Using Cross-Media Correlation for Scene Detection in Travel Videos

Wei-Ta Chu
Department of CSIE
National Chung Cheng University
Taiwan
wtchu@cs.ccu.edu.tw

Che-Cheng Lin
Department of CSIE
National Chung Cheng University
Taiwan
john72831@yahoo.com.tw

Jen-Yu Yu
Info. and Comm. Research Labs
Industrial Technology Research Inst.
Taiwan
KevinYu@itri.org.tw

ABSTRACT

Focusing on travel videos taken in uncontrolled environments and by amateur photographers, we exploit correlation between different modalities to facilitate effective travel video scene detection. Scenes in travel photos, i.e., content taken at the same scenic spot, can be easily determined by examining time information. For a travel video, we extract several keyframes for each video shot. Then, photos and keyframes are represented as a sequence of visual word histograms, respectively. Based on this representation, we transform scene detection into a sequence matching problem. After finding the best alignment between two sequences, we can determine scene boundaries in videos with the help of that in photos. We demonstrate that we averagely achieve a purity value of 0.95 if the proposed method is combined with conventional ones. We show that not only features of visual words aid in scene detection, but also cross-media correlation does.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing –*Indexing methods.* I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding – *Video analysis.*

General Terms

Algorithms, Experimentation, Performance.

Keywords

Cross-media correlation, scene detection, sequence matching, travel video.

1. INTRODUCTION

Large amounts of home videos have been created and disseminated due to the popularity of consumer appliances and low-cost video capturing and sharing technologies. People get used to record daily life and travel experience by digital cameras and camcorders. Although it's fun to lively capture the moments of your life, editing or browsing lengthy home videos, especially in the conditions of bad photography skills, is never a funny experience.

Massive home videos captured in uncontrolled environments are often suffered from annoying effects, such as overexposure/underexposure and hand shaking. In addition, there is no clear structure in home videos, which is different from widely-studied news and sports videos. Therefore, organization and analysis of unconstrained home videos pose great challenges.

Recently, relatively fewer researches have been conducted for home videos. In contrast to flourishing studies of multimedia information retrieval for news and sports videos, home video analysis still focuses on video segmentation, clustering, object detection, and summarization [1-8]. Development of a practical home video management system is still impeded by uncontrolled and high-variation content. In this paper, we focus on the videos captured in journeys, and address one of the most essential problems in video analysis – scene detection. Results of accurate scene detection provide important foundations for advanced analysis and applications.

It's worth to clearly define *a scene in travel videos*. Generally, a scene in a video is a cluster of video shots that correspond to the same semantic concept. During a journey, we visit scenic spots according to a travel plan, and take photos and videos sequentially. The most appropriate unit for managing travel media is by scenic spots. Therefore, in this work, a scene is a travel video means *a cluster of video shots that correspond to a scenic spot*.

Although scene detection is an age-old problem and has been studied in various media [9][10], the methods specially designed for structured videos often can't succeed for segmenting travel videos. For the methods that cluster video segments based on visual features, noises caused by bad lighting and hand shaking significantly degrade clustering performance. In this paper, we take advantage of the correlation between videos and photos captured in journeys and achieve accurate scene detection for unstrained travel videos.

Correlation between data in different modalities can be found in many cases. For example, there are often similar photos and video segments for the same news story [20]. During a journey, people often capture travel experience by still cameras and camcorders alternately. The content stored in photos and videos contain similar information, such as landmarks, thus there is close correlation between two modalities. Moreover, we can easily cluster photos based on some mature techniques, such as

CIVR'09, July 8–10, 2009, Santorini, Greece.

clustering based on photos' shooting time. With these properties, we can exploit cross-media correlation and the results of photo clustering to achieve accurate scene detection for travel videos. We design a matching scheme based on bags of visual words to find the correspondence between two modalities. The reported performance shows the superiority of the proposed scheme, and further confirms the benefits of utilizing multi-modality context information.

Contributions of this work are summarized as follows:

- We introduce the idea of using cross-media correlation to perform scene detection. Cross-media alignment based on approximate string matching has been designed to find the optimal matching between two modalities.
- We perform comprehensive study about the performance of the proposed scheme and conventional scene detection approaches.

