Memory Matrix: A Novel User Experience for Home Video

Qianqian Xu[1], Zhipeng Wu[1], Guorong Li[1], Lei Qin[2,3], Shuqiang Jiang[2,3], Qingming Huang[1,2,3]
[1]Graduate University, Chinese Academy of Sciences, Beijing, 100049, China
[2]Key Lab of Intell. Info. Process, Chinese Academy of Sciences, Beijing, 100190, China
[3]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China
{qqxu, zpwu, grli, lqin, sqjiang, qmhuang}@jdl.ac.cn

Abstract
Nowadays, various efforts have sprung up aiming to automatically analyze home videos and provide users satisfactory experiences. In this paper, we present a novel user experience for home video called Memory Matrix, which could facilitate users to re-experience the joy of their memories, travelling along not only the time axis but also the space axis. In other words, the video clips (sub-shots) are organized both by taken times and taken locations, which further allows the user to browse home videos taken at similar locations. Moreover, given a specific query in Memory Matrix (row, column), it can also provide the user optional summaries along the time axis or space axis. The summarization scheme in this paper is based on a top-down interest score generation algorithm which automatically propagates the pre-labeled video level interest scores to sub-shot level interest scores. Firstly, the user is asked to provide interest scores to all the video sequences in the home video collection. Then, the video sequences are decomposed into sub-shots which are represented by key-frames. Consequently, we employ multi-scale spatial saliency analysis to remove the foregrounds and model the background scenes based on histogram of visual words. Finally, the interest scores are propagated from video level to sub-shot level by using gradient descent algorithm. Experimental results demonstrate the effectiveness, efficiency, and robustness of our framework.

Categories and Subject Descriptors
H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems-video.

General Terms
Algorithms, Design, Human Factors, Experimentation.

Keywords
Memory Matrix, Background Scene Modeling, Interest Score Propagation.

1. INTRODUCTION
Nowadays, with the rapid development and wide application of digital media devices, home videos have become more and more popular in daily lives. However, since they are usually taken by amateur users, there exist lots of redundant information and lower quality segments compared with professional videos (e.g. movie, TV series). In order to effectively organize home videos and make them as readable as professional productions, many solutions have been proposed. Existing commercial video editing systems, such as Adobe Premiere and Ulead Video Studio are of great help for editing video. But to make good use of these systems is time-consuming and tedious while requiring significant editing skills. Therefore, in recent years, various efforts have sprung up aiming to automatically analyze home videos and provide users satisfactory experiences.

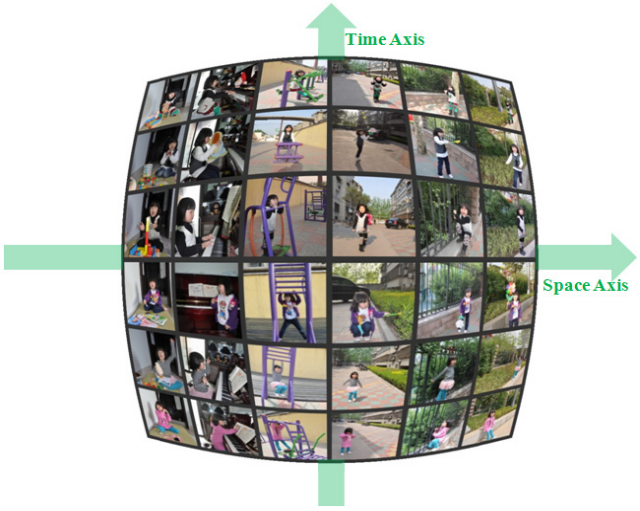

Figure 1. Memory Matrix.

Related works can be roughly divided into three categories: 1) From the viewpoint of understanding video content, aiming to provide users "excited" or "attractive" segments. In [3], Automated Video Editing (AVE) is proposed as an automatic solution for highlight extraction and video-music matching. In [9], it presents a generic framework of user attention model which can be used for video content importance ranking. 2) From the viewpoint of understanding video quality, aiming to provide users "high visual quality" segments [2, 10]. Besides, video quality can also be used to detect and correct the lighting and shaking artifacts of home videos [14]. 3) From the viewpoint of video transformation effects, such as fast/slow motion, thresholding, binarization and watercolor, which facilitate generating more compelling and interesting results [4]. Although all of them have provided convincing results for home video analysis, it still leaves room for a more convenient and interesting user experience:

1. Most of the automatic analysis and selection schemes for home video recommendation are based on objective assessments (e.g. attention, quality, transformation effects). However, we believe that user is the ultimate receiver for home video and there is a need for personalized home video analysis which is fully oriented to the user's individual taste.

