are no more than 0.05.

Fig. 6 visualizes the video clips generated from the "Singapore Flyer," which is described by eight views and each row representing one. We use the keyframes selection module to identify five keyframes of each view. The rank of keyframes are calculated with the constraints of interview and intraview smoothness.

5) Overall Evaluation: In this experiment, we evaluate the overall performance of the proposed approach. Four popular video summarization algorithms are adopted as baselines: RS, uniform sampling (US), the keyframe-based video summarization tool (KBVST) [26], and the superframe-based video summarization (SFVS) [27].

1) *RS:* It arbitrarily selects three scenic spots to generate video clips. For each one, a fixed number of keyframes are acquired randomly for the users.
2) *US:* Different from RS, US selects keyframes for each scenic spot by sampling with an equal interval.

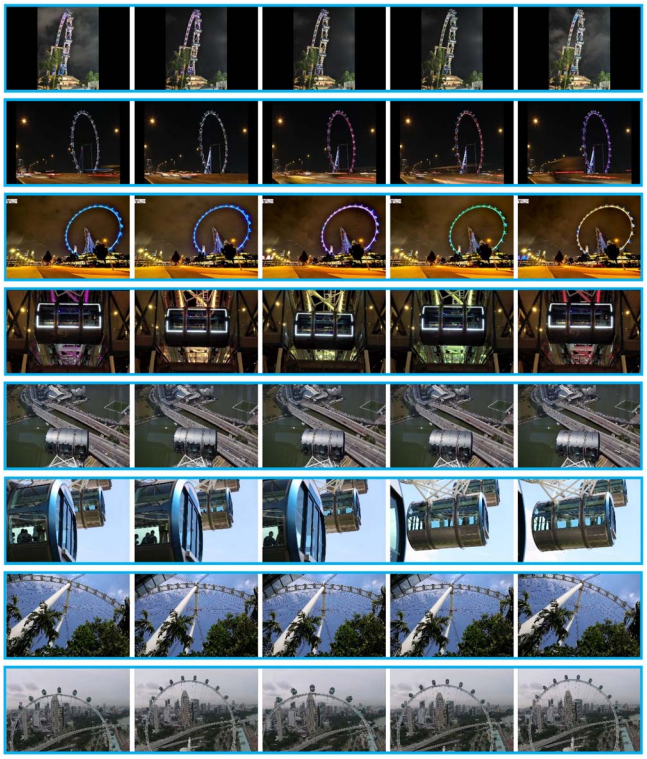


Fig. 6. Visual illustration of different scenic spots after SnapVideo.

TABLE V
AUDIENCE-BASED OVERALL EVALUATION

	Beijing	Singapore	Beijing+Singapore	P-value
RS	2.42	2.44	2.43±0.19	< 0.05
US	2.67	2.97	2.82±0.90	< 0.05
KBVST	2.94	3.03	2.99±0.68	< 0.05
SFVS	2.81	2.89	2.85±0.97	< 0.05
Ours	**3.61**	**3.81**	**3.71±0.28**	--

TABLE VI
INTERAGREEMENT EVALUATION AMONG VOLUNTEERS IN TERMS OF KAPPA METRIC. EACH VOLUNTEER LABELS 30 VIDEOS GENERATED BY THE FIVE METHODS WITH THREE IDENTITIES

Case #	Category #	Volunteer #	Overall Agreement	Fixed-marginal Kappa	Free-marginal Kappa
30	2	12	72.3%	23.4%	36.9%

3) *KBVST:* It generates a summary by clustering visually similar frames based on the user-selected visual features. This method selects appropriate visual features. We employ color and edge directivity descriptor [41] for the experiment. As KBVST is a single-video summarization algorithm, we combine single videos into a long one. A maximum of 21 frames per scenic spot is adopted.
4) *SFVS:* It first segments videos using superframe segmentation and then estimates their interestingness. Afterward, it selects an optimal subset of superframes to generate a descriptive summary.

