A MID-LEVEL SCENE CHANGE REPRESENTATION VIA AUDIOVISUAL ALIGNMENT

Jinqiao Wang[1,2], *Lingyu Duan*[2,3], *Hanqing Lu*[1], *Jesse S. Jin*[3], *Changsheng Xu*[2]

[1] National Lab of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China
{jqwang, luhq}@nlpr.ia.ac.cn
[2] Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
{lingyu, xucs}@i2r.a-star.edu.sg
[3] The School of Design Communication and Information Technology,
University of Newcastle, NSW 2308, Australia
{Jesse.Jin}@newcastle.edu.au

ABSTRACT

Scene is a series of semantic correlated video shots. An effective scene detection depends on domain knowledge more or less. Most existing approaches try to directly detect various scene changes by applying clustering or supervised learning methods to low level audiovisual features. However, robustly detecting diverse scene changes derived from complex semantic meanings is still a challenging problem. In this paper we are focused on the association of visual signal changes (e.g. cuts, fade-in, fade-out, etc.) and audio signal changes (e.g. speaker change, background music change, etc.) to propose a mid-level scene change representation, which is meant to locate candidate scene change points by characterizing temporally uncorrelated properties of audio and visual track in the case of scene change happening. By incorporating domain knowledge, enhanced features can be further extracted to complement this representation to bridge semantic gap towards scene change detection. We utilize a camera motion estimation algorithm to detect visual signal changes. Such visual change positions are selected as time-stamp points. An alignment is performed to search for candidate audio signal change positions by multi-scale Kullback-Leibler(K-L) distance computing. Both metric-based K-L distance approach and model-based HMM are applied to determine true audio signal changes. The associated visual and audio signal changes are considered as the mid-level scene change representation. This representation has been successfully applied to detect boundaries of individual commercial in TV broadcast stream with an accuracy of around 95%. Particularly the systematic alignment approach can be utilized in video summarization.

1. INTRODUCTION

Currently, Scene change is used more often to segment and classify all types of video data. In news video, for example, scene change may be referred to as the transitions among programs of news, weather, sports, and commercial. News program can be segmented into different stories. Saraceno *et al*. [1] proposed a rule based approach to classify video scenes into: dialogs, stories, actions, and generic. Huang *et al*. [2] used HMM-based classifiers to segment and classify scene according to predefined scene classes; in film video, Sundaram *et al*. [3] proposed a listener model and defined a correlation function that determines the correlation with past data to decide scene changes; in sports video, scenes may be classified according to events such as goal, foul, attack, etc.

The challenges of scene change detection lie in these aspects: (a) Since a scene is a series of semantic correlated shots, scene change is required to indicate semantic inconsistency between adjacent scenes, whereas it is difficult to generally represent the semantic scene change directly with low level audio-visual features; (b) Scene change depends on the predefinition of scene classes at different semantic levels. For example, the scene classes of news, weather, commercial and sports in [2]; the scene classes of play and break in soccer video in [4]; (c) Editing effects may bias the similarity computing between different scenes; (d) Visual signal change may not synchronize with audio signal change, this also add the difficulty of scene change detection; (e) The judgement of semantic correlation between scenes is subjective more or less.

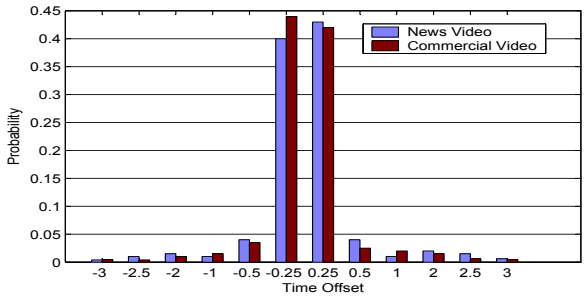

Fig. 1. Statistics of time offsets between an audio signal change and its associated visual signal change in news program and commercial (a positive offset value in sec indicates a delayed audio signal change).

According to extensive observation of news video and commercial video, the video signal changes often occur ahead of audio signal changes in news video whereas it is the inverse in commercial video as indicated in Fig. 1. In news video, the camera shot first switch to the speaker, then speaking starts. A response time cause the delay. The time offset is said to result from the production procedure. In order to attract attention, commercial video uses more editing effects. The voice often comes first, then the speaker shot fades in. The time offset is said to result from post-editing effects. For editing effects, such as fade in and fade out, the visual change position is located at the middle of the shot transition while the audio change position is often delayed to the end of the shot transition. According to the

statistics in Fig. 1, more than 91% time offsets lie in the range of less then 0.5 second in both kinds of video.

