A NEW APPROCH TO AUTOMATIC MUSIC VIDEO SUMMARIZATION

Xi Shao[#], Changsheng Xu[#], Mohan S Kankanhalli**

[#]Institute for Infocomm Research, 21 Heng Mui Keng Terrace Singapore 119613
{shaoxi,xucs}@i2r.a-star.edu.sg
*School of Computing, National University of Singapore
mohan@comp.nus.edu.sg

ABSTRACT

In this paper, a new automatic summarization approach for music videos is presented. The proposed method detects and recognizes lyric captions appearing commonly in Karaoke music video and uses the captions to analyze music video structure and identify the most salient music part. The summary of music video is created based on the salient part. The experiment result shows our proposed method is promising.

1. INTRODUCTION

Nowadays, many music companies are putting their music video products on their websites and customers can purchase them on line. But from the customer point of view, they would prefer to watch the highlights of a music video before they make a decision on whether to purchase or not. On the other hand, from the music company point of view, they would be glad to invoke the buying interests of the music fans by showing the highlight of a music video rather than the whole one. Although there are summaries in some websites, they are generated manually, which needs expensive manpower and is time-consuming. Therefore, it is crucial to come up with an automatic summarization approach for music videos.

There are a number of approaches proposed for automatically creating summaries for text, music and video. Most summarization techniques start with an analysis of the structure or semantics of the source material. The work on statistical text summarization uses term frequency/inverse document frequency (tf/idf) to select sentence, or key phrases that are both representative of the document and differentiate it from other documents [1]. Music summarization techniques typically use a segmentation phase followed by a audio structure analysis process that extracts the most salient part of a whole song [2,3]. The existing video summarization methods can be classified into two categories: key-frame extraction [4] and highlight creation [5]. Key frames can help the user identify the desired shots of video, but they are insufficient to help the user obtain a general idea of whether the created summary is relevant or not. Video highlight creation method can provide an impression of the entire video content or contain only the most interesting video sequences. It has been applied to sports video [6], news video [7], home video [8] and movies [9], but music video analysis and summarization has been neglected amongst the existing work. We proposed a music video summarization method [10]. The music video is separated into music and visual tracks. For music track, we perform music summarization, while for visual track shot are detected and clustered. The final summary is created by aligning the music summary and clustered video shots. An obvious drawback for this method is discontinuous in the boundary of different music segments.

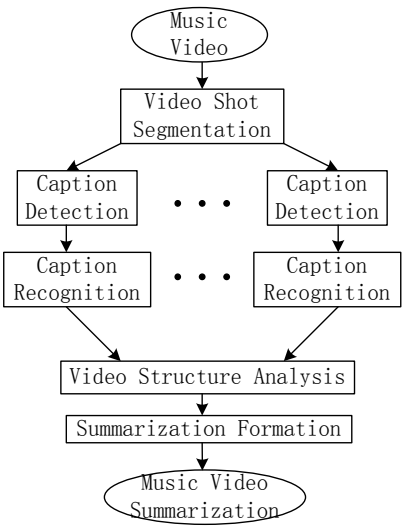

Figure 1 Proposed summarization system block diagram

In this paper, a new automatic summarization approach for music videos is presented. The proposed method makes full use of lyric captions appearing commonly in

0-7803-8554-3/04/$20.00 ©2004 IEEE.

Karaoke Music Video. Figure 1 is the block diagram of the proposed approach. For input music video, we first segment video into shots. For each shot, we select a key frame as the representative frame for that shot and perform caption detection and recognition. Then, in the video structure analysis module, we use Dynamic Programming algorithm to find the most repeated caption as the most salient part of the music video. Finally, summarization is created based on the most salient part being detected.

2. APPROACH

This section explains the proposed approach including four steps.

2.1. Shot Segmentation

First, we need to turn the raw video sequence into a structured data set where boundaries of all the camera shots are identified. After the structuring process the original video sequence can be described by the shot set $S= \{s_1, s_2, \ldots, s_n\}$ where n is the number of shot detected.
For each shot s_i, we choose a key frame f_i as the representative frame of that shot. To detect the most salient caption appearing steadily in the shot, the representative frame f_i is selected in the middle of the shot, other than at the two ends of the shot boundary, because the shot boundaries commonly contain transition frames which will blur the lyric caption.
Now, we have the representative frame set $F= \{f_1,f_2,\ldots,f_n\}$, which can be used to represent the characteristics of shot set S.

