Audio-Based Video Editing with Two-Channel Microphone

Tetsuya Takiguchi
Organization of Advanced Science and Technology
Kobe University, Japan
takigu@kobe-u.ac.jp

Jun Adachi
Graduate School of Science and Technology
Kobe University, Japan
j-adachi@me.cs.scitec.kobe-u.ac.jp

Yasuo Ariki
Organization of Advanced Science and Technology
Kobe University, Japan
ariki@kobe-u.ac.jp

Abstract

Audio has a key index in digital videos that can provide useful information for video editing, such as capturing conversations only, clipping only talking people, and so on. In this paper, we are studying about video editing based on audio with a two-channel (stereo) microphone that is standard equipment on video cameras, where the video content is automatically recorded without a cameraman. In order to capture only a talking person on video, a novel voice/non-voice detection algorithm using AdaBoost, which can achieve extremely high detection rates in noisy environments, is used. In addition, the sound source direction is estimated by the CSP (Crosspower-Spectrum Phase) method in order to zoom in on the talking person by clipping frames from videos, where a two-channel (stereo) microphone is used to obtain information about time differences between the microphones.

1. Introduction

Video camera systems are becoming popular in home environments and they are often used in our daily lives to record family growth, small home parties, and so on. In home environments, the video contents, however, are greatly subjected to restrictions due to the fact that there is no production staff, such as a cameraman, editor, switcher, and so on, as with broadcasting or television stations.

When we watch a broadcast or television video, the camera work helps us to not lose interest in or to understand its contents easily owing to the panning and zooming of the camera work. This means that the camera work is strongly associated with the events on video and the most appropriate camera work is chosen according to the events. Through the camera work in combination with event recognition, more interesting and intelligible video content can be produced [4].

Audio has a key index in the digital videos that can provide useful information for video retrieval. In [10], audio features are used for video scene segmentation, in [3, 2], they are used for video retrieval, and in [5], multiple microphones are used for detection and separation of audio in meeting recordings. In [9], they describe an automation system to capture and broadcast lectures to online audiencs, where a two-channel microphone is used for locating talking audience members in a lecture room. Also, there are many approaches possible for the content production system, such as generating highlights, summaries, and so on [7, 1, 12] for home video content.

In this paper, we are studying about home video editing based on audio. In home environments, since it may be difficult for one person to record video continuously (especially for small home parties: just two persons), it will require the video content to be automatically recorded without a cameraman. However, it may result in a large volume of video content. Therefore, this will require digital camera work which uses virtual panning and zooming by clipping frames from hi-resolution images and controlling the frame size and position [4].

In this paper, we propose a method of video editing based on audio, such as voice/non-voice events and sound source direction, from video content that is recorded without a cameraman. This system can automatically capture only conversations using a voice/non-voice detection algorithm based on AdaBoost. In addition, this system can clip and zoom in on a talking person only by using the sound source direction estimated by CSP, where a two-channel (stereo) microphone is used.

One of the advantages of the digital shooting is that the

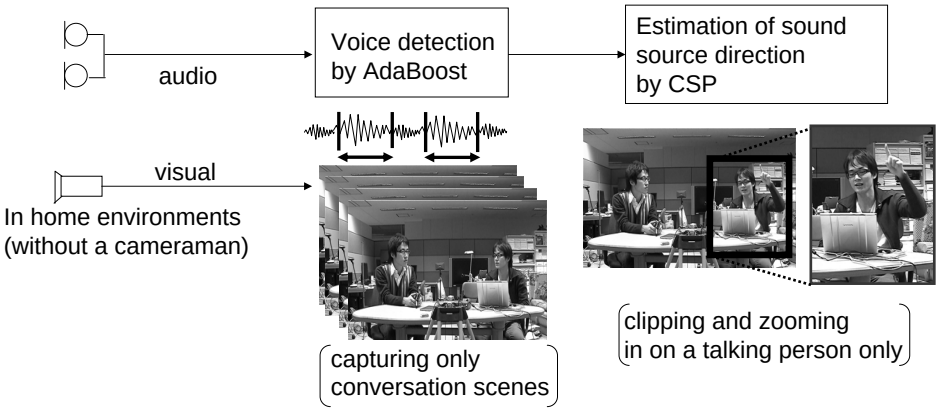

Figure 1. Video editing system by audio-based digital camera work.