MM'10, October 25–29, 2010, Firenze, Italy.

2. Existing methods like video summarization can help users better organize video clips taken at one specific moment, which we call "memories along the time axis". However, users may also want to organize the video clips taken at similar locations, which we call "memories along the space axis". By combining the time axis and space axis together, we can form a matrix-like representation for home video.

Motivated by these observations, this paper presents a novel user experience called Memory Matrix which could facilitate users to re-experience the joy of their memories, travelling along not only the time axis but also the space axis. This is the first contribution of this paper. Memory Matrix is a matrix-like organization for video sub-shots (the reason why we choose sub-shot as our basic unit can be found in section 2.1). As illustrated in figure 1, the X-coordinate represents the space axis while Y-coordinate denotes the time axis. The Memory Matrix automatically organizes home video sub-shots according to the time stamps and background scenes. Moreover, given a specific query, it can also provide the user optional summaries along the time axis or space axis. Since home video collection usually consists of a great number of sub-shots taken at different locations, Memory Matrix is actually a sparse matrix. To avoid this, we eliminate locations with low occurrence frequencies.

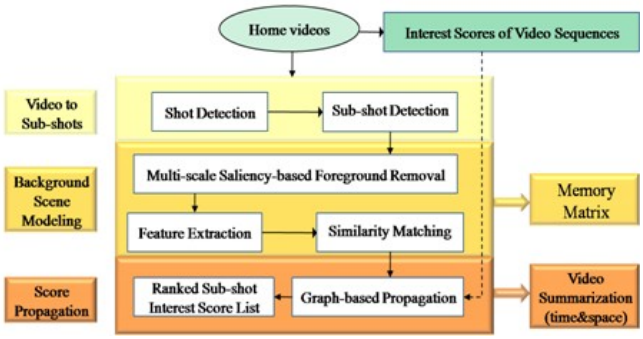

Figure 2. Flowchart of our proposed scheme.

The other contribution is a user-oriented video sub-shot rating method, which adopts a top-down propagation mechanism. The user only has to provide interest scores to the input video sequences and then the sub-shot level interest scores can be automatically obtained based on a propagation algorithm. Figure 2 illustrates the flowchart of our scheme. Firstly, the user is asked to provide interest scores (Bad-1, Poor-2, Fair-3, Good-4, and Excellent-5) to all the video sequences in the collection. Then, the video sequences are decomposed into sub-shots which are represented by key-frames. We employ multi-scale spatial saliency analysis to remove the foregrounds and model the background scenes based on histogram of visual words. The Memory Matrix can thus be established according to the taken time and background scenes of the sub-shots. Finally, we propagate interest scores from video level to sub-shot level by using gradient descent algorithm. Here, we have the following hypothesis:

If two sub-shots are taken at similar locations by the same camera with the same resolution, they are likely to stand on the same interest score level which indicates the degree how user likes them. The difference, if exists, is tenuous.

The third contribution is a new method for home video summarization (time & space) which is based on the user's interest scores of the sub-shots. The experiments in section 3 further show the effectiveness of the proposed Memory Matrix for video summarization.

2. THE PROPOSED APPROACH

2.1 Video to Sub-Shots

Firstly, the input videos are segmented into shots using the shot boundary detection method proposed in [7]. As most of the shot boundaries in raw home videos are simple cuts, they are much easier to detect correctly in comparison with professionally edited videos. Each shot is then decomposed into sub-shots by a reliable motion threshold-based algorithm [6]. The reason why we choose sub-shot as our basic unit is that a shot in raw home videos usually lasts a relatively long time and contains inconsistent content, while sub-shot usually has consistent camera motion and self-contained semantics.

2.2 Background Scene Modeling

As mentioned above, Memory Matrix organizes sub-shots both by their time stamps and background scenes. In that case, we have to remove the influence of foregrounds and classify the sub-shots according to their backgrounds. Background modeling and subtraction is a common but difficult task in the field of computer vision. It can be achieved via classical methods like Gaussian Mixture Modeling (GMM) [13]. However, obtaining real time results is difficult for the existence of problems like "ghost" which is due to the slow updating speed of background. In this paper, as a simplified version designed for Memory Matrix, we select one key-frame to represent the whole sub-shot and employ saliency analysis to remove foregrounds and obtain background maps.