We present the evaluation results by volunteers in Table V. It is observable that our method outperforms the baselines by a large margin. One reason is that the other methods fail to generate video summaries that satisfy tastes of different audiences. To better reconstruct the semantics of a sightseeing trip, our method arranges frames from multiple views into an aesthetically pleasing and smooth video clip. We utilize the Kappa method [52] to evaluate the intervolunteer agreement. The Kappa metric reflects the degree of intervolunteer agreement. A Kappa result of 1 indicates a perfect agreement while that of 0 indicates an agreement by random. A higher Kappa value corresponds to a more uniform agreement. Typically, Kappa value over 0.7 means that the agreement is highly uniform. In this paper, we employ the online Kappa calculator tool[6] to obtain the Kappa values. Table VI shows the interagreement analysis.[7] As shown in the table, there are enough intervolunteer agreements for our task.

Due to the subjectiveness of video quality assessment, a user study is conducted to evaluate the quality of video clips produced by different algorithms. In our experiment, paired comparison is carried out to evaluate the effectiveness of the compared algorithms. It is worth emphasizing that both rating and ranking are not suitable here because both are too complex for an observer to duly provide and would be an unnatural task for an observer. Paired comparison is to present each subject with a pair of video clips calculated using two different algorithms. Volunteers are then required to indicate a preference, for one of the two synthesized results. We select ten scenic spots from both Beijing and Singapore to evaluate the results, and the 213 volunteers vote on each scenic spot. We spent tremendous time and resources to invited 213 volunteers as well as their family members (at least two for each volunteer) to the user study and fill the preference matrix. Besides, we also online invited about 70–80 persons with different occupations whom are strangers to these 213 volunteers to participate the user study. Most of them are master's/Ph.D. students from the art and engineering departments. Each of the 213 volunteers watched both the original video clips or input photos and the generated video clip. As shown in Fig. 7, the entity 1423 in row "US" and column "KBVST" reflects that 1423 votes prefer US than KBVST. Statistic results demonstrated that our method achieves the best performance on the ten scenic spots from both Beijing and Singapore, respectively.

	RS	US	KBVST	DTS	VSUMM	SFVS	Ours	Score
RS	--	1121	1343	1454	1721	1611	102	7352
US	1009	--	1432	1121	1654	655	216	6087
KBVST	787	698	--	1196	895	578	119	4273
DTS	676	1009	934	--	1723	1512	321	6175
VSUMM	409	476	1235	407	--	1137	214	3878
SFVS	519	1475	1552	618	993	--	1110	6267
Ours	2028	1914	2011	1809	1916	1020	--	10698

Preference matrix from the video clips generated in Beijing (filled by 213 volunteers)

	RS	US	KBVST	DTS	VSUMM	SFVS	Ours	Score
RS	--	956	1114	1543	1478	1854	186	7131
US	1174	--	1354	1265	1589	564	347	6293
KBVST	1016	776	--	1217	858	612	203	4682
DTS	587	865	913	--	1658	1438	274	5735
VSUMM	652	541	1272	472	--	1264	327	4528
SFVS	276	1566	1518	692	866	--	1177	6095
Ours	1944	1783	1927	1856	1803	953	--	10266

Preference matrix from the video clips generated in Singapore (filled by 213 volunteers)

Fig. 7. User studies conducted on 213 volunteers.

V. CONCLUSION

SnapVideo is a useful technique in computer vision [10], [28]–[30], [37], [38] and multimedia [3], [4], [9], [32]–[34]. This paper presented a novel SnapVideo technique which intelligently transforms a user's photo album recording a trip into a comprehensive, well-aesthetic, and coherent video clip. We introduced an scenic spot identification model as the first module to discover scenic photos preferred by multiple audiences as prior queries. The view generation module then describes a scenic spot comprehensively by clustering the relevant video keyframes into an optimal number of views. To preserve the semantics of a sightseeing trip, a probabilistic model is proposed to intergrade frames from multiple views into an aesthetically pleasing and smooth video clip. Thorough experiments demonstrate that our method creates beautiful video clips covering tourists' sightseeing experiences and meeting the taste of different audiences.

[6] http://justusrandolph.net/kappa/

[7] In Table VI, the number of category is set to 2. Scores ranging from 0 to 2 represent unsatisfied results, and scores between 3 and 5 mean satisfied results.

