TravelMedia: An Intelligent Management System for Media Captured in Travel

Wei-Ta Chu, Cheng-Jung Li, and Sheng-Chun Tseng
National Chung Cheng University, Taiwan
wtchu@cs.ccu.edu.tw, zoneli1987@gmail.com, bittertea0503@gmail.com

Abstract

A media management system exploiting characteristics of travel media is designed to facilitate efficient management and browsing. According to travel schedules, travel media often have implicit thematic structure. Correlation between different modalities also provides implicit cues to media analysis. In this system, we exploit techniques of near-duplicate detection to select representative photos, and determine region-of-interest in photos to enhance browsing experience. For face-name association, a face clustering module based on visual language models is constructed. To systematically segment travel videos of bad visual quality and significant motion, we explore correlation between photos and videos based on approximate visual word histogram matching. Experimental results demonstrate the effectiveness of the proposed approaches and show that they are practical functions.

Keywords: Travel media management; representative selection, region of interest; face clustering; video scene detection.

1. Introduction

Traveling has been one of the most important activities in the modern age. It not only relaxes tense life, but is also a symbol of vogue or taste. People treasure travel experience, put it into memory, and want to efficiently manage or manipulate it. Massive digital content, therefore, demands a set of analysis and management tools that are specially designed by considering characteristics of travel media. Although there have been many studies on analyzing news, sports, and TV programs, these methodologies are either too general (such as video concept detection [1][2]) such that finer and practical analysis for travel media is hardly achieved, or too specific (such as heuristic rules or specific models for sports events [3][4]) to a limited domain that has significantly different characteristics from travel media. In this work, we target on media collected in journeys, especially photos and videos, and discover how temporal/visual characteristics and implicit correlations between them facilitate development of intelligent analysis techniques.

Different from other consumer media, travel media have special characteristics that may be conducive or cumbersome to practical technique development.

- According to a pre-arranged travel schedule, travelers visit scenic spots and capture photos/videos sequentially. Photos and videos are taken frequently at a scenic spot, and are taken rarely during transportation from one spot to another. The temporal grouping characteristic facilitates appropriate segmentation of travel media.
- At the same scenic spot, the famous landmarks or buildings are often captured many times. This characteristic provides clues for detecting the most representative photos and objects.
- Content captured at the same scenic spot would have significantly different appearances, which destroys conventional clustering methods for image clustering or video scene detection.
- Scenic spots are visited sequentially, and various media are taken alternately or simultaneously in the same temporal order. Different media may thus be correlated. For example, photos and videos may be visually and temporally correlated.

With the characteristics described above, we study travel media management from various perspectives. From modalities being processed, travelers may take photos to capture scenes or objects, and may take videos to capture dynamic evolution of events (e.g. artist performance and animal moving). Text-based metadata, such as GPS (Global Position System) and time information, are automatically stored in file headers. From targets being browsed and managed, people used to ask *what* happened in the media, *when* and *where* events/objects occurred/appeared, and *who* were in the media. From the perspective of required functions, people may need to annotate, browse, retrieve, and manipulate various media. From the perspective of media correlation, single modality or multiple modalities may be processed. Finally, from the perspective of access devices, people may access travel media via PCs, TVs, mobile phones, PDAs, etc.

Although there may be unlimited viewpoints to conduct intelligent management, studies in this article focus on travel experience browsing and management. Figure 1 shows our work from the perspectives of system functionalities. For browsing efficiency, we examine each photo's degree of representative and select the most important one to represent a cluster of photos. An ROI determination module determines the most prominent region of a photo, in order to facilitate efficient display on resolution-limited devices [5]. In addition to landmark and building, human faces in media are always the most attractive features. A face clustering module is developed to cluster the same individual's faces together so that users can easily browse and achieve face-name association [20]. At the last but not the least, we explore correlation between videos and photos to perform video scene detection [18]. Although scene detection is a widely studied problem, great challenges are especially drawn by travel videos because of unstructured and unlimited content. Overall, Table 1 shows the correspondence between these modules and the discussed issues.

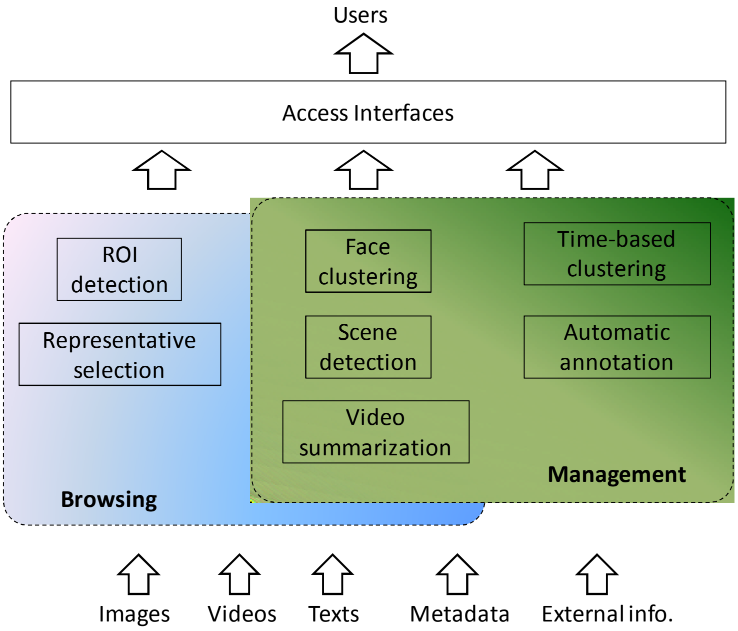

Figure 1. Overview of the proposed travel media management system.

Table 1. Relationships between the proposed modules and research perspectives.

	Modalities	Facets	Functions	Correlation	Access manners
Representative selection	Video, photo	What, where	Browsing	Single modality	PC, PDA, mobile phone
ROI determination	Video, photo	What, where	Browsing	Single modality	PC, PDA, mobile phone
Face clustering	Video, photo	Who	Annotation, browsing, retrieval	Single modality	PC, PDA, mobile phone
Video scene detection	Video, photo	Where	Annotation, browsing	Multiple modalities	PC

Relationships between the proposed modules and travel media can also be described as a vector space model, as illustrated in Figure 2. An entity captured in journeys can be viewed as a vector in a space, which is constituted by the orthogonal bases *who*, *where*, *what*, and *when*. The four properties of each entity are not always known, and the developed modules don't simultaneously consider all properties. For example, the representative selection module and ROI determination module conceptually project data into the two-dimensional space constituted by *what* and *where*. This is a one-to-one mapping, while not all data can be successfully processed.

On the other hand, the face clustering process conceptually project data into the one-dimensional space constituted by *who*. With this projection, data containing similar faces would be projected into proximity. One photo may contain many faces, and thus this projection is one-to-many. In this article, the proposed modules proceed in parallel, while some of the evaluation data can be processed by more than one module.

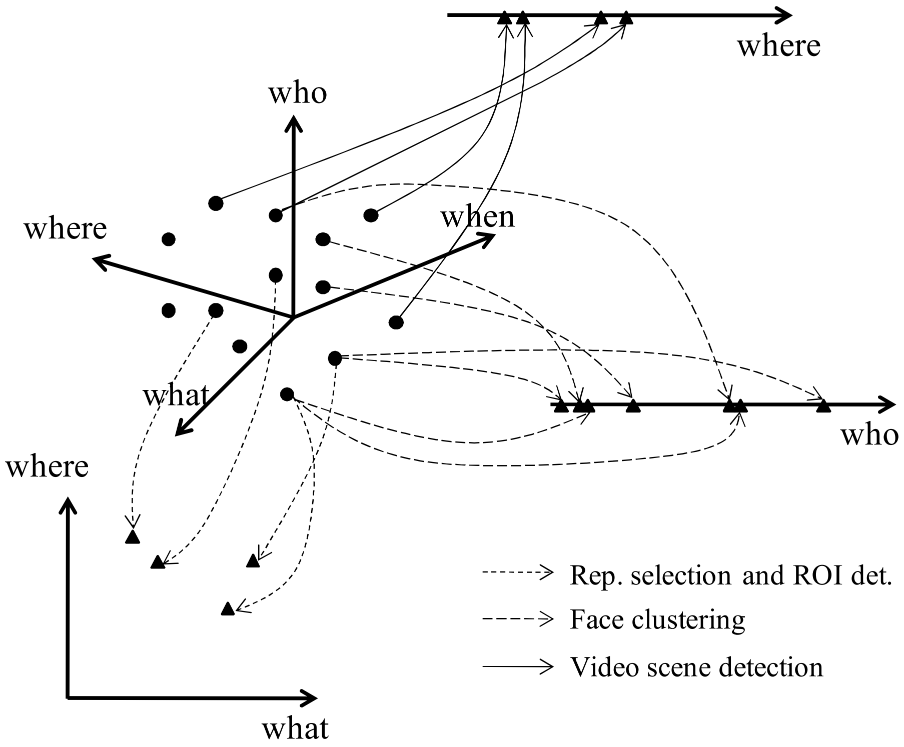

Figure 2. Vector space model of travel media management.