2.2. Caption Detection and Recognition

Given the representative frame set $F= \{f_1,f_2,\ldots,f_n\}$, text detection is performed for each representative frame f_i, using the method described in [11].
Several heuristic rules that are common for lyric caption of Karaoke music video are listed as following, which will be used to facilitate the detection process.

a) Lyric caption always appears in the lower half part of the frame.
b) Lyric caption is a bar whose width is larger than height.

The frames that contain the lyric caption will be kept to form the caption frame set F', where $F'=\{f'_1,f'_2,\ldots,f'_m\} \subseteq F$. For each frame in caption frame set F', the content of each caption should be recognized. The low resolution of video (typically 72 dpi) is a major source of problems in text recognition. Today's OCR (Optical Character Recognition) systems have been designed to recognize text in documents, which were scanned at a resolution of at least 200dpi to 300dpi resulting in a minimal text height of at least 40 pixels. In order to obtain good results with standard OCR system, it is necessary to enhance the resolution of segmented text lines. In our experiment, we use cubic interpolation to rescale the text height (normally about 20 pixels) into 40 pixels while keep preserving the aspect ratio.
It should be noted that although there is no OCR software can achieve 100% accuracy, it will not affect the final result much, as the error can be supplemented by the following approximate string matching operation.
After text recognition, the recognition results are saved in the caption set $C=\{c_1,c_2,\ldots,c_m\}$. Each element c_i in this set corresponding to the text content of frame f'_i in caption frame set F'.

2.3. Video Structure Analysis

The aim of video structure analysis is to find the most salient part of a music video. We assume the most salient part of a music video happens in the most salient music part (i.e. chorus). Although what makes a music part distinguished among a music work is not clear, current research typically assumes it to be the most repeated part.
Generally, chorus of a song contains most repeated music phrases. In this paper, music phrase is defined as a short musical passage, which is similar to linguistic sentence in the speech.
Considering the caption set C obtained in previous step. Since a music phrase lasts for several shots which may correspond to several continuous captions in caption set C. We need to merge these continuous captions into one to represent the music phrase corresponding to it. After merging process, music phrase set $P= \{p_1,p_2,\ldots,p_t\}$ has been formed.
Given the music phrase set P, we use Dynamic Programming [12] to match each caption (i.e., p_i) with the caption sequence starting from this caption (i.e.p_i $p_{i+1}\ldots p_t$), since it has been proven efficient for string matching that allows errors ,or called approximate string matching . Suppose we need to match the caption p_i (denoted as X) with the caption sequence starting from this caption (denoted as Y), we should fill a edit distance matrix $\boldsymbol{D_i}$ (X,Y), which is defined as minimum cost of a sequence of modification (insertion, deletions and substitution) that transforms X into Y. In the matrix, the element $D_i(k,l)$ represents the minimum number of modifications that are needed to match $X_{1..k}$ to $Y_{1..l}$. The algorithm can be described as following:

Initial: $D_i\,(k,0)=k\,; D_i\,(0,l)=0;\ 1\le k\le |X|,\ 1\le l\le |Y|$
Recurrence:

$$D_i(k,l)=\min\begin{cases} D_i(k-1,l-1)+\delta(X_k,Y_l) & 1\le k\le |X| \\ D_i(l-1,k)+1 & 1\le l\le |Y| \\ D_i(l,k-1)+1 & \end{cases} \quad (1)$$

where $\delta(X_k,Y_l)$ =0 if X_k=Y_l and 1 otherwise. $|X|$ and $|Y|$ denote the length of string X and Y respectively.

The rationale for the above formula is as follows. $D_i(k,0)$ and $D_i(0,l)$ represent the edit distance between a string of length k or l and the empty string. For $D_i(k,0)$, clearly k deletions are needed on the nonempty string. While for $D_i(0,l)$, because we must allow that any text position in Y is the potential start matching point, we set the first row of the matrix to zeros, which means the empty pattern matches with zero errors at any text position.

