AUDIO AND VIDEO COMBINED FOR HOME VIDEO ABSTRATION

Ming Zhao Jiajun Bu Chun Chen

School of Computer Science, Zhejiang University, Hangzhou, 310027, P.R.China
Contact: zhaoming1999@hotmail.com, {bjj, chenc}@cs.zju.edu.cn

ABSTRACT

With the increasing number of people who can afford to make videos to record their lives, home videos play more and more important role in multimedia. Video abstraction is an efficient way to help review such a huge amount of home videos. In this paper, a home video abstraction technique combining audio and video features is presented. The audio contents are firstly classified as silence, pure speech, non-pure speech, music and background sound using SVMs. Then non-pure speech is further classified into song and other non-pure speech using SVM, and background sound is classified into laughter, applause, scream and others using Hidden Markov Models (HMMs). For video contents, motion level and blur degree are acquired. Finally, video segments containing special effects, such as speech, laughter, song, applause, scream, and specified motion level and blur degree, are extracted as the main parts of the abstract. The remaining parts of the abstract are generated using key frame information. The experimental results show that the proposed algorithm can extract desired parts of home video to generate satisfactory video abstracts.

1. INTRODUCTION

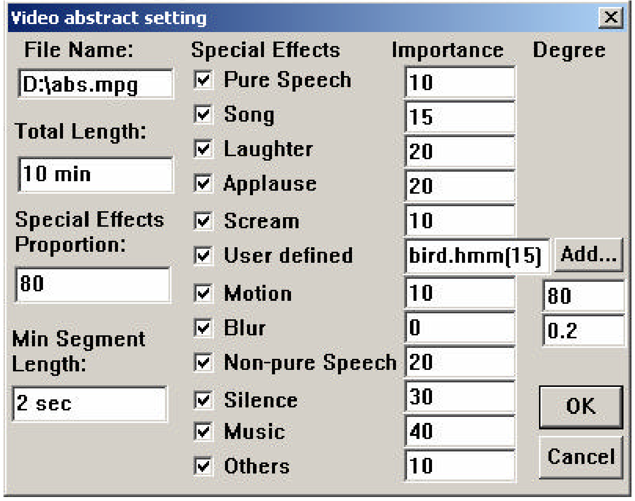
Video abstraction techniques enable a quick browse of a large collection of video data and to achieve efficient content access and representation. Home videos usually add up to many hours of material, which makes it inconvenient for people to review them. And the raw video material is unedited, and therefore long-winded and lacking appealing things. Although video editing would help, it is still too time-consuming. And video editing is inflexible and cannot adjust to the viewers' various needs. However, a system capable of abstracting raw videos into shorter ones automatically can not only offer appealing things but also the flexibility for different purpose.

In the VAbstract system developed by the University of Mannheim, Germany [1]. It mainly used visual feature for the abstraction. Although some audio features are also considered, it was very limited and simple. The Informedia Project at Carnegie Mellon University [2] aims to create a very short synopsis of the original video by extracting the significant audio and video information. Satisfying results may not be achievable using such a text-driven approach on other videos with a soundtrack containing more complex audio contents, which is usually true for home videos from which there are nearly no text keywords to be extracted. In the work reported by A. Hanjalic and H.J. Zhang [3], audio contents are neglected. R. Lienhart [4] proposed an automatic video abstracting method for home video. It paid more attention to visual features. Although audio features are used, it was fairly simple and did not utilize the abundant audio contents effectively.

After extensive investigation, we believe that audio features are especially important for home video abstraction. Usually, people are the focus of home videos. Different sounds made by people can indicate different important events. So, audio contents are very important clues for home video abstraction. It is easier to detect most important events and appealing things by audio features. For example, if there happen to be any interesting things during traveling, people will talk or laugh or cry out. So these events can be detected by audio contents easily, while it is very difficult or even impossible for visual contents to do so. Therefore, audio feature should be paid more attention than before.

Based on the above observation, we attempt an abstraction technique using both audio and visual contents. The audio special effects such as speech, song, laughter, applause, scream are extracted by audio segmentation and classification using SVMs and HMMs. Video special effects (motion level and blur degree) are also acquired. Finally, video segments containing these audio and video special effects are extracted as the main parts of the abstract. As home video abstracts could be used for various purposes, the video abstraction system should possess flexible properties. In this paper, flexibility is achieved through a flexible interface, which is shown in Figure 1. Through this interface, the user can not only specify their interest for different special effects, but also add other effects.