Contributions of this work are summarized as follows.

- To the best of our knowledge, this work is one of the first studies on sketching the contours of specially designed travel media analysis techniques, in terms of modalities, facets, functions, and access manners.
- Part-based models and related statistics are newly adopted to be integrated with our previous work [5]. The integrated approach eliminates the limitation of our previous work and improves its feasibility.
- A new feature matching scheme is specially designed to adjust the scheme originally designed in [13]. This extension mitigates the problem of our previous work [20], in which few feature points can be detected and matched on faces.

The rest of this paper is organized as follows. Section 2 surveys works related to the proposed modules. Section 3 gives details of each developed components. Comprehensive evaluation on each component is given in Section 4, followed by concluding remarks in Section 5.

2. Related Work

- Representative selection

One early example of selecting a representative for a group is keyframe selection from a video shot. Uniform and non-uniform samplings are both adopted for years. Girgensohn and Boreczky [7] propose a hierarchical cluster algorithm to dynamically determine number of keyframes, with extra emphasis on specific frames such as close-ups. A classical study on related issues can be found in [33]. Recently, researchers conduct representative selection in terms of semantics rather than simply visual appearance. Hsieh et al. [34] cluster image search results and select multiple canonical images based on a context graph constructed from visual features and text-based metadata. Jing et al. [41] discovers impact of local features and find a canonical image from commonly searched products. Local features, or the so-called part-based model, are demonstrated to be effective in capturing semantics in images.

Selecting representative or canonical samples for a group is also widely known as the centrality problem in social network analysis [42]. By modeling relationships between data as a network, the node with the largest centrality value is viewed as the representative. In our previous work [5], we exploit near-duplicate detection to describe relationships between photos as a graph, and then discover the most centric one as the representative photo. In this article, this work is extended so that photo collections without clear near-duplicate properties can be processed.

- ROI determination

Due to popularity of mobile device and advance of perceptual coding, automatic extraction of ROI has attracted much attention in recent years. A large amount of studies base on the computational user attention model [11][12]. A saliency map that integrates information of intensity, orientation, and color contrast is constructed, and the region that covers the largest saliency value is selected as ROIs. This idea is also extended to videos, with the consideration of motion information [43].

In our work, we determine ROIs based on the results of representative selection. In the representative photo, the region that covers the feature points used to determine near-duplicate properties is selected as the ROI. This method is conceptually similar to common pattern discovery [38][40]. Yuan and Wu [38] randomly partition each image as subimages of different sizes, and then compare them to find one or multiple common patterns between images. Liu and Yan [40] do the same job by comparing images based on local features and adopt characteristics of spatial layout. These works assume common patterns certainly exist between images. However, in our case, not all photos taken at the same scenic spot contain the most prominent landmark. In addition, these works equally treat the processed images, while our work elaborately finds the most prominent region in the most representative photo.

- Face clustering

The problem of face-name association has attracted computer vision researchers for years. Instead of reviewing rich literature on face recognition, we simply survey a few important studies on annotating people in consumer photos. Zhang et al. [28] extract features from the upper part of body, face, and eyes, and use a Bayesian framework to predict identification of each face in family albums. Zhao et al. [29] propose a graphical model to integrate face and clothes information, in which clothes information is used to eliminate identification errors. Gallagher and Chen [30] especially investigate grouping characteristic of people in family albums. Instead of treating each face individually, social context and its related features are modeled to facilitate face annotation.

Another media urgently demanding face-name association is news videos. Satoh et al. [26] propose a prestigious system to detect face sequences from videos, extract names from transcript and caption, and evaluate co-occurrence information between different modalities. Based on a million news pictures and captions from Yahoo! News, Berg et al. [39] consider variety of faces, such as illumination changes and pose variation, and propose a sequence of clustering methods to achieve face clustering. Recently, The Pham et al. [27] adopt an EM algorithm to link faces and names, based on assigning a name to a face, and assigning a face to a name. Ozkan and Duygulu [36] adopt part-based models to compare faces. By describing relationship between faces as a graph, they transform the problem of recognizing a query face as finding the highly connected sub-graph. In our work, we focus on personal photo collections in which text-based annotation is not available. Part-based models as in [36], with the systematic description by visual language models, are developed.

- Video scene detection

For video scene detection, Yeung and Yeo [23] propose a classical work called scene transition graph to describe relationships between video shots, and achieve scene detection by analyzing links of the graph. For movies, Hanjalic et al. [24] investigate context between video shots, and determine boundaries of logical story units such as dialogue and action scenes. Sundaram and Chang [25] take film-making rules and psychology of audition into account to build a computational scene model, which mimics characteristics of human's short-term and long-term memory. Rasheed and Shah [37] develop a two-pass algorithm based on motion, shot length, and color properties, to find semantics-related scenes in movies and TV shows. More recently, Chasanis et al. [21] estimate appropriate number of keyframes for each video shot based on a spectral clustering approach, and then determine scene boundaries by sequence alignment techniques. Due to unequivocal importance of video scene detection, integrated framework such as [31] and systematic evaluation method such as [32] have been proposed. The TRECVID benchmark [8] also issues the "story segmentation" task for years, while it focuses only on news and TV programs.

- Other Aspects

It's obvious that travel media can be analyzed from unlimited perspectives other than our proposed works. For example, a number of studies detect events in consumer photos and videos. Based on bag-of-feature representation, Jiang and Loui [44] detect visual concepts and thus model semantic events, such as birthday, wedding, and picnic. Semantic events here are highly correlated to time (*when*) and space (*where*) domains as we described in Figure 2. Actually, time information of photos has been exploited to cluster photos into events for years [45][9]. Automatically detecting events and event-based browsing for consumer photos/videos have continuously attracted researchers' attention [46][47].

With easy creation and sharing of geo-tagging, some studies organize photos based on geographic locations, which makes much sense and is also described as the *where* axis in Figure 2. Naaman et al. [48] construct location and event hierarchies based on time and location information embedded in images. Ahern et al. [49] analyze tags associated with geo-referenced Flickr images to find representative tags and visualize locations, photos, and associated tags via a map interface. Some studies compute the viewpoint of each photo and thus construct a 3D scene, so that users can interactively browse unconstructed photos taken at the same scene [50].

Many web-based sharing platforms notice the explosive usage of consumer photos and urgent demand of management functionalities, and provide functions such as event/object/geographic tagging in Flickr. Google Picasa's [6] name-tag function allows users to easily annotate faces based on preliminary face clustering results. Facebook, as the biggest social network website in the world, currently provides face tagging and may become the largest face database with associated names. Although some works have been conducted to boost face recognition based on such social context [51], related ideas just emerge and still need to be verified from various aspects. As compared to our proposed works, the platforms described above put more attention on images than videos, and serve for general tagging/clustering without special consideration about travel media characteristics. The goal of our work is to appropriately utilize characteristics derived from journeys and develop functions well matching with user's requirement.

3. Travel Media Management

3.1 Representative Selection

To efficiently present or manage photo collections, small amounts of representative photos that compactly present massive data are necessary. For example, when sharing photos on web albums, users often select the most canonical view of a scenic spot or the most important landmark to provide album visitors a way of grasping a folder of photos at a glance. Travelers used to capture the most canonical view several times, especially when many people in the same group tour want to take photos with it. In our previous work [5], we adopt a near-duplicate detection technique [10] to find near-duplicate photos in a set of photos taken at the same scenic spot. Duplicate properties between photos are described as a graph (the top part of Figure 3(a)), and then we exploit a social network analysis technique to discover the most centric photo (Figure 3(b)). This process is applied to each scenic spot so that a compact representation of the whole journey, in terms of representative photos, can be made.

Two assumptions were implicitly made in [5]: 1) the most important landmark or building are captured more than one time, and 2) photos containing the same object are successfully detected as near-duplicate. The first assumption is dependent on human's photographing habit, and is not true for every traveler. Although we don't limit to any specific near-duplicate detection technique, to our best knowledge, none of the state-of-the-art near-duplicate detection methods have perfect detection performance. Both factors destroy the methods proposed in [5]. To complement the shortage, in this article we further compare images based on part-based matching. Given a set of photos taken at the same scenic spot, we detect feature points on images and then quantize them into visual words. A visual word histogram is then collected to represent each image. Also motivated by social network analysis, the one that is most similar to all others is considered to be the most centric (important) role.