REFERENCES

[1] T. Bühler and M. Hein, "Spectral clustering based on the graph p-Laplacian," in *Proc. Int. Conf. Mach. Learn.*, Montreal, QC, Canada, 2009, pp. 81–88.

[2] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li, "Learning to rank: From pairwise approach to listwise approach," in *Proc. Int. Conf. Mach. Learn.*, Corvallis, OR, USA, 2007, pp. 129–136.

[3] Q. Wang, Y. Yuan, P. Yan, and X. Li, "Saliency detection by multiple-instance learning," *IEEE Trans. Cybern.*, vol. 43, no. 2, pp. 660–672, Apr. 2013.

[4] L. Liu, L. Shao, X. Zhen, and X. Li, "Learning discriminative key poses for action recognition," *IEEE Trans. Cybern.*, vol. 43, no. 6, pp. 1860–1870, Dec. 2013.

[5] X. Chen and D. Cai, "Large scale spectral clustering with landmark-based representation," in *Proc. AAAI Conf. Artif. Intell.*, San Francisco, CA, USA, 2011, pp. 313–318.

[6] F. R. K. Chung, *Spectral Graph Theory*. Providence, RI, USA: Amer. Math. Soc., 2011.

[7] Y. Cong, J. Yuan, and J. Luo, "Towards scalable summarization of consumer videos via sparse dictionary selection," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 66–75, Feb. 2012.

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

[9] L. Zhang *et al.*, "Feature correlation hypergraph: Exploiting high-order potentials for multimodal recognition," *IEEE Trans. Cybern.*, vol. 44, no. 8, pp. 1408–1419, Aug. 2014.

[10] C. Hou, F. Nie, X. Li, D. Yi, and Y. Wu, "Joint embedding learning and sparse regression: A framework for unsupervised feature selection," *IEEE Trans. Cybern.*, vol. 44, no. 6, pp. 793–804, Jun. 2014.

[11] F. Dorey, "In brief: The P value: What is it and what does it tell you?" *Clin. Orthopaed. Related Res.*, vol. 468, no. 8, pp. 2297–2298, 2010.

[12] Y. Fu, L. Wang, and Y. Guo, "Multi-view learning for multi-view video summarization," *CoRR arXiv:1405.6434*, 2014.

[13] Y. Fu *et al.*, "Multi-view video summarization," *IEEE Trans. Multimedia*, vol. 12, no. 7, pp. 717–729, Nov. 2010.

[14] J. Hamm and D. D. Lee, "Grassmann discriminant analysis: A unifying view on subspace-based learning," in *Proc. Int. Conf. Mach. Learn.*, Helsinki, Finland, 2008, pp. 376–383.

[15] E.-H. Han, G. Karypis, V. Kumar, and B. Mobasher, "Hypergraph based clustering in high-dimensional data sets: A summary of results," *IEEE Data Eng. Bull.*, vol. 21, no. 1, pp. 15–22, 1998.

[16] M. T. Harandi, C. Sanderson, S. Shirazi, and B. C. Lovell, "Graph embedding discriminant analysis on Grassmannian manifolds for improved image set matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, 2011, pp. 2705–2712.

[17] L. Herranz and J. M. Martínez, "A framework for scalable summarization of video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 9, pp. 1265–1270, Sep. 2010.

[18] R. Hong *et al.*, "Beyond search: Event-driven summarization for Web videos," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 7, no. 4, 2011, Art. no. 35.

[19] Y. Huang, Q. Liu, S. Zhang, and D. N. Metaxas, "Image retrieval via probabilistic hypergraph ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, Jun. 2010, pp. 3376–3383.

[20] P. Indyk and R. Motwani, "Approximate nearest neighbors: Towards removing the curse of dimensionality," in *Proc. ACM Symp. Theory Comput.*, Dallas, TX, USA, 1998, pp. 604–613.

[21] A. K. Jain and A. Vailaya, "Image retrieval using color and shape," *Pattern Recognit.*, vol. 29, no. 8, pp. 1233–1244, 1996.