camera work, such as panning and zooming, is adjusted to user preferences. This means that the user can watch his/her own video produced by his/her own virtual editor, cameraman, and switcher based on the user's personal preferences. The main point of this paper is that home video events can be recognized using a microphone-array technique and then used as the key indices to retrieve the events and also to summarize the whole home video.

The organization of this paper is as follows. In Section 2, the overview of the video editing system based on audio is presented. Section 3 describes voice detection with AdaBoost in order to capture conversation scenes only. Section 4 describes the estimation of the talker's direction with CSP in order to zoom in on the talking person by clipping frames from the conversation scene videos. Section 5 describes the digital camera work.

2 Overview of the System

Figure 1 shows the overview of the video editing system using audio-based digital camera work. The system is composed of two steps. The first step is voice detection with AdaBoost, where the system identifies whether the audio signal is a voice or not in order to capture conversation scenes only. When the captured video is a conversation scene, the system performs the second step. The second step is estimation of the sound source direction using the CSP (Crosspower-Spectrum Phase) method, where a two-channel microphone is used. Using the sound source direction, the system can clip and zoom in on a talking person only.

3 Voice Detection with AdaBoost

In automatic production of home videos, a speech detection algorithm plays an especially important role in capture of conversation scenes only. In this section, a speech/non-speech detection algorithm using AdaBoost, which can achieve extremely high detection rates, is described.

"Boosting" is a technique in which a set of weak classifiers is combined to form one high-performance prediction rule, and AdaBoost [6] serves as an adaptive boosting algorithm in which the rule for combining the weak classifiers adapts to the problem and is able to yield extremely efficient classifiers.

Figure 2 shows the overview of the voice detection system based on AdaBoost. The audio waveform is split into a small segment by a window function. Each segment is converted to the linear spectral domain by applying the discrete Fourier transform (DFT). Then the logarithm is applied to the linear power spectrum, and the feature vector is obtained.

The AdaBoost algorithm [6] uses a set of training data,

$$\{(X(1), Y(1)), \ldots, (X(N), Y(N))\}, \tag{1}$$

where $X(n)$ is the n-th feature vector of the observed signal and Y is a set of possible labels. For the speech detection, we consider just two possible labels, Y = {-1, 1}, where the label, 1, means voice, and the label, -1, means noise. Next, the initial weight for the n-th training data is set to

$$w_1(n) = \begin{cases} \dfrac{1}{2m}, & Y(n) = 1 \ \ (\text{voice}) \\ \dfrac{1}{2l}, & Y(n) = -1 \ \ (\text{noise}) \end{cases}$$

where m is the total voice frame number and l is the total noise frame number.

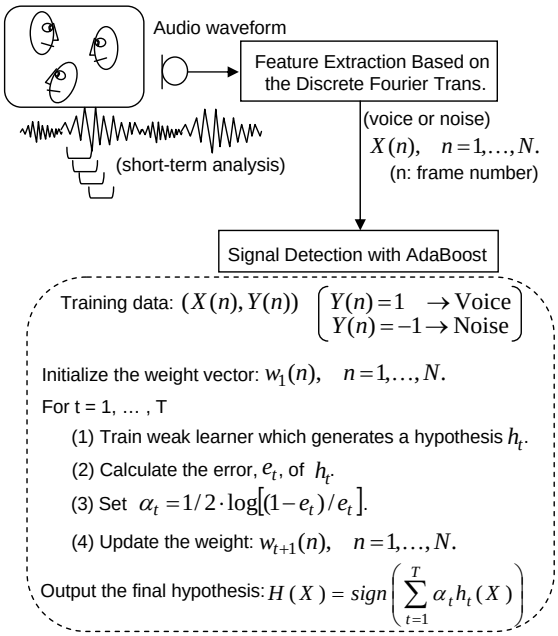

Figure 2. Voice detection with AdaBoost.

