21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Integrated Personalized Video Summarization and Retrieval

Hessamoddin Shafeian, Bir Bhanu
Center for Research Intelligent Systems, University of California at Riverside
hshaf001@ucr.edu, bhanu@cris.ucr.edu

Abstract

We propose an integrated and personalized video retrieval and summarization system. We estimate and impose appropriate preference values on affinity propagation graph of the video frames. Then, our system produces the summary which is useful for the user in her/his relevance feedback and for the retrieval module for comparing video pairs. The experiments confirm the effectiveness of our approach for various query types.

1. Introduction

Providing relevance feedback for video retrieval systems takes much more time than their image retrieval peers. This is one of the reasons why video retrieval is less understood than image retrieval. Since video summarization provides a quick overview of video content to the user, its integration with retrieval can help to solve this problem.

Currently, the relevance feedback's role in video retrieval is limited to the estimation of the weight parameters of different feature vectors [1], or the weight of the different data modalities [9], or in its more advanced usage, to learn user's preference from the interactions or some other basic functions. Beside other shortcomings such as the training requirement, there is very few systems in which the retrieval is assisted by personalized summarization.

The first type of personalized video summarizations have no interaction with the user and learn her/his interest either by inspecting the user's profile [8] or her/his browsing behavior [2], hence they are deprived from information provided by the user including potential changes in her/his information need. The second type of such systems are interactive. Many interactive video summarization systems demand technical details from the user, which is not desirable. For example they might ask about the summary duration [11] or to generate exemplars [4].

In response to the mentioned shortcomings, the contributions of this paper are: (1) We propose personalized video summarization that helps a user to quickly comprehend a video and give relevance feedback, (2) We devise an integrated system in which the interactive video retrieval and video summarization modules collaborate, (3) We propose a method to estimate the user's preference, (4) We enable the system to capture and accumulate the user's experience.

2. Technical Approach

The block diagram of our integrated search system is shown in Figure 1. The search process begins when the user submits a query video and the system returns some of the most similar videos to the query known as *top videos*. At the end of each iteration, the summary of the top videos are displayed and relevance feedbacks are given by the user. The process terminates when the user refrains from giving relevance feedback. We then use this preference to generate personalized storyboard summaries for the retrieval module. Retrieval uses these updated key frames for the sake of comparing videos with the query.

2.1 Affinity Propagation for Summarization

2.1.1 Affinity Propagation

Recently, an interesting factor graph-based clustering approach called *Affinity Propagation* (AP) is introduced [3] which gains higher fitness value than the others. Soon afterwards, AP was employed as a tool for clustering video frames to generate storyboard summaries [6].

• **AP Graph:** AP algorithm operates by simultaneously considering all data points of the underlying factor graph as potential exemplars and exchanging messages between data points until a representative set of exemplars and clusters emerges. For our summarization application we consider the video frames as data points and key frames are the exemplars. The storyboard sum-

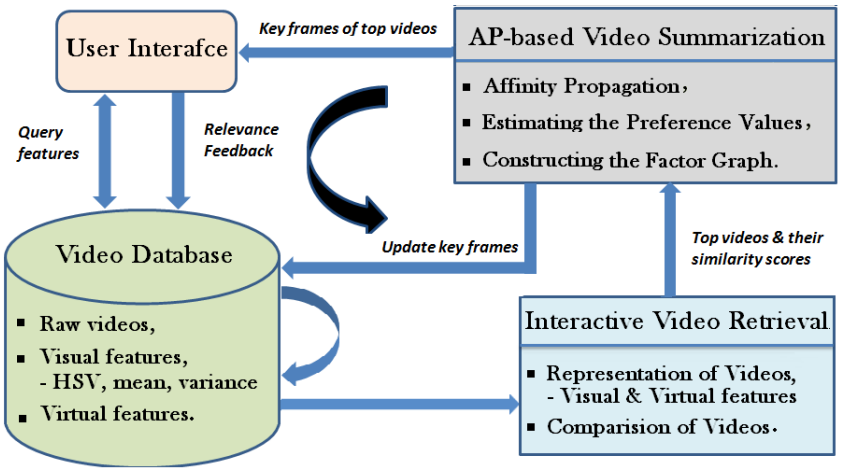
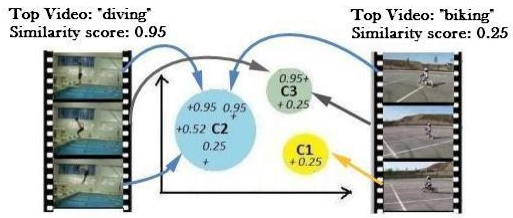
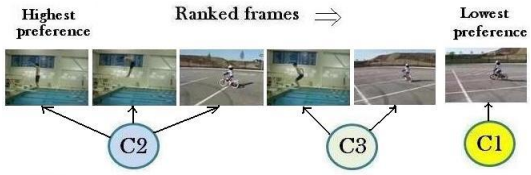