The rest of this paper is organized as follows. Section 2 reviews related works. We describe the proposed cross-media alignment framework in Section 3. To represent photos and video clips, bags of visual words are extracted in Section 4. The scene detection task is transformed into an approximating sequence matching problem in Section 5. Section 6 givens experimental results and discussion, and Section 7 concludes this paper.

2. Related Work

Because there is no standard benchmark and evaluation metrics for home video analysis, studies in this field are diverse and rise from different perspectives. We will briefly review the works on home video structuring, automatic editing, browsing, and intention analysis. Then, studies especially about scene detection and cross-media correlation are reviewed as well.

Although there is no restriction in capturing home videos, Gatica-Perez et al. [1] cluster video shots based on visual similarity, duration, and temporal adjacency, and therefore find hierarchical structure of videos. On the basis of motion information, Pan and Ngo [2] decompose videos into snippets, which are then used to index home videos. For the purpose of automatic editing, temporal structure and music information are extracted, and subsets of video shots are selected to generate highlights [3] or MTV-style summaries [4]. Peng et al. [5] further take media aesthetics and editing theory into account to perform home video skimming. In [6], a system called Hyper-Hitchcock is developed to semi-automatically edit videos and to facilitate hyperlink properties. From the perspective of intention analysis on home videos, [7] and [8] model user intention for video repurposing and browsing.

Especially for scene detection, which is the main task of this paper, Yeung and Yeo [9] proposed a classical work called scene transition graph to describe relationships between video shots, and achieved scene detection by analyzing links in the graph. For movies and TV shows, Rasheed and Shah [10] developed a two-pass algorithm based on motion, shot length, and color properties. For the purpose of systematic evaluation, a method was proposed in [11].

Most scene detection algorithms utilize convention in video production, such as shot length in movies or TV shows, or visual appearance in video capturing, such as motion and color properties. However, there is no convention in producing home videos, and motion and color information may make no sense in scene detection. For example, drastic changes in motion activity don't imply scene changes, because motion may be caused by hand shaking. Drastic color changes don't necessarily imply scene changes, because color may significantly change due to bad lighting conditions or motion blur.

In this paper, we exploit multimodality correlation between photos and videos captured in the same journey, and achieve video scene detection with the aid of the results of image clustering. Photos are often taken in much better quality, and achieving accurate scene change detection is much easier than that in corresponding video clips.

3. The Proposed Framework

3.1 Essence of the Idea

Due to low cost and popularity of video capturing devices, many people used to capture travel experience by both cameras and camcorders. This behavior is no longer limited to professional photographers. Amateurs often do so by using digital cameras that are able to capture videos, or even cell phones equipped with both photo and video capturing functions. For example, to obtain high-quality data, people often capture famous landmarks or human faces by still cameras. To capture evolution of an event, people prefer to record them by videos. The content in two modalities often has high correlation. Therefore, the idea of this work is that we want to utilize the correlation so that we can succeed the works that are harder to be conducted in videos, but are easier to be done in photos.

Scene detection for photos is much easier than that in videos. For travel photos, we can accurately cluster photos taken at the same place together by checking time information. As the example illustrated in Figure 1, this journey can be separated into three scenes based on photo clustering. Photos in the same cluster usually present the content of the same scenic spot. To perform scene detection in videos, we extract several keyframes for each video shot, and find the optimal matching between photo and keyframe sequences to find the correspondence between photo scenes and video scenes. The essence of correlation between travel photos and videos, and the elaborate design of visual-based matching between image sequences constitute the major contribution of this work.

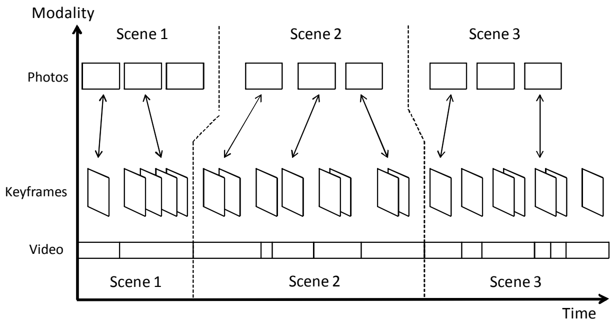

Figure 1. The idea of scene detection based on cross-media alignment.