As shown in Figure 2, the weak learner generates a hypothesis h_t: $X \rightarrow \{-1, 1\}$ that has a small error. In this paper, single-level decision trees (also known as decision stamps) are used as the base classifiers. After training the weak learner on t-th iteration, the error of h_t is calculated by

$$e_t = \sum_{n:h_t(X(n)) \neq Y(n)} w_t(n). \tag{2}$$

Next, AdaBoost sets a parameter α_t. Intuitively, α_t measures the importance that is assigned to h_t. Then the weight w_t is updated.

$$w_{t+1}(n) = \frac{w_t(n) \exp\{-\alpha_t \cdot Y(n) \cdot h_t(X(n))\}}{\sum_{n=1}^{N} w_t(n) \exp\{-\alpha_t \cdot Y(n) \cdot h_t(X(n))\}} \tag{3}$$

The equation (3) leads to the increase of the weight for the data misclassified by h_t. Therefore, the weight tends to concentrate on "hard" data. After T-th iteration, the final hypothesis, $H(X)$, combines the outputs of the T weak hypotheses using a weighted majority vote.

In home video environments, speech signals may be severely corrupted by noise because the person speaks far from the microphone. In such situations, the speech signal captured by the microphone will have a low SNR (signal-to-noise ratio) which leads to "hard" data. As the AdaBoost trains the weight, focusing on "hard" data, we can expect that it will achieve extremely high detection rates in low SNR situations. For example, in [11], the proposed method has been evaluated on car environments, and the experimental results show an improved voice detection rate, compared to that of conventional detectors based on the GMM (Gaussian Mixture Model) in a car moving at highway speed (an SNR of 2 dB).

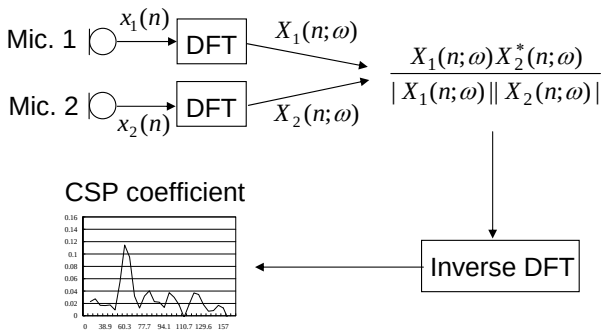

Figure 3. Estimation of sound source direction by CSP.

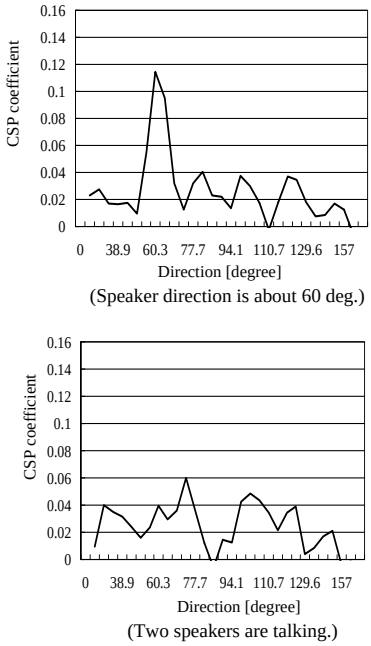

Figure 4. CSP coefficients.

4 Estimation of Sound Source Direction with CSP

The video editing system is requested to detect a person who is talking from among a group of persons. This section describes the estimation of the person's direction (horizontal localization) from the voice. As the home video system may require a small computation resource due to its limitations in computing capability, the CSP (Crosspower-Spectrum Phase)-based technique [8] has been implemented on the video-editing system for a real-time location system.

