Travel Photo and Video Summarization with Cross-Media Correlation and Mutual Influence

Wei-Ta Chu[1], Che-Cheng Lin[1], and Jen-Yu Yu[2]

[1] Department of Computer Science and Information Engineering,
National Chung Cheng University, Chiayi, Taiwan
wtchu@cs.ccu.edu.tw, john72831@yahoo.com.tw
[2] Information and Communication Research Lab,
Industrial Technology Research Institute, Hsinchu, Taiwan
KevinYu@itri.org.tw

Abstract. This paper presents how cross-media correlation facilitates summarization of photos and videos captured in journeys. Correlation between photos and videos comes from similar content captured in the same temporal order. We transform photos and videos into sequences of visual word histograms, and adopt approximate sequence matching to find correlation. To summarize photos and videos, we propose that the characteristics of correlated photos can be utilized in selecting important video segments into video summaries, and on the other hand, the characteristics of correlated video segments can be utilized in selecting important photos. Experimental results demonstrate that the proposed summarization methods well take advantage of the correlation.

Keywords: Cross-media correlation, photo summarization, video summarization.

1 Introduction

Recording daily life or travel experience by digital videos or photos has been widely accepted in recent years, due to popularity and low cost of digital camcorders and cameras. Large amounts of videos and photos are especially captured in journeys, in which people are happy to capture travel experience at will. However, massive digital content burdens users in media management and browsing. Developing techniques to analyze travel media thus has drawn more and more attention.

There are at least two unique challenges in travel media analysis. First, there is no clear structure in travel media. Unlike scripted videos such as news and movies, videos captured in journeys just follow the travel schedule, and the content in video may consist of anything people willing or unwilling to capture. Second, because amateur photographers don't have professional skills, the captured photos and videos often suffer from bad quality. The same objects in different photos or video segments may have significant appearance. Due to these characteristics, conventional image/video analysis techniques cannot be directly applied to travel media.

People often take both digital cameras and digital camcorders in journeys. They usually capture static objects such as landmark or human faces by cameras and capture

S. Boll et al. (Eds.): MMM 2010, LNCS 5916, pp. 577–587, 2010.

evolution of events such as performance on streets or human's activities by camcorders. Even with only one of these devices, digital cameras have been equipped with video capturing functions, and on the other hand, digital camcorders have the "photo mode" to facilitate taking high-resolution photos. Therefore, photos and videos in the same journey often have similar content, and the correlation between two modalities can be utilized to develop techniques especially for travel media.

In our previous work [1], we investigate content-based correlation between photos and videos, and develop an effective scene detection module for travel videos. The essential idea of this work is to solve a harder problem (video scene detection) by first solving an easier problem (photo scene detection) accompanied with cross-media correlation. In this paper, we try to further take advantage of cross-media correlation to facilitate photo summarization and video summarization. We advocate that summarizing a media can be assisted by other media's characteristics and the correlation between them.

Contributions of this paper are summarized as follows.

- We explore cross-media correlation based on features resisting to significant visual variations and bad quality. Two-level cross media correlations are investigated to facilitate the targeted tasks.
- We advocate that the correlated video segments influence selection of photos in photo summaries, and in the opposite way, the correlated photos influence selection of video segments in video summaries.

The rest of this paper is organized as follows. Section 2 gives literature survey. Section 3 describes the main idea of this work and the components developed for determining cross-media correlation. Photo summarization and video summarization are addressed in Section 4. Section 5 gives experimental results, and Section 6 concludes this paper.

2 Related Works

We briefly review works on home video structuring and editing. Then, studies especially about highlight generation and summarization are reviewed as well. Gatica-Perez et al. [2] 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 [3] 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 [4] or MTV-style summaries [5]. Peng et al. [6] further take media aesthetics and editing theory into account to perform home video skimming.

For summarizing videos, most studies exploit features such as motion and color variations to estimate the importance of video segments. However, different from scripted videos, drastic motion changes in travel videos don't imply higher importance, because motion may be caused by hand shaking. Similarly, drastic color changes may be from bad lighting conditions or motion blur. In this paper, we exploit correlation between photos and videos to define importance of photos and video segments.