The rest of this paper is organized as follows. Audio segmentation and classification is presented in section 2 and video content processing is presented in Section 3. Section 4 gives the abstraction algorithm. And experimental results are presented in section 5. We conclude this paper in section 6.



Video abstract setting

File Name: D:\abs.mpg

Total Length: 10 min

Special Effects Proportion: 80

Min Segment Length: 2 sec

Special Effects	Importance	Degree
☑ Pure Speech	10	
☑ Song	15	
☑ Laughter	20	
☑ Applause	20	
☑ Scream	10	
☑ User defined	bird.hmm(15)	
☑ Motion	10	80
☑ Blur	0	0.2
☑ Non-pure Speech	20	
☑ Silence	30	
☑ Music	40	
☑ Others	10	

Figure 1 input interface of the video abstraction system

2. AUDIO SEGMENTATION AND CLASSIFICATION

In order to achieve both the accuracy and flexibility of the audio classification, a two-level hierarchy method is used in this paper. In the top/first level, SVMs are used for accuracy and HMMs are used for flexibility in the bottom/second level. It will be detailed at section 3.2.

2.1 Audio feature selection

Feature selection is an important step of audio classification. Different features should be used in different methods and different applications. In this paper, we use different features for SVMs and HMMs.

In our data, all audio clips divided into non-overlapping sub-clips. A sub-clip is of 1 second duration and is further divided into forty 25ms-long frames. The segmentation is performed based on the classification of these one-second sub-clips.

For SVM-based classification, we consider the features, which are used in [5], including 8-order MFCCs, short time energy (STE), zero crossing rates (ZCR), sub-band powers distribution, brightness, bandwidth, the pitched ratio, Spectrum Flux (SF), Linear Spectrum Pair (LSP) divergence shape, Band periodicity (*B*P). All these features are combined as a feature vector of a frame. The mean and standard deviations of these feature vectors over all forty frames are computed, and these statistics compose a new feature vector. Finally, the feature vector is normalized by dividing each feature component by its standard deviation calculated from the ensemble of the training data. The normalized feature vector is considered as the final representation of this one-second sub-clip.

For HMM-base classification, two types of information are contained in the HMMs, i.e. timbre and rhythm. Timbre is generally defined as the quality which allows one to tell the difference between sounds of the same level and loudness when made by different musical instruments or voices. Each kind of timbre is denoted by one state of HMM, and represented with the Gaussion mixture density. Rhythm is the quality of happening at regular periods of time. Here, it is extended to represent the change pattern of timbres in a sound segment. The rhythm information is denoted by transition and duration parameters in HMM. We refer to [6] for details.