The crosspower-spectum is computed through the short-term Fourier transform applied to windowed segments of the signal $x_i[t]$ received by the i-th microphone at time t:

$$CS(n;\omega) = X_i(n;\omega)X_j^*(n;\omega), \tag{4}$$

where $*$ denotes the complex conjugate, n is the frame number, and ω is the spectral frequency. Then the normalized crosspower-spectrum is computed by

$$\phi(n;\omega) = \frac{X_i(n;\omega)X_j^*(n;\omega)}{|X_i(n;\omega)||X_j(n;\omega)|} \tag{5}$$

that preserves only information about phase differences between x_i and x_j. Finally, the inverse Fourier transform is computed to obtain the time lag (delay) corresponding to the source direction.

$$C(n;l) = \mathcal{F}^{-1}\phi(n;\omega) \tag{6}$$

Given the above representation, the source direction can be derived. If the sound source is non-moving, $C(n;l)$ should consist of a dominant straight line at the theoretical delay. In this paper, the source direction has been estimated averaging angles corresponding to these delays. Therefore, a lag is given as follows:

$$\hat{l} = \underset{l}{\operatorname{argmax}} \left\{ \sum_{n=1}^{N} C(n;l) \right\}, \tag{7}$$

where N is the total frame in a voice interval which is estimated by AdaBoost. Figure 3 shows the overview of the sound source direction by CSP.

Figure 4 shows the CSP coefficients. The top is the result for a speaker direction of 60 degrees, the middle is that for 105 degrees and the bottom is that for two speakers' talking. As shown in Figure 4, the peak of the CSP coefficient (in the top figure) is about 60 degrees, where the speaker is located at 60 degrees.

When only one speaker is talking in a voice interval, the shape peak is obtained. However, plural speakers are talking in a voice interval, a sharp peak is not obtained as shown in the bottom figure. Therefore, we set a threshold, and the peak above the threshold is selected as the sound source direction. In the experiments, the threshold was set to 0.08. When the peak is below the threshold, a wide shot is taken.

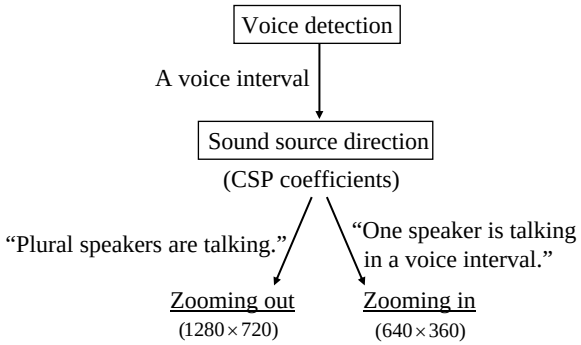

Figure 5. Processing flow of digital zooming in and out.

5 Camera work mudule

In the camera work module, the only one digital panning or zooming is controlled in a voice interval. The digital panning is performed on the HD image by moving the coordinates of the clipping window and the digital zooming is performed by changing the size of the clipping window.

5.1 Zooming

Figure 5 shows the processing flow of the digital camera work (zooming in and out). After capturing a voice interval by AdaBoost, the sound source direction is estimated by CSP in order to zoom in on the talking person by clipping frames from videos.

As described in Section 4, we can estimate that one speaker is talking or plural speakers are talking in a voice interval. In the camera work, when plural speakers are talking, a wide shot (1280×720) is taken. On the other hand, when one speaker is talking in a voice interval, the digital camera work zooms in the speaker. In this paper, the size of the clipping window (zooming in) is fixed to 640×360.

5.2 Clipping position (Panning)

The centroid of the clipping window is selected according to the face region estimated by using the OpenCV library. If the centroid of the clipping window is changing frequently in a voice interval, the video becomes not intelligible so that the centroid of the clipping window is fixed in a voice interval.