3.2 Overview of Framework

Figure 2 shows the proposed cross-media scene detection framework. For scene detection in photos, we check shooting time of temporally adjacent photos, and determine the boundaries of scenes by a dynamic threshold [12]. This approach is not only efficient (easy) but also effective (accurate) for photo clustering, which have been shown in many studies.

For video data, we first perform shot change detection, and then extract several keyframes for each shot based on a global k-means algorithm [13]. Based on keyframes, we then filter out the ones that are suffered from significant quality degradation, such as motion blur. The remaining keyframes and the photos in the clusters described above are then represented as feature vectors, in which each dimension denotes a visual word derived from clusters of SIFT (Scale-Invariant Feature Transform) feature points [14]. A sequence of visual-word-based feature vectors then represents photos and keyframes, respectively. In this work, finding the correspondence between two sequences is then taken as an approximate sequence matching problem. We apply a matching algorithm based on the dynamic programming scheme to determine the correspondence between two modalities. Because scene boundaries of photo have been determined, we can accordingly determine the scene boundaries of videos.

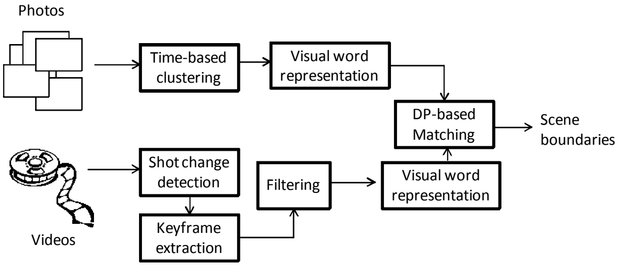

Figure 2. The proposed cross-media scene detection framework.

4. Preprocessing

4.1 Scene Detection for Photos

People often frequently take photos at the same place, and there are large time gaps between photos in different scenic spots because of transportation. This characteristic can be easily utilized to cluster photos, because we just have to check the shooting time of temporally adjacent photos. Let us denote the time difference between the *i*th photo and the (*i*+1)-th photo as g_i:

$$g_i = t_{i+1} - t_i. \quad (1)$$

According to the dynamic thresholding proposed in [12], a scene boundary is claimed to occur between the *n*th and (*n*+1)-th photos if

$$\log(g_N) \geq K + \frac{1}{2d+1}\sum_{i=-d}^{d}\log(g_{N+i}), \quad (2)$$

where K is an empirical threshold, and d is the size of a sliding window, which is for characterizing a range of photos. According to the suggestion in [12], we set K as 17 and set d as 10 in this work.

4.2 Keyframe Extraction

For video data, we first find shot boundaries based on color histogram difference between adjacent frames. For each frame, a 16-bin HSV normalized histogram is used, in which 8 bins are for hue, and 4 bins are for saturation and value, respectively.

To efficiently represent each video shot, one or more keyframes are extracted. One approach for this task is to uniformly sample several frames or just select the first or the middle frame as the keyframes. This approach is not appropriate for travel videos, because a shot may have large scale of visual variation due to handshaking or rapid changes of visual content. Advanced approaches such as unsupervised clustering method [15] dynamically extract keyframes according to visual variations of shots. However, selection of the number of targeted clusters is still a problem.

In this work, we adopt the approach proposed in [16], which automatically determines the most appropriate number of keyframes based on an unsupervised global k-means algorithm [13]. Global k-means is an incremental deterministic clustering algorithm that iteratively performs k-means clustering while increasing k by one at each step. The clustering process ends until the clustering results converge.

By this algorithm, we overcome the initialization problem of conventional k-means algorithm, and adaptively determine appropriate number of clusters for each shot. Frames of a video shot are clustered into groups. Then the frame closest to the centroid of each group is selected as a keyframe.

4.3 Filtering

Amateur photographers often shake the camera or move too fast. Severe motion blur causes, and thus damages the following visual-based matching process. Therefore, we would like to filter out keyframes with severe blurred effects at this stage.

