Acta Technica *62* (2017), No. 6B, 59–66

Key technologies of digital radio and TV video production in the new media environment

Teng Liu[1]

Abstract. One news video recognition method based on audio and visual template matching has been proposed in this paper. In the process of template establishment, audio template will be extracted from theme music of news video titles, visual template has been extracted from extended face region in anchorperson shot, these two compose audio and visual template together; audio template matching has been made firstly for TV video stream in the recognition process, and then locate to corresponding video shot through candidate time passed from matching. In the following, matching for extended face region in shot has been made through visual template, and then confirmed anchorperson shot and accomplished news video recognition finally. Experimental results have shown that this method is with high calculation efficiency, easy to operate and with good practical value.

Key words. News video, TV play, Video production, Video retrieval, New media.

1. Introduction

TV video stream consists of rich program categories, including news, TV series, music, advertisement and cartoon etc. With the rapid increase of TV channels, video data amount increases rapidly. For users who are only interested in news program, the realization of automatic recognition of news video for TV video stream is of important significance [1]. It can not only narrow the selection range and improve the efficiency of video collection, but also be taken as input of video retrieval to establish digital system from source of video collection to analysis, query and storage of video content.

News video recognition based on content is actually a process combining video segmentation and classification together. In other words, it needs to locate news program in continuous video streams. Researchers have done a lot of work in video segmentation and video classification. In the field of video semantic analysis, Rasheed

[1]Centre of News, Chongqing University, Chongqing , 400044, P. R. China

and others [2] have divided films into several types of tragedy, action, drama and horror with only four visual characteristics (average shot length, color difference, sports content and lighting) by combining the characteristics of films. Liu and others [3] have extracted a series of time-frequency characteristics based on statistics from audio and classified report, sports and advertising programs with neural network. Liang Liuhong and others [4] have extracted the titles and trailers of programs automatically through searching repeated video clips in multi-days video, which has set up visual program template for one period of time for specific TV channel and then accomplish the segmentation of program. Wang Jinqiao and others [5] have proposed one video program segmentation method with multimodal feature fusion, which discovers space-time characteristics at boundary of TV program through visual, audio, audio and text information, and then set up model with support vector machine and realize program segmentation through two-element classification for candidate boundary points produced by shot inspection. However, these methods are designed for extensive program types and do not consider about the differences of news video with other program types in structure; moreover, traditional program classification methods set up classifier model to accomplish recognition based on bottom feature extraction, which will cause shortages of high calculation complexity and slow speed inevitably. Therefore, connecting video segmentation and video segmentation simply can't meet practical application demand.

Wang Jinqiao and others [5] have further pointed out that video and audio characteristics at the boundary of programs can be used to describe specific programs for quick browse and locate program. Inspired by this, structural differences between news video and other videos have been found out through analyzing and extracting structural characteristics of news video in program boundary and then make recognition for news video. Therefore, on the basis of deep analysis of structural characteristics of news video, one news video recognition method based on audio and visual template matching has been proposed in this paper. This method considers about theme music and anchorperson shot characteristics in news video comprehensively, which compose audio template and visual template respectively. It realizes news video recognition effectively by combining hierarchical template matching method.

2. News video recognition method based on audio and visual template matching

It can be found from a lot of analysis and statistics for various kinds of new programs that the extraction of characteristics of theme music and anchorperson shot is an effective way of full depiction of news video and non-news video. Considering about that the broadcasting time for news programs in specific channels are fixed, if the start of the program is detected, it can start to record based on prior knowledge of broadcasting time until the ending of the program. Therefore, starting from practicability, this paper has utilized structural characteristics appeared after theme music and anchorperson shot and proposed a news video recognition method based on audio and visual template matching.

2.1. Algorithm process

First, extract off-line theme music template and anchorperson shot template, in which theme music template is composed of segment feature composition based on MFCC parameter; anchorperson shot is composed of block HSV color histogram in extended face region; second, make video and audio separation for TV video stream and make matching for theme music template first; third, locate from the candidate time passed matching to corresponding segmented video shot, make use of anchorperson shot template to match extended face region detected in shot, determine if anchorperson shot appears based on matching results; if it appears, it means the starting of news video and then start recording until the end and accomplish news video recognition. Schematic diagram of method is as shown in picture 1.