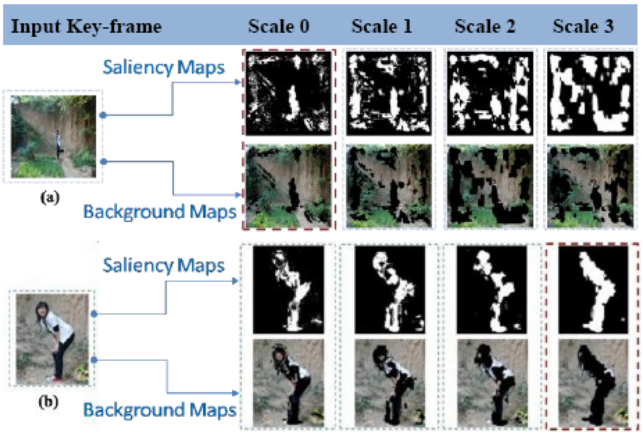

Figure 3. Visualized example for foreground removing.

Intuitively, the regions with large saliency values in key-frame are regarded as foregrounds. However, the proportion of foreground areas varies dramatically between different key-frames that makes the single-scale saliency map appropriate only for specific cases. Motivated by the existence of this problem, we propose a background scene modeling scheme based on multi-scale saliency analysis. Firstly, four spatial scales are created using dyadic Gaussian pyramids, which progressively yield horizontal and vertical image-reduction factors ranging from 1:1 (scale 0) to 1:8 (scale 3). Then center-surround differences are calculated to obtain the saliency maps [5] in accordance with the four scales. We simply regard the regions with high saliency values as foregrounds and remove them to obtain four background maps for

each key-frame. Figure 3 shows a visualized example of our foreground removing algorithm. Noted that the proportions of foreground areas for two input key-frames are different, it demonstrates the reason why we have to use multi-scale spatial saliency analysis in background modeling. According to this figure, scale 0 is suitable for key-frame (a) which has low foreground proportion, while scale 3 is better for the key-frame (b) which has relatively high foreground proportion.

After obtaining the background maps, it comes to the stage of background similarity matching. In practice, the illumination conditions for the videos taken at similar locations could change dramatically (e.g. daytime, weather, switching light). In such circumstances, global features (e.g. global color histogram) may lose its discriminative power. Alternatively, the proposal of local features such as SIFT [8] provides us a much more promising orientation. In our approach, SIFT is first detected and quantized into visual words [12]. The similarities between key-frames' background maps are calculated by the histogram intersection for all the visual words in four background maps. We further construct background scene models for all the key-frames extracted from the sub-shots by clustering them with their near neighbors and output the cluster center as the models. Based on the background scene modeling scheme, supposing we have video sequences taken at *m* different times (*m* rows along the time axis in Memory Matrix), and we further cluster the sub-shots into *n* scenes (*n* columns along the space axis in Memory Matrix), the $m \times n$ Memory Matrix is established. A typical Memory Matrix is illustrated in figure 1.

2.3 Propagation from Video Score to Sub-Shot Score

Given a small number of interest scores on the video level, we aim to obtain the interest scores for sub-shots by a propagation algorithm. The interest scores of sub-shots are used for summary generation, and the algorithm is detailed as follows:

Input: Video sequences labeled with *m* time stamps (*m* rows along the time axis in Memory Matrix) and *n* background scene labels (*n* cluster centers from background scene modeling).

Output: Interest Score (*IS*) for all elements in the Memory Matrix. (E.g. $IS(i,j)$ related to sub-shot $S(i,j)$ located at the i^{th} row and j^{th} column of Memory Matrix).

Step 1: Construct the graph $G =< S,W >$, where *S* is the sub-shot set derived from the input video sequences. Edge $w_{ij,uv}$ represents the pair-wise similarity between sub-shot $S(i,j)$ and $S(u,v)$. Then, we can obtain a symmetric similarity matrix *W*.