In this work, we examine edge characteristics to detect blur. A wavelet-based method [21] is used to detect occurrence and extent of blur. Keyframes with severe blur degradation are eliminated from the following processes. This process not only reduces the time of cross-media matching, but also eliminates the influence of bad-quality images.

4.4 Visual Word Representation

Determination of the correspondence between photos and keyframes lies on matching between these two image sequences. Due to uncontrolled environments, bad photography skills, and different capture devices, we have to represent data by the features that are more robust than conventional color and texture information. In this work, we adopt the idea of bag of visual words to characterize images. This representation describes *what* are in images, and is proven to be effective in image matching and video concept detection [17][18][24][25].

We apply the difference-of-Gaussian (DoG) detector to detect feature points in keyframes and photos, and use SIFT to describe each point as a 128-dimensional feature vector [14]. SIFT-based feature vectors from training data are then clustered by a k-means algorithm, and feature points in the same cluster are claimed to belong to the same *visual word*. For a keyframe or a photo, each SIFT-based feature point in it is categorized as a visual word, and the distribution of visual words in an image is described as a normalized visual word histogram. Therefore, we finally transform keyframes and photos as a sequence of normalized visual word histograms, respectively.

The essence of this representation is that we view each image as a document, and each document is composed of some basic visual words. Each visual word conceptually represents a basic visual element, such as a corner of a building, tips of a tower, tips of leaves, and etc. We then estimate the similarity between two images (documents) based on the distribution of visual elements.

5. Approximate Sequence Matching

5.1 Visual Word Histogram Matching

Because the content in videos and photos are not consistently the same, we find the correspondence between two normalized visual word sequences by approximate matching techniques. We find the optimal correspondence between two sequences by determining the longest common subsequence between them.

Given two visual word histogram sequences, $X = \langle \boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_m \rangle$ and $Y = \langle \boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_n \rangle$, which are corresponding to photos and keyframes, respectively. Each item in these sequences is a visual word histogram, i.e., $\boldsymbol{x}_i = h[j]$, $0 \leq j \leq N-1$, where N is the number of visual words. The longest common subsequence between two subsequences X_m and Y_n is described as follows.

$$LCS(X_m, Y_n) = \begin{cases} LCS(X_{m-1}, Y_{n-1}) + 1, & \text{if } x_m = y_n, \\ \max(LCS(X_{m-1}, Y_n), LCS(X_m, Y_{n-1})), & \text{otherwise}, \end{cases} \tag{3}$$

where X_i denotes the *i*th prefix of X, i.e., $X_i = \langle \boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_i \rangle$, and $LCS(X_i, Y_j)$ denotes the length of the longest common subsequence between X_i and Y_j. This recursive structure facilitates usage of the dynamic programming approach to find the global optimal solution.

Based on visual word histograms, the equality in Eqn. (3) occurs when the following criterion is met:

$$x_i = y_j \text{ if } \textstyle\sum_{k=0}^{N-1} |(h_i(k) - h_j(k))| < \delta, \tag{4}$$

where h_i and h_j are the visual word histograms corresponding to the images x_i and y_j. According to this measurement, if the distribution of visual words is similar between a keyframe and a photo, we claim that they are conceptually "common" and contain similar content in images.

The algorithm mentioned above is a general sequence matching formulation. However, it's possible to design a formulation that better fits the problem we meet. In a journey, we sequentially visit different scenic spots, and take photos and videos in the same time order. Because there is consistent time order in travel photos and videos, we can impose time constraints in the matching process.

Assume that there are N_g scenes in the photo set $S = \{(s_1, \Delta t_1), ..., (s_{Ng}, \Delta t_{Ng})\}$, where Δt_i denotes the number of photos in the scene s_i. According to the proportion of the number of a scene to that of all scenes, only the photo x_i and the keyframe y_j meet the following time constraint are processed in the matching algorithm.

