QUERY SENSITIVE DYNAMIC WEB VIDEO THUMBNAIL GENERATION

Chunxi Liu[1], Qingming Huang[1,2], Shuqiang Jiang[2]
[1]Graduate University of Chinese Academy of Sciences, Beijing, 100190, China
[2]Key Lab of Intell. Info. Process., Inst. of Comput. Tech., CAS, Beijing, 100190, China

ABSTRACT

With the fast rising of the video sharing websites, the online video becomes an important media for people to share messages, interests, ideas, beliefs, *etc*. In this paper, we propose a novel approach to dynamically generate the web video thumbnails according to user's query. Two issues are addressed: the video content representativeness of the selected video thumbnail, and the relationship between the selected video thumbnail and the user's query. For the first issue the reinforcement based algorithm is adopted to rank the frames in each video. For the second issue the relevance model based method is employed to calculate the similarity between the video frames and the query keywords. The final video thumbnail is generated by linear fusion of the above two scores. Compared with the existing web video thumbnails, which only reflect the preference of the video owner, the thumbnails generated in our approach not only consider the video content representativeness of the frame, but also reflect the intention of the video searcher. In order to show the effectiveness of the proposed method, experiments are conducted on the videos selected from the video sharing website. Experimental results and subjective evaluations demonstrate that the proposed method is effective and can meet the user's intention requirement.

Index Terms— video thumbnail, intention gap

1. INTRODUCTION

With the rapid development of the video sharing websites [1][2][3][4], online video becomes an important way for us to share interests, ideas, and comments with others. Compared with the traditional videos services, videos on the web have their unique characteristics. They are uploaded by different users and the length is relatively short. There is a lot of useful context information available, such as the video title, video description and the tags. In order to represent the video content properly and provide the user a vivid preview about the video content, video thumbnail is the commonly used technology for video-user interactive. As indicated by [5], the video thumbnail image has strong influence over user's browsing behavior, and studies have shown that more representative thumbnails greatly improve the performance of video search and retrieval tasks and user satisfaction [6]. Usually, a key frame selected by the video owner is used as the video thumbnail. However, many video owners are reluctant to select the video thumbnail among the video frames carefully, but to use the first frame of the video as the default thumbnail. For the users, in order to find their favorite videos from the web, they usually have to provide their query keywords and search through the search engine. After the retrieved video list is obtained, users can choose their favorite videos according to their preference. Among the video related contexts, the thumbnail is the most straightforward way for users to know the content of the video. However, as described above the existing video thumbnails generated for the web video is set as the first frame of the video or selected by the video owner when uploading the videos, which only reflect the preference of the video owner. The thumbnail generated by the video author may not meet the requirement of the other users. Therefore, there exists an intention gap between the thumbnails generated by the owner and the queries provided by the users, as shown in Figure 1.

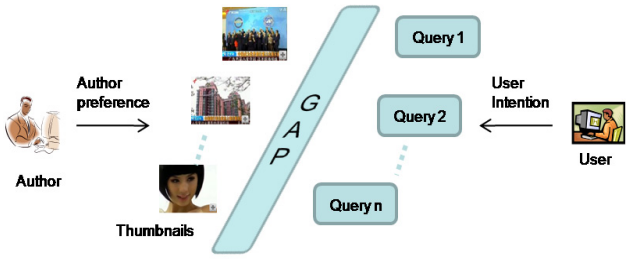

Figure 1. The intention gap between the author generated video thumbnail and the user's query.

In the multimedia domain, many web video based applications have been proposed, such as online video recommendation, web video retrieval, and near duplicate web video search result elimination, *etc*. Some video recommendation methods have been proposed to aggregate related topics according to user demand and guide the users into their interested areas [7] [8]. For web video retrieval, lot of work has been proposed. For example K. Takada *et. al.* [9] propose to use the earth mover's distance by integrating color, motion and sound for web video retrieval. Wu [10] proposes an interesting work to address the near duplicate video removal problem in the web video search result. For video thumbnail generation, there are also some works have been proposed. The works in [11] [12] employ the color, spatial and motion information to select key frame in the shot. DuFaux [13] incorporates high level features such as face detection into the thumbnail generation process. A

recent work in [5] proposes to use the thematic criteria to rank the key frames for thumbnail generation. The selected video thumbnail covers the semantic thematic of the surrounding text. Although some approaches have been proposed for video thumbnail generation, most of them either utilize low-level and high-level visual features of the video sequence to select key frames [11-13], which only consider the quality of the key frame and the intra video content, or utilize the image [5], which better covers the semantic of the surrounding text, as video thumbnail. None of them addresses the user intention gap problem described above. In this paper, we propose a novel personalized approach to dynamically generate the web video thumbnails according to user's query, which not only consider the video visual content, but also meet the requirement of the user. The proposed query sensitive web video thumbnail generation approach is shown in Figure 2.

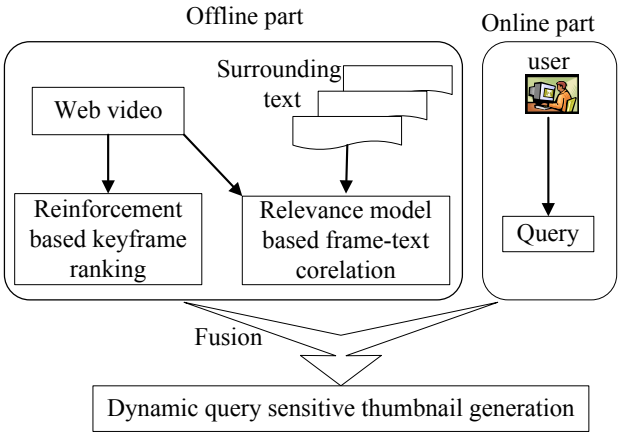

Figure 2. The proposed personalized query sensitive video thumbnail generation approach.

Our approach consists of two parts: offline part and online part. For offline part, the video is first cut into shot. Then, a key frame is selected from each shot to represent the shot content. After that the frames are ranked using the reinforcement based algorithm. And the frame-text correlation is calculated by the relevance model based method. For online part, the web user provides his query words and the query is fused with offline analysis to generate the query sensitive video thumbnail.

The rest of the paper is organized as follows. Section 2 describes the reinforcement based ranking algorithm. Section 3 describes the relevance model based frame-text correlation mining. The final fusion for dynamic query sensitive thumbnail generation is achieved in section 4. Finally, the experimental results are provided in section 5.

2. REINFORCEMENT BASED FRAME RANKING

For key frame ranking the first step is to cut the video into shot. Video shot detection is deemed as the first step toward the semantic video analysis and has been well studied. In our approach we adopt a coarse-to-fine approach [14], which is simple yet effective. After that the middle frame of each shot is selected as the key frame to represent the shot content.

After the key frames are selected, we propose to use the mutual reinforcement based algorithm [15] to rank these frames and assign each key frame with a content score, which reflects the video content representativeness of the key frame. The reinforcement ranking algorithm, which utilizes the mutual information between images, is proposed by Joshi [15] to select representative images from the image set. Figure 3 shows the idea of the mutual reinforcement, where r_i is the rank of image i, and s_{ij} is the similarity between image i and j.

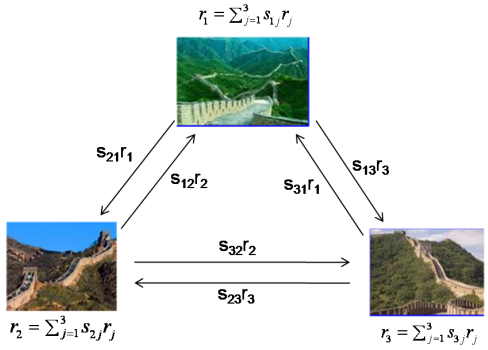

Figure 3. Mutual reinforcement is depicted by the arrows for the three images.

Algorithm1: The reinforcement based ranking

1. Initialize $\vec{r}^0 = (r_1^0, r_2^0, ..., r_N^0)$ for the key frames randomly such that $\sum_{i=1}^{N} r_i^0 = 1$ and $r_i^0 > 0 \ \forall i$.
2. Set $t = 1$.
3. $r_i^t = \sum_{j=1}^{N} s_{ij} r_j^{t-1} \quad \forall i \in 1, ..., n.$
4. $r_i^t = r_i^t / \| \vec{r}_i^t \|$, $\| \vec{r}_i^t \| = \sum_{i=1}^{N} r_i^t$
5. $t = t + 1$.
6. Repeat steps 3 to 5 till convergence.
7. Select the ranking score r as the video content representativeness score for the key frames.

The objective of our key frame ranking is similar to that of [15], which is to select a content representative frame from the keyframe set. Therefore, the algorithm is suitable for our application. The reinforcement based ranking algorithm is described as below. Let $I_1^c, I_2^c, ..., I_n^c$ represent the set of key frames in the cth video. We define the rank of key frame I_i^c as r_i^c, which is the solution to the equation:

$$r_i^c = \sum_{j=1}^{n} s_{ij} r_j^c \qquad (1)$$

where s_{ij} represents the similarity between key frame i and j. The above equation can be solved iteratively by using

Algorithm 1. After the algorithm convergences, we use the ranking scores to represent the content representativeness of the key frames. More detailed discussion of the reinforcement based algorithm can be found in [15].

3. RELEVANCE MODEL BASED IMAGE-TEXT CORRELATION

In this subsection, we propose to adopt the information retrieval based method to calculate the relationship between the key frame and the keyword. Our approach is inspired by the dual cross media relevance model [16], which has achieved encouraging performance in image annotation. In order to calculate the relationship between the frame and the keywords, we estimate the joint distribution of the word and image, which is $p(w,k)$. For traditional relevance model based algorithms [16], a training database with high quality is required. The joint distribution can be computed as

$$\begin{aligned} w^* &= \arg\max_{w\subset V}\{P(w \mid I_u)\} = \arg\max_{w\subset V}\{P(w,I_u)\} \\ &= \arg\max_{w\subset V}\{\sum_{J\in T} P(w,I_u \mid J)P(J)\} \end{aligned} \quad (2)$$

where J is an image in the training set T, w is a word or a set of words in the annotation set V, and I_u is an untagged image. In the dual cross-media relevance model, with the assumption that the word w and the image I_u, is mutually independent given word v, the above function is rewritten as

$$\begin{aligned} w^* &= \arg\max_{w\subset V}\{P(w,I_u)\} \\ &= \arg\max_{w\subset V}\sum_{v\in V} P(I_u / v)(w \mid v)P(v) \end{aligned} \quad (3)$$

In our approach, there are not any training image data set and word set. Therefore, we adopt the simplified version of the dual relevance mode based method. Instead of using a set of words to model the relationship between the word and key frame, we directly calculate the similarity between the key frame and word. Thus the joint distribution of the word and image can be rewritten as

$$\begin{aligned} P(w,I_u) &= P(w)P(I_u / w)P(w / w) \\ &= P(w)P(I_u / w) \end{aligned} \quad (4)$$

where $P(w)$ denotes the probability of the word w, and indicates the importance of the word. In our approach the number of query keywords is limited and we set the word probabilities to be uniform. $P(I_u / w)$ denotes the probability of a key frame given word w, and $P(I_u / w)$ models how the image I_u is relevant to the given word w. This modeling is just consistent with the target of the keyword-based image retrieval. In our approach, after the word and key frame are obtained, we use the word as query to search in the image search engine. Assume the top retrieved n image for word w is $\{v_1,v_2,\ldots,v_n\}$. Then the similarity between the word and the key frame can be calculated as

$$\mathrm{S}(I_u,w)=P(I_u,w) \approx P(I_u / w) = \frac{1}{n}\sum_{i\in n} s(I_u,v_i) \quad (5)$$

Although the retrieved result may contain noise, these data can reflect the main visual aspect of the word.

4. LINEAR FUSION FOR WEB VIDEO THUMBNAIL GENERATION

The web video thumbnails generated by the video owner only reflect the preference of the author, and the existing methods for thumbnail generation only consider the video context or content coverage. However these thumbnails are served for the web users. Therefore, there exists an intention gap between the current thumbnail generation methods and the web video user demand. We propose to combine the representativeness of the key frames in the video and the frame to query keyword relationship together to bridge the intention gap described above. Assume the representativeness of key frame i is r_i, and the number of keywords is K. The final ranking of the key frame is calculated as

$$\begin{aligned} R_i &= \alpha_0 r_i + \sum_{k=1}^{K}\alpha_k s_{ki} \\ s.t. \quad & \alpha_0 + \sum_{k=1}^{K}\alpha_k = 1,\ \mathrm{i} = 1,\ldots,N \end{aligned} \quad (6)$$

where R_i is the final ranking score of key frame i, r_i is the reinforcement based ranking score of frame i, s_{ki} is the similarity between keyword k and frame i, α_0 and α_k are the weighting parameters used to modulate the key frame representativeness and text-to-key frame relation. After the final ranking score R is obtained, we select the frame with the biggest ranking score as the video thumbnail. It is worth noting that our method is extensible as any other ranking features can be easily integrated.

5. EXPERIMENTAL RESULT

In this paper 8 query topics from the web video sharing website Youku [1] are used as our experimental data. The 8 query topics are: 1) *Qinghai earthquake*; 2) *Shanghai World Expo*; 3) *China southwest drought*; 4) *Kyrgyz riot*; 5) *Brother sharp*; 6) *Tiger spring festival gala*; 7) *Morakot*; 8) *Jiabao Wen*. These topics are selected for the reason that they are the top hot search topics at the moment. For each topic the top 5 videos are retrieved. These queries are analyzed and used as keyword to search in the *google image* [17] to retrieve the query related images.

After these videos are obtained, they are cut into shots and key frames are extracted to represent the shot content. For key frame/image content representation, the bag-of-visual word approach [18] is adopted. Firstly, the Scale-Invariant Feature Transform (*SIFT*) based local feature descriptors are detected in each frame/image. Then the local feature descriptors are clustered, a vocabulary tree is generated and the leaf nodes (cluster centers) are considered as the visual words. Finally, we build a histogram for each image based on the visual codebook. In this paper the size of the codebook is set as 1000 according to experience. In

order to measure the similarity between two images/frames, the cosine distance is adopted. After the data and features are obtained, we utilize the method described above to generate the query sensitive video thumbnails. In order to show the effectiveness of the proposed approach, we compare the generated video thumbnails with the original thumbnails in the video sharing website, which are generated by the video owners. A subjective user study is conducted to compare these two thumbnails. 10 users, aging from 23 to 30, are invited to take part in the evaluation. These users are familiar with the web videos and they are used to browsing web video through the web video sharing web site every day. Three scores are defined: better, the same and worse, which mean the thumbnails generated by our method are better than, the same with or worse than the original thumbnails according to the query respectively. The final evaluation results for the 8 topics are shown in Figure 4.

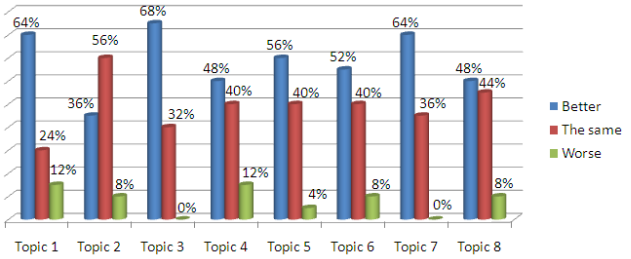

Figure 4. Subjective evaluation results.

From Figure 4 we can see that the thumbnails generated by our approach are generally better than the original thumbnails. In order to show the result more clearly the average evaluation results are shown in Figure 5. From Figure 5 we can see that 54.5% of the users think the thumbnails generated in our approach are better than the original thumbnail, 39% of the user think the two thumbnails are comparable and only 6.5% of the users think the original thumbnails are better. In summary, compared with the original thumbnails, the thumbnails generated by our approach fit better with the user's queries, which show that the proposed video thumbnail generation approach is effective.

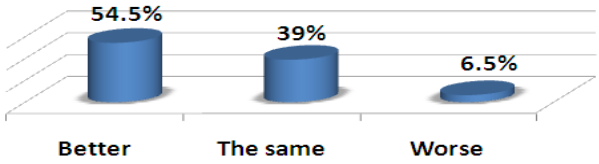

Figure 5. The average evaluation results for the 8 topics.

6. CONCLUSION

In this paper, we propose a personalized approach to dynamically generate the web video thumbnails according to user's query to meet the user's intention requirement. Compared with the existing web video thumbnails, which only reflect the preference of the video owner, the thumbnails generated by our approach not only consider the video content representativeness of the frame, but also reflect the intention of the video searcher. The experimental results and the subjective evaluations demonstrate that the proposed method is effective and can meet the user's requirement. In future we will try to explore more effective ranking algorithm and test more videos to verify the effectiveness of the method. In addition, we will try to fuse more effective features besides the bag-of-visual-word feature to further improve the thumbnail generation result.

7. ACKNOWLEDGMENT

This work was supported in part by National Natural Science Foundation of China: 61025011, 60833006 and 61070108, and in part by Beijing Natural Science Foundation: 4092042.

8. REFERENCES

[1] http://www.youtube.com/.
[2] http://vids.myspace.com/.
[3] http://video.yahoo.com/.
[4] http://www.youku.com/.
[5] Y. Gao, T. Zhang, and J. Xiao, "Thematic Video Thumbnail Selection," ICIP, 2009.
[6] M. Christel, "Evaluation and user studies with respect to video summarization and browsing," SPIE Conference on Multimedia Content Analysis, Management, and Retrieval, January 2006.
[7] C. Liu, S. Jiang, and Q. Huang, "Personalized online video recommendation by neighborhood score propagation Based Global Ranking," ICIMCS, 2009.
[8] B. Yang, T. Mei, X.-S. Hua, L. J. Yang, S.-Q. Yang, and M. J. Li, "Online vide recommendation based on multimodal fusion and relevance feedback," ACM CIVR, pp.73-78, 2007.
[9] K. Takada and K. Yanai, "Web Video Retrieval Based on the Earth Mover's Distance by Integrating Color. Motion and Sound," ICIP, 2008.
[10] X. Wu, C.-W. Ngo, A. G. Hauptmann, and H.-K. Tan, "Real-Time Near-Duplicate Elimination for Web Video Search with Content and Context," IEEE Transactions on Multimedia, volume 11, issue 2, pp. 196-207, February 2009.
[11] Y. Gong, X. Liu, "Generating video summaries," ICIP, 2000.
[12] X. Hua, S. Li, and H. Zhang, "Video booklet," ICME, 2005.
[13] F. DuFaux, "Key frame selection to represent a video," ICIP, 2000.
[14] C. Liu, H. Liu, S. Jiang, Q. Huang, and Y. Zheng, "JDL at TRECVID 2006 shot boundary detection," TRECVID workshop, 2006.
[15] D. Joshi, J. Wang, and J. Li, "The story picturing engine: finding elite images to illustrate a story using mutual reinforcement," ACM SIGMM workshop on MIR, 2004.
[16] J. Liu, B. Wang, M. Li, Z. Li, W.-Y. Ma, H. Lu, S. Ma, "Dual cross-media relevance model for image annotation," ACM Multimedia, 2007.
[17] http://images.google.com/.
[18] S. Zhang, Q. Tian, G. Hua, Q. Huang, and S. Li, "Descriptive visual words and visual phrases for image applications," ACM Multimedia, 2009