[22] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2. San Diego, CA, USA, 2005, pp. 524–531.

[23] H. Li, K. N. Ngan, and Q. Liu, "FaceSeg: Automatic face segmentation for real-time video," *IEEE Trans. Multimedia*, vol. 11, no. 1, pp. 77–88, Jan. 2009.

[24] Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma, "A survey of content-based image retrieval with high-level semantics," *Pattern Recognit.*, vol. 40, no. 1, pp. 262–282, 2007.

[25] H. Mansour, P. Nasiopoulos, and V. Krishnamurthy, "Rate and distortion modeling of CGS coded scalable video content," *IEEE Trans. Multimedia*, vol. 13, no. 2, pp. 165–180, Apr. 2011.

[26] M. Lux, O. Marques, K. Schöffmann, L. Böszörmenyi, and G. Lajtai, "A novel tool for summarization of arthroscopic videos," *Multimedia Tools Appl.*, vol. 46, nos. 2–3, pp. 521–544, 2010.

[27] M. Gygli, H. Grabner, H. Riemenschneider, and L. V. Gool, "Creating summaries from user videos," in *Proc. Eur. Conf. Comput. Vis.*, Zürich, Switzerland, 2014, pp. 505–520.

[28] L. Nie *et al.*, "Disease inference from health-related questions via sparse deep learning," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 8, pp. 2107–2119, Aug. 2015.

[29] L. Nie *et al.*, "Beyond doctors: Future health prediction from multimedia and multimodal observations," in *Proc. ACM Multimedia*, Brisbane, QLD, Australia, 2015, pp. 591–600.

[30] Y. Liu, L. Zhang, L. Nie, Y. Yan, and D. S. Rosenblum, "Fortune teller: Predicting your career path," in *Proc. AAAI*, Phoenix, AZ, USA, 2016, pp. 201–207.

[31] L. Nie, Y.-L. Zhao, M. Akbari, J. Shen, and T.-S. Chua, "Bridging the vocabulary gap between health seekers and healthcare knowledge," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 2, pp. 396–409, Feb. 2015.

[32] Y. Yang *et al.*, "A multimedia retrieval framework based on semi-supervised ranking and relevance feedback," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 723–742, Apr. 2012.

[33] Y. Yang, H. T. Shen, Z. Ma, Z. Huang, and X. Zhou, "*l*2 1-norm regularized discriminative feature selection for unsupervised learning," in *Proc. IJCAI*, Barcelona, Spain, 2011, pp. 1589–1594.

[34] Y. Yang, Y.-T. Zhuang, F. Wu, and Y.-H. Pan, "Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval," *IEEE Trans. Multimedia*, vol. 10, no. 3, pp. 437–446, Apr. 2008.

[35] A. K. Moorthy, P. Obrador, and N. Oliver, "Towards computational models of the visual aesthetic appeal of consumer videos," in *Proc. Eur. Conf. Comput. Vis.*, Heraklion, Greece, 2010, pp. 1–14.

[36] Y. Pang, S. Wang, and Y. Yuan, "Learning regularized LDA by clustering," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 12, pp. 2191–2201, Dec. 2014.

[37] L. Shao, X. Zhen, D. Tao, and X. Li, "Spatio-temporal Laplacian pyramid coding for action recognition," *IEEE Trans. Cybern.*, vol. 44, no. 6, pp. 817–827, Jun. 2014.

[38] L. Chen, D. Xu, I. W.-H. Tsang, and X. Li, "Spectral embedded hashing for scalable image retrieval," *IEEE Trans. Cybern.*, vol. 44, no. 7, pp. 1180–1190, Jul. 2014.

[39] S. Pongnumkul, J. Wang, and M. Cohen, "Creating map-based storyboards for browsing tour videos," in *Proc. ACM Symp. User Interface Softw. Technol.*, Monterey, CA, USA, 2008, pp. 13–22.

[40] M. K. Saini, R. Gadde, S. Yan, and W. T. Ooi, "Movimash: Online mobile video mashup," in *Proc. ACM Int. Conf. Multimedia*, Nara, Japan, 2012, pp. 139–148.