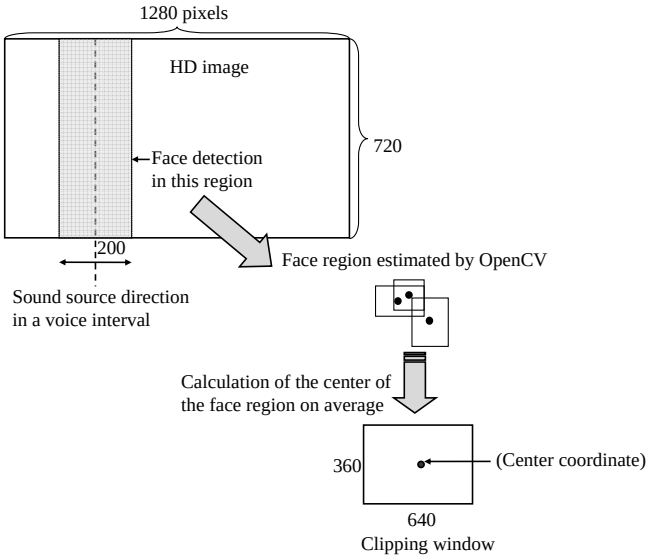

Figure 6. Clipping window for zooming in.

The face regions are detected within the 200 pixels of the sound source direction in a voice interval as shown in Figure 6. Then the average centroid is calculated in order to decide that of the clipping window.

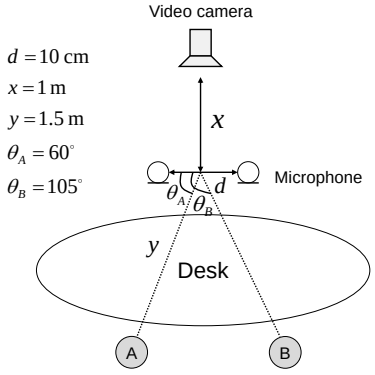

Figure 7. Room used for the experiments. A two-person conversation is recorded.

6 Experiments

Preliminary experiments were performed to test the voice detection algorithm and the CSP method in a room. Figure 7 shows the room used for the experiments, where

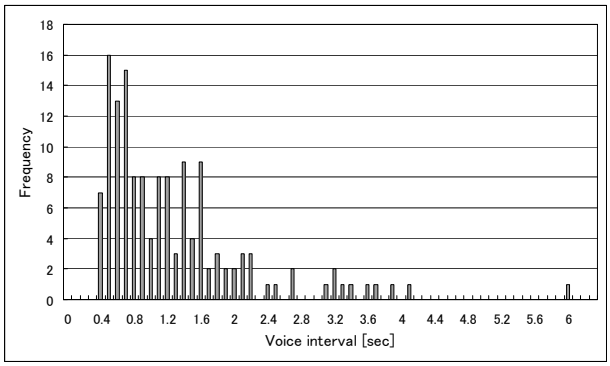

Figure 8. Interval of conversation scene that was estimated by AdaBoost.

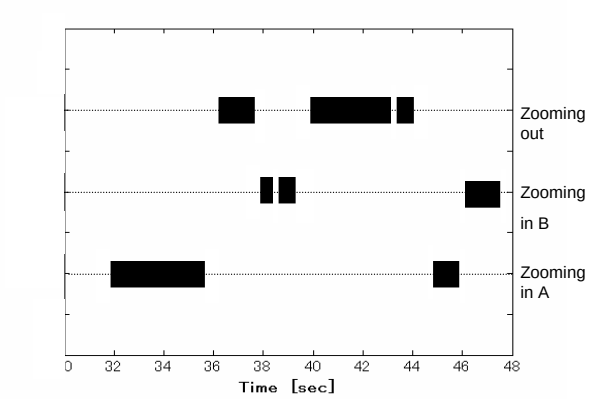

Figure 9. Example of time sequence for zooming in and out.

a two-person conversation is recorded. The total recording time is about 303 seconds.

In the experiments, we used a Victor GR-HD1 Hi-vision camera (1280×720). The focal length is 5.2 mm. The image format size is 2.735 mm (height), 4.864 mm (width) and 5.580 mm (diagonal). From these parameters, we can calculate the position of a pixel number corresponding to the sound source direction in order to clip frames from high-resolution images. (In the proposed method, we can calculate the horizontal localization only.)