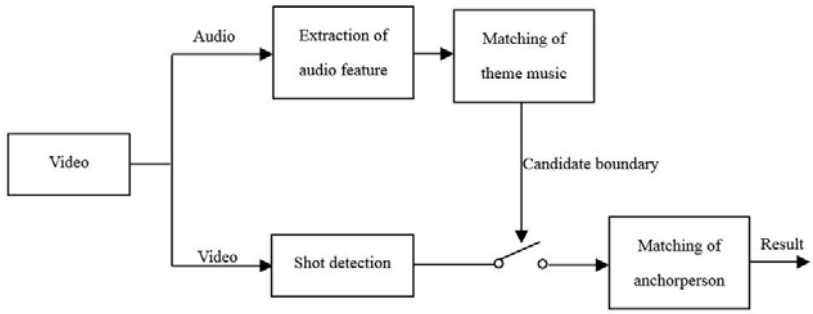

Fig. 1. Schematic diagram of news video recognition method

2.2. Establishment and matching of theme music template

Audio template is theme music template. In the matching of audio template, the selected characteristics should depict important classification characteristics of audio, which is with robustness for environmental change.

Characteristics based on spectrum have strong ability of distinguishing audio and music signals. Therefore, I select Mel frequency cepstral coefficient as the characteristic of audio frame [6]. First, it takes frame as unit to calculate MFCC parameters of 12 modals, frame length is 20ms, frame shift is 10ms, and then calculate the mean value of MFCC parameter within 1 second audio segment and take it as segment characteristic. The second dimensional probability distribution in 12 modals dimensional segment characteristic parameters of 30 minutes of voice, music and background music is as shown in figure 2.

Mean value of second dimensional MFCC parameter is with good distinctiveness for speech, music and environmental background sound. This paper also makes statistics for the probability distribution of other dimensional MFCC parameter mean values under three audios. The results have shown that MFCC parameter mean can well distinguish these three kinds of audios.

The play time of theme music is constant. Assume the time is N seconds, and then theme music template is composed of N slice features in time order, which is

expressed as $T = (t_1, t_2, ..., t_N)$, in which t_i, $1 \leq i \leq N$ are No. i segmental feature.

The play speed of theme music is constant, so there is no need for complicated dynamic matching method on template matching. Under this situation, it is proper to calculate linear correlation, which has been verified in following experiment. Set $T = (t_1, t_2, ..., t_N)$ as target theme music template, $O = (o_1, o_2, ..., o_M)$ is segment feature sequence extracted from video stream.

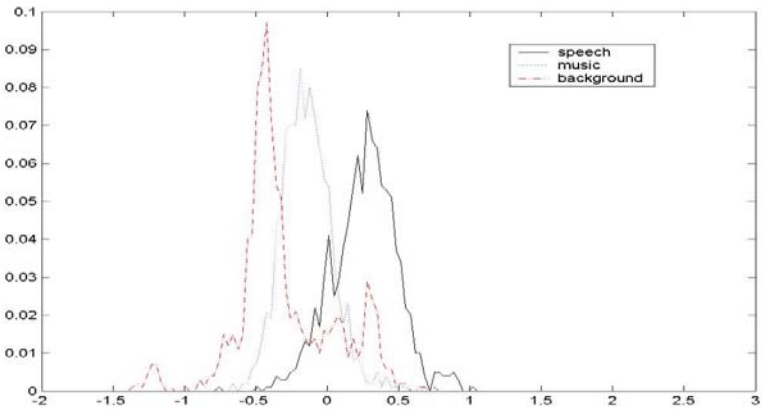

Fig. 2. Distribution map for the second dimensional MFCC parameter mean probability

In which, t_i and o_i are corresponding features extracted from No. i second. M and N are numbers of segments of these two sequences. It needs to point out that the length of video stream is infinite in theory; here use M to express the length for the demand of description and it can be understood that M is infinitely great. Make theme music template slide on segment feature sequence of video stream and sliding length is one second. Similarity of No. n segment between segment feature sequence and template can be defined as:

$$S(n) = \frac{\sum_{i=1}^{N} (t_i - \bar{t}) \cdot (o_{i+n} - \bar{o}_n)}{\sqrt{\sum_{i=1}^{N} \|t_i - \bar{t}\|^2} \sqrt{\sum_{i=1}^{N} \|o_{i+n} - \bar{o}_n\|^2}}, \quad n = 0, ..., M - N\,. \tag{1}$$

In which, $\bar{t}$ is the mean of template feature vector, $\bar{o}_n$ is the mean of segment feature sequence $o_n, ..., o_{n+N}$, $\|*\|$ is the norm. If local maximum point of $S(n)$ is bigger than the set threshold value α, and then determine it as the starting point of theme music. Once the theme music is located, anchorperson shot can be confirmed through fixed length positioning.