Time constraint:

The photo $x_i \in s_p$ and the order of the keyframe y_j ranges from $n \times \frac{\sum_{k=0}^{p-1} \Delta t_k}{m} - \frac{n}{d}$ to $n \times \frac{\sum_{k=0}^{p} \Delta t_k}{m} + \frac{n}{d}$, where $\Delta t_0 = 0$, and there are totally m photos and n keyframes in the dataset. The value $\frac{n}{d}$ is a tolerance parameter that slightly relaxes the interpolation-based time constraint. In this work, the value of d is empirically set as the number of photo scenes.

With this idea, we reformulate the criterion defined in Eqn. (4) as

$$x_i = y_j \text{ if } \textstyle\sum_{k=0}^{N-1} |(h_i(k) - h_j(k))| < \delta$$
$$\text{and the pair } (x_i, y_j) \text{ meets the time constraint.} \tag{5}$$

5.2 Postprocessing

Because the scene boundaries in photos have been determined by the time-based clustering method, we can accordingly estimate the scene boundaries in videos by utilizing the correspondence between photo and keyframe sequences. For the keyframe f_i, if it is matched with a photo p_j, denoted by $f_i \approx p_j$, then the scene of f_i is determined as that corresponds to p_j. If the keyframe f_i doesn't match with any photo, find the two nearest keyframes $f_{i-\ell}$ and f_{i+r} that are matched with some photos. We check the matched situations as follows.

1) If $f_{i-\ell} \approx p_j$ and $f_{i+r} \approx p_{j+t}$, and both p_j and p_{j+t} belong to the same scene s_k, then f_i is claimed to be in the scene s_k.
2) If $f_{i-\ell} \approx p_j$ and $f_{i+r} \approx p_{j+t}$, but p_j belongs to the scene s_k, and p_{j+t} belongs to the scene s_{k+n}.
   - If f_i and $f_{i-\ell}$ are in the same video shot, f_i is claimed to be in the scene s_k.
   - If f_i and f_{i+r} are in the same video shot, f_i is claimed to be in the scene s_{k+n}.
   - If f_i don't fall into the same shot as $f_{i-\ell}$ and f_{i+r}, the scene corresponding to f_i is determined by interpolation. The time differences between f_i and $f_{i-\ell}$, and that between f_i and f_{i+r}, are calculated. The estimated scene $s_{k+\Delta}$ between s_k and s_{k+n} is claimed to contain f_i.

After this postprocessing, scene boundaries of travel videos are determined.

6. Experiments

6.1 Evaluation Data

We evaluate the proposed method based on five data sets. Each dataset includes a video clip and a set of photos. Lengths of these video clips range from eight to fifteen minutes, and these videos are stored in MPEG-1 format and 352 x 240 resolution. There are 20 to 126 corresponding photos in different datasets, stored as at most 400 x 300 resolution. Photos are rescaled to smaller sizes due to the efficiency of feature points processing and visual word construction. Videos and photos are captured by different amateur photographers, with different capturing devices. Figure 3 shows snapshots of videos and some corresponding photos. We see that these data are unconstrained and contain wide range of content. Table 1 shows the detailed information of evaluation data, including the number of extracted keyframes for each video clip.

Table 1. Detailed information of evaluation data

	# scenes	length	# of extracted keyframes	# of corresponding photos
Video 1	6	12:57	176	101
Video 2	4	10:20	113	20
Video 3	3	15:07	73	41
Video 4	5	8:29	74	46
Video 5	5	11:03	127	126

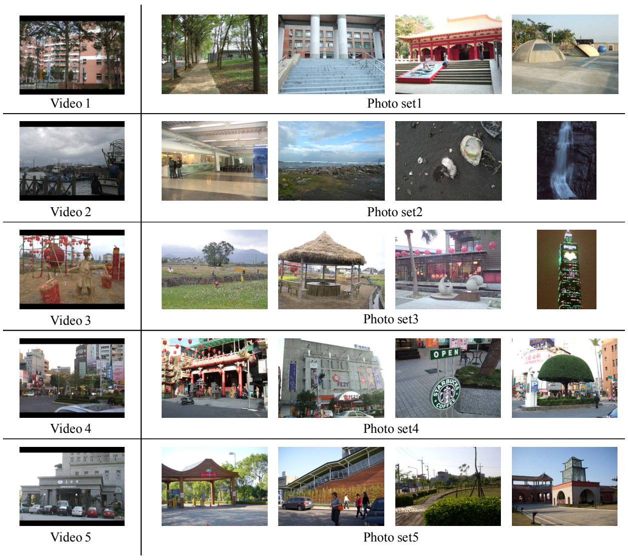

Figure 3. Snapshots of videos and corresponding photos.