The last row of Matrix $\boldsymbol{D_i}(X,Y)$ is defined as function $h_i(r)$,r=1.. $|Y|$. It measures how well the string X matches with different locations shifted by r in the string Y. Figure 2 plots out one caption repetition detection result for the music video "Five Hundred Miles".

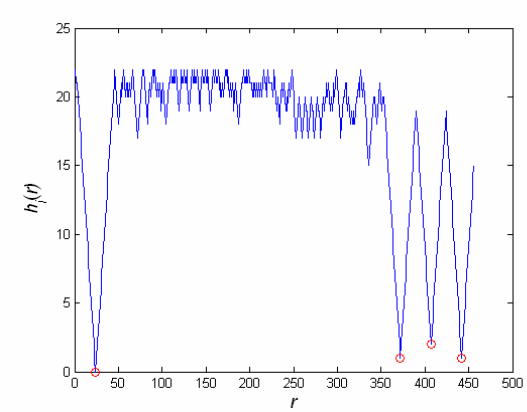

Figure2. One caption repetition detection result

From the figure above, we can see that except for the p_i itself (the first local minimum denoted with circle in the figure), there are other three matching points, also denoted with circles. The three matching points do not ideally equal to zero because of the OCR mistakes. We can set a threshold to find the local minimum of function $h_i(r)$. In our implementation, the threshold is set to 2*(1-OCR accuracy) multiplying the length of text p_i.

Thus, the task to find the salient part of music can be converted to the task to find the most repeated music phrase in the set P. The detailed algorithm is described below:

1) Take the first element in set P, and use Dynamic Programming to find the repeated music phrases in set P.

2) Select the first element in set P, together with its repeated music phrases found to construct a subset R_j. Meanwhile, delete these music phrases in set P. Increase j. Repeat step 1) and 2) until P is empty.

The set R=$\{R_1,\cdots R_j,\cdots,R_k\}$ contains the k subsets, each subset R_j represents a cluster containing the same music phrase in set P.

By counting the number of element for each subset R_j in set R, we can find the subset containing the most repeated music phrase, denoted as R^*_{opt}.

2.4. Music Video Summary Formation

The final music video summarization is based on the most salient music phrase found in previous steps. Since most salient part repeated itself at the different position of the music, we select the first appearance of it as the basis to create our music video summary. The summary creation progress is illustrated in Figure 3.

First, for the subset R^*_{opt}, we select the first music phrase in this subset by time order, link it back to the caption set C where there are several continuous captions related to this music phrase. Then, we link back these captions to shot set S, where there is one shot corresponding to one caption. The music video summary can be created by merging these shots together. If the summary is shorter than the required length, it can grow forward or backward to the previous or next music phrase boundary until it reaches the required length. The previous or next music phrase boundary can be obtained by linking the elements in other subset of set R back to the shot set S.

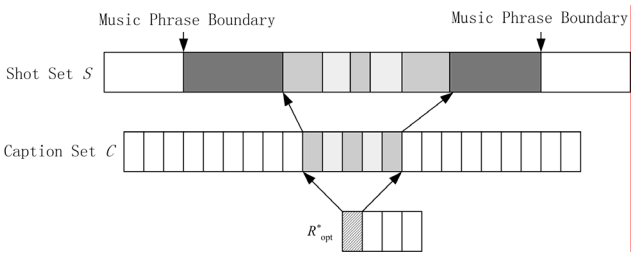

Figure 3. Summarization generating progress

3. EXPERIMENT AND EVALUATION

Since there is no absolute measure of summarization quality available today to evaluate the quality of a music video summary, we employed a subjective user study [13] to evaluate the performance of our music video summarization method. The basic idea of this user study method is to use appropriate attributes to access the users' perception of systems. For an ideal music video summary, the following attributes are considered.

a. *Clarity*: This pertains to the clearness and comprehensibility of the music video summary.
b. *Conciseness*: This pertains to the terseness of the music video summary and how much of the music video captures the essence of the music video.

c. *Coherence*: This pertains to the consistency and natural drift of the segments in the music video summary.

We have evaluated our proposed method on a test set of about 100 general Karaoke music videos. All are English pop music video. The length of music video testing samples is from 2m52s to 4m27s. The length of the summary for each sample is set to 30s.