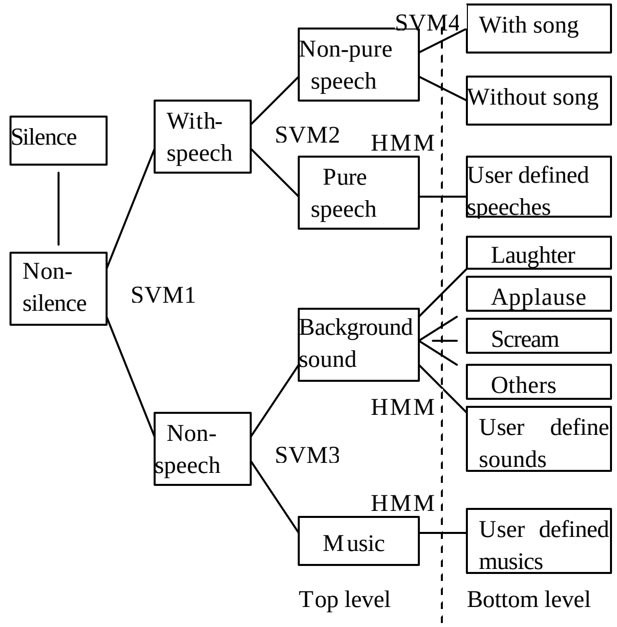


Figure 2 Audio Classification tree using SVMs and HMMs

2.2 Audio classification using SVM and HMM

SVM and HMM are two popularly used classification method. They perform well in difference situations. SVM is a discriminant model, and it is good at classification [7]. HMM is a generative model, and it is good at recognition. Also, it is easy to train HMMs and add more HMMs. So HMMs are easier to be extended to contain more classes. This is very important for the end user to use HMMs to add their defined classes. In order to achieve both the accuracy and flexibility of the audio classification, a two-level hierarchy method is used in this paper. In the first/top level, audio segments are categorized into 5 classes: silence, pure speech, non-pure speech, music, background sound (or special sound) using support vector machines (SVMs) [6]. In the second/bottom level, non-pure speech is further classified into song and other non-pure speech using SVM; and background sound (or special sound) is further classified into 4 classes, including laughter, scream, applause and other background sound using Hidden Markov Models (HMMs). As we can't define all the types of sound in which the users are interested, HMMs provide an extensible method is for the users to add their types. The users can train their sound type with HMMs using the input interface, as shown in Figure 1, and then add them to the abstraction system. Figure 2 illustrates our classification scheme.

In the top/first level, we use the method proposed by [6] to classify audio into five classes: silence, music, background sound, pure speech, non-pure speech, which includes speech over music and speech over noise. The input audio is first classified into silence and non-silence clip. And then, for those non-silence sub-clips, the left 4 classes are classified using SVM classifiers. We use a simpler scheme and construct a bottom-up binary tree for classification, as shown in left part of Figure 2. By comparison between each pair, a unique class label will appear on the top of the tree.

In the bottom/level, non-pure speech is further classified into song and other non-pure speech using SVM; and background sound (or special sound) is further classified into 4 classes, including laughter, scream, applause and other background sound using Hidden Markov Models (HMMs).

After the above classification process, all the sub-clips with the length of one second belong to one audio class. So we merge the consecutive sub-clips with the same class type into an audio segment. And noise filtering process is applied too.

3. VIDEO SHOT DETECTION AND FEATURE EXTRATION

The video contents will be first segmented into video shots. Then key frames, motion level and blur degree are acquired for each shot.

A shot designates a video sequence which was recorded by an uninterrupted camera operation. In our abstraction system, we used the twin comparison algorithm [8]. And the key-frame selection algorithm described in literature [9] is used.

Motion level is measured by color variance. The temporal variance of mean color over all frames in a shot is used as an indicator of the scope of temporal content changes within the shot. The temporal variance is an effective feature to distinguish shots of high-activity level from those of low-activity level.

We used a blur-detection algorithm, proposed by Xavier Marichal et al[10]. It is computational efficient and its result is fairly well.

4. AUDIO AND VIDEO COMBINED ABSTRATION ALGORITHM

After the audio segmentation and classification stage, audio contents are divided into segments and each of them belongs to one of the following types: silence, song, non-pure speech without song, pure speech, laughter, applause, scream, music and others. But song, pure speech, laughter, applause, scream are more important audio contents for home video abstraction. We called them audio special effects. The users can select some of them for their different use. And they can also add their audio special effects with HMMs through the interface of Figure 1. For video contents, the users can specify the motion level and blur degree. We call them video special effects. Now the task of video abstraction is how to use these audio and video special effects to generate video abstract according to the user's various requirements.

The abstraction algorithm starts with a user interaction: the users need to provide the target length Len of the abstract, select the audio special effects and video special effects including video motion level and blur degree or add other audio effects. Then the following steps are taken to generate the video abstract.

1) The length of abstract containing the audio and video special effects is obtained by a proportion a of the target length. In this paper, the default value is 80%. And the user can set this value to his mind for different purpose.
2) If blur degree is specified by the user, each of the other segments containing the special effects are check to see whether their blur degrees meet the user's need. For those which do not meet the user's need, they're discarded.
3) The numbers of segments containing each special effect are calculated separately. Let $N_i \quad (i \in [1..K])$ denote the number of them, where K is the number of special effects including those user selected and added. And the total number of all the segments is $N = \sum_{i=1}^{K} N_i$.
4) The length $L_j \quad (j \in [1..N])$ of each extracted special effects segment is calculated using the results of the above steps. In the default manner, all the lengths L_j are the same, i.e. $L_j = a * Len / N \quad (j \in [1..N])$. Due to the fact that the user will generate video abstract for different purpose, special effect contents have different significance in video abstracts. So in this paper, the user can specify the importance of different special

effects, which is used as weights $w_i \quad (i \in [1..K])$ for the length of each type of special effect contents. They can be seen in the user interface of Figure 1 as “Importance”. Then for every segment $j \in [1..N]$ belonging to special effect type $i \in [1..K]$, its length is $L_j^i = \frac{w_i a * Len}{\sum_{i=1}^{K} w_i * N_i} \quad (j \in [1..N])$. Generally speaking, L_j^i should not be shorter than a minimum value, so that every extracted segment can give the user a clear meaning. The user can specify this value through the user interface as “Min Segment Length” in Figure 1.