3 Cross-Media Correlation

In travel media, a photo scene or a video scene means a set of photos or video shots that were captured in the same scenic spot. To faithfully represent a journey by a photo summary or a video summary, we have to consider important parts of media and fairly select data from each scene to generate summaries. Scene boundaries of photos and videos are therefore important clues to the proposed summarization modules. We will determine cross-media correlation first, and briefly review video scene detection utilizing correlation [1].

Figure 1 shows the flowchart of finding cross-media correlation between a photo set and a video. Note that all video segments captured in the same journey are concatenated as a single video stream according to the temporal order.

- Photo Scene Detection

There are large time gaps between photos in different scenic spots because of transportation. This characteristic can be utilized to cluster photos into several scenes. We check time gaps between temporally adjacent photos, and claim a scene boundary exists between two photos if their time gap exceeds a dynamic threshold [7]. The method proposed in [7] has been widely applied in photo clustering, and has been proven very effective. After this time-based clustering, photos taken at the same scenic spot (scene) are clustered together.

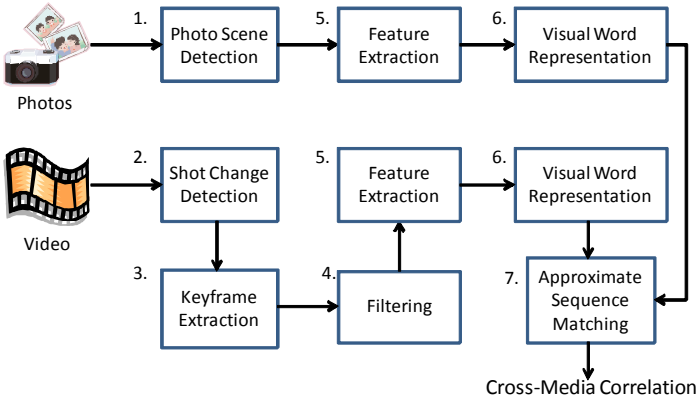

Fig. 1. Flowchart for finding cross-media correlation

- Keyframe Extraction

For the video, we first segment it into shots based on difference of HSV color histograms in consecutive video frames. To efficiently represent each video shot, one or more keyframes are extracted. We adopt the method proposed in [8], which automatically determines the most appropriate number of keyframes based on an unsupervised global k-means algorithm [9]. The global k-means algorithm 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.

- Keyframe Filtering

Video shots with blurred content often convey less information, and would largely degrade the performance of correlation determination. To detect blurred keyframe, we check edge information in different resolutions [10]. The video shots with blurred keyframes are then put aside from the following processes.

Video shot filtering brings two advantages to the proposed work. First, fewer video shots (keyframes) are needed to be examined in the matching process described later. Moreover, this kind of filtering reduces the influence of blurred content, which may cause false matching between keyframes and photos.

- Visual Word Representation

After the processes above, correlation between photos and videos is determined by matching photos and keyframes. Image matching is an age-old problem, and is widely conducted based on color and texture features. However, especially in travel media, the same place may have significantly different appearance, which may be caused by viewing angles, large camera motion, and overexposure/underexposure. In addition, landmarks or buildings with apparent structure are often important clues for image matching. Therefore, we need features that resist to luminance and viewpoint changes, and are able to effectively represent local structure.

We extract SIFT (Scale-Invariant Feature Transform) features [11] from photos and keyframes. The DoG (difference of Gaussian) detector is used to locate feature points first, and then orientation information around each point is extracted to form 128-dimenional feature descriptors.

SIFT features from a set of training photos and keyframes are clustered by the k-means algorithm. Feature points belong to the same cluster are claimed to belong to the same *visual word*. Before matching photos with keyframes, SIFT features are first extracted, and each feature point is quantized into one of visual words. The obtained visual words in photos and keyframes are finally collected as visual word histograms. Based on this representation, the problem of matching two image sequences has been transformed into matching two sequences of visual word histograms. According to the experiments in [1], we present photos and keyframes by 20-bin visual word histograms.

Conceptually, each SIFT feature point represents texture information around a small image patch. After clustering, a visual word presents a concept, which may correspond to corner of building, tip of leaves, and so on. The visual word histogram presents what concepts compose the image. To discover cross-media correlation, we would like to find photos and keyframes that have similar concepts.

- Approximate Sequence Matching

To find the optimal matching between two sequences, we exploit the dynamic programming strategy to find the longest common subsequence (LCS) 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 correspond 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.