2.3. Establishment and matching of anchorperson shot template

Visual template is anchorperson shot template. We extract extended face region (EFR in short) of anchorperson as general template. Extended face region is upper body area attained from downward extension of face region in proportion. The EFR of anchorperson shot includes anchorperson face information, color and style information of clothes of anchorperson, which is irrelevant to studio background, anchorperson position and captions. It can also be distinguished from EFR similar EFR of anchorperson shot in live report, which provides powerful evidence and guarantee for detecting anchorperson shot in news video.

Extraction and matching of EFR is established based on shot detection, which process key frames of shot. Detailed process of this algorithm is as shown in literature [7].

3. Experimental results and performance analysis

To verify the effectiveness of this method, this paper has collected TV programs of CCTV1, CCTV2, British BBC and Taiwan's ETTV. Considering about that the news recognition method in this chapter only detects the starting of news video, so the news programs are not complete and they only include the parts which can detect the starting parts. At the same time, it combines TV programs belonging to the same TV station together manually, which compose test video stream of this TV station. These data is from different times and there are 40 hours in total, covering various programs of news, TV series, music and cartoon etc, in which there are 56 starting points of news programs. Before the testing, the starting points of all TV programs have been marked manually, which will be taken as standard reference of method detection results.

Based on the method process in chapter 2, audio and visual templates have been set up respectively for news programs of different TV stations and then make hierarchical template matching. Experiment has shown that threshold value α is 0.9 and β is 0.1.

To make comparison, this chapter has adopted template matching method based on theme music (audio method in short)[8], template matching method based on anchorperson shot (visual method in short) [4] as well as method proposed in this paper to process above experimental data and then make news video recognition. Commonly used two indexes of precision ratio (accuracy ratio) and recall ratio (recall ratio) have been adopted to evaluate the performance of news video recognition algorithm. Detection results of three methods for video stream of each TV station are as shown in table 3. At the same time, average recall ratio and accuracy ratio have been listed in figure 3.

It can be seen from figure 3 that the average recall ratio of this method reaches 96.85% and the average recall ratios of the other two methods are higher than the method mentioned in this paper, which are 98.63% and 98.23%. This is because news video is with two characteristics of theme music and anchorperson shot. Moreover,

Table 1. Comparison of detection results of different detection algorithms

Video program	Detection algorithm	Number of news video	Detection number	Error number	Missing number	Recall ratio	Precision ratio
CCTV1	Audio method	12	15	3	0	100%	80%
	Visual method		13	1	0	100%	92.3%
	Method in this paper		12	0	0	100%	100%
CCTV2	Audio method	12	16	4	0	100%	75%
	Visual method		14	2	0	100%	85.7%
	Method in this paper		12	0	0	100%	100%
BBC	Audio method	18	23	6	1	94.5%	73.9%
	Visual method		22	4	0	100%	81.8%
	Method in this paper		19	2	1	94.5%	89.5%
Ettoday	Audio method	14	18	4	0	100%	77.8%
	Visual method		15	2	1	92.9%	86.7%
	Method in this paper		14	1	1	92.9%	92.9%

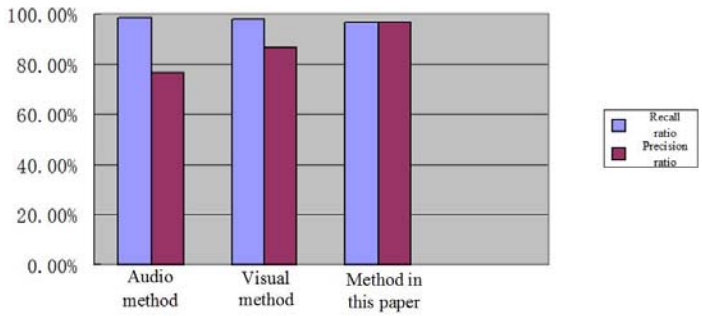

Fig. 3. Average recall ratios and precision ratios of various methods

all anchorperson shots appeared after theme music are all sitting in the studio, there are not too many interferences of background and other factors, which makes the