5) Different sounds imply the different time for the events to take place. As for laughter, applause and scream, the attractive events indicated by them usually take place before them. While for speech and song, appealing events often happens at the middle of them.

6) The remaining contents are extracted from those segments containing audio contents of silence, non-pure speech without song, music, and others. In this paper, the key-frame selection algorithm described in literature [9] is used. Then the contents near the key-frames are extracted.

5. EXPERIMENTAL RESULTS

There is no absolute measure of the quality of a home video abstract, because different people have different aims. And even people with the same aim will give different evaluation. So the practical method is asking test persons for their evaluation. We test our algorithm in the following way.

We generated 5 abstracts of 1 to 5 minutes from five home videos ranging from 2 to 3 hours in the default way. First, 10 test persons watched each abstract. Then they browsed the original video. In the end, they were asked to give his evaluation of the abstract on a scale of 0 to 5, corresponding to total disagreement and total agreement respectively. The average evaluation is 4.1, 4.2, 4.3, 4.7, 4.7. Then we let the test persons generate the home video at their will and give their evaluation. In this way, the evaluation is 4.3, 4.5, 4.7, 4.8, 4.9.

From the experimental results, we can see that the abstract algorithm works well in home videos. And with the user’s interaction, we can achieve even better results.

6. CONCLUSION

This paper presents a home video abstraction technique combining audio and video features. A two-level hierarchy audio classification method using SVMs and HMMs is used to segment and classify the audio contents. For video contents, motion level and blur degree are acquired. The main part of the abstract is extracted from audio and video special effects. Experimental results show that this method can generate satisfactory home video abstracts. Future work involves improving ways to combine audio and visual contents, studying the use of face detection and recognition in the current system.

7. REFERENCES

[1] S. Pfeiffer, R. Lienhart, S. Fischer and W. Effelsberg, “ Abstracting Digital Movies Automatically”, Journal of Visual Communication and Image Representation, vol. 7, no. 4, pp.345-353, Dec. 1996.

[2] M. A. Smith and T. Kanade, “ Video Skimming and Characterization through the Combination of Image and Language Understanding Techniques”, CVPR’97, pp. 775-781, 1997.

[3] A. Hanjalic and H. J. Zhang, “ An Integrated Scheme for Automated Video Abstraction based on Unsupervised Cluster-validity Analysis”, IEEE Transactions on Circuits and Systems for Video Technology, vol. 9, no. 8, pp.1280-1289, Dec. 1999.

[4] R. Lienhart, “ Abstracting Home Video Automatically”, Proc. ACM Multimedia 99, pp.37-40, Orlando, FL, October 1999

[5]L. Lu, S.Z. Li, and H.J. Zhang, "Content-Based Audio Segmentation Using Support Vector Machines", ICME 2001, Waseda University, Tokyo, Japan, August 22-25, 2001.

[6] T. Zhang and C.-C.J. Kuo, “ Hierarchical Classification for Audio Data for Archiving and Retrieving,” Proc. ICASSP, Phoenix, vol. 6, pp. 3001-3004, Mar, 1999

[7] G.D. Guo, H.J. Zhang, and S.Z. Li. "Boosting for Content-Based Audio Classification and Retrieval: An Evaluation". ICME. 2001

[8] H.J. Zhang, A. Kankanhalli and S.W. Smoliar. Automatic partitioning of full-motion video. Multimatia Systems, 1(1):10--28, 1993.

[9] H.J. Zhang, J.H. Wu, D. Zhong, S.W. Smoliar, ”an Integrated System for Content-based Video Retrieval and Browsing”, Pattern Recognition,Vol.30,No.4,pp.643-658,1997

[10] Xavier Marichal, Wei Ying Ma and H.J. Zhang. Blur Determination in the Compressed Domain Using DCT Information. ICIP'99, Kobe, September 1999, Proc. Vol. II, pp. 386-389.