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

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.

Based on visual word histograms, the equality in eqn. (1) occurs when the following criterion is met:

$$x_i = y_j \text{ if } \sum_{k=0}^{N-1} |(h_i(k) - h_j(k))| < \delta, \quad (2)$$

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 visual word distributions are similar between a keyframe and a photo, we claim that they are conceptually "common" and contain similar content.

- Video Scene Detection

Figure 2 shows an illustrated example to conduct video scene detection based on cross-media correlation. The double arrows between photos and keyframes indicate matching determined by the previous process, and are representation of the so-called cross-media correlation. If a video shot's keyframe matches the photo in the *i*th photo scene, this video shot is assigned as in *i*th video scene as well. For those video shots without any keyframe matched with photos, we apply interpolation and nearest neighbor processing to assign them. Details of visual word histogram matching and scene detection processes please refer to [1].

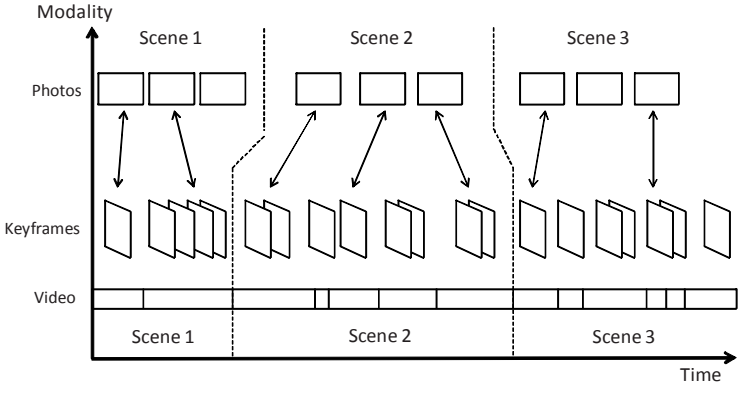

Fig. 2. Illustration of video scene detection

4 Photo Summarization and Video Summarization

On the basis of photo scenes and video scenes, each of which corresponds to a scenic spot, we develop summarization modules that consider characteristics of the correlated media. In a word, how video content evolves affects the selection of photos in photo summary. On the other hand, how photo being taken affects the selection of video segments in video summary. This idea is totally different from conventional approaches, such as attention modeling in photo summarization and motion analysis in video summarization.

4.1 Local Cross-Media Correlation

Matching based on visual word histogram and the LCS algorithm comes from two factors: First, the matched photos and keyframes contain objects with similar concepts, e.g., both images contain large portion of grass and tree, or both images contain artificial objects. Second, the matched images were taken in the same temporal order, i.e., a photo at the beginning of a journey unlikely matches with a keyframe at the end of a journey.

Correlation determined by this process suffices for scene boundary detection. However, to select important data as summaries, finer cross-media correlation is needed to define importance value of each photo and keyframe. In this work, we call the correlation described in Section 3 *global cross-media correlation*, which describes matching in terms of visual concepts. In this section, we need further analyze *local cross-media correlation* to find matching in terms of objects.

For the photos and keyframes in the same scene, we perform finer matching between them by the SIFT matching algorithm [11]. Let s_i denote a feature point in the photo p_m, we calculate the Euclidean distance between s_i and each of the feature points in the keyframe k_n, and find the feature point t_{j^*} that are nearest to s_i. That is,

$$j^* = \arg\min_{j=1,2,\ldots,J} d(s_i, t_j). \tag{3}$$

Similarly, we can find the second nearest feature point $t_{j^\dagger}$ to s_i. The feature point s_i is claimed to match with the point t_{j^*} if

$$\frac{d(s_i, t_{j^*})}{d(s_i, t_{j^\dagger})} < \gamma, \tag{4}$$

where the threshold γ is set as 0.8 according to the suggestion in [11].

For the photo p_m and the keyframe k_n, we claim they contain the same object, such as a building or a statue, if the number of matched feature points exceeds a predefined threshold τ. This threshold can be adjusted dynamically according to the requirements of users. The experiment section will show the influence of different thresholds on summarization performance.

We have to emphasize that local cross-media correlation is determined based on SIFT feature matching rather than visual word histograms. Visual word histograms describe global distribution of concepts (visual words), while feature points describe local characteristics that more appropriate whether two images have the same building or other objects.

