Multi-modal Interview Concept Detection for Rushes Exploitation

Anan Liu[1, 2]**, Sheng Tang** [2]**, Yongdong Zhang** [2]**, Jintao Li** [2]**, Zhaoxuan Yang** [1, 2]

1. School of Electronic Engineering,
Tianjin University
Tianjin, 300072, China

2. Institute of Computing Technology,
Chinese Academy of Sciences
Beijing, 100080, China
liuanan@ict.ac.cn

Abstract

According to the concepts of Large-Scale Concept Ontology for Multimedia (LSCOM) and requirement of the 4th task in the 2006 TRECVID, i.e., rushes exploitation, the “interview” concept is an important semantic concept for rushes content analysis. The paper presents the shot-level “interview” concept detection method. Face detection and audio classification are implemented to detect “face” and “speech” concepts for each shot. By integrating audiovisual information, “interview” concept is finally detected. The utilization of the method will definitely benefit the video edit. Large-scale experimental results strongly demonstrate the accuracy and effectiveness of the proposed method.

1. Introduction

The TREC conference series is sponsored by the National Institute of Standards and Technology (NIST) with additional support from other U.S. government agencies. The goal of the conference series is to encourage research in information retrieval（Guidelines, 2006）. In the 2006 TRECVID, there are three system tasks and one exploratory task: shot boundary determination, high-level feature extraction, search and rushes exploitation. In broadcasting and filmmaking industries, “rushes” is a term for raw footage, which is used for productions such as TV programs and movies. Usually up to 40 hours of raw footage is converted into one hour of TV program (P. Allen, 2005). In rushes, there are a lot of static scenes, redundant episodes and out of focus fragments. Rushes’ soundtracks can be noisy and indecipherable for automatic speech recognition. Moreover, there is no caption and textual information is rather sparsely available for rushes content analysis (P. Allen, 2005). Due to the characteristics of rushes, the content analysis on it is different form current work on edited video, for example, movies, news video and sports video. Consequently, it is a challenging and promising work to develop novel data mining techniques for rushes.

In the 2006 TRECVID, about 50 hours of rushes is provided by the BBC Archive for rushes exploitation. The main content of them are interview scene, person activity scene, natural scene and some redundant shots. In our report for Rushes exploitation in TRECVID 06 (Tang, 2006), we have generally presented our work for the items mentioned above. Obviously, the interview scene is the most useful part for news program production. Compared to the previous work only focusing on the specific person identification and domain knowledge based video content indexing on the edited news video (Kuo, 2005 ; Albiol, 2003), the interview concept detection aims to extract integrated semantic episodes for video edit on the raw material. Therefore, the paper detailedly presents a shot-level “interview” concept detection method.

The rest of the paper is organized as follows: Section 2 specifically illustrated the shot-level “interview” detection method. Section 3 provides the experiment results and detailed analysis. In section 4, concluding remarks and future advanced work are presented.

2. Shot-level interview concept detection

According to the 330th concept of LSCOM in (LSCOM ; Naphade, 2006), interview shots mainly mean those shot on the special location out of the studio. Generally speaking, interview shots can be classified into two kinds, monologue as shown in Fig.1 (a) and dialogue as shown in Fig.1 (b).

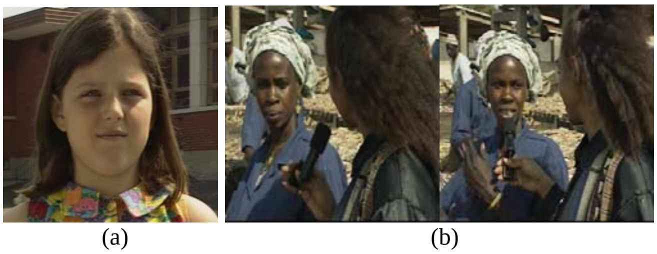

Fig. 1. Two kinds of interview scenes: (a) Monologue and (b) Dialogue

“Interview” can be seen as a high level semantic concept containing both “face” and “speech” information. In the fusion method, “face” and “speech” concepts are integrated to detect “interview”. Fig. 2 shows the framework that consists of four major modules: (1) shot boundary detection and key frame extraction, (2) “speech” concept detection, (3) “face” concept detection, (4) “interview” concept detection based on audiovisual cues. As shown in Fig.2, an input video is firstly divided into audio and visual streams. Then shot boundary detection and key frame extractions are performed on the visual stream. With the shot boundary information, the visual stream and audio stream are divided into corresponding shot-level clips that separately consist of the visual sequence and the audio sequence. For each visual clip, face detection is implemented on the key frames to judge the shot-level “face” concept. At the same time, each audio clip is classified into four kinds: silence, speech, music and background and the shot-level “speech” concept is determined. Finally, with the fusion of both concepts, shot-level “interview” concept is detected.

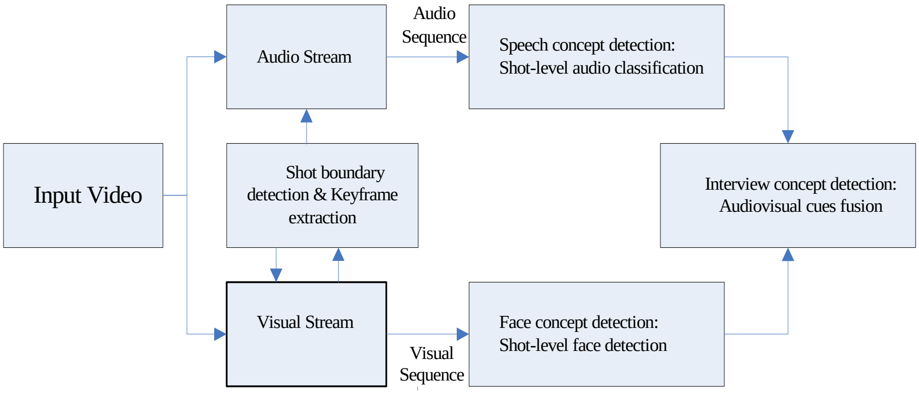
Audio Stream
Audio Sequence
Speech concept detection: Shot-level audio classification
Input Video
Shot boundary detection & Keyframe extraction
Interview concept detection: Audiovisual cues fusion
Visual Stream
Visual Sequence
Face concept detection: Shot-level face detection

Fig. 2. Framework of shot-level “interview” concept detection method

2.1 Shot boundary detection and keyframe extraction

Structuralization for video content analysis includes shot boundary detection and keyframe extraction. The method for video structuring is presented in our report for Rushes exploitation in the TRECVID 2006 (Tang, 2006).

2.2 “Speech” concept detection: Shot-level audio classification

2.2.1 Audio feature extraction

The audio stream is firstly segmented into nonoverlapping 20-ms short time frame (ST frame). Then five frame-level audio features, namely, ***Short-Time Energy Function***, ***Short-Time Zero-Crossing Rate***, ***Mel-frequency cepstral coefficients***, ***Frequency energy and Sub-band energy ratio***, are extracted to represent the character of each ST frame. The specific definitions and methods of calculating these features are elaborated in (Bai, 2005).

2.2.2 ST frame –level audio classification based on SVM

A Support Vector Machine (SVM) is a supervised binary classifier that constructs a linear decision boundary or a hyperplane to optimally separate two classes. Since its inception, the SVM has gained wide attention due to its excellent performance on many real-world problems. It is also reported that SVMs can achieve a generalization performance that is greater than or equal to other classifiers, while requiring significantly less training data to achieve such an outcome (Wan, 2000). Because rushes are unedited raw material, rushes' soundtracks are usually noisy and indecipherable. Therefore, rule-based methods for audio classification are not applicable and the SVM-based audio classification is applied. Depending on the need of rushes content analysis, three binary SVM classifiers are trained to classify each ST frame into either one of the four kinds: silence, speech, music and background.

2.2.3 Shot-level audio classification

After each ST frame is classified, the speech frame is labeled with "1" and the non-speech frame is labeled with "0". Then three steps mentioned below are implemented to classify shot-level audio clip.

Step 1: 50 continuous ST frames in temporal domain are set into one group. The group is labeled with decision rule (1), where Th1 denotes the threshold determined in the experiments.

$$\begin{cases} Ratio_1 = \dfrac{Speech\_frame\_number}{Group\_frame\_number} \\ Ratio_1 > Th_1, label = 1; Ratio_1 <= Th_1, label = 0 \end{cases} \tag{1}$$

Step 2: A finite state machine (FSM) is defined as:

$$A = (Q, \sum, \sigma, q_0, F) \tag{2}$$

$$\Sigma : \left\{ \begin{array}{l} C_1 : Label = 1; C_2 : Label = 0; \\ C_3 : P(speech \mid S_3) < Th_2 \,\&\, (Counter++) < Th_3; \\ C_4 : P(speech \mid S_3) >= Th_2 \,\&\, (Counter++) < Th_3; \\ C_5 : P(speech \mid S_3) < Th_2 \,\&\, (Counter++) >= Th_3; \\ C_6 : Label = 0; C_7 : Label = 1; \\ C_8 : P(nonspeech \mid S_4) < Th_2 \,\&\, (Counter++) < Th_3; \\ C_9 : P(nonspeech \mid S_4) >= Th_2 \,\&\, (Counter++) < Th_3; \\ C_{10} : P(speech \mid S_4) < Th_2 \,\&\, (Counter++) >= Th_3 \end{array} \right\} \tag{3}$$

$$Q : \left\{ \begin{array}{l} S_1 : Speech, S_3 : Transition\_SpeechToNonspeech, \\ S_2 : Non-Speech, S_4 : Transition\_NonspeechToSpeech \end{array} \right\} \tag{4}$$

$$q_0 \, and \; F : \{ I_1 : Speech, I_2 : Non-Speech \} \tag{5}$$

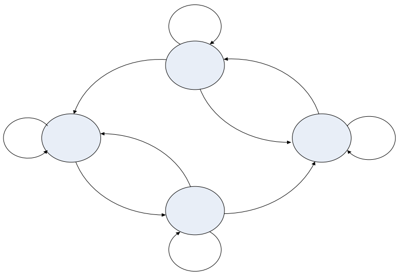

Fig. 3. Finite state machine for smoothing

where Q is a set of states in the FSM, σ is the set of transitions, $\sum$ is the set of the conditions for the transitions, q_0 is the initial state, and F is the set of accepting (final) states (Zhai, 2005). Because the continuity of audio stream, it is impossible that there appear abruptly and frequently changed audio groups in the continuous audio stream. FSM is used to smooth the labeling results in step 1.It can be described in Eq. (3) (4) (5) where Label means the classification of the audio group, P(ratio|S3) and P(ratio|S4) denotes the probability of the "speech" groups over the total groups in a shot on the conditions of State 3 and State 4, Th2 and Th3 are the thresholds determined in the experiments and counter is the frame counter. The finite state transition diagram is shown in Fig.3.

Step 3: After smoothing the labeling results, the number of groups with label "1" (Positive Num) and the total group number (Total Num) are got in one shot and the shot-level "speech" concept is detected with decision rule (6) where Th_4 is an experienced threshold.

$$\begin{cases} Ratio_2 = \dfrac{Positive\ Num}{Total\ Num} \\ Ratio_2 > Th_4, Speech;\ Ratio_2 <= Th_4, Nonspeech \end{cases} \quad (6)$$

2.3 "Face" concept detection: Shot-level face detection

2.3.1 Face detection based on improved AAM

Active Appearance Models (AAM) is very powerful to extract facial features for success of applications such as face recognition, expression analysis and face animation. It is composed of two parts, AAM subspace model and AAM search. The superiority of AAM mentioned in (Zhao, 2004) is that an approach for optimizing the parameterization of the AAM subspace model according to the search procedure is proposed while in the conventional methodology, the two sections are treated separately. In (Zhao, 2004), the subspace error is decomposed into one of subspace reconstruction and one of AAM search. Then a method to optimize the AAM subspace model based on the subspace error decomposition is developed. The novelty is that the eigenvectors for the subspaces is selected considering both the subspace modal and the search procedure. The experiment results demonstrate that more accurate and faster results can be obtained.

2.3.2 Shot-level face detection

With the improved AAM method presented above, shot-level face detection is executed with the following three steps.

Step 1: The AAM method is only implemented on the keyframes in each shot to save computational time. The frame containing face concept is defined as positive frame. Otherwise, it is a negative frame.

Step 2: The ratio of the positive frame number over the total keyframe number in one shot are calculated in Eq. (7).

$$Ratio_3 = \frac{Positive\ frames\ Number}{Total\ Keyframes\ Number} \quad (7)$$

Because of the influences of light, background, the gesture of the interviewee and so on in the raw footage, the low recall directly affects the accuracy of Ratio$_3$. There are two typical cases.

(1) The interview shot shown in Fig.1 (a) usually lasts for a long time and includes few keyframes because the interviewee is almost still and the visual content changes lightly. In this circumstance, most of the keyframes may contain face information. However, if missing detection is severe, it will affect the detection result.

To improve the recall in case 1, Convergence Degree (CD) is defined in Eq. (8). CD denotes the average importance of one frame. If the total frame number is larger and the keyframe number is smaller, CD is larger, which means each keyframe has stronger ability to represent other similar frames. Although the missing detection makes Ratio$_3$ smaller, CD can give Ratio$_3$ a weight in Eq. (9) to avoid the influence.

$$CD = \frac{Total\_frame\_Number}{Keyframe\_Number} \quad (8)$$

$$Ratio_4 = Ratio_3 * CD \quad (9)$$

(2) The interview shot shown in Fig.1 (b) usually lasts for a long time and includes many keyframes. However the keyframes containing face information became fewer because it is difficult to detect face when the interview progresses with the movement of the person, the gestures of persons are multiple and even sometimes interviewees are only back to the camera.

In this circumstance, missing detection cannot be solved using visual information. To improve the recall in the shots lasting for a long time which is likely to be an interview shot, Time Factor (TF) is defined in Eq. (10) and the rectified Ratio$_5$ is defined in Eq. (11). Although it may degrade the precision, the problem can be solved by integrating audio cue.

$$TF = \frac{Shot\_time}{Video\_time} \quad (10)$$

$$Ratio_5 = Ratio_3 * TF \quad (11)$$

Step 3: With Ratio$_3$, Ratio$_4$ and Ratio$_5$, three decision values (DV) can be got for shot-level face detection. The decision rule can be expressed as:

$$\begin{cases} \textit{If } Ratio_3 > threshold_1, DV_1=1; \textit{ otherwise, } DV_1=0; \\ \textit{If } Ratio_4 > threshold_2, DV_2=1; \textit{ otherwise, } DV_2=0; \\ \textit{If } Ratio_5 > threshold_3, DV_3=1; \textit{ otherwise, } DV_3=0; \\ \textit{If } DV_1 \,||\, DV_2 \,||\, DV_3 =1, \textit{ the shot contains "face" concept;} \\ \quad \textit{otherwise, the shot does not contain "face" concept.} \end{cases} \qquad (12)$$

2.4 Shot-level "Interview" concept detection: audiovisual cues fusion

It is perceivable that the interview shot contains "face" concept and "speech" concept. Depending on the characteristics of Rushes, for the multimodal fusion scheme for interview detection, we apply the AND operation on audiovisual cues. If the shot contains both concepts, the shot is considered as an interview shot. Otherwise, it is not an interview shot.

3. Experimental results

3.1 Results of audio classification

3.1.1 Data preparation

There are 48 videos in both rushes development data and test data in the 2006 TRECVID. Each video lasts for about 30 minutes. In the experiment, the audio features are extracted firstly and then three groups of train data and three groups of test data are prepared by randomly sampling. The training data and test data are shown in Table1.

Training data	Positive sample (minute)	Negative sample (minute)
Silence/Non silence	10	10
Speech/Non speech	20	7
Music/Non music	3.5	28
Test data	**Sample (minute)**	
Silence/Non silence	40	
Speech/Non speech	270	
Music/Non music	80	

Table 1. Training data and test data

Test results	Accuracy
Silence	95.5%
Speech	88.3%
Music	94.5%

Table 2. Classification results

3.1.2 Classification Results

The classification results are shown in Table 2.The accuracy is defined as the ratio of correctly classified samples over all test data of the class.

3.1.3 Results of shot-level "speech" detection

The audio streams of 48 videos in rushes test data are all used to detect the shot-level "speech" concept. The results are shown in Table 3.

Speech detection	Labeled shots with"speech"	Detected shots with "speech"	Missing detectionshots	Error detetionshots
Total shots	864	919	114	169
Criteria	Precision		Recall	
Result	81.6 %		86.8 %	

Table 3. Results of shot-level speech detection

3.2 Results of shot-level "face" detection

The visual streams of 48 videos in rushes test data are all used to detect the shot-level "face" concept. The results are shown in Table 4.

Face detection	Labeled shots with"face"	Detected shots with "face"	Missing detectionshots	Mistaking detetionshots
Total shots	394	368	60	34
Criteria	Precision		Recall	
Result	90.8%		84.8%	

Table 4. Results of shot-level face detection

3.3 Results of shot-level "interview" detection

48 videos in rushes test data are all used to detect the shot-level "interview" concept. The results by using audio cue, visual cue and fusion method are compared in Table 5.

"Interview"detection	Precision	Recall
Audio cue	30.8%	98.3%
Visual cue	62.3%	78.9%
Fusion method	84.2 %	77.2%

Table 5. Comparison of shot-level interview detection by using audio cue, visual cue and fusion method

Comparison in Table 5 clearly shows the superiority of integrating audiovisual cues. It is very probable that in an interview shot both speech and face concepts cannot be detected simultaneously. Therefore the recall of fusion method is worse than those of other two methods. However, both concepts, face and speech, explain the meaning of interview well and can be detected by using low level audiovisual features. As a result, for precision the fusion method strongly outperforms other two ways. The later statistical analysis shows that only five videos, in which the background and complexion severely affect the precision of face detection, have great influence on the recall by using fusion method. Excluding the five videos, precision is 81.9% and recall is up to 87.7%.

4. Conclusion and future work

The paper presents a shot-level "interview" concept detection method based on the fusion of "speech" and "face" concepts. The experimental results show the superiority and efficiency of the proposed method. With the satisfactory results, the method can be used to analyzing the rushes content and has promising application in producing TV program, movie, broadcast news and so on.

Moreover, "interview" can be further classified into monologue/dialogue, indoor-interview/outdoor-interview. As the future work, speaker recognition technique and other LSCOM visual concepts can be used for fine-level interview classification. On the basis of the related work, an ideal rushes content mining system can be founded for raw footage analysis, video production and so on.

Acknowledgements

This work was supported by Beijing Science and Technology Planning Program of China (D0106008040291).

Reference

Guidelines for the TRECVID 2006 Evaluation: http://www-nlpir.nist.gov/projects/tv2006

Bradley P. Allen, Valery A. Petrushin, Searching for Relevant Video Shots in BBC Rushes Using Semantic Web Techniques, In Proc. TRECVID Workshop, 2005. http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html

Jin-Hau Kuo, Jen-Bin Kuo, A hierarchical and multi-modal based algorithm for lead detection and news program narrative parsing, Proc. of 19th International Conference on Advanced Information Networking and Applications, 2005, vol.2, pp: 511-514

Albiol, A., Torres, L., Delp, E.J., The indexing of persons in news sequences using audio-visual data , Proc. of ICASSP '03, vol.3, pp: III- 137-40

LSCOM Lexicon Definitions and Annotations, http://www.ee.columbia.edu/ln/dvmm/lscom/.

Naphade, M., Smith, J.R., et al, Large-scale concept ontology for multimedia, IEEE MultiMedia, vol. 13, no. 3, pp. 86-91, 2006.

Sheng Tang, Yong-Dong Zhang, Jin-Tao Li et al. TRECVID 2006 Rushes Exploitation By CAS MCG. In Proc. TRECVID Workshop, 2006. http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html

Bai Liang, Hu Yaali, Lao Songyang, et al. Feature analysis and extraction for audio automatic classification. Proc. of IEEE International Conference on System, Man and Cybernetics, 2005.

Wan, M. Campbell, "Support vector machines for speaker verification and identification," Proc. of the IEEE Signal Processing Society Workshop on Neural Networks, 2000.

Zhai Y., Rasheed Z., "Semantic classification of movie scenes using finite state machines", IEE Proc of Vision, Image and Signal Processing, vol. 152, pp. 896-901, 2005.

Zhao Ming,Chen Chun,Li S Z,et al. Subspace analysis and optimization for AAM based face alignment [A].In Proc. of Sixth IEEE International Conference on Automatic Face and Gesture Recognition [C].Seoul,South Korea,2004.290-295.