6.2 Evaluation Metric

To faithfully evaluate performance of scene detection, we consider overlaps between detected video scenes and ground truths (photo scenes), rather than counting the number of scene boundaries to calculate precision and recall values. In this work, we evaluate scene detection results in terms of purity [19]. 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

$$\rho = \left(\sum_{i=1}^{Ng} \frac{\tau(s_i)}{T} \sum_{j=1}^{Nv} \frac{\tau^2(s_i, s_j^*)}{\tau^2(s_i)} \right) \cdot \left(\sum_{j=1}^{Nv} \frac{\tau(s_j^*)}{T} \sum_{i=1}^{Ng} \frac{\tau^2(s_i, s_j^*)}{\tau^2(s_j^*)} \right), \tag{6}$$

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

6.3 Performance of Scene Detection

We first study performance variations when we represent images by different numbers of visual words, with different similarity thresholds δ's used in Eqn. (4). We calculate purity values based on 20, 50, 100, 200, and 500 visual words, with different thresholds, for the Video 3. Figure 4 shows the experimental results. It's expectable that the best performance occurs in different settings for different visual words. Overall, the best purity values with different experiment settings are slightly higher than 0.9.

We need more computation time in generating more visual words. Therefore, by considering computation efficiency and scene detection performance, we use 20 visual words to represent each photo or keyframe. In the following experiments, the reported purity values are the best performance we can obtain from the representation of 20 visual words, with the most appropriate threshold. For example, from Figure 4, we obtain the best performance when the similarity threshold is set as 0.4.

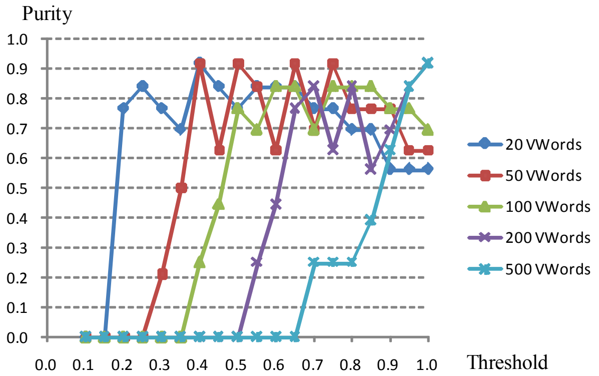

Figure 4. For Video 3, performance in terms of purity based on different numbers of visual words, with different similarity thresholds.

To verify the effectiveness of the proposed method, we compare performance in terms of purity based on four approaches:

1) The proposed sequence matching approach with features of visual word histograms.

2) The proposed sequence matching approach with features of HSV histograms. We replace the histogram used in Eqn. (4) by 16-bin HSV color histograms (8 bins for hue, and 4 bins for saturation and values, respectively).

3) A naïve method, in which video scene boundaries are determined by interpolation, based on the scene boundaries in photos.

4) The proposed sequence matching approach with features of the concatenation of a visual word histogram and an HSV histogram. This approach combines the effects of two types of features.

Figure 5 shows performance comparison in four different approaches. Because different video clips have drastically different content, we may need to find the most appropriate thresholds for the approaches 1, 2, and 3. We experimented on a wide range of thresholds, and find the best performance for each video clip and report them in Figure 5.

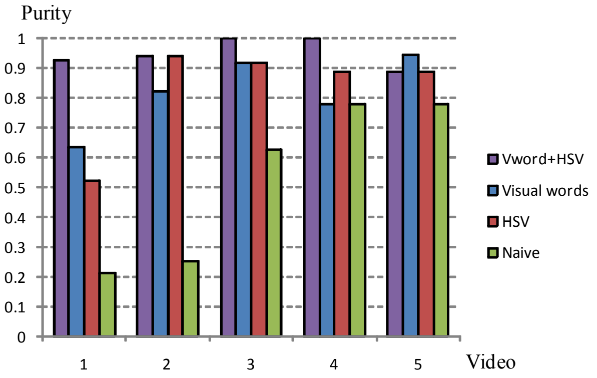

Figure 5. Performance based on four different scene detection approaches.