recall ratio very high. In addition, the method mentioned in this paper is the combination of the other two methods, which combines missing situations of the other two methods inevitably. But the average precision ratio of the method mentioned in this paper reaches 95.6%, which has been increased by 19.5% and 9.1% when compared with audio method and visual method and the effect is very distinctive. Table 1 lists comparisons for recognition performance of video programs in different TV stations. Audio method makes matching with adoption of theme music template, so it is with fast recognition speed but there are more error detections. However, visual method is superior to audio method in precision ratio, but error detection happens easily for face shot appeared in non-news video. At the same time, because it needs to make face detection for each shot and matching for following anchorperson shot, its recognition speed is slow. This method combines audio method and visual method together to recognize, which makes full use of advantages of the two, highlights the differences between news video and non-news video, attains good effects for different types of TV programs and is with good practical value.

4. Conclusions

News video recognition for TV video stream solves problem of data source of news video screening, which is of important significance for its practicality. Traditional research methods focus on structuring of video stream and there are few researches on program recognition of special type of news video. On the basis of full study of structural differences between news video and other types of programs, one news video recognition method based on audio and visual template matching has been proposed in this paper. Experimental results have shown that this method is with high calculation efficiency and easy to operate, which has attained 96.8% recall ratio and 95.6% precision ratio and is with good practical value.

References

[1] C. Wang, Q. Duan, W. Gong, et al.: *An evaluation of adaptive surrogate modeling based optimization with two benchmark problems*[J]. Environmental Modelling & Software *60* (2014), No. 76, 167–179.

[2] J. Revaud, M. Douze, C. Schmid, et al.: *Event Retrieval in Large Video Collections with Circulant Temporal Encoding*[C]// Computer Vision and Pattern Recognition. IEEE (2013), 2459–2466.

[3] J. Revaud, M. Douze, C. Schmid, et al.: *Event Retrieval in Large Video Collections with Circulant Temporal Encoding*[J]. *9* (2013), No. 4, 2459–2466.

[4] S. Jones, L. Shao: *Content-based retrieval of human actions from realistic video databases*[J]. Information Sciences *236* (2013), No. 1, 56–65.

[5] F. Metze, D. Ding, E. Younessian, et al.: *Beyond audio and video retrieval: topic-oriented multimedia summarization*[J]. International Journal of Multimedia Information Retrieval *2* (2013), No. 2, 131–144.

[6] M. Halvey, D. Vallet, D. Hannah, et al.: *ViGOR: a grouping oriented interface for search and retrieval in video libraries.*[J]. College of Science and Engineering > School of Computing Science (2017), 87–96.

[7] K. Liao, G. Liu, L. Xiao, et al.: *A sample-based hierarchical adaptive K -means clustering method for large-scale video retrieval*[J]. Knowledge-Based Systems (2013), No. 49, 123–133.

[8] K. Schoeffmann: *A User-Centric Media Retrieval Competition: The Video Browser Showdown 2012-2014*[J]. IEEE Multimedia *21* (2014), No. 4, 8.

[9] Y. Li, R. Wang, Z. Huang, et al.: *Face video retrieval with image query via hashing across Euclidean space and Riemannian manifold*[C]// Computer Vision and Pattern Recognition. IEEE (2015), 4758–4767.

[10] J. Tang, W. Bian, N. Yu, et al.: *Editorial: Intelligent processing techniques for semantic-based image and video retrieval*[J]. Neurocomputing *119* (2013), No. 06, 1–2.

[11] S. Schmiedeke, P. Xu, Ferran, et al.: *Blip10000: a social video dataset containing SPUG content for tagging and retrieval*[C]// Multimedia Systems (2013), 96–101.

[12] Y. Cai, L. Yang: *Large-Scale Near-Duplicate Web Video Retrieval: Challenges and Approaches*[J]. Multimedia IEEE *20* (2013), No. 2, 42–51.

[13] M. Bendersky, L. Garcia-Pueyo, J. Harmsen, et al.: *Up next: retrieval methods for large scale related video suggestion*[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM (2014), 1769–1778.

Received May 7, 2017