Step 2: Propagate the scores to each sub-shot by using video scores and the graph constructed in the first step. Interest scores of the sub-shots are initialized with the corresponding input sequences' score. Based on our hypothesis in section 1, similar sub-shots should have similar interest scores. In other words, the more similar two sub-shots are, the smaller the difference is between their interest scores .

$$|IS(i,j) - IS(u,v)| = f(w_{ij,uv}) \quad (1)$$

where $f(\cdot)$represents a monotonic decreasing function and $w_{ij,uv}$ denotes the pair-wise similarity between sub-shot $S(i,j)$ and $S(u,v)$. We further constrain:

$$IS(i,j) \in [1,5] \quad (2)$$

which indicates that interest scores for each sub-shot should range from 1 to 5. Besides, we hope $IS(i,j)$ could be as close to its initial score $IS(i,j)^0$ as possible, that is:

$$IS(i,j) = IS(i,j)^0 \quad (3)$$

By applying penalty function, (1), (2), (3) can be transformed into an unconstrained optimization problem:

$$\min\{\sum_{ij,uv}\lambda_{ij,uv}[(IS(i,j)-IS(u,v))^2 - f(w_{ij,uv})^2]^2 + \alpha\sum_{ij}[\max\{0,1-IS(i,j)\}]^2$$
$$+\beta\sum_{ij}[\max\{0,IS(i,j)-5\}]^2 + \gamma\sum_{ij}(IS(i,j)-IS(i,j)^0)^2\} \quad (4)$$

For simplicity, here we choose $f(w_{ij,uv}) = \frac{1}{w_{ij,uv}}$, $\lambda_{ij,uv} = 1$, $\alpha = 10$, $\beta = 10$, and $\gamma = 1$. By adopting the gradient descent method, we could propagate scores from video level to sub-shot level using similarity matrix by an iterative procedure. The performance of such a graph-based propagation technique relies on the similarity matrix *W*, penalty coefficients $\lambda_{ij,uv}, \alpha, \beta, \gamma$, and the number of iterations in gradient descent algorithm.

3. EXPERIMENTS

We evaluate the effectiveness of Memory Matrix by generating video summaries both along the time axis and space axis. The goal for time axis evaluation is actually the same with traditional video summary generation schemes, and we compare the proposed approach with two baselines presented in [10] and [11]. For the task of space axis evaluation, we show several examples.

3.1 Time Axis Evaluation

We generate video summaries by selecting sub-shots with high interest scores from the original video. According to [10], we set the skim ratio for summarization as 5%, 10%, and 15%. Six volunteers providing their home video collections to our data set are invited to do this user study. Firstly, given the users' video collections, we decompose the sequences and generate six Memory Matrices respectively. Then the video summaries are generated based on the interest scores of the sub-shots. Details of the test data can be found in table 1.

Table 1. Test data

User	#Sequences	Total Time	Shot	Sub-shot
1	7	01:21:38	65	256
2	11	02:10:10	78	225
3	9	00:56:20	45	189
4	7	02:20:19	92	547
5	10	01:13:40	57	468
6	9	01:46:23	83	677

We compare the proposed approach with other video summarization schemes, including quality based [10] and attention based [11]. According to [10], given the skim ratio, the strategies of sub-shot selection should follow three criteria: 1) the total duration of selected sub-shots equals given duration; 2) the overall interest/quality/attention scores of selected sub-shots are the maximum; 3) selected sub-shots are uniformly distributed.

The owners of the video sequences are invited to provide their preference to all the resulting summaries generated by the three approaches. Here, we choose a paired-comparison test scheme [1] which asks a participant to compare two summaries simultaneously and vote which one is more preferable. Figure 4 shows the experimental results. The colors in figure 4 stand for different

skim ratios. The bin value denotes the proportion of our results which are more preferable than the baseline. For most of the cases, the proposed approach receives relatively higher preference score (preference > 0.5) than the other two, which implies the effectiveness of our method.

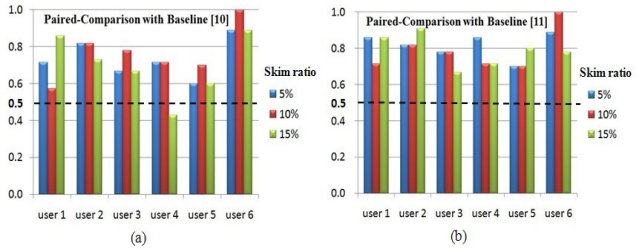

Figure 4. Comparison results of video summarization.