Comparing with the effectiveness of different features, visual word histograms have better performance than HSV histograms in Videos 1 and 5. Based on visual words, we describe *what* are in an image. However, color information is eliminated in extracting SIFT feature points, thus HSV histograms still work better in Videos 2 and 4. We achieve the best scene detection performance when visual words are combined with color information. We averagely obtain a purity value of 0.95, which is very promising in scene detection, especially in the uncontrolled travel media.

To show effectiveness of the sequence matching approach, we compare the proposed method with a naïve approach, in which no visual information is considered, and no matching is performed between different modalities. From Figure 5, we can see that the proposed method is much superior to the naïve one. The performance of the naïve method depends on the consistency between the distributions of number of photos and number of keyframes. The naïve method has satisfactory performance in Video 5, because the number of photos and keyframes in each scene are proportionally similar.

To further show that the proposed method is more appropriate to be applied in travel video, we compare it with the method proposed in [16]. 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 2. 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 [16] and our method (visual word + HSV), respectively. For example, in the second columns for Video 1, (4,1) means that the second scene in Video 1 is segmented into 4 scenes and one scene by the method in [16] and our approach, respectively. From Table 2, we see the effect of over-segmentation is severe in the unsupervised clustering approach, and our approach works much better from this perspective.

Table 2. Over-segmentation situations in different videos.

	S1	S2	S3	S4	S5	S6
Video 1	(1,1)	(4,1)	(7,2)	(3,1)	(9,2)	(3,1)
Video 2	(2,2)	(8,1)	(1,1)	(1,1)		
Video 3	(6,1)	(3,1)	(1,1)			
Video 4	(1,1)	(1,1)	(1,1)	(3,1)	(2,1)	
Video 5	(1,1)	(2,2)	(1,1)	(5,2)	(1,1)	

The major strength of the method in [16] is that there is no need to predefine the number of targeted shot clusters, which are then manipulated by some postprocesses to form the results scene detection. Because there may be drastically different visual appearance in the same scenic spot, the approach that just takes color information into account often overly segments scenes. On the other hand, we automatically determine the number of scenes from the photos' time information, and then appropriately segment videos into scenes by cross-media alignment. We conclude that the proposed method takes advantages of the prior knowledge of travel media, and is more appropriate to facilitate effective management of travel photos and videos.

6.4 Discussion

Correlation between different modalities is not limited in travel photos and videos. Similar ideas for reranking the results of visual search has been proposed in [20], while they use conventional visual features such as texture and color. Moreover, Chu and Chen [22] exploited correlation between different teaching media, such as HTML pages, teacher's sound, and navigation objects. Therefore, we argue that the same approach is possible to be extended to other domains, such as topic tracking or detection in news media.

Although we have demonstrated that visual word histogram performs well, it's possible to improve more if we take the most recent variations of visual word representation, such as that in [18] and [23]. In our work, we describe an image by a visual word histogram, which is a representation derived from "hard" quantization of SIFT descriptors. In [18], the authors describe characteristics of SIFT descriptors by Gaussian mixture model. The work in [23] poses similar idea that they describe images by "softly" describing a SIFT feature point in terms of several visual words. We would like to apply the same concept to construct more flexible representation for images in the future.

7. Conclusion

We have presented a video scene detection method that is based on the correlation between different modalities. A scene in travel photos and videos represents a scenic spot, which can be easily determined by the time information of photos. We segment travel videos into shots, and extract keyframes for each shot by a global k-means algorithm. After representing keyframes and a photo set by a sequence of visual word histograms, we transform scene detection into a sequence matching algorithm. By using a dynamic programming approach, we find optimal matching between two sequences, and then determine video scene boundaries with the help of photo scene boundaries. We experiment on five different travel videos, with different parameter settings, and demonstrate that the proposed method achieves very promising performance for unconstrained travel videos. This result shows that using correlation between different modalities is an interesting and effective approach in analyzing consumer multimedia content, especially travel videos and photos.