[41] S. A. Chatzichristofis, K. Zagoris, Y. S. Boutalis, and N. Papamarkos, "Accurate image retrieval based on compact composite descriptors and relevance feedback information," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 24, no. 2, pp. 207–244, 2010.

[42] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1427–1441, Jun. 2010.

[43] J. She, D. Wang, and M. Song, "Automatic image cropping using sparse coding," in *Proc. 1st Asian Conf. Pattern Recognit.*, Beijing, China, 2011, pp. 490–494.

[44] S. J. Sheather and M. C. Jones, "A reliable data-based bandwidth selection method for kernel density estimation," *J. Roy. Stat. Soc. B*, vol. 53, no. 3, pp. 683–690, 1991.

[45] M.-L. Shyu, Z. Xie, M. Chen, and S.-C. Chen, "Video semantic event/concept detection using a subspace-based multimedia data mining framework," *IEEE Trans. Multimedia*, vol. 10, no. 2, pp. 252–259, Feb. 2008.

[46] J. Tang, X.-S. Hua, G.-J. Qi, Y. Song, and X. Wu, "Video annotation based on kernel linear neighborhood propagation," *IEEE Trans. Multimedia*, vol. 10, no. 4, pp. 620–628, Jun. 2008.

[47] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[48] B. T. Truong and S. Venkatesh, "Video abstraction: A systematic review and classification," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 3, no. 1, Feb. 2007, Art. no. 3.

[49] J. Venkatanathan *et al.*, "Who, when, where: Obfuscation preferences in location-sharing applications," INST Softw. Res. Internat, Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep. CMU-ISR-11-110, 2011.

[50] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.

[51] U. Von Luxburg, "A tutorial on spectral clustering," *Stat. Comput.*, vol. 17, no. 4, pp. 395–416, 2007.

[52] M. J. Warrens, "Inequalities between multi-rater kappas," *Adv. Data Anal. Classification*, vol. 4, no. 4, pp. 271–286, 2010.
[53] C. Xu and Z. Su, "Identification of cell types from single-cell transcriptomes using a novel clustering method," *Bioinformatics*, vol. 31, no. 12, pp. 1974–1980, 2015.
[54] H.-H. Yeh, C.-Y. Yang, M.-S. Lee, and C.-S. Chen, "Video aesthetic quality assessment by temporal integration of photo-and motion-based features," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1944–1957, Dec. 2013.
[55] J. Yu, D. Tao, and M. Wang, "Adaptive hypergraph learning and its application in image classification," *IEEE Trans. Image Process.*, vol. 21, no. 7, pp. 3262–3272, Jul. 2012.
[56] L. Zhang *et al.*, "Probabilistic skimlets fusion for summarizing multiple consumer landmark videos," *IEEE Trans. Multimedia*, vol. 17, no. 1, pp. 40–49, Jan. 2015.
[57] W.-L. Zhao, C.-W. Ngo, H.-K. Tan, and X. Wu, "Near-duplicate keyframe identification with interest point matching and pattern learning," *IEEE Trans. Multimedia*, vol. 9, no. 5, pp. 1037–1048, Aug. 2007.
[58] Z. Ma, Y. Yang, N. Sebe, and A. G. Hauptmann, "Knowledge adaptation with partially shared features for event detection using few exemplars," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 9, pp. 1789–1802, Sep. 2014.
[59] Y. Yang, D. Xu, F. Nie, S. Yan, and Y. Zhuang, "Image clustering using local discriminant models and global integration," *IEEE Trans. Image Process.*, vol. 19, no. 10, pp. 2761–2773, Oct. 2010.
[60] D. Zhou, J. Huang, and B. Schölkopf, "Learning with hypergraphs: Clustering, classification, and embedding," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2006, pp. 1601–1608.
[61] S. Wei, D. Xu, X. Li, and Y. Zhao, "Joint optimization toward effective and efficient image search," *IEEE Trans. Cybern.*, vol. 43, no. 6, pp. 2216–2227, Dec. 2013.
[62] M. Gygli, H. Grabner, and L. V. Gool, "Video summarization by learning submodular mixtures of objectives," in *Proc. CVPR*, 2015, pp. 3090–3098.
[63] J. Zhu, S. Liao, and S. Z. Li, "Multicamera joint video synopsis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 6, pp. 1058–1069, Jun. 2016.
[64] L. Zhang *et al.*, "Weakly supervised photo cropping," *IEEE Trans. Multimedia*, vol. 16, no. 1, pp. 94–107, Jan. 2014.
[65] J. Tompkin, K. I. Kim, J. Kautz, and C. Theobalt, "Videoscapes: Exploring sparse, unstructured video collections," *ACM Trans. Graph.*, vol. 31, no. 4, pp. 1–12, 2012.
[66] F. Wang and B. Merialdo, "Multi-document video summarization," in *Proc. ICME*, New York, NY, USA, 2009, pp. 1326–1329.
[67] X. Zhu, J. Fan, A. K. Elmagarmid, and X. Wu, "Hierarchical video content description and summarization using unified semantic and visual similarity," *Multimedia Syst.*, vol. 9, no. 1, pp. 31–52, 2010.
[68] Y. Cong, J. Yuan, and J. Luo, "Towards scalable summarization of consumer videos via sparse dictionary selection," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 66–75, Feb. 2012.
[69] L. Zhang *et al.*, "Feature correlation hypergraph: Exploiting high-order potentials for multimodal recognition," *IEEE Trans. Cybern.*, vol. 44, no. 8, pp. 1408–1419, Aug. 2014.
[70] X. Wang, W. Bian, and D. Tao, "Grassmannian regularized structured multi-view embedding for image classification," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2646–2660, Jul. 2013.
[71] X. Chang, Y. Yang, E. P. Xing, and Y. L. Xu, "Complex event detection using semantic saliency and nearly-isotonic SVM," in *Proc. ICML*, Bonn, Germany, 2005.