This paper proposes a mid-level scene change representation to bridge the semantic gap between low-level features and high-level scene change. Like audio and video keywords used in sports video, the mid-level scene change representation is proposed as one of mid-level features. It can be combined with other enhanced features to accomplish video segmentation, classification and retrieval at the level of semantic scenes. Our proposed scene change representation has been successfully applied to individual TV commercial boundary detection [5]. Scene change representation is meant to indicate temporal inconsistency of audio and visual content that human naturally perceive. For visual and audio signal change indicate the uncorrelated property on audio and video separately, an alignment problem is studied. Since visual and audio signal changes may not occur synchronously, an integrated consideration of these two changes can well characterize video content's temporal correlated or uncorrelated properties. A representation of jointly considering these two kinds of changes can capture the influence of auditory and visual continuity on human perception.

This paper is organized as following: Section 2 gives an overview of our proposed mid-level scene change representation. Section 3 briefly introduces visual signal change detection. Section 4 discusses audio signal change detection. An audiovisual alignment is studied in Section 5. Experimental results are given in Section 6. Section 7 draws a conclusion. Finally Section 8.

2. SYSTEM OVERVIEW

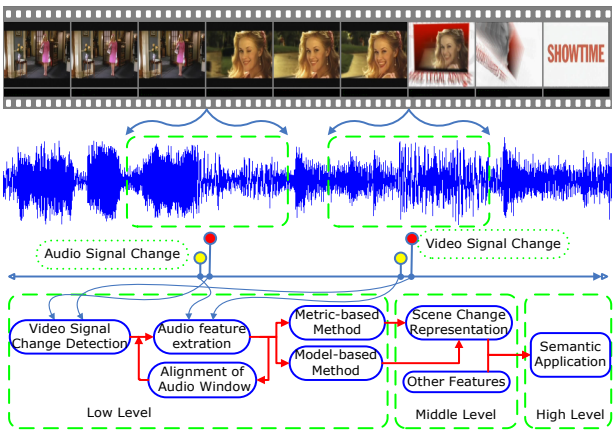

Fig. 2. System Framework

As illustrated in Fig. 2, we are focused on the association of visual signal changes and audio signal changes to propose a mid-level scene change representation. At the low-level, we detect visual signal change by camera motion analysis; by using the visual change position as a time-stamp point, we extract audio features and shift the audio window at a certain range to search for the audio signal change. Each pair of associated visual and audio signal change is considered as the scene change representation. At the mid-level stage, this representation can be combined with other domain dependent features to achieve semantic scene segmentation and classification, such as individual TV commercial boundary detection, wherein each commercial corresponds to a scene.

3. VISUAL SIGNAL CHANGE

In our approach, camera motion estimation algorithm is applied to detect the visual signal changes. We use the support associated to the estimated dominant camera motion to detect visual change positions. At each time instant, we choose the robust multi-resolution motion estimation algorithm described in [6] to estimate camera motion. The support $\mathcal{S}_d$ is the set of point satisfying $\omega(p) \geq \nu$, where $\omega(p)$ is the weight indicating whether or not the point belongs to part of the frame under the dominant motion and ν is a predefined threshold. Hinkley test in parallel is employed to look for the downwards and upwards visual signal changes:

$$S_k = \sum_{t=0}^{k} (\zeta_t - m_0 + \frac{\delta_{min}}{2}) \qquad (k \geq 0) \tag{1}$$

$$T_k = \sum_{t=0}^{k} (\zeta_t - m_0 - \frac{\delta_{min}}{2}) \qquad (k \geq 0) \tag{2}$$

$$M_k = \max_{0 \leq i \leq k} S_i; \qquad N_k = \max_{0 \leq i \leq k} T_i$$

in which ζ_t is the normalized support; the mean m_0 is the support when no change takes place; δ_{min} is the minimal change magnitude. If $M_k - S_k > \alpha$ or $T_k - N_k > \alpha(\alpha$ is a threshold), the visual signal change occurs. Some of the results are illustrated in Fig. 3.