4.2 Photo Summarization

The idea of defining each photo's importance comes from two perspectives. The first factor directly comes from the determined local cross-media correlation. When a view or an object is both captured in photos and videos, the captured content must attract people more and is likely to be selected into summaries. We propose the second factor by considering the characteristics of videos to define photo's importance. When people take a closeup shot on an object, this object must attract people more and is likely to be selected into summaries. Therefore, a photo's importance is set higher if it matches with a keyframe that is between a zoom in action and a zoom out action, or is between a zoom in action and camera turning off.

To detect zoom in and zoom out actions, we first find motion vectors and motion magnitudes based on the optical flow algorithm. A keyframe is equally divided into four regions, i.e., left-top, right-top, left-bottom, and right-bottom regions. If motion vectors in all four regions point to the center of the keyframe, a zoom out action is detected. If all motion vectors diverge from the center of the keyframe, a zoom in action is detected.

Two factors defining importance values can be mathematically expressed as follows.

- Factor 1:

The first importance value of the photo p_m is defined as

$$PT_{1,m} = \frac{I_{1,m}}{\max_{i=1,2,\ldots,M} I_{1,i}}, \tag{5}$$

where *M* is the number of photos in this dataset. The value $I_{1,m}$ is calculated as

$$I_{1,m} = \begin{cases} L_1(h_{p_m}, h_{k_n}), & \text{if the photo } p_m \text{ matches with the keyframe } k_n, \\ 0, & \text{otherwise,} \end{cases}$$

where h_{p_m} and h_{k_n} are visual word histograms of the photo p_m and the keyframe k_n, respectively. The value $L_1(\cdot,\cdot)$ denotes L_1-distance between two histograms.

- Factor 2:

The second importance value of the photo p_m is defined as

$$PT_{2,m} = \frac{I_{1,m} \times ZoomIn(p_m)}{\max_{i=1,2,\ldots,M} I_{1,i} \times ZoomIn(p_i)}, \tag{6}$$

where $ZoomIn(p_m) = 1$ if the keyframe k_n that matches with p_m locates between a zoom in action and a zoom out action, or between a zoom in action and camera turning off. The value $ZoomIn(p_m) = 0$, otherwise.

Note that two importance values are normalized, and then integrated by linear weighting to form the final importance value of p_m:

$$PT_m = \alpha \times PT_{1,m} + \beta \times PT_{2,m}. \tag{7}$$

Currently, the values α and β are set as 1.

In photo summarization, users can set the desired number of photos in summaries. To ensure the generated summary contain photos of all scenes (scenic spots), we first pick the most important photo in each scene to the summary. After the first round, we sort photos according to their corresponding importance values in descending order, and pick photos sequentially until the desired number is achieved.

According to the definitions above, only photos that are matched with keyframes have importance values larger than zero. If all photos with importance values larger than zero are picked but the desired number hasn't achieved, we define the importance value of a photo p_i not picked yet by calculating the similarity between p_i and its temporally closest photo p_j that has nonzero importance value, i.e.,

$$T_i = L_1(h_{p_i}, h_{p_j}). \tag{8}$$

We sort the remaining photos according to these alternative importance values in descending order, and pick photos sequentially until the desired number is achieved.

4.3 Video Summarization

Similar to photo summarization, we advocate that photo taking characteristics in a scene affect selection of important video segments in video summaries. Two factors are also involved with video summary generation. The first factor is the same as that in photo summarization, i.e., video shots whose content also appears in photos are more important. Moreover, a video shot in which many keyframes match with photos is relatively more important. Two factors defining importance values can be mathematically expressed as follows.

- Factor 1:

The first importance value of a keyframe k_m is defined as

$$KT_{1,m} = \frac{I_{1,m}}{\max_{i=1,2,\ldots,M} I_{1,i}}, \tag{9}$$

where *M* is the number of keyframes in this dataset. The value $I_{1,m}$ is calculated as

$$I_{1,m} = \begin{cases} L_1(h_{k_m}, h_{p_n}), & \text{if the keyframe } k_m \text{ matches with the photo } p_n, \\ 0, & \text{otherwise,} \end{cases}$$

where h_{k_m} and h_{p_n} are visual word histograms of keyframe k_m and the photo p_n, respectively.