For this experiment, we invited 20 participants with music experience to evaluate the music video summaries. Before the tests, the subjects could watch each testing sample for as many times as needed till he/she grasped the theme of the sample. Then the subjects watched summaries generated from test samples and rated the summaries in four categories (Clarity, and Conciseness, Coherence, and Overall Quality) on a scale of 1-5, corresponding to worst and best respectively. We employ the overall quality of the video as an attribute to evaluate a summary because it pertains to the general perception of the users to the video summaries. The average grade of summaries from all subjects is the final grade. In order to make comparison, we also asked the subjects to rate the summaries generated using our previous summarization method [10]. Table 1 shows the average scores of the user evaluation using proposed method and our previous method respectively.

From the test results, it can be seen that the summaries using proposed method performed quite well, especially in the coherent attribute, compared with our previous method. This is because our previous method just focus on the most frequent music segments which may occur in different places in a song, and the summary is created by concatenating these segments together. As a result, the discontinuity will happen either in the summarized segment beginning from the middle of music phrase or in the boundary of two different summarized segments. These two problems are avoided in our proposed method as we made the music video summary based on the continuous music phrases. Video summarization examples can be viewed at http://www.comp.nus.edu.sg/~shaoxi/Vsum/vsum2.htm.

Table 1 Results of user evaluation

	Clarity	**Conciseness**	**Coherent**	**Overall Quality**
Proposed Method	4.4	4.7	4.9	4.7
Previous Method	4.3	4.5	4.1	4.2

4. CONCLUSION AND FUTURE WORK

We have presented a new approach to summarize the Karaoke music video. In this approach, the most salient part of the video is located by video structure analysis according to the lyric captions detected and recognized. The video summary is created based on the salient part.

The future work we wish to investigate is to extend this method to Karaoke music videos of other languages such as Chinese and Japanese. Compared with English, these languages are more complex to segment and recognize.

5. REFERENCES

[1] K.Zechner, Fast generation of abstracts from general domain text corpora by extracting relevant sentences, In *Proc.of the International Conference on Computational Linguistics*, 1996

[2] W.Cai and B.L.Vercoe, Music Thumbnailing via Structural Analysis, In *Proc. Of ACM International Multimedia Conference 2003*, p223-p226

[3] B.Logan and S.Chu, Music Summarization using Key phrases, In *Proc.of International Conference of Acoustic, Speech, Signal Processing*,2000.

[4] B.Gunsel, and A.M.Tekalp, Content-based video abstraction, In *Proceedings of IEEE International Conference on Image Processing*, Chicago, IL, 1998.

[5] J.Assfalg, M.Bertini, A.DelBimbo, W.Nunziati and P.Pala , Soccer highlights detection and recognition using HMMs, In *Proceedings of IEEE International Conference on Multimedia and Expo*, Lausanne, Switzerland, 2002.

[6] D.Yow, B.L.Yeo, M.Yeung and G.Liu, Analysis and presentation of soccer highlights from digital video, In *Proceedings of Asian Conference on Computer Vision*, Singapore, 1995.

[7] Y.Nakamura and T.Kanade, Semantic analysis for video contents extraction – spotting by association in news video, In *Proceedings of ACM International Multimedia Conference*, 1997.

[8] Y.Gong, X.Liu and W.Hua , Summarizing video by minimizing visual content redundancies, In *Proceedings of IEEE International Conference on Multimedia and Expo*, pp. 788-791, Tokyo, Japan, 2001.

[9] S.Pfeiffer, R.Lienhart, S.Fischer and W.Effelsberg, Abstracting digital movies automatically, *Journal of Visual Communication and Image Representation*, 7(4): 345-353, 1996.

[10] X.Shao, C.Xu and M.S.Kankanhalli, Automatically generating summaries for musical video, In *Proceedings of IEEE International Conference on Image Processing,* 2003

[11] X.S,Hua, X.R.Chen, W.Liu, H.J.Zhang, Automatic location of text in video frames. *3rd International Workshop on Multimedia Information Retrieval*, Ottawa, 2001

[12] G.Navarro, A guided tour to approximate string matching, *ACM Computing Surveys*,Vol.33,NO.1, March 2001,pp.31-88

[13] John .P.Chin , Virginia A. Diehl and Kent L.Norman, Development of an instrument measuring user satisfaction of the human-computer interface,*Proceedings of SIGCHI'88*.pp.213-218 ,New York,1988