8. ACKNOWLEDGEMENTS

This work was partially supported by the National Science Council of the Republic of China under grants NSC 97-2221-E-194-050.

9. REFERENCES

[1] Gatica-Perez, D., Loui, A., and Sun, M.-T. 2003. Finding structure in home videos by probabilistic hierarchical clustering. IEEE Transactions on Circuits and Systems for Video Technology, vol. 13, no. 6, 539-548.

[2] Pan, Z. and Ngo, C.-W. 2004. Structuring home video by snippet detection and pattern parsing. In Proc. of ACM International Workshop on Multimedia Information Retrieval, 69-76.

[3] Hua, X.-S., Lu, L., and Zhang, H.-J. 2004. Optimization-based automated home video editing system, vol. 14, no. 5, 572-583.

[4] Lee, S.-H., Wang, S.-Z., and Kuo, C.C.J. 2005. Tempo-based MTV-style home video authoring. In Proc. of IEEE International Workshop on Multimedia Signal Processing.

[5] Peng, W.-T., Chiang, Y.-H., Chu, W.-T., Huang, W.-J., Chang, W.-L., Huang, P.-C., and Hung, Y.-P. 2008. Aesthetics-based automatic home video skimming system. In LNCS 4903, 186-197.

[6] Shipman, F., Girgensohn, A., and Wilcox. L. 2008. Authoring, viewing, and generating hypervideo: an overview of Hyper-Hitchcock. ACM Transactions on Multimedia Computing, Communications, and Applications, vol. 5, no. 2, Article no. 15.

[7] Achanta, R.S.V., Yan, W.-Q., and Kankanhalli, M.S. 2006. Modeling intent for home video repurposing. IEEE Multimedia, vol. 13, no. 1, 46-55.

[8] Mei, T. and Hua, X.-S. 2005. Intention-based home video browsing. In Proc. of ACM Multimedia, 221-222.

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

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

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

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

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

[14] Lowe, D. 2004. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60, 2, 91-110.

[15] Zhuang, Y., Rui, Y., Huang, T.S., and Mehrotra, S. 1998. Adaptive key frame extraction using unsupervised clustering. In Proc. of IEEE International Conference on Image Processing, 866-870.

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

[17] Wang, F., Jiang, Y.-G., and Ngo, C.-W. 2008. Video event detection using motion relativity and visual relatedness. In Proc. of ACM Multimedia, 239-248.

[18] Zhou, X., Zhuang, X., Yan, S., Chang, S.-F., Hasegawa-Johnson, M., and Huang, T.S. 2008. SIFT-bag kernel for video event analysis. In Proc. of ACM Multimedia, 229-238.

[19] Vinciarelli, A. and Favre, S. 2007. Broadcast news story segmentation using social network analysis and hidden Markov models. In Proc. of ACM Multimedia, 261-264.

[20] Hsu, W.H., Kennedy, L., and Chang, S.-F. 2007. Reranking methods for visual search. IEEE Multimedia, vol. 14, no. 3, 14-22.

[21] Tong, H., Li, M., Zhang, H.-J., and Zhang, C. 2004. Blur detection for digital images using wavelet transform. In Proc. of IEEE International Conference on Multimedia & Expo, 17-20.

[22] Chu, W.-T. and Chen, H.-Y. 2005. Towards better retrieval and presentation by exploring cross-media correlations. Multimedia Systems, vol. 10, no. 3, 183-198.

[23] Philbin, J., Chum, O., Isard, M., Sivic, J., and Zisserman, A. 2008. Lost in quantization: improving particular object retrieval in large scale image databases. In Proc. of IEEE Conference on Computer Vision and Pattern Recognition.

[24] Sivic, J. and Zisserman, A. 2008. Efficient video search for objects in videos. Proceedings of the IEEE, 96, 4, 548-566.

[25] Dumont, E. and Merialdo, B. 2007. Video search using visual dictionary. In Proc. of International Workshop on Content-based Multimedia Indexing, 315-322.