For an image I_i, we calculate the distance between it to another image I_j based on visual word histograms: $d(I_i, I_j) = \|H_i - H_j\|_2$, where H_i and H_j are visual word histograms for I_i and I_j, respectively. We then sum up all distances between I_i to all other images as $D_i = \sum_{j \setminus i} d(I_i, I_j)$. The accumulated distances are calculated for every image in the given photo set. The rank of representative is finally determined by sorting the accumulated distances in ascending order. The image that has the minimal accumulated distance to others is the most canonical image. For the image I_i, we denote its corresponding rank of representative as r_i^v, in which the superscript v denotes that this rank is determined based on visual word histogram.

The same ranking approach is adopted to rank images based on degree of centrality calculated in [5]. For the image I_i, the corresponding rank of representative is denoted as r_i^u, in which the superscript u means that this rank is determined based on near-duplicate detection. Given the set of photos $P = \{I_1, I_2, ..., I_N\}$, two ranks are combined to determine the rank of representative r_i for any image I_i:

$$r_i = (1-w)r_i^u + wr_i^v, \tag{1}$$

where the parameter $w = [0,1]$ controls the impact between two ranking methods. If some near-duplicate photos can be found in this photo set, the rank r_i^v serves as an enhancement factor, and the weight w is set as 0.3. If no near-duplicate photo can be found (due to miss detection of the near-duplicate detection module or no duplicate photos in fact), the rank r_i^v determines everything, and the weight w is set as 1.

Corresponding to the issues in Table 1, this module works mainly for photos, though the same technique can be applied to find representative keyframes in travel videos. It addresses the "what" and "where" problems, because canonical views or objects are found. It mainly contributes to efficient browsing, although it may help improve efficiency of annotation as well. Only visual modality (photo or video) is considered in processing, and various access manners as in Figure 3(c) are benefited by this module.

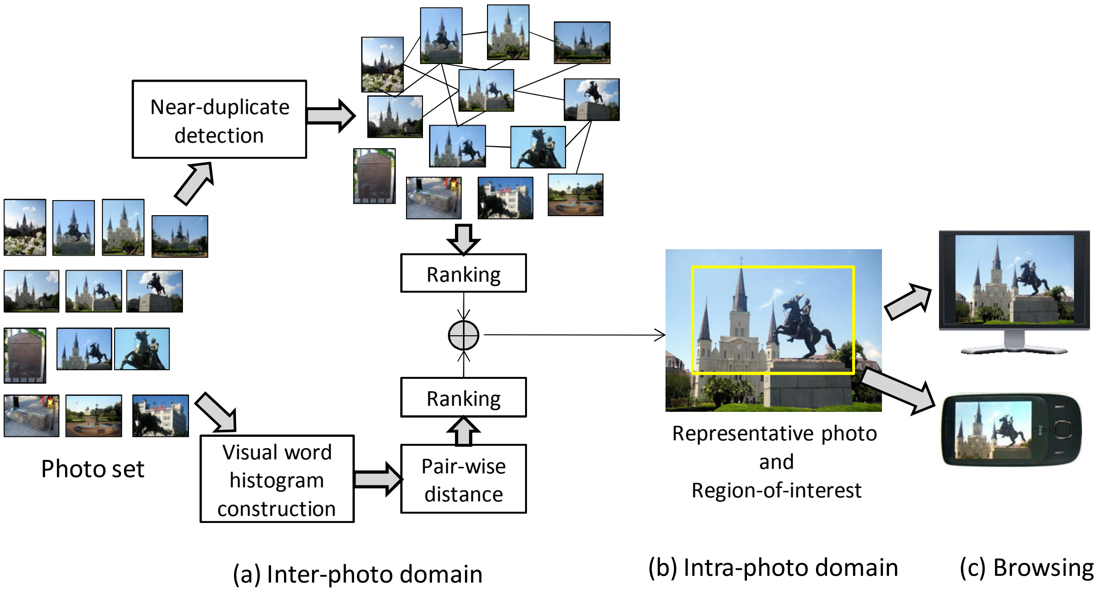

Figure 3. Representative selection in the inter-photo domain and ROI determination in the intra-photo domain.

3.2 ROI determination

After selecting the representative photo, we take a further step to find the most representative region in it. In this system, we don't conduct this issue by a saliency-based approach [11], which constructs a saliency map and determines one or more ROIs by finding the regions that cover the largest saliency values [12]. Although the saliency-based approach is designed according to characteristics of the human vision system, it doesn't account for semantic meanings of objects. To tackle with travel photos containing buildings, landmarks, or statues, we take advantage of SIFT-based (Scale-Invariant Feature Transform) matching [13] and important visual words to determine ROIs. In our previous work [5], spatial information of features used to identify the representative photo as near-duplicate to others is adopted to find an ROI. Because artificial objects that have specific geometry often give rise to clear orientation information, matched SIFT feature points are usually located inside or on the surface of objects. In [5], we showed that promising ROIs can be obtained if the near-duplicate detection module works well. However, as we mentioned in Section 3.1, the near-duplicate detection module would fail if duplicate objects have significantly different appearances. In this case, the ROI determination method proposed in [5] would fail, either.

In Section 3.1, there are two cases in representative selection: 1) selection based on both near-duplication detection and visual word histogram; 2) selection simply based on visual word histogram. In the first case, we determine ROI as the method in [5], due to its promising results and higher robustness of near-duplicate properties (if they can be found). In this article, we further develop an ROI determination method corresponding to the second case.

Assume that M visual words $\{w_1, w_2, ..., w_M\}$ are used to characterize images. If the representative photo I_r is selected solely based on visual word histogram, we evaluate the importance value of the visual word w_i with respective to this image as

$$t_i = \frac{n_{w_i}}{n_{I_r}} \log \frac{N}{N_{w_i}}, \tag{2}$$

where n_{w_i} denotes the number of feature points in I_r that are quantized into w_i, and n_{I_r} denotes the total number of feature points in I_r. The value N denotes the number of photos in this scenic spot, and the value N_{w_i} denotes the number of photos that contain feature points quantized into w_i. To determine an ROI in I_r, we identify feature points in I_r that are quantized into visual words in the set T, which contains w_i with the ten largest importance values:

$$F = \{f_j | Q(f_j) = w_k \wedge w_k \in T\}. \tag{3}$$

The minimum bounding box that covers all features in F is then found to form the ROI. With this newly proposed method, the limitation of successful near-duplication detection in [5] is removed.

Based on two methods described above, we determine the most important region in the intra-photo domain, which benefits browsing experience on resolution-limited devices as shown in Figure 3(c). For the processed modalities discussed in Table 1, this module is applied to photos only. Although the same method can be used to find ROIs for video keyframes, bad quality video frames diminish the values of ROI displaying. ROIs address the "what" and "where" problem, which is the main characteristic of the proposed module superior to conventional saliency-based approaches. It obviously brings advantages to browsing on various devices, and no cross-media correlation is used to determine ROIs.

3.3 Face clustering

To provide a face-name association function, it's straightforward to consider face recognition. However, poor lighting and side-view faces drive significant challenges to recognition processes. Many practical systems such as Google Picasa [6] instead make a compromise that they first clusters faces of the same individual together, and then provide a friendly interface to assist face tagging. In contrast to face recognition, face clustering techniques answer "which faces present the same individual?" rather than "who is this individual?"

We previously proposed a face clustering method to determine whether two faces present the same individual [20]. Two ideas form the foundation of this method: 1) part-based methods that extract features from facial areas are superior to methods using global features, when faces are varied in poses [35]. 2) We do not describe characteristics of faces, but describe matching situations between faces. Faces are divided into three regions, which respectively represent the part of eyebrows and eyes, the part of nose and cheek, and the part of mouth and chin. Pairs of faces are matched based on SIFT descriptors. To systematically describe matching situations, we transform them into the so-called visual sentences, and respectively construct a visual language model (VLM) M_1 [15] describing matching situations between faces of the same individuals, and a model M_2 describing matching situations between two distinct individuals. Given the visual sentence $VS(f_p, f_q) = (v_1 v_2 v_3)$ representing the matching situation between two faces f_p and f_q, we determine whether two faces are similar as:

$$f_p \text{ and } f_q \text{ are } \begin{cases} \text{similar} & \text{if } p(VS(f_p, f_q)|M_1) \geq p(VS(f_p, f_q)|M_2), \\ \text{not similar} & \text{otherwise.} \end{cases} \tag{4}$$

This method has been demonstrated to have promising performance when feature-based matching can be found in face pairs [20]. However, most facial areas are smooth and few features can be found. It's often the case that no features can be matched between faces of the same individual, especially when they are in significantly different poses. The original matching criterion in [13] considers the ratio between the best match and the second-best match. To prevent insufficient feature matching, we have to relax this constraint but appropriately remain robustness of part-based matching. In this work, we extend feature matching based on the idea proposed in [36], followed by describing matching situations by VLMs.