- Factor 2:

The second importance value of the keyframe k_m is defined as

$$KT_{2,m} = \frac{I_{2,m}}{\max_{i=1,2,\ldots,M} I_{2,i}}, \tag{10}$$

where the value $I_{2,m}$ is the sum of visual word histogram similarities between keyframes at the same shot as k_m and their matched photos. That is,

$$I_{2,m} = \sum_{j=1}^{J} L_1(h_{k_j}, h_{p_{j^*}}). \tag{11}$$

This expression means there are J keyframes in the shot containing k_m, and the notation p_{j^*} denotes the photo matched with the keyframe k_j.

These two importance values are integrated by linear weighting to form the final importance value of k_m:

$$KT_m = \alpha \times KT_{1,m} + \beta \times KT_{2,m}. \tag{12}$$

In video summarization, users can set the desired length of video summaries. To ensure the generated summary contain video segments of all scenes (scenic spots), we first pick the most important keyframe of each scene. Assume that the keyframe k_i is selected, we determine length and location of the video segment S_i corresponding to k_i as

$$S_i = \left(\frac{t(k_{i-1}) + t(k_i)}{2}, \frac{t(k_i) + t(k_{i+1})}{2} \right), \tag{13}$$

where $t(k_i)$ denotes the timestamp of the keyframe k_i, and k_{i-1} and k_{i+1} are two nearest keyframes that are before and after k_i, and with nonzero importance values. Two values in the parentheses respectively denote the start time and end time of the segment S_i.

We pick keyframes and their corresponding video segments according to keyframe's importance values until the desired length of video summary is achieved. If all keyframes with nonzero importance values are picked but the desired length hasn't achieved, we utilize a method similar to that in eqn. (8) to define remaining keyframe's importance, and pick appropriate number of keyframe accordingly.

5 Experimental Results

We collect seven sets of travel media captured in seven journeys for performance evaluation. Each dataset includes a video clip and many photos. The video is encoded in MPEG-1 format, and the resolution is 480×272. Each photo is normalized into 400×300 in experiments. The first five columns of Table 1 show information about the evaluation data.

To objectively demonstrate summarization results, we ask content owners of these data to manually select subset of keyframes and photos as the ground truth of summaries. In generating video summaries or photo summaries, we set the number of keyframes or photos in manual summaries as the targeted number to be achieved. For example, in generating the video summary for the first dataset, 98 keyframes and their corresponding video segments should be selected in the video summary. We measure summarization results by precision values, i.e.,

$$\text{Precision} = \frac{\#\text{ correctly selected keyframes}}{\#\text{ selected keyframes}}. \tag{14}$$

Note that precision and recall rates are the same due to the selection policy.

Figure 3(a) shows precision/recall rates of video summarization for seven datasets under different matching thresholds τ, while Figure 3(b) shows precision/recall rates of photo summarization. In Section 4.1, a photo is claimed to has local cross-media correlation with a keyframe if matched SIFT points is larger than the threshold τ. Generally,

we see that using five or ten matched points as the threshold we can obtain better summarization results, i.e., looser thresholds draw slightly better performance. The second dataset has the worst performance, because photos and videos in this dataset don't have high content correlation as that in others, and the content in them is involved with large amounts of natural scenes such that local cross-media correlation based on SIFT matching cannot be effectively obtained. This result conforms that cross-media correlation really impacts on the proposed photo and video summarization methods.

We also conduct subjective evaluation by asking content owners to judge summarization results. They give a score from five to one, in which a larger score means higher satisfaction. Table 2 shows results of subjective evaluation. Overall, both video and photo summarization achieves more than 3.7. The worse performance on the second dataset also reflects in this table.

Table 1. Information of evaluation data

Dataset	# scenes	length	#kf	#photos	#kf in manual sum.	#photos in manual sum.
S1	6	12:57	227	101	98	48
S2	4	15:07	153	30	32	12
S3	5	8:29	98	44	71	11
S4	5	11:03	176	62	97	21
S5	3	16:29	136	50	103	15
S6	2	5:34	67	23	43	12
S7	6	15:18	227	113	112	32

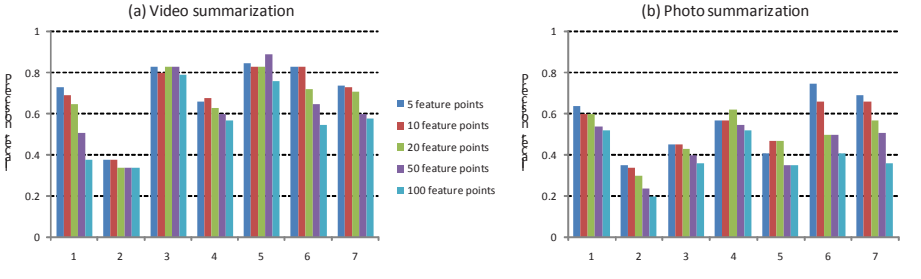

Fig. 3. Performance of (a) video summarization and (b) photo summarization

Table 2. Subjective evaluation on summarization results

	S1	S2	S3	S4	S5	S6	S7	Overall
Video sum.	4	2	4	3	5	4	4	3.7
Photo sum.	5	2	4	3	4	5	4	3.8

6 Conclusion

Two novel ideas have been presented in this paper. Because photos and videos captured in journeys often contain similar content, we can find correlation between them based

on an approximate sequence matching algorithm. After that, we first solve an easier problem (photo scene detection), and then solve a harder problem (video scene detection) by consulting with the correlation. To summarize photos and videos, we further exploit cross-media correlation and propose that photo summarization is influenced by the correlated video segments, and contrarily video summarization is influenced by the correlated photos. We respectively consider two factors based on correlation to conduct summarization, and demonstrate the effectiveness of the proposed methods.

In the future, extensive experiments will be conducted to demonstrate the effectiveness and limitation of the proposed methods. For those datasets with less content correlation, more elaborate techniques should be integrated to accomplish the targeted tasks. Moreover, we will try to extend this work to other domains, in which different modalities have high content correlation.

Acknowledgments

This work was partially supported by the National Science Council of the Republic of China under grants NSC 98-2221-E-194-056 and NSC 97-2221-E-194-050.

References

1. Chu, W.-T., Lin, C.-C., Yu, J.-Y.: Using Cross-Media Correlation for Scene Detection in Travel Videos. In: ACM International Conference on Image and Video Retrieval (2009)
2. Gatica-Perez, D., Loui, A., Sun, M.-T.: Finding Structure in Home Videos by Probabilistic Hierarchical Clustering. IEEE Transactions on Circuits and Systems for Video Technology 13(6), 539–548 (2003)
3. Pan, Z., Ngo, C.-W.: Structuring Home Video by Snippet Detection and Pattern Parsing. In: ACM International Workshop on Multimedia Information Retrieval, pp. 69–76 (2004)
4. Hua, X.-S., Lu, L., Zhang, H.-J.: Optimization-based Automated Home Video Editing System. IEEE Transactions on Circuits and Systems for Video Technology 14(5), 572–583 (2004)
5. Lee, S.-H., Wang, S.-Z., Kuo, C.C.J.: Tempo-based MTV-style Home Video Authoring. In: IEEE International Workshop on Multimedia Signal Processing (2005)
6. Peng, W.-T., Chiang, Y.-H., Chu, W.-T., Huang, W.-J., Chang, W.-L., Huang, P.-C., Hung, Y.-P.: Aesthetics-based Automatic Home Video Skimming System. In: Satoh, S., Nack, F., Etoh, M. (eds.) MMM 2008. LNCS, vol. 4903, pp. 186–197. Springer, Heidelberg (2008)
7. Platt, J.C., Czerwinski, M., Field, B.A.: PhotoTOC: Automating Clustering for Browsing Personal Photographs. In: IEEE Pacific Rim Conference on Multimedia, pp. 6–10 (2003)
8. Chasanis, V., Likas, A., Galatsanos, N.: Scene Detection in Videos Using Shot Clustering and Symbolic Sequence Segmentation. In: IEEE International Conference on Multimedia Signal Processing, pp. 187–190 (2007)
9. Likas, A., Vlassis, N., Verbeek, J.J.: The Global K-means Clustering Algorithm. Pattern Recognition 36, 451–461 (2003)
10. Tong, H., Li, M., Zhang, H.-J., Zhang, C.: Blur Detection for Digital Images Using Wavelet Transform. In: IEEE International Conference on Multimedia & Expo., pp. 17–20 (2004)
11. Lowe, D.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)