Figure 1. Block diagram of the proposed interactive video summarization and retrieval system.



(a) All frames of all top videos are clustered in visual space.

(b) All clusters are ranked based on the total frame scores.



(c) All frames inside the clusters are ranked based on the individual similarity scores, resulting in a long ranked list of frames from all top videos.

Figure 2. Process of estimating preference value of the video frames.

mary is created by arranging exemplar frames and eliminating the rest.

• **Similarity Matrix and Preference Value:** AP algorithm takes a similarity matrix S as input. $S(i,k)$ indicates how well-suited is data point k serves as an exemplar for data point i. For our video summarization algorithm, we compute this value by visual similarity of the frames (data points). The diagonal elements of this matrix are known as *preference* and their role is to impose the preference for the data points to be selected as exemplars. It is suggested by Frey and Dueck [3] to set the preference value for all points as median of the rest of elements in matrix S. But for personalization, since the different frames have different importances, we set higher (or lower) preference values for those frames that are more (or less) important for the user.

2.1.2 Estimating Preference Values

In Figure 2 we displayed the process of estimating the preference value of each frame of the top videos. The top videos and their similarity scores from the retrieval module (see section (2.2.2)) are the inputs to this process. The score for all the frames in a top video is the similarity score of the video to the query.

(1) Initially all the frames of all top videos are clustered according to their visual features (Figure 2(a)). Each resultant cluster is a mix of different frames from potentially different videos along with their similarity score. The weight of each cluster is the sum of its frame scores.

(2) The higher the weight of a cluster, the more important the cluster is since it includes a more com-

mon visual pattern. Next, the clusters are ranked by their weights in descending order (Figure 2(b)).

(3) Since each ranked cluster is an amalgam of frames from different top videos with different importance (similarity scores), we rank the frames within each cluster as well. This leads to a big ranked list of all frames in top videos (Figure 2(c)). The frames in higher similarity scores within the higher ranked clusters appear in higher ranks.

(4) Since a higher rank for a frame indicates its importance for the user, we convert it to a value by $p_i = 1 - \frac{\textit{Rank of frame i-1}}{\textit{Total \# of frames}}$.

2.1.3 Constructing the Factor Graph

(1) Now, for each top video, we construct a graph in which the top videos' frames are its nodes (data points) and the edges connecting those nodes are weighted by the visual affinity of the corresponding frames (elements of S matrix). Visual affinity is the negative of the Euclidean distance.

(2) $preference$ values of the frames (nodes) which belong to the top videos are set. For frame i this value is set to $K' \times p_i$ where K' is a constant.

(3) AP algorithm is executed and the outcome are the exemplars and their corresponding nodes.

(4) Storyboard summary of the top video is generated by sorting the exemplars (key frames) according to their order in the original video. The weight of each key frame (used in Equation (1)) is proportional to the number of frames (nodes) it serves as their exemplar.

2.2 Interactive Video Retrieval

2.2.1 Representation of Videos

• **Visual Features:** In offline stage, for every frame of all videos residing in the dataset, we extract HSV color space and compute the mean and variance of its channels $[H_{1-5}, S_{1-5}, V_{1-5}, \mu_H, \sigma_H^2, \mu_S, \sigma_S^2, \mu_V, \sigma_V^2]$.

• **Virtual Features:** In order to enable the video search system to capture and accumulate the user's experience, we adapt virtual features from image search domain [10]. Contrary to the visual features, virtual features are generated during the online stage and accumulate and expand as the user gives the relevance feedback. Virtual feature's format is as follows $VF(T) = c_1^{e_1} \otimes c_2^{e_2} \otimes ... \otimes c_m^{e_m}$, subject to $0 < e_1 \leq e_2 \leq ... \leq e_m \leq 1, \sum_{i=1}^{m} e_i^2 = 1$, where T is a video clip in the dataset, c_m is the m^{th} concept of the virtual features and e_m is its weight. The operator $\otimes$ aggregates the distinct concepts within a virtual feature.

2.2.2 Comparison of the Videos

Visual similarity between query video Q and the target video T is:

$$Sim_{visual}(Q,T) = 1 - \frac{\min(d(Q,T), d(T,Q))}{\min(d(Q,T), d(T,Q)) + K} \quad (1)$$

where $d(Q,T) = \sum_{i=1}^{m} w_i^q w_j^t \min_j \|q_i - t_j\|$. q_i is i^{th} key frame of query Q and t_j is j^{th} key frame of the target video T. w_i^q and w_j^t are the weights of q_i and t_j in their videos. K is a constant for normalization.

The similarity between virtual feature vectors of query (Q) and target (T) video is their dot product [10]. The total similarity score of T with respect to Q is the multiplication of their similarity on visual (Equation (1)) and virtual feature (high-level) spaces.

3. Experimental Results

3.1 Video Data, Parameters and Metric

We carried out the experiments on YouTube Action Dataset [7] which contains 1581 videos in 11 different realistic action categories. The display number or the number of top videos displayed in GUI is 10. K in Equation (1) is set to 1000, K' in section (2.1.3) is 10 and the length of the virtual features vector is equal to 200. Since our video retrieval and summarization modules are coupled with each other and mutually beneficial, we take the retrieval precision as a measure of the performance of both modules.

3.2 Experiments

To evaluate our system in different aspects, we conduct four different experiments for three iterations - i.e., EXP1, EXP2, EXP3 and EXP4 - and display the results in Table 1.

• **Effect of Summarization:** According to the Table 1, at the third iteration in ten out of eleven query types the system with summarization (EXP1) outperforms the one with no summarization (EXP2). For the case with summarization (EXP1), the precision at that iteration varies from 39.53% ('biking/cycling') to 74.81% ('horseback riding'), while at the same time, in the absence of key frames (EXP2) it ranges from 24.00% ('volleyball spiking') to 64.00% ('diving'). The other effect of summarization is the efficiency we gain by comparing only the key frames of the videos.

• **Comparison with Tiny Video [6]:** We add our own retrieval module to the summary generated by Tiny Videos (shown as EXP3). Except 'biking/cycling' in all other categories our system outperforms Tiny Video.

	Precision at iteration I				*Precision at iteration II*				*Precision at iteration III*			
Query types	EXP1	EXP2	EXP3	EXP4	EXP1	EXP2	EXP3	EXP4	EXP1	EXP2	EXP3	EXP4
basketball shooting	30.57	**39.17**	21.72	39.19	31.02	**39.58**	22.56	39.40	**40.60**	39.58	22.56	40.53
biking/cycling	24.16	33.20	**35.70**	35.11	31.44	40.00	**47.80**	40.01	39.53	42.80	**47.80**	41.97
diving	38.69	52.00	24.00	**56.79**	38.02	62.40	28.06	**68.13**	68.32	64.00	30.40	**69.26**
golf swinging	29.20	33.33	30.41	**38.03**	35.31	35.83	34.37	**41.40**	**44.28**	37.50	34.50	42.83
horseback riding	**44.09**	43.60	28.20	40.99	**67.88**	44.00	32.50	45.26	**74.81**	44.80	33.10	46.47
soccer juggling	44.94	51.20	43.81	**52.88**	**65.33**	51.20	50.37	53.74	**72.64**	55.60	50.50	56.71
swing	**43.45**	43.33	30.62	38.08	**69.62**	44.58	35.26	39.75	**73.83**	45.83	35.26	40.75
tennis swinging	46.32	**50.00**	30.90	48.64	**53.76**	50.00	31.15	48.64	**64.64**	62.40	33.10	60.40
trampoline jumping	40.24	46.00	30.33	**48.01**	**54.47**	47.20	32.60	49.14	**63.85**	48.80	34.22	50.60
volleyball spiking	33.26	23.20	15.59	**37.00**	**41.54**	23.20	16.33	37.26	**48.96**	24.00	16.33	37.74
walking w/a dog	**33.57**	27.20	23.11	30.73	**53.12**	32.40	32.40	34.28	**57.64**	33.20	24.93	34.87

Table 1. Precision (in %) for three iterations of various experiments, *EXP1:* Our system, *EXP2*: Our system with no summarization, *EXP3:* Tiny Videos [6], *EXP4:* Evenly subsampled by 3. Bold font is used to highlight the best results among all the experiments for a specific action.

Our personalization step is the reason for this difference. In the summary generated by Tiny Video system [6], the preference values of all frames are the same.

• **Effect of Non-uniform Subsampling:** One of the basic types of summarization is uniform subsampling of the original video, but our method is based on non-uniform sampling. To evaluate the effect of the non-uniform sampling, we compare the precisions of our system (EXP1 in Table) and the integrated systems in which the key frames are extracted via subsampling by 3 (EXP4). As the user continues to give relevance feedback, our system better understands her/his need and finally outperforms the subsampled in nine out of eleven query types.

3.3 Conclusions

Experiments have shown that the summarization not only saves time for the user interaction, but also boosts the retrieval performance. Our method surpassed its rival methods in most action categories. Indeed the accumulated user history through virtual features improves the video retrieval with more accurate similarity scores and these scores in turn are used to improve preference value estimations and hence video personalization.

3.4 Acknowledgments

This work was supported in part by NSF grants 0905671 and 0641076.

References

[1] L.-H. Chen, K.-H. Chin, and H.-Y. Liao. An integrated approach to video retrieval. In *Proceedings of the 9th Conference on Australasian Database*, pages 49 –55, 2007.

[2] A. Ferman and A. Tekalp. Two-stage hierarchical video summary extraction to match low-level user browsing preferences. *IEEE Transactions on Multimedia*, 5:244 – 256, 2003.

[3] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315:972 –976, 2007.

[4] B. Han, J. Hamm, and J. Sim. Personalized video summarization with human in the loop. In *IEEE Workshop on Applications of Computer Vision*, pages 51 –57, January 2011.

[5] A. Karpenko and P. Aarabi. Tiny videos: a large data set for nonparametric video retrieval and frame classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(3):618 –630, March 2011.

[6] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos "in the wild". In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1996 –2003, June 2009.

[7] N. Nitta, Y. Takahashi, and N. Babaguchi. Automatic personalized video abstraction for sports videos using metadata. *Multimedia Tools and Applications*, 41:1 – 25, January 2009.

[8] B. Yang, T. Mei, X.-S. Hua, L. Yang, S.-Q. Yang, and M. Li. Online video recommendation based on multimodal fusion and relevance feedback. In *Proceedings of the 6th ACM International Conference on Image and Video Retrieval*, pages 73 –80, July 2007.

[9] P.-Y. Yin, B. Bhanu, K.-C. Chang, and A. Dong. Long-term cross-session relevance feedback using virtual features. *IEEE Transactions on Knowledge and Data Engineering*, 20(3):352 –368, 2008.

[10] X. Zhu and X. Wu. Sequential association mining for video summarization. In *Proceedings of International Conference on Multimedia and Expo*, volume 3, pages 333–6, July 2003.