Figure 8 shows the interval of the conversation scene that was estimated by AdaBoost. The average interval is 1.32 sec., the max is 6.07 sec., and the minimum is 0.46 sec. The total number of conversation scenes detected by AdaBoost is 149 (186.4 sec) and the detection accuracy is 94.6%.

After capturing conversations only, the sound source di-

Table 1. Total time of zooming in and out.

	correct time	estimated time
zooming in A	63.0	67.3
zooming in B	41.0	55.6
zooming in another direction	0.0	0.5
zooming out	81.8	63.0

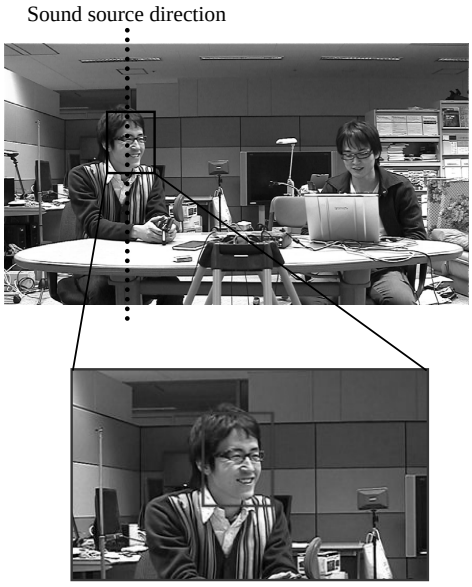

Figure 10. Example of digital shooting (zooming in).

rection is estimated by CSP in order to zoom in on the talking person by clipping frames from videos. The clipping accuracy is 65.5% in this experiment. Some conversation scenes cause a decrease in the accuracy of clipping. This is because two speakers are talking in one voice (conversation) interval estimated by AdaBoost and it is difficult to set the threshold of the CSP coefficient. Figure 9 shows an example of time sequence for zooming in and out, and Table 1 shows the results of the digital camera work (zooming in and out). Figure 10 shows an example of the digital shooting (zooming in). In this experiment, the clipping size is fixed to 640×360. In the future, we need to automatically select the size of the clipping window according to each situation.

7 Conclusions

In this paper, we investigated about home video editing based on audio with a two-channel (stereo) microphone, where the video content is automatically recorded without a cameraman. In order to capture a talking person only, a novel voice/non-voice detection algorithm using AdaBoost, which can achieve extremely high detection rates in noisy environments, is used. In addition, the sound source direction is estimated by the CSP (Crosspower-Spectrum Phase) method in order to zoom in on the talking person by clipping frames from videos, where a two-channel (stereo) microphone is used to obtain information about time differences between the microphones. Our proposed system can not only produce the video content but also retrieve the scene in the video content by utilizing the detected voice interval or information of a talking person as indices. To make the system more advanced, we will develop the sound source estimation and emotion recognition in future, and we will evaluate the proposed method on more test data.

References

[1] B. Adams and S. Venkatesh. Dynamic shot suggestion filtering for home video based on user performance. In *ACM Int. Conf. on Multimedia*, pages 363–366, 2005.

[2] K. Aizawa. Digitizing personal experiences: Capture and retrieval of life log. In *Proc. Multimedia Modelling Conf.*, pages 10–15, 2005.

[3] T. Amin, M. Zeytinoglu, L. Guan, and Q. Zhang. Interactive video retrieval using embedded audio content. In *Proc. ICASSP*, pages 449–452, 2004.

[4] Y. Ariki, S. Kubota, and M. Kumano. Automatic production system of soccer sports video by digital camera work based on situation recognition. In *Eight IEEE International Symposium on Multimedia (ISM)*, pages 851–858, 2006.

[5] F. Asano and J. Ogata. Detection and separation of speech events in meeting recordings. In *Proc. Interspeech*, pages 2586–2589, 2006.

[6] Y. Freund and R. E. Schapire. A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence*, 14(5):771–780, 1999.

[7] X.-S. Hua, L. Lu, and H.-J. Zhang. Optimization-based automated home video editing system. *IEEE Transactions on circuits and systems for video technology*, 14(5):572–583, 2004.

[8] M. Omologo and P. Svaizer. