THEMATIC VIDEO THUMBNAIL SELECTION

Yuli Gao, Tong Zhang, Jun Xiao

Hewlett-Packard Laboratories, Palo Alto, CA USA

ABSTRACT

Existing techniques to automatically create image thumbnail(s) for videos are mostly based on low-level feature analysis of the video frames, such as color and motion information. However, these approaches do not contain semantic models of the underlying theme of the video, and as a result, the selected frames may not be semantically representative. To address this problem, we propose a theme-based keyframe selection algorithm that explicitly models the visual characteristics of the underlying video theme. This thematic model is constructed by finding the common features of relevant visual samples, which are obtained by querying a visual database with keywords associated with the video. Our initial testing on a set of videos shows promising results of our video thumbnail image selection method.

Index Terms— video thumbnail, video keyframe extraction, image similarity measure, video search and browsing, video theme

1. INTRODUCTION

In recent years we have witnessed an explosion of video sharing as a new killer internet application. The most successful site, YouTube, is now ingesting 15 hours of video content every minute. While the amount of online video content available to people is rapidly growing, the visual and temporal nature of video raises well-known issues in abstraction and description, making it hard for people to effectively browse and search the video collection.

On the visual side the possibilities for abstraction are still images (single frame or storyboard) and video highlights, with the former one used more often for internet videos due to performance reasons. A single 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 [1]. However, current approaches mainly utilize low-level visual features of the video sequence to assess the quality of video keyframes. Typically, they use color, spatial and motion information within the shot as their selection criteria [2][3]. Some have incorporated higher level features such as face detection into the analysis [4].

We believe that such "quality-based" thumbnail generation algorithms may not be adequate to provide satisfying thumbnail image(s) that are semantically representative of the concept or theme of the video. There are several published works on learning video semantic models by observing the video frames [5][6]. In contrast, our method goes beyond that and tries to leverage the wealth of information existing on the Internet.

Overall, our approach composes of two steps. First, candidate keyframes are extracted from the video clip with an intelligent keyframe extraction algorithm which analyzes the video content based on color, motion, face, audio, image quality and video segmentation information. The resulting keyframe set contains a number of high quality video frames that cover different views of the scene and different people/objects in the video.

Secondly, based on relevant textual information of the video clip, such as tags and descriptions available from the video sharing website, we obtain sample images that are representative of the video concept by searching against a visual database. For example, if the video clip is tagged "Christmas", images returned by Google's image search may contain objects such as Christmas tree, Christmas wreath, Santa Clause, etc. Given these visual samples of the concept, we then construct a visual theme model to capture their commonality. Then the candidate keyframes are evaluated against the visual theme model. Keyframes that are similar to the theme model are considered as good semantic representations of the video clip. Such matched keyframes are ranked by similarity to the theme model and can be selected as thumbnail(s) of the video.

The rest of the paper is organized as follows. The intelligent keyframe extraction algorithm and the theme-based keyframe ranking method are described in sections 2 and 3, respectively. Experimental results are presented in section 4, followed by conclusions and future work in section 5.

2. INTELLIGENT KEYFRAME EXTRACTION

A set of representative keyframes are automatically selected from the video clip. We developed two versions of keyframe extraction algorithm, for edited video and unedited video, respectively, due to their different characteristics. Professionally edited video usually contains frequent shot cuts with each shot having relatively short length and single scene. On the contrary, unedited video, especially home video, usually has much longer shots, with plenty of camera motions and dynamic scenes in each shot. Therefore, we focus on tracking changes within a shot in unedited video; while for edited video we pay more attention to detecting shot cuts and selecting frames with the best image quality.

2.1. Keyframe extraction from unedited video

There are two steps for selecting keyframes from unedited video.

2.1.1. Generating Candidate Keyframes

In the first step, a number of candidate keyframes are automatically selected which reveal different views and important people/objects of the video. Two color features, the accumulative color histogram difference and the accumulative color layout difference [7], are computed and compared against empirically determined thresholds in selecting candidate keyframes to ensure that these frames are diversified in color histogram and color layout and therefore represent different views of the scene. Next, to detect video high-lights, camera motions (panning, zooming,

focus, etc.) are tracked in order to estimate user intent. For example, if there is a period of focus after a period of panning or zooming motion, it might indicate a scene or object of interest, thus a candidate key-frame is selected during the focus period. In contrast, fast camera motions suggest content of no interest, and no candidate keyframes are chosen during fast motion periods. Another way to find video highlights is to detect and trace moving objects in the video so that a candidate keyframe can be selected with a moving object of the right size and position in the frame. For details of tracking camera motion and object motion, please refer to [8]. Furthermore, a face detector is used to obtain the number of faces and their positions and sizes in video frames, and this information is used in selecting candidate keyframes containing featured people of the video. Finally, video highlights are also identified by analyzing the audio content and recognizing events such as laughter and speech.

2.1.2. Selecting Final Keyframe Set

In the second step, candidate keyframes are grouped into a number of clusters. The number of clusters can be either pre-determined by the user or a variable that is determined by the video content, i.e. the more diversity detected in the scene, the more clusters. In the former case the K-means algorithm is used; and for the latter case, the adaptive sample set construction method [9] is used. In both cases, the color histogram is used as the basis for clustering.

An importance score is computed for each candidate keyframe based on camera motions, human faces, the size and position of moving object(s) and audio events detected in the frame. The image quality of each candidate keyframe is evaluated, including sharpness, brightness and contrast. One representative frame is then selected from each cluster according to the importance score, closeness to the cluster center, as well as the image quality.

Fig.1 shows keyframes obtained with our method (right) from an 11-second long home video clip which contains one single shot, in comparison with keyframes evenly sampled over time (left). Apparently, our result shows the video theme much better because more keyframes were extracted during the video highlight.

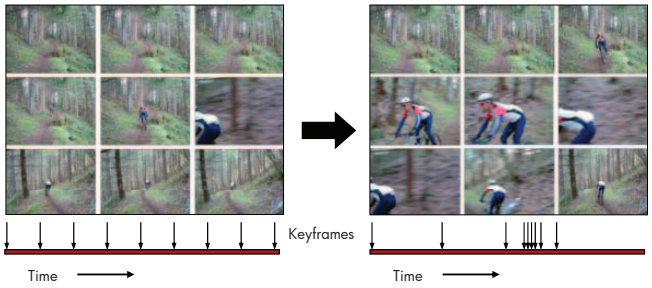

Fig.1. Keyframe extraction from a video about mountain biking.

2.2. Keyframe extraction from edited video

For a professionally edited video, it is first segmented into shots [10]. The keyframe extraction has three modes:

1. One keyframe from each shot: the frame with the best image quality in each shot is selected.
2. Fixed number of keyframes for the entire video: regardless of the length and content of the video, a fixed number of key-frames are extracted.
3. Fixed keyframe rate: one keyframe is selected every fixed period of time (e.g. one keyframe every two seconds).

2.2.1. Keyframe allocation to shots

When mode 2 is selected, a fixed number of keyframes N are to be extracted regardless of the number of shots M. When $N>M$, more than one keyframes are selected from some shots. The shots are ranked according to shot length and a score measuring content changes in the shot [10]. Then, the shot with the highest rank is split into two segments at the frame that has the largest histogram change in the shot. Then, the two segments are treated as two new shots and all the shots are ranked again. This procedure continues until $N=M$. When $N<M$, one keyframe is selected from each shot first. Then, keyframes with low image quality or those from very short shots are removed, until $N=M$. When mode 3 is used, a fixed keyframe rate is selected by the user. For every fixed period of time, one keyframe is chosen based on shot boundaries, the image quality and changes of the content.

2.2.2. Keyframe selection from a shot

To select a keyframe from a shot or video segment, four features are computed for each frame: entropy, sharpness, brightness and contrast. The brightness and sharpness features are used to detect very bright or very dark frames, and to avoid such frames from being included in the keyframe set. Then we choose the frame with the best entropy, sharpness and contrast features. We also skip the beginning and the ending portions of a shot/segment to avoid choosing keyframes from the transitional part of a gradual shot change (e.g. fade-in/fade-out, dissolve), as well as to reduce redundancy among keyframes.

Fig.2 shows some keyframes extracted from an edited video clip about a family Christmas celebration. These keyframes contain objects that obviously indicate the Christmas theme, including the Christmas tree, cookies, stockings, house with lights, etc.

Fig.2. Keyframes extracted from an edited video about Christmas.

3. THEME-BASED KEYFRAME RANKING

Given a set of keyframes extracted from a short video clip, our next task is to rank these frames based on how well they represent the video theme. We assume that each video is associated with a central theme, which can be covered by one or more keywords. For example, for birthday party videos, thematic keywords can include “birthday”, “cake”, “candle”, etc. These keywords can be obtained from the video title, through explicit Q/A, or by other automatic methods. For instances, keywords can be extracted from a) speech from the audio channel; b) tags from video sharing websites; or c) OCR results directly from video frames.

Once we have these keywords, a “contextual-visual link” can be established through keyword search of a visual database to get representative visual samples of the underlying theme. A model

can then be built to capture the "commonality" of these visual samples. Finally, we compared every keyframe of the video against the model for a "semantic" ranking. The frames with highest rankings are the ones to be used as video thumbnail(s).

3.1. Theme Modeling

The core problem here is how to model the essence of a theme given some visual samples obtained from querying a visual database. Since the advent of local keyponit features such as SIFT [11], the bag-of-keypoints approach has become very popular in the computer vision community. It was proved to be effective in solving challenging image categorization problems like the Caltech 101 [12]. However, we observed that the performance of this method degrades drastically as the appearance variation increases within the visual theme.

In our case, we use the Google Image search engine to obtain image samples of a visual theme for a set of relevant keywords. Although the Google-returned image samples are mostly relevant to the queried concept, they are very diverse in terms of visual appearance. An example is shown in Fig.3, which consists of the top ranked images returned from Google with query keyword of "Halloween". The theme of "pumpkin" is very prominent in these images. However, using bag-of-keypoint features such as SIFT to model this theme did not result in satisfactory performance in our experiments, mainly because the appearances of these pumpkins differ not only in size, orientation and view point (which the local features can handle decently), but also in their shape and texture (which the local features can not handle well).

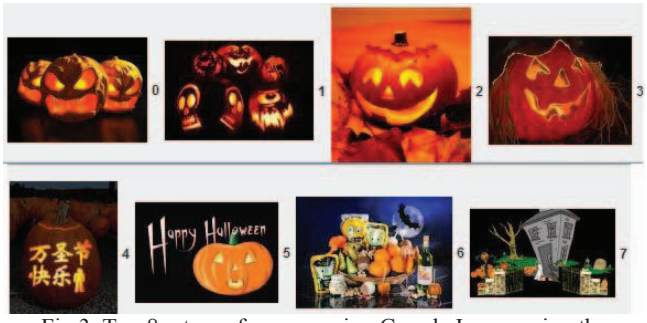

Fig.3. Top 8 returns from querying Google Image using the keyword "Halloween".

However, one key observation is that despite the diversity in appearance the orange pumpkin color stays somewhat "constant". This motivates us to develop a model to capture the "principle theme colors" to work with images that have unique color themes like those taken in Christmas and Halloween. The task of modeling other types of themes remains as future work.

3.2. Principle Thematic Color Components

Given a set of images $S = \{I_1, I_2, \ldots I_n\}$ with the same color theme T, we aim to extract the common principle color components D_{pc} for the entire image set. An algorithm is devised to achieve this goal, and is outlined as the following:

1. For each image I_i in S, cluster its pixels to obtain major color clusters and their corresponding weights: $D_i = \{(c_{i1}, w_{i1}), (c_{i2}, w_{i2}), \ldots (c_{im}, w_{im})\}$, where c_{ij} represents the mean color of the cluster and w_{ij} represents its relative size.
2. Compute the joint color distribution D_s of the entire image set S by concatenating the color clusters obtained from each image I_i in S, i.e.
$$D_s = \bigcup_{I_i \in S} D_i(I_i)$$
3. Iteratively merge color clusters in D_s with color distances smaller than a threshold T_1, until all the remaining color clusters in D_s are mutually distant from each other according to the threshold.
4. Rank the color components in D_s by their weights and obtain D_{pc} as the top K color components in D_s such that the sum of their weights exceeds a percentage threshold T_2 of the sum of total color component weights in D_s.

As an example, the extracted principle color components of the top 20 returned images from Google using the query keyword "Halloween" are shown in Fig.4. It is clear that the algorithm is effective in capturing the prominent "Halloween" thematic colors such as black (representing the night) and various orange colors (representing the pumpkins).

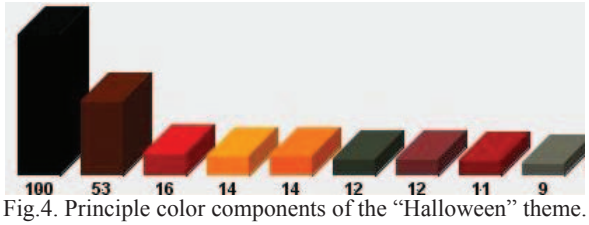

Fig.4. Principle color components of the "Halloween" theme.

Now that we have a color theme model, the extracted keyframes can be ranked by their "visual distance" to the model. Specifically, major color components of each keyframe are first extracted as in step 1 of the algorithm mentioned above. Then the standard Earth Mover Distance (EMD) is adopted [13] as the measure to determine the distance between each keyframe and the common color theme model.

To illustrate the effectiveness of this ranking approach, the ranked keyframe sequence (extracted from a Halloween home video) is shown in Fig.5. Note that the top ranked images are the ones with the thematic orange colors, which contain objects that clearly indicate the Halloween theme. For comparison, the ranked sequence generated by existing selection algorithm based on image quality is shown in Fig. 6.


Fig.5. Keyframes ordered by their distance to the theme model. Top left is the best, bottom right is the worst.

Fig.6. Keyframes ordered by image quality only.

Fig.7. Keyframes ordered by a study participant.

4. TEST RESULTS

Evaluating video thumbnail selection algorithms is a difficult task, as the subjective preferences over what makes up a representative and informative thumbnail image can confound the study result. Furthermore, the video clips used in many traditional information retrieval studies, such as TRECVID [14], are mostly news video, TV programs, movie trailers, advertisements, which are of much higher quality than typical user generated videos shared on sites such as YouTube.

We carried out the experimental tests on a set of short video clips sampled from the web with maximum length of 5 minutes, including themes of holiday e.g. Christmas, event e.g. pumpkin patch ride, place of interest e.g. Stanford campus, and action e.g. skateboarding. Four study participants were asked to view the videos and manually choose video keyframes and order them based on how the frames match the description and the theme of the videos. Notably, the participants were instructed to carefully choose three most relevant and three least relevant keyframes. Fig.7 shows the keyframe sequence of the Halloween home video ranked by one subject.

We then compared the ranked keyframe sequence from the human subjects with the outputs from our algorithm. On average, there were 59% overlapping of the top three keyframe candidates. This implies that our algorithm agrees with human subjects' top choices almost 2 out of 3. Manual inspection of the human subjects' choices found that blurry images, unflattering images e.g. faces being cut off, and images that are too busy or bland, and extreme close ups are often chosen as least relevant keyframes. This indicates that introducing a stricter image quality measure into our keyframe selection algorithm may further reduce the false positive rate of the algorithm.

5. CONCLUSIONS AND FUTURE WORK

This paper is a step towards automatic video thumbnail creation that reflects the theme of the video content. We show that by utilizing the wealth of extra information available on the Internet, and by combining video content analysis with visual theme modeling, we are able to achieve thumbnail results that are more semantic to the video theme. Furthermore, as the process relies on measures easily computed from videos and images, the computational complexity is kept low for large scale applications.

In the future, we would like to explore other models to capture the variation among other types of common visual themes. We also plan to incorporate other image measures, such as quality measures, into the thumbnail selection procedure to generate more visually pleasing as well as more descriptive thumbnails.

6. REFERENCES

[1] M. G. Christel, "Evaluation and user studies with respect to video summarization and browsing," *Proceedings of the IS&T/SPIE Conference on Multimedia Content Analysis, Management, and Retrieval*, January 2006, San Jose, CA.

[2] Y. Gong, X. Liu, "Generating video summaries," *Proc. IEEE Conf. on Multimedia and Expo*, vol.3, p.1559-62, July 2000.

[3] X. Hua, S. Li, H. Zhang, "Video booklet," *IEEE Int'l Conf. on Multimedia and Expo*, July 2005.

[4] F. DuFaux, "Key frame selection to represent a video," *Proc. of ICIP'00*, vol II, p.275-278.

[5] A. Ghoshal, P. Arcing, S. Khudanpur, "Hidden markov models for automatic annotation and content-based retrieval of images and video", *Proc. ACM SIGIR*, 2005.

[6] R. Yan, M. Naphade. "Semi-supervised cross feature learning for semantic concept detection in videos", *Proc. CVPR*, 2005.

[7] T. Zhang, "Intelligent keyframe extraction for video printing," *ITCom'04 Conference on Internet Multimedia Management Systems*, vol.5601, p.25-35, Philadelphia, Oct. 2004.

[8] Y. Wang, T. Zhang, D. Tretter, "Real time motion analysis toward semantic understanding of video content," *Proc. of International Conference on Visual Communications and Image Processing,* pp.740-751, Beijing, July 2005.

[9] S. Bow, *Pattern Recognition*, Marcel Dekker Inc., 1984.

[10] T. Zhang, "Key-frame extraction from edited video," *US Patent Application*, No. 200504967.

[11] D. Lowe, "Object recognition from local scale-invariant features". *Proc. of Int'l Conf. on Computer Vision, 1999.*

[12] F. Li, R. Fergus and P. Perona, "Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories," *IEEE CVPR Workshop on Generative-Model Based Vision*, 2004.

[13] Y. Rubner; C. Tomasi, L. Guibas, "A metric for distributions with applications to image databases". *Proc. of ICCV 1998*, p.59-66.

[14] TRECVID. http://www-nlpir.nist.gov/projects/trecvid