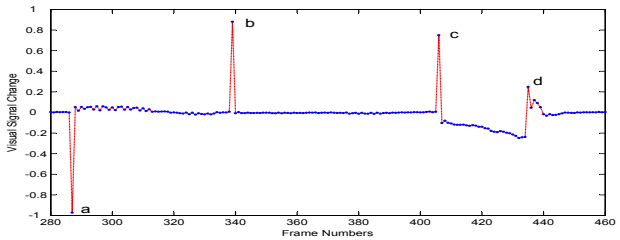

Fig. 3. Visual Signal Change Detection. a, b, c are cuts, a is a down jump, b and c are up jumps; d is a jump from zoom in to zoom out.

4. AUDIO SIGNAL CHANGE

4.1. Audio Feature Selection

Our audio signal change approach considers 43-dimensional audio features comprising Mel-frequency cepstral coefficients (MFCCs) and its delta values and acceleration values, (36 features), mean and variance of short time energy log measure (STE) (2 features), mean and variance of short-time zero-crossing rate (ZCR) (2 features), short-time fundamental frequency (or Pitch) (1 feature), mean of the spectrum flux (SF) (1 feature), and harmonic degree (HD) (1 feature) [7, 8]. As a result of the dynamic nature of complex sounds, we divide the audio signal into many successive 20 ms analysis frames obtained by shifting a 20 ms sliding window with an interval of 10 ms. Those features are computed for each frame. Within each 20 ms analysis frame, we compute the features of STE, ZCR, SF and Harmonic peaks once every 50 samples at an input sampling rate of 22,050 samples/s wherein the sliding window duration is set to 100 samples. Means and variances of STE and ZCR are calculated for 7 results from 7 overlapped frames while mean of SF is calculated for 6 results from 7 neighbor frames. HD is the ratio of the number of frames that have harmonic peaks to the total number of 7. Pitch and MFCCs are computed directly from each 20 ms frame.

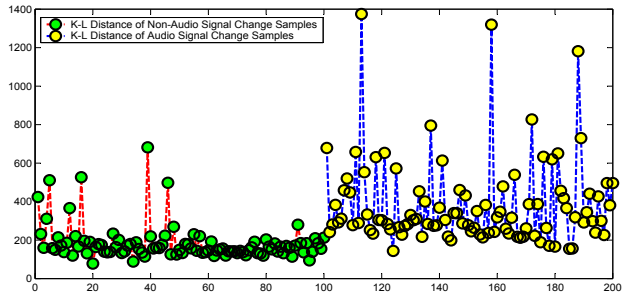

Fig. 4. A series of K-L distance to show feature effectiveness for audio signal change detection. Most of the K-L distances between non-audio signal change samples are 160; while most of the K-L distances between audio signal change samples are 285.

MFCCs furnish a more efficient representation of speech spectra, which are widely used in speech recognition. STE provides a basis for discriminating between voiced speech components and unvoiced speech components, speech and music, audible sounds and silence. Compared with speech, ZCR of music features a much lower variance and average amplitude. ZCR is also useful for distinguishing environmental sounds. Pitch determines the harmonic property for audio signals. Voiced speech components are harmonic while unvoiced speech components are non-harmonic. Sounds from most musical instruments are harmonic while most environmental sounds are non-harmonic. General speaking, SF values of speech are higher than those of music but less than those of environmental sounds.

To show the effectiveness of selected 43-dimensional audio features, we choose 200 non-audio signal change samples and 200 audio audio signal change samples. The duration of each audio sample is 2 second. The audio data cover diverse audio classes such as speech, different kind of music, speech with music background, environment sound, silence, speech with noise background, etc. Fig. 4 illustrates the difference of K-L distances [9] calculated from non-audio signal change samples and audio signal change samples.

4.2. Audio Signal Change Detection

Our task is to find whether there is an audio signal change at a candidate audio signal change position. We apply a metric based approach and a model based approach respectively to examine audio signal change. For the metric based approach, by the candidate audio signal change position as a time-stamp point, we choose two consecutive audio segments of 2 second, one is before the time-stamp point, the other is after the time-stamp point. The change detection is accomplished by examining the K-L distance between these two consecutive audio segments. The normal density function is currently employed to estimate the probability distribution of 43-dimensional audio features for each segment.