Assume that two faces f_p and f_q are represented as sets of SIFT feature points $\{t_1^p...t_M^p\}$ and $\{t_1^q...t_N^q\}$, respectively. For a point t_i^p in f_p, the feature t_j^q in f_q is claimed as matching with t_i^p if the Euclidean distance from t_i^p to t_j^q is the minimum of all distances between t_i^p to any feature point in f_q. This scheme enforces each feature point in f_p to match with one of features in f_q, and therefore generates many false matches. Two constraints are applied to eliminate inappropriate matchings – geometric constraint and unique-match constraint. In [36], the normalized spatial distance is calculated, and the matched feature points with distance larger than a threshold are eliminated. Because we have detected the locations of eyes, noses, and mouths on faces and divided them into three regions, matched feature points that are located in different regions are eliminated. As for the unique-match constraint, we apply two-way assignment to guarantee that only t_j^q can be matched with t_i^p, and only t_i^p can be matched with t_j^q.

Figure 4 shows three sample results of feature matchings based on criteria in [13] and the newly proposed one. We clearly see that conventional matching scheme has significantly fewer matching between faces. On the other hand, our method has more matches, though some of them are more false matches. Based on visual language models, we prefer more matches to facilitate effective representation of visual sentences.

Corresponding to Table 1, this module is concerned with photos and videos, although the function is not emphasized for videos right now. This module addresses the "who" problem, which often plays the most important role in managing and browsing travel media. It facilitates development of annotation, browsing, and retrieval, and only the visual modality is considered in processing. Finally, accessing with different devices can be benefited by this module.

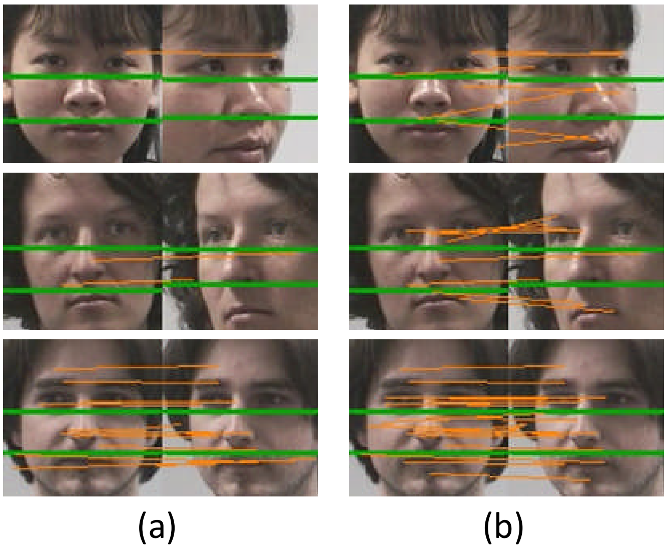

Figure 4. (a) Feature matching based on Lowe's criteria [13], and (b) based on the method modified from [36].

3.4 Video Scene detection

Because most travelers don't capture videos with a tripod and don't have good photography skills, videos often suffer poor lighting conditions and severe camera shaking. Moreover, people may be excited to capture large amounts of videos, which are often too lengthy and include many things that are thought meaningless afterward. Therefore, a systematic process to facilitate effective browsing is desired for most amateur photographers and travelers. One of the most fundamental processes is shot and scene change detection. This topic has been widely studied in the past two decades and is deemed being solved (especially shot change detection) in news, sports, movies, and other structured videos [17]. However, the characteristics of travel videos give rise to new challenges for scene detection. A scene in travel videos is defined as a set of shots that contains content captured in the same scenic spot. It doesn't make much sense to assume that content taken in the same scenic spot would have similar visual appearance because poor lighting and unstable motion make appearance significantly different even the same things were captured.

The essential idea of developing this module is to exploit cross-media correlation between photos and videos [18]. Finding scene boundaries for videos is harder, but determining scene boundaries for photos is much easier. Because the content conveyed in photos and videos is often correlated, we first determine photo scene boundaries, discover the correlation between photos and videos, and then determine video scene boundaries with the clues of photo scene boundaries.

Scene detection for photos can be easily accomplished by performing time-based clustering [9]. For videos, we first segment videos into shots, and extract appropriate number of keyframes for each shot by the global k-means algorithms [19]. We then extract SIFT feature points [13] on both photos and keyframes, quantize feature points into visual words, and then construct a visual word histogram for each photo and keyframe. Two sequences of visual word histograms, which are constructed from two sequences of photos and keyframes sorted in temporal order, are then generated. Note that histograms of visual words [14] are invariant to scale, orientation, and some degree of viewpoint changes. By this process, the problem of finding the correlation between photos and videos is transformed into finding correspondence between two visual word histogram sequences, which can be modeled as an approximate sequence matching problem and can be solved by a dynamic programming approach [18].

Because scene boundaries of photos have been determined in advance, we can infer boundaries of video scenes by exploiting the correspondence between photos and video keyframes, as shown in Figure 5. Corresponding to Table 1, this module is a classical example of multimodal processing. It addresses the "where" problem, because different scenes (scenic spots) are discriminated. It facilitates the development of annotation, browsing, and other editing functions. The major contribution of this module is that the correlation between photos and videos is elaborately used. Recently, results of video scene detection and extended developments are accessed by PCs.

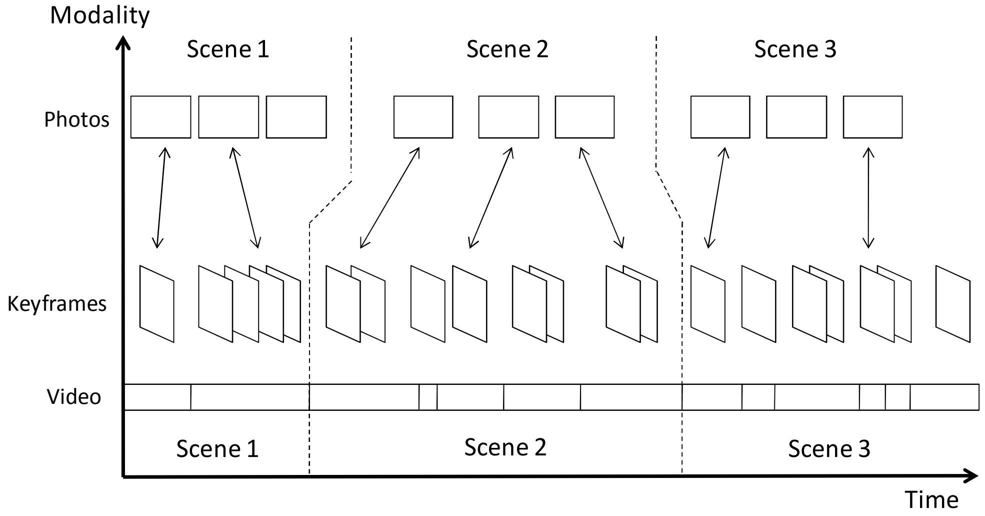

Figure 5. Cross-media correlation between photos and the corresponding video.

3.5 Corporation of Modules

As an image processing software that provides several modules, such as cropping, resizing, and de-blur, the proposed modules serve as a tool set for travel media management. Unlike homogeneous inputs, i.e. image, to an image processing software, our proposed modules process heterogeneous inputs (i.e. photo, video, and time information). Not every data item can be processed by all proposed modules. On the other hand, the proposed modules don't process the same entity as a pipeline. As shown in Figure 2, different modules work as different transformations that allow users to view part of the data corpus from different perspectives.

Although there is no necessary causality between the proposed modules, results of some modules may be beneficial to other modules. For example, a travel video can be first segmented into scenes, followed by the representative selection modules to find the most canonical keyframe from each scene. The face clustering module can also works on video frames. In addition to linear browsing and scene-based nonlinear browsing, we can browse videos by scenes that contain the same person.