Luming Zhang (M'14) received the Ph.D. degree in computer science from Zhejiang University, Hangzhou, China.

He is currently a Faculty Member with the Hefei University of Technology, Hefei, China. His current research interests include visual perception analysis, image enhancement, and pattern recognition.

Peiguang Jing (M'16) is currently pursuing the Ph.D. degree with Tianjin University, Tianjin, China.

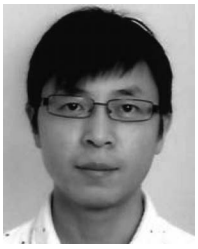
His current research interests include multimedia and computer vision.

Yuting Su received the B.Eng. and Ph.D. degrees from Tianjin University, Tianjin, China.

He was a Visiting Scholar with Case Western Reserve University, Cleveland, OH, USA. He is currently a Professor with the School of Electronic Engineering, Tianjin University. His current research interests include multimedia content analysis and security.

Chao Zhang received the B.S. and M.S. degrees from Zhejiang University, Hangzhou, China. He is currently pursuing the Ph.D. degree with the Computer Science Department, University of Illinois at Urbana–Champaign, Champaign, IL, USA.

His current research interests include spatiotemporal data mining, social media analysis, and applied machine learning.



Ling Shao (M'09–SM'10) received the B.Eng. degree in electronic and information engineering from the University of Science and Technology of China (USTC), Anhui, China, the M.Sc. degree in medical image analysis and the Ph.D. (D.Phil.) degree in computer vision at the Robotics Research Group from the University of Oxford, Oxford, U.K.

He is a Professor with the Department of Computer Science and Digital Technologies, Northumbria University, Newcastle upon Tyne, U.K. He was a Senior Lecturer with the Department of Electronic and Electrical Engineering, University of Sheffield, Sheffield, U.K., from 2009 to 2014, and a Senior Scientist with Philips Research, Eindhoven, The Netherlands, from 2005 to 2009. His current research interests include computer vision, image/video processing and machine learning.

Dr. Shao is an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON CYBERNETICS, and several other journals. He is a fellow of the British Computer Society and the Institution of Engineering and Technology.