For the model based approach, The temporal information is incorporated with left to right Hidden Markov Model(HMM). Unlike traditional audio content analysis methods, which classify audio signals into different categories and detect the transition between different categories [8, 2], we select a 4 second window centered at the candidate audio signal change position and classify the 4 second time series data into audio signal changes or non-audio signal changes by supervised learning. The candidate audio signal change position is an important time-stamp point where the hidden state may change from one to another. For the audio signal change HMM, this hidden state change represents variations from one kind of audio to another;

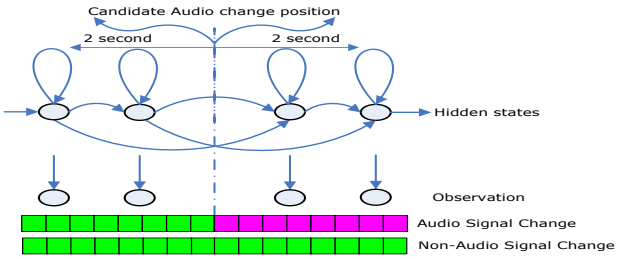

Fig. 5. An HMM to model the audio signal change/non-audio signal change

while in non-audio signal change HMM, this hidden state change represents variance within the same kind of audio. Once the model topology and observation vectors are determined, Baum-welch algorithm is used for parameter estimation. Fig. 5 illustrates the structure of an HMM prototype used in our framework.

5. AUDIOVISUAL ALIGNMENT

As discussed in Section 1, we know audiovisual alignment is an important procedure for video content analysis. For each visual signal change position, we choose overlapping sliding windows to carry out alignment to search candidate audio signal change positions. As illustrated in Fig. 6, K-L distance metric is used to evaluate the changes between successive audio analysis windows. Window size is critical to good modeling. The difference curves in Fig. 7 have indicated different locations of change peaks in the case of different window sizes, which means the candidate audio signal change position is not the same with different window sizes. Since one does not know a priori what sound one is analyzing, a multi-scale difference computing is used. That is, we first make use of different window sizes to yield a set of difference sequences; each difference sequence is then normalized to [0, 1] through dividing difference values by the maximum of each sequence; the most likely audio signal change is determined by locating the highest accumulated difference values derived from the set of difference sequences. The probability $p(\omega_i)$ of each window position being the candidate audio signal change position can be calculated as follow:

$$p(\omega_i) = \frac{1}{N} \sum_{scale=1}^{N} \left(\frac{Distance_{K-L}(i)}{\max\limits_{1 \le i \le M} (Distance_{K-L}(i))} \right) \quad (3)$$

$$p(\theta) = \max_{i}(p(\omega_i)) \qquad (i = 1, \ldots, M) \quad (4)$$

where M is the total window position number, the position corresponding to the probability $p(\theta)$ is the candidate audio signal change position. As shown in Fig. 7, a set of uniform difference peaks associated with the true audio signal change has been located with around 240 ms delay. According to offset statistics in Fig. 1, the shift of adjusted change point is currently confined to the range of [-500ms, 500ms] for balancing the advantage and disadvantage of time shift. For the minimum window of 500 ms we have in total 499 samples of 20 ms unit with a 10 ms overlap. At the sliding window level an overlap of 100 ms has been uniformly employed for multi-scale computing as shown in Fig. 6.

Once we get the candidate audio signal change positions, we further extract audio feature and arrange audio features within adjusted 4 second windows. For comparing the metric based approach and the model based approach, K-L distance metric and HMM model

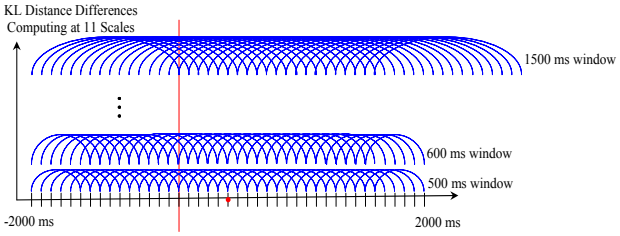

Fig. 6. A Kullback-Leibler distance based alignment of audio and visual signal changes

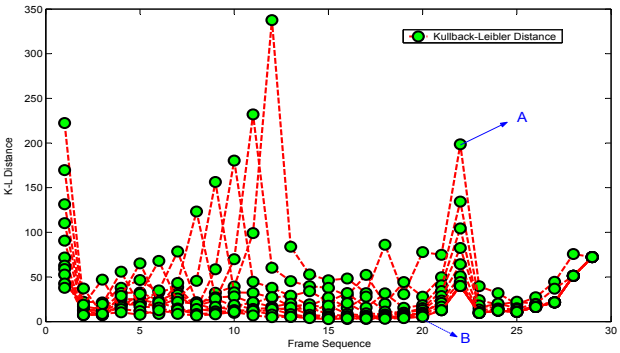

Fig. 7. 11 Curves of Kullback-Leibler distance differences between two successive windows with an overlap of 100 ms wherein 11 different window sizes are applied. A is a set of uniform difference peaks associated with the true audio signal change, around 240 ms delayed; B is The initial audio signal change point(visual signal change position) before adjustment.

are both used to examine audio signal changes and non-audio signal changes. The visual and audio signal changes are finally associated to form the scene change representations.