Basically ROI determination approaches can be categorized into two types: bottom-up and top-down. The method described in Section 3.2 is a bottom-up approach, because only spatial information of visual features is used. The ROI determination method is also limited to landmark or building. With the help of face clustering, we can extend the proposed ROI determination method to a top-down approach, by finding important faces in photos. After face clustering, the clusters that consist of many faces provide hints of important members in this journey. We can find regions that cover faces of important members to be ROIs.

4. Evaluation

- Representative Selection and ROI Determination

We collect photos captured by amateur photographers in their journeys, with totally 1024 photos in 52 scenic spots. Resolution of each photo is normalized into 320×240 or 240×320. Note that "photos taken at the same scenic spot" doesn't mean all photos include the main landmark or view. Many irrelevant objects, such as pedestrians or stores, would be captured.

Due to space limitation, we just show six sample results in Figure 6. These photos are automatically selected from a collection of photos taken at the same scenic spot, and the bounding boxes illustrate the determined ROI. The bottom row of this figure shows some photos taken in "Statue of Liberty." By comparing the original collection with the selected representative, we can clearly see that the inter-photo processing effectively picks appropriate one to represent the collection.

To evaluate performance of representative selection, we asked seven observers to judge each photo that is determined near-duplicate to someone else, or is determined as the top-ranked photo according to similarity analysis based on visual word histogram. A brief guideline for giving scores is shown in Table 2. For each photo, the degree of representativeness is calculated by averaging the scores given by observers. Table 3 shows the results based on [5], solely based on visual words, and combination of two methods, respectively. We see that the best performance can be obtained by combining the influence of two features. The average score 3.98 means that we can find the most canonical landmark or view in almost every scenic spot, which is largely better than the result in [5] (3.63).

Table 2. Guidelines for scoring a photo.

Score	Description
5	This photo shows the most representative object you know for this scenic spot.
4	Although it's not good in shooting angles or lighting conditions, the most representative object shows on the image.
3	This photo doesn't contain the most representative object, but some other buildings or specific objects are shown.
2	There are objects without specific topic in this photo, e.g. a sign, or the quality of the photo is bad.
1	You totally don't know the purpose of this photo, e.g. crowd or grass.

Table 3. Average scoring results of representative selection.

	[5]	VW	[5]+VW
Average score	3.63	3.30	3.98

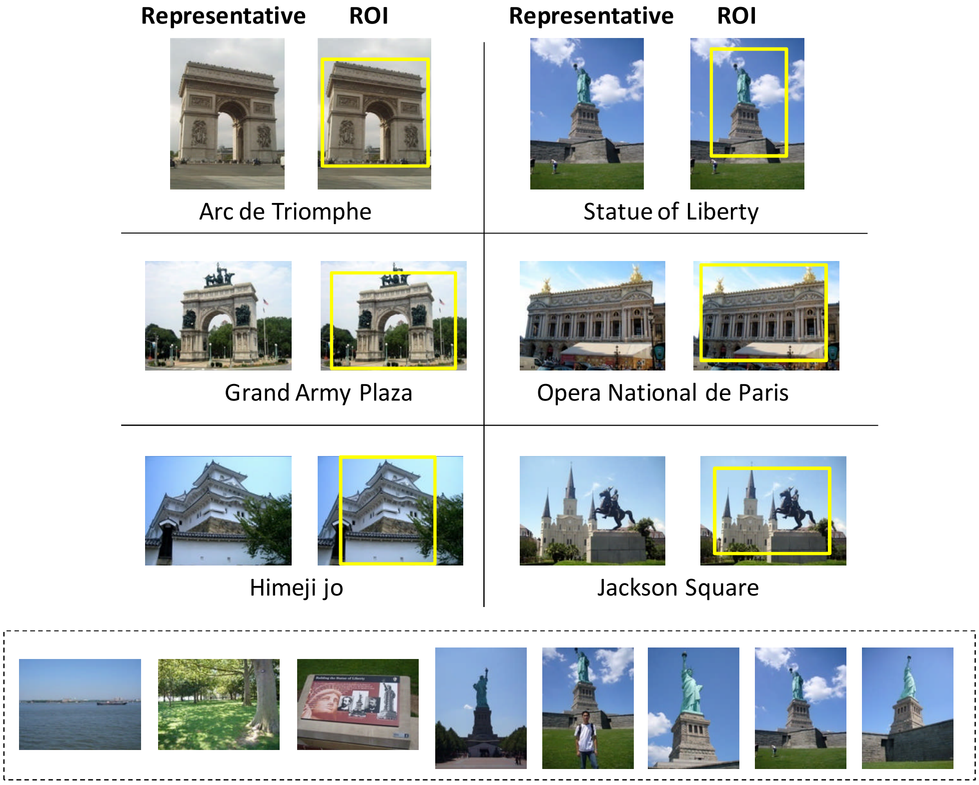

Figure 6. Results of representative selection and ROI determination.

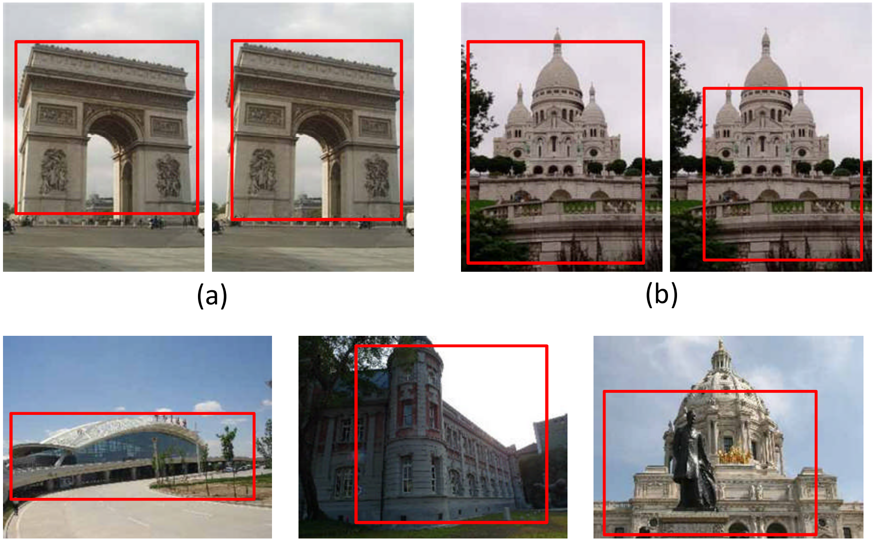

Figure 7. Performance comparison of ROI determination.

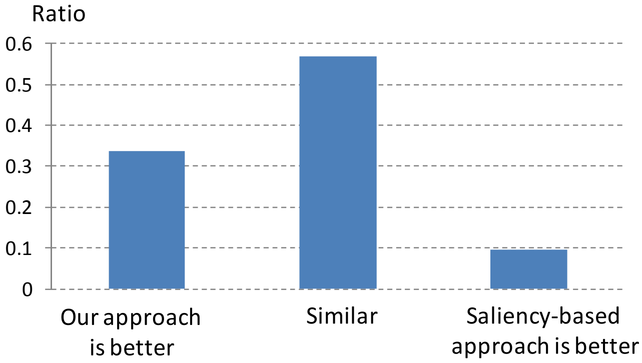

Figure 8. Performance comparison of ROI determination, in terms of subjective preference.

Figure 7 shows performance variations of ROI determination based on near-duplicate information [5] and visual words. The left side of Figure 7(a) shows the bounding box determined by near-duplication information, and the right side shows the bounding box covering the top-ranked visual words. We see that both methods work well in this case. In Figure 7(b), the ROI determined by near-duplicate information is better. Top of the building is not covered by the box in the right side, which shows that solely exploiting importance of visual words sometimes doesn't catch semantics. Figure 7(c) shows three ROIs that cannot be successfully detected by the method in [5], and thus only visual word information can be used. Because both quality of visual words and distributions of feature points affect performance of this approach, we see performance variations in different cases. We have satisfactory performance in the left and middle images. In the right image of Figure 7(c), fewer feature points are on the tip of the building, and thus worse ROI determination result is obtained.

Figure 8 shows results of subjective evaluation on ROI determination. Observers were asked to judge ROIs in each scenic spot by telling (1) the ROI determined by our approach is better; (2) by saliency-based method [11] is better; or (3) similar. Overall, observers think that our determined ROIs are better in 33.6% of scenic spots, while the saliency-based methods works better only in 9.6% of scenic spots. ROIs in over half of scenic spots (56.8%) are considered similar, since ROIs are similar in either way when the determined ROI occupies most area of the original photo. Relative to the saliency-based approach, our method more elaborately finds contour of important objects and determines more accurate ROIs.