3.2 Space Axis Example

After obtaining the similarity matrix *W* and the interest scores of each sub-shot, summarization along space axis for home video becomes an easy task. Some of the results are shown in figure 5. The key-frames in each column denote the sub-shots which are clustered into the same background scene model and we notice that the proposed method can effectively retrieve similar locations in home video. Multi-scale saliency-based foreground removal ensures the precision and robustness of the background similarity matching. Graph-based propagation technique enables the interest score assignment reasonable and efficient.

Figure 5. Visualized example for travelling along space axis.

4. CONCLUSION

This paper presents a novel user experience called Memory Matrix for home video. The main contributions are:

1. We establish a matrix-like representation for home video, which could facilitate users to re-experience the joy of their memories, travelling along not only the time axis but also the space axis.
2. We aim to provide a user-oriented summarization scheme for home video. Intuitively, considering the large number of sub-shots in video sequences and the lack of consecutive plots in sub-shots, it is tedious to manually assess the interest scores for the sub-shots by observers. In this paper, the user only need to provide interest scores to the input video sequences and the sub-shot level interest scores can be automatically obtained based on a propagation algorithm.
3. Based on Memory Matrix, a new method for home video summarization (time & space) is proposed. Experimental results on time axis demonstrate the advantage of our proposed method. Besides, the summarization on space axis is novel and of practical value for home video users.

The proposed matrix representation can also be implemented for home photo album. In future, we will extend the two-dimension matrix representation into a cube (time-space-character, three-dimension). In addition, other video resources such as TV series can also be taken into consideration.

5. ACKNOWLEDGMENTS

This work was supported in part by National Natural Science Foundation of China: 60833006 and 60702035, and in part by Beijing Natural Science Foundation: 4092042.

6. REFERENCES

[1] K-T. Chen, C-C. Wu, Y-C. Chang, and C-L. Lei. A Crowdsourceable QoE Evaluation Framework for Multimedia Content. Proc. ACM MM, pp.491-500, 2009.

[2] A. Girgensohn, J. Boreczky, P. Chiu, J. Doherty, J. Foote, G. Golovchinsky, S. Uchihashi, and L. Wilcox. A Semi-automatic Approach to Home Video Editing. Proc. UIST, pp.81-89, 2000.

[3] X-S. Hua, L. Lu, and H-J. Zhang. Ave – automated home video editing. Proc. ACM MM, pp.490-497, 2003.

[4] X-S. Hua, and H-J. Zhang. Content and Transformation Effect Matching for Automated Home Video Editing. Proc. ICIP, pp.1613-1616, 2004.

[5] L. Itti, C. Koch, and E. Niebur. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol.20, No.11, pp.1254-1259, 1998.

[6] J-G. Kim, H. S. Chang, J. Kim, and H-M. Kim. Efficient Camera Motion Characterization for Mpeg Video Indexing. Proc. ICME, pp.1171-1174, 2000.

[7] C. Liu, H. Liu, S. Jiang, Q. Huang, Y. Zheng, and W. Zhang. JDL at Trecvid 2006 Shot Boundary Detection. TRECVID 2006 Workshop.

[8] D. G. Lowe. Distinctive Image Features from Scale-invariant Keypoints. International Journal of Computer Vision, Vol.60, No.2, pp.91-110, 2004.

[9] Y-F. Ma, L. Lu, H-J. Zhang, and M. Li. A user attention model for video summarization. Proc. ACM MM, pp.533-542, 2002.

[10] T. Mei, X-S. Hua, C-Z. Zhu, H-Q. Zhou, and S. Li. Home Video Visual Quality Assessment with Spatiotemporal Factors. IEEE Trans. Circuits and Systems for Video Technology, Vol.17, No.6, 2007.

[11] X. Qiu, S. Jiang, Q. Huang, and H. Liu. Spatial-Temporal Video Browsing for Mobile Environment Based on Visual Attention. Proc. ICME, pp.1282-1285, 2009.

[12] J. Sivic, and A. Zisserman. Video Google: A Text Retrieval Approach to Object Matching in Videos. Proc. ICCV, pp.1470-1477, 2003.

[13] C. Stauffer, and W. E. L. Grimson. Adaptive Background Mixture Models for Real-Time Tracking. Proc. CVPR, pp.246-252, 1999.

[14] W-Q. Yan, and M. S. Kankanhalli. Detection and Removal of Lighting & Shaking Artifacts in Home Videos. Proc. ACM MM, pp.107-116, 2002.