6. EXPERIMENT

The proposed algorithm is tested on TRECVID 2005 video data. The data set is taken from 6 general channels: CNN, NBC, MSNBC, LBC, CCTV4, NTDBC and includes 6 genres: news, commercial, movie, sports, MTV, animation. Ground truth for audio signal changes and visual signal change was manually labeled. Half is used for training and half for testing.

In our experiment, the number of hidden states in the two HMM models is 8, each state's observation distribution is modeled by 12 Gaussian mixtures with 43 dimensional mean and 43 by 43 diagonal variance. Table 1 lists the result of visual signal change; Table 2 lists the result of metric based method and model based method, including the comparisons before and after using audiovisual alignment. The audio segments comprise 2394 non-audio signal change samples and 1932 audio signal change samples.

Table 1. Experiment results of visual signal change detection

	precision	recall	Accuracy
Visual Signal Change	78.2%	83.5%	86.1%

After alignment, the results are improved by above 4 percent both for metric based approach and model based approach. Compared with metric base symmetrized K-L distance, model based HMM approach yield better performance.

Table 2. Experiment results of audio signal change detection

	alignment	precision	recall	F1	Accuracy
K-L	No	72.8%	76.6%	74.6%	79.8%
K-L	Yes	76.7%	81.8%	79.2%	84.0%
HMM	No	76.1%	80.5%	78.2%	83.6%
HMM	Yes	79.5%	84.9%	82.1%	87.9%

To evaluate the mid-level scene change representation, we has combined this representation with the so-called FMPI (Image Frames Marked with Production Information) frames [5], black frames, silence to detect individual commercial's boundaries, A good detection accuracy of around 95 percent has been achieved.

7. CONCLUSION

We have proposed a mid-level scene change representation to bridge the gap between low level audiovisual features and high level semantic scene change. It is suggested to deal with audio signal change and visual signal change jointly. The associated visual and audio signal changes are considered as the scene change representation. When the mid-level feature is combined with other features to detect individual commercial boundaries, we get a promising accuracy of 95 percent, which proves its effectiveness. In future work, we will apply this mid-level scene representation to story segmentation in news video and scenes classification in films.

8. ACKNOWLEDGEMENT

This work is supported by National Natural Science Foundation of China (Grant No. 60475010 and 60121302).

9. REFERENCES

[1] C. Saraceno and R. Leonardi, "Audio as a support to scene change detection and characterization of video sequence," *Proc. ICASPP'97*, vol. 4, pp. 2597–2600, April 1997.

[2] Jincheng Huang, Zhu Liu, and Yao Wang, "Jiont scene classification and segmentation based on hidden markov model," *IEEE Trans. Multimedia*, vol. 7, no. 3, June 2005.

[3] H. Sundaram and S.F. change, "Audio scene segmentation using multiple features, models and time scales," *Proc. ICASPP'00*, June 2000.

[4] L. Xie, S.F. Chang, A. Divakaran, and H. Sun, "Unsupervised discovery of multilevel statistical video structures using hierarchical hidden markov models," *Proc. ICME'03*, July 2003.

[5] Lingyu Duan, Jinqiao Wang, Jesse S. Jin, and Hanqing Lu, "Fusing multi-model features to detect boundaries of individual commericals in video streams," *submitted to Proc. ICIP'06*, 2006.

[6] P. Bounthemy, M. Gelgon, and F. Ganansia, "A unified approach to shot change detection and camera motion characterization," *IEEE Trans. Circuits and System for Video Technolgy*, vol. 9, no. 7, October 1999.

[7] S. Srinivasan, D. Petkovic, and D. Ponceleon, "Towards robust features for classifying audio in the cuevideo system," *Proc. ACM Multimedia'99*, pp. 393–400, Nov 1999.

[8] Tong Zhang and C.-C. Jay Kuo, "Audio content analysis for online audiovisual data segmentation and classification," *IEEE Trans. on Speech ans Audio Processing*, vol. 9, no. 4, May 2001.

[9] J. P. Campbell and JR, "Speaker recognition: A tutorial," *Proceedings of the IEEE*, pp. 1437–1642, 1997.