- Face Clustering

To evaluate clustering performance, we conducted experiments consisting of 150 rounds (called *Exp1*). In each round, we randomly select five individuals from the FERET database [16], and randomly select four images from each individual's pool. The *Exp1* experiments adopt the matching scheme proposed in [13]. Based on the same datasets, 150 rounds are also conducted based on the matching scheme in [36], called *Exp2*. To evaluate performance in consumer photos, respectively based on two methods, *Exp3* and *Exp4* are conducted based on photos collected from twenty-one journeys including 105 individuals with totally 1071 face images.

Figure 9 shows the average clustering accuracy in these experiments, in which the I-shaped bar indicates standard deviations. Comparing results of *Exp1* with *Exp2*, we verify that the matching scheme with looser constraints improves performance. More matches can be found, at the same time matching robustness is maintained, and thus visual language models work better. Performance of clustering faces in consumer photos is worse (*Exp 3* and *Exp4*). It is reasonable because faces in consumer photos have significant pose, lighting, and expression variations. While the performance of *Exp4* is slightly better than that in *Exp3*, clustering performance in *Exp4* is much stable, which again demonstrates impact of the new matching scheme.

The potential drawback of the new matching scheme is false matching. Falsely matched features would be quantized into incorrect visual words, and thus incorrectly describe matching situations. To investigate this issue, we experiment two matching schemes based on the data used in *Exp1*, in which each face image is scaled into two different sizes, i.e. 75×75 and 150×150. Table 4 shows performance variations for faces of different sizes. When size of face decreases, fewer false matches are obtained, and clustering based on the matching scheme modified from [36] achieves clear performance gain. On the other hand, clustering based on the conventional matching scheme has similar performance in both cases.

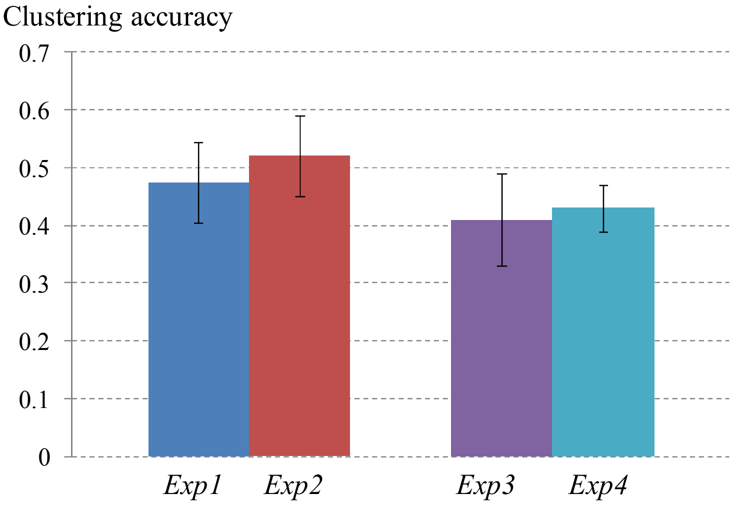

Figure 9. Clustering accuracies in different datasets.

Table 4. Performance variations based on different sizes of images.

Methods	[13]		Modified from [36]	
Image sizes	150 × 150	75 × 75	150 × 150	75 × 75
Avg. accuracy	0.48	0.475	0.48	0.52

To further justify the proposed method, we compare it with a classical eigenface approach. In the eigenface approach, we project faces into the eigenspace, transform each face as a vector that represents weights on each eigenface, and then perform k-means clustering based on these vectors [20]. A face clustering system should not only achieve high clustering accuracy, but also limit over-clustering. When the targeted number of cluster increases, clustering accuracy generally (but not strictly) increases as well. In the extreme case, we can achieve perfect clustering result if one cluster contains only one face. From this perspective, we find the least targeted number of clusters needed to make clustering accuracy no less than 80%. Figure 10 shows ten sample results from subsets of *Exp2*. From the fifth result in Figure 10(a), for example, the VLM method is able to achieve at least 80% accuracy when the targeted number of cluster is set as eight (K=8). If we also group faces into eight clusters by the eigenface approach, only 66.7% accuracy can be achieved. In this figure, the VLM method achieves better performance in most subsets.

On the other hand, we evaluate the numbers of clusters least required for a method to achieve 80% accuracy. Recall that all subsets in Figure 10 actually contain only five individuals. The VLM method needs to over-cluster the fifth subset into eight clusters to achieve 80% accuracy, while the eigenface approach needs ten clusters to do so. Smaller number of the needed targeted clusters means slighter over-clustering. A ratio is calculated to quantify this effect: $R = |F_{eig}|/|F_{VLM}|$, where $|F_{eig}|$ and

$|F_{VLM}|$ are the least numbers of clusters needed by the eigenface approach and the VLM method to achieve 80% accuracy, respectively. Based on *Exp2*, we obtain the average ratio $\bar{R}$=1.45, which means that the eigenface approach over-clusters 1.45 times than the proposed VLM approach. This shows the proposed method more effectively clusters faces when a targeted accuracy is required.

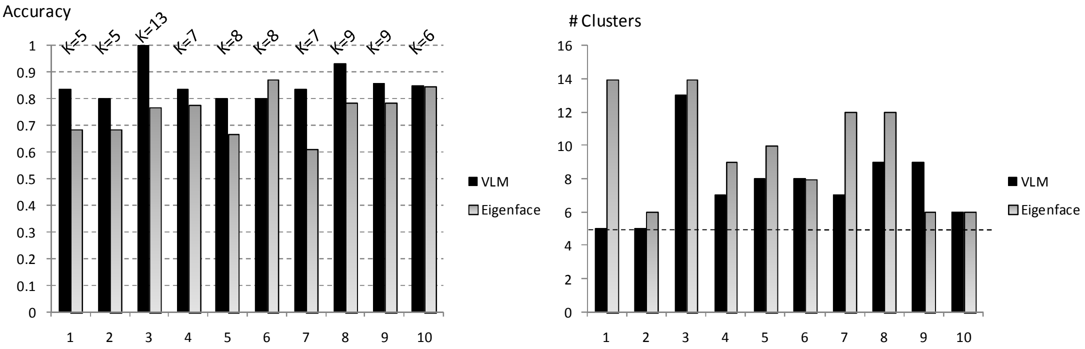

Figure 10. (a) Accuracies of two methods when a targeted number of cluster is set; (b) Number of cluster least required to achieve 80% accuracy.

- Video Scene Detection

Five data sets are used in video scene detection. Each dataset includes a video clip and a set of corresponding photos. Lengths of these video clips range from eight to fifteen minutes, and there are 20 to 126 corresponding photos. We first evaluate the performance of keyframe selection, by comparing the method [21] adopting the global k-means algorithm [19] with a naïve method. If the global k-means method determines that there are four keyframes in a video shot, the naïve method uniformly selects four frames for this shot. In this experiment, we manually define ground truths, and calculate precision of keyframe selection. Overall, the global k-means method achieves 0.76 precision, while the naïve method achieves 0.57 precision. This result confirms the superiority of this method claimed in [21].

To evaluate performance of scene detection, we consider overlaps between the detected video scenes and the ground truths, and evaluate performance in terms of the purity value [22]. Given the ground truth of scenes $S=\{(s_1,\Delta t_1),...,(s_{Ng},\Delta t_{Ng})\}$ and the results of scene detection $S^*=\{(s_1^*,\Delta t_1^*),...,(s_{Nv}^*,\Delta t_{Nv}^*)\}$, a purity value ρ is defined as

$$\rho=\left(\sum_{i=1}^{Ng}\frac{\tau(s_i)}{T}\sum_{j=1}^{Nv}\frac{\tau^2(s_i,s_j^*)}{\tau^2(s_i)}\right)\cdot\left(\sum_{j=1}^{Nv}\frac{\tau(s_j^*)}{T}\sum_{i=1}^{Ng}\frac{\tau^2(s_i,s_j^*)}{\tau^2(s_j^*)}\right), \tag{5}$$

where $\tau(s_i,s_j^*)$ is the length of overlap between the scenes s_i and s_j^*, $\tau(s_i)$ is the length of the scene s_i, and T is the total length of all scenes. The values of "length" are measured by number of shots. The first parenthesis indicates the fraction of the current evaluated scene, and the second one indicates how much a given scene is split into smaller scenes. A purity value ranges from 0 to 1. Larger purity value means that the result is closer to the ground truth.

We compare performance in terms of purity based on four approaches:

- The proposed sequence matching approach based on visual word histograms.
- The proposed sequence matching approach with features of 16-bin HSV histograms.
- The proposed sequence matching approach with features of the concatenation of a visual word histogram and an HSV histogram.
- A naïve method, in which timestamps of video scene boundaries are proportional to corresponding timestamps of photo scene boundaries. For example, if the second photo scene starts at the one third of duration of the photo sequence, the second video scene also starts at the one third of time duration of the video sequence.

Figure 11 shows detection performance of four different approaches. Visual word histograms have better performance than HSV histograms in Videos 1 and 5. However, color information is eliminated in extracting SIFT feature points, and thus HSV histograms work better in Videos 2 and 4. The best scene detection performance is achieved when visual words are combined with color information. The average purity value is 0.95, which is very promising in scene detection for travel media.

To show the proposed method is more appropriate to be applied in travel videos, we compare it with the method proposed in [21]. One of the major challenges in scene detection is the over-segmentation problem. We measure this effect in two methods and list the results in Table 5. In each cell of this table, the value (*m*, *n*) denotes that the corresponding scene is segmented into *m* and *n* scenes, by the method in [21] and our method, respectively. For example, in the second columns for Video 1, (4,1) means that the second scene in Video 1 is segmented into four scenes by the method in [21], and only one scene by ours. From these results, we see the effect of over-segmentation is severe in the results of [21], which is originally designed for structured videos. Our method achieves perfect results for Videos 2 and 4, and also has much better results in other data.

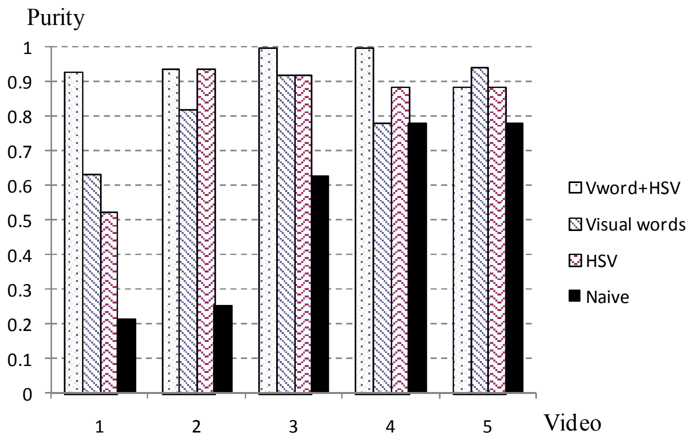

Figure 11. Performance based on four different scene detection approaches.

Table 5. Over-segmentation situations in different videos.

	scene 1	scene 2	scene 3	scene 4	scene 5	scene 6	Overall
V1	(1,1)	(4,1)	(7,2)	(3,1)	(9,2)	(3,1)	(27,8)
V2	(6,1)	(3,1)	(1,1)	(1,1)			(11,4)
V3	(4,2)	(2,2)	(3,1)				(9,5)
V4	(1,1)	(1,1)	(1,1)	(3,1)	(2,1)		(8,5)
V5	(1,1)	(2,2)	(1,1)	(5,2)	(1,1)		(10,7)

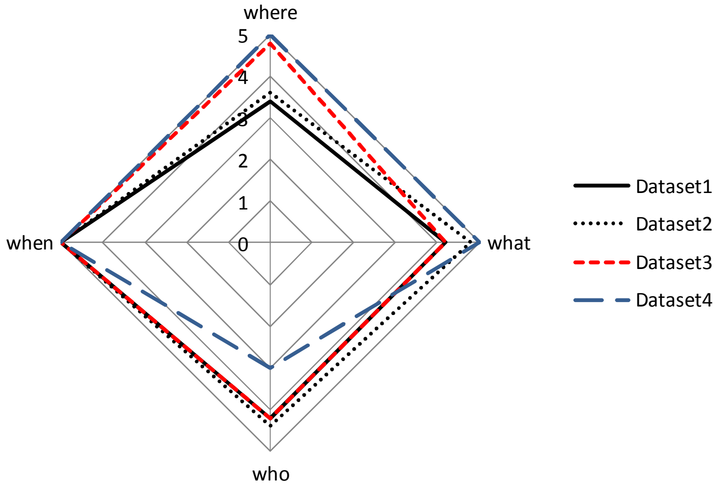

Figure 12. Performance of satisfaction in terms of where, what, who, and when.

- Overall User Study

To evaluate overall impact of the proposed system, we collect photos in four journeys and respectively conduct representative selection, ROI determination, and face clustering. Based on these results, ten observers were asked to judge their satisfaction in terms of four axes described in Figure 2. Note that these observers are unaware of owners of these datasets, and don't know where they go or who they are in advance.

According to how easily the observer can recognize where, what, and who were captured from the selected representative and ROI, he/she can give a score from 1 to 5, in which a larger score means higher satisfaction.

Figure 12 shows the overall results. Because we display timestamps of data, observers easily recognize when this data were captured. The good performance in the *what* axis indicates that appropriate representative photos are selected and appropriate regions are selected as ROIs. Comparing to the *what* information, observers feel tougher to recognize *where* the selected representative photos were captured. How easily an observer recognizes the presented place may depend on his life or travel experience. We have especially worse performance in the *who* axis for Dataset3 because half of faces in this dataset appears only once, and the face clustering method erroneously cluster different individuals into the same cluster. We averagely obtain 4.2, 4.55, 3.95, and 5 for *where*, *what*, *who*, and *when*, respectively. The relatively less satisfaction for *who* information seems to indicate that face clustering in unconstrained images is still a very challenging problem, or human beings are most sensitive to analytical results of human faces.

5. Conclusion

We first describe the requirements of designing a travel media management in terms of processing modalities, access facets, active functions, correlation between different modalities, and access manners. Corresponding to these issues, we develop a system that mainly addresses browsing and management requirements. The representative selection module exploits near-duplicate detection or visual word analysis to transform relationships between photos into a graph. By analyzing graph structure, the most representative photo is determined. Spatial locality characteristics of SIFT matched points and statistics-based importance of visual words are further exploited to find the most important region in a photo. In contrast to saliency-based methods, this approach is able to find semantically meaningful objects. Visual language models are introduced to characterize SIFT matching situations between faces. Matched feature points conceptually constitute a visual sentence, and this representation is then used to identify whether two faces present the same individual. For video scene detection, finding correlation between videos and photos is transformed as an approximate sequence matching problem, which is then solved by a dynamic programming strategy. With the clues of photo scene boundaries, video scene boundaries are determined with the help of cross-media correlation.

Though some issues about cooperation between different modules are provided in this article, more elaborate investigation on jointly processing heterogeneous data in travel media is needed in the future. Furthermore, intelligent techniques of analyzing more media captured in journeys, such as audio, GPS information, and maps, remains to be studied to construct a more comprehensive travel media management system.

Acknowledgement

The authors would like to thank Chia-Hung Lin, Ya-Lin Lee, and Che-Cheng Lin for their efforts on preliminary results of this work. The authors would also like to thank anonymous reviewers for giving valuable comments. The work was partially supported by the National Science Council of Taiwan, Republic of China under research contract NSC 99-2221-E-194-036 and NSC 98-2221-E-194-056.

Reference

[1] C.G.M. Snoek, B. Huurnink, L. Hollink, M. de Rijke, G. Schreiber, and M. Worring, “Adding semantics to detectors for video retrieval,” IEEE Transactions on Multimedia, vol. 9, no. 5, pp. 975-986, 2007.

[2] M. Naphade, J.R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. “Large-scale concept ontology for multimedia,” IEEE Multimedia, pp. 86-91, 2006.

[3] A. Ekin, A.M. Tekalp, and R. Mehrota. “Automatic soccer video analysis and summarization,” IEEE Transactions on Image Process, vol. 12, no. 7, pp. 796–807, 2003.

[4] D.A. Sadlier and N.E. O’connor. “Event detection in field sports videos using audio-visual features and a support vector machine, “ IEEE Transactions on Circuits and Systems for Video Technology, vol. 15, no. 10, pp. 1225–1233, 2005.

[5] W.-T. Chu and C.-H. Lin, “Consumer photo management and browsing facilitated by near-duplicate detection with feature filtering,” Journal of Visual Communication and Image Representation, vol. 21, no. 3, pp. 256-268, 2010.

[6] Google Picasa, http://picasa.google.com/index.html

[7] A. Girgensohn and J. Boreczky, “Time-constrained keyframe selection technique,” Multimedia Tools and Applications, vol. 11, no. 3, pp. 347-358, 2000.

[8] TREC Video Retrieval Evaluation, http://www-nlpir.nist.gov/projects/trecvid/

[9] J.C. Platt, M. Czerwinski, and B.A. Field, “PhotoTOC: automating clustering for browsing personal photographs,” Proceedings of IEEE Pacific Rim Conference on Multimedia, pp. 6-10, 2003.

[10] W.-L. Zhao, C.-W., Ngo, H.-K. Tan, and X. Wu, “Near-duplicate keyframe identification with interest point matching and pattern learning,” IEEE Transactions on Multimedia, vol. 9, no. 5, pp. 1037-1048, 2007.

[11] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 11, pp. 1254–1259, 1998.

[12] V. Gopalakrishnan, Y. Hu, and D. Rajan, “Salient region detection by modeling distributions of color and orientation,” IEEE Transactions on Multimedia, vol. 11, no. 5, pp. 892-905, 2009.

[13] D. Lowe, “Distinctive image features from scale-invariant keypoints,” International Journal of Computer Vision, vol. 60, no. 2, pp. 91-110, 2004.

[14] J. Sivic and A. Zisserman, “Efficient visual search of videos cast as text retrieval,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 4, pp. 591-606, 2009.

[15] L. Wu, M. Li, Z. Li, W.-Y. Ma, and N. Yu, “Visual language modeling for image classification,” Proceedings of ACM International Workshop on Multimedia Information Retrieval, pp. 115-124, 2007.

[16] The color FERET database, http://face.nist.gov/colorferet/

[17] A. Hanjalic, “Shot-boundary detection: unraveled and resolved?” IEEE Transactions on Circuits and Systems for Video Technology, vol. 12, no. 2, pp. 90–105, 2002.

[18] W.-T. Chu, C.-C. Lin, and J.-Y. Yu, “Using cross-media correlation for scene detection in travel videos,” Proceedings of ACM International Conference on Image and Video Retrieval, 2009.

[19] A. Likas, N. Vlassis, and J.J. Verbeek, “The global k-means clustering algorithm,” Pattern Recognition, vol. 36, 451-461, 2003.

[20] W.-T. Chu, Y.-L. Lee, and J.-Y. Yu, “Visual language model for face clustering in consumer photos,” Proceedings of ACM Multimedia Conference, pp. 625-628, 2009.

[21] V. Chasanis, A. Likas, and N. Galatsanos, “Scene detection in videos using shot clustering and symbolic sequence segmentation,” Proceedings of IEEE International Conference on Multimedia Signal Processing, pp. 187-190, 2007.

[22] A. Vinciarelli and S. Favre, “Broadcast news story segmentation using social network analysis and hidden Markov models,” Proceedings of ACM Multimedia, 261-264, 2007.

[23] M. Yeung and B.-L. Yeo, “Segmentation of video by clustering and graph analysis,” Computer Vision and Image Understanding, vol. 71, no. 1, pp. 94-109, 1998.

[24] A. Hanjalic, R.L. Lagendijk, and J. Biemond, “Automated highlevel movie segmentation for advanced video retrieval system,” IEEE Transactions on Circuits and System for Video Technology, vol. 9, no. 4, pp. 580-588, 1999.

[25] H. Sundaram and S.-F. Chang, "Determining computable scenes in films and their structures using audio-visual memory models," Proceedings of ACM Multimedia, pp. 95-104, 2000.

[26] S. Satoh, Y. Nakamura, and T. Kanade, "Name-It: naming and detecting faces in news videos," IEEE Multimedia, vol. 6, no. 1, pp. 22-35, 1999.

[27] P. The Pham, M.-F. Moens, and T. Tuytelaars, "Cross-media alignment of names and faces," IEEE Transactions on Multimedia, vol. 12, no. 1, pp. 13-27, 2010.

[28] L. Zhang, L. Chen, M. Li, and H.J. Zhang, "Automated annotation of human faces in family albums," Proceedings of ACM Multimedia, pp. 355-358, 2003.

[29] M. Zhao, Y.W. Teo, S. Liu, T.-S. Chua, and J. Ramesh, "Automatic person annotation of family photo album," Proceedings of ACM International Conference on Image and Video Retrieval, pp. 163-172, 2006.

[30] A.C. Gallagher and T. Chen, "Understanding image of groups of people," Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2009.

[31] J. Wang and T.-S. Chua, "A framework for video scene boundary detection," Proceedings of ACM Multimedia, pp. 243-246, 2002.

[32] J. Vendrig and M. Worring, "Systematic evaluation of logical story unit segmentation," IEEE Transactions on Multimedia, vol. 4, no. 4, pp. 492-499, 2002.

[33] H.-J. Zhang, J. Wu, D. Zhong, and S.W. Smoliar, An integrated system for content-based video retrieval and browsing," Pattern Recognition, vol. 30, no. 4, pp. 643-658, 1997.

[34] L.-C. Hsieh, K.-T. Chen, C.-H. Chiang, Y.-H. Yang, G.-L. Wu, C.-S. Ferng, H.-W. Hsueh, A. C.-R. Tsai, and W.H. Hsu, "Canonical image selection and efficient image graph construction for large-scale Flickr photos," Proceedings of ACM Multimedia, pp. 1121-1122, 2009.

[35] B. Heisele, P. Ho, J. Wu, and T. Poggio, "Face recognition: component-based versus global approaches," Computer Vision and Image Understanding, vol. 91, pp. 6-21, 2003.

[36] D. Ozkan and P. Duygulu, "Interesting faces: a graph-based approach for finding people in news," Pattern Recognition, vol. 43, pp. 1717-1735, 2010.

[37] Z. Rasheed and M. Shah, "Scene detection in Hollywood movies and tv shows," Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 343-348, 2003.

[38] J. Yuan and Y. Wu, "Spatial random partition for common visual pattern discovery," Proceedings of IEEE International Conference on Computer Vision, 2007.

[39] T.L. Berg, A.C. Berg, J. Edwards, M. Maire, R. White, Y.-W. Teh, E. Learned-Miller, and D.A. Forsyth, "Names and faces in the news," Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004.

[40] H. Liu and S. Yan, "Common visual pattern discovery via spatially coherent correspondences," Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2010.

[41] Y. Jing, S. Baluja, and H. Rowley, "Canonical image selection from the web," Proceedings of ACM International Conference on Image and Video Retrieval, pp. 280-287, 2007.

[42] J. Scott, Social network analysis: a handbook. Newbury Park, 1991.

[43] W.-H. Cheng, W.-T. Chu, and J.-L. Wu, "A visual attention based region-of-interest determination framework for video sequences," IEICE Transactions on Information and Systems Journal, vol. E-88D, no. 7, pp. 1578-1586, 2005.

[44] W. Jiang and A. Loui, "Semantic event detection for consumer photo and video collections," Proceedings of IEEE International Conference on Multimedia & Expo, pp. 313-316, 2008.

[45] M. Cooper, J. Foote, A. Girgensohn, and L. Wilcox, "Temporal event clustering for digital photo collections," ACM Transactions on Multimedia Computing, Communications and Applications, vol. 1, no. 3, pp. 269-288, 2005.

[46] J. Chen and S. Hibino, "Reminiscing view: event-based browsing of consumer's photo and video-clip collections," In Proceedings of IEEE International Symposium on Multimedia, pp. 23-30, 2008.

[47] W.-H. Cheng, Y.-Y. Chuang, Y.-T. Lin, C.-C. Hsieh, S.-Y. Fang, B.-Y. Chen, and J.-L. Wu, "Semantic analysis for automatic event recognition and segmentation of wedding ceremony videos," IEEE Transaction on Circuits and Systems for Video Technology, vol. 18, no. 11, pp. 1639-1650, 2008.

[48] M. Naaman, Y.J. Song, A. Paepcke, and H. Garcia-Molina, "Automatic organization for digital photographs with geographic coordinates," Proceedings of ACM/IEEE-CS joint conference on Digital libraries, pp. 53-62, 2004.

[49] S. Ahern, M. Naaman, R. Nair, and J. Yang, "World explorer: visualizing aggregate data from unstructured text in geo-referenced collections," Proceedings of ACM/IEEE-CS joint conference on Digital libraries, pp. 1-10, 2007.

[50] N. Snavely, S.M. Seitz, and R. Szeliski, "Photo tourism: exploring photo collections in 3d," Proceedings of ACM SIGGRAPH, pp. 835-846, 2006.

[51] Z. Stone, T. Zickler, and T. Darrell, "Toward large-scale face recognition using

social network context," Proceedings of the IEEE, vol. 98, no. 8, pp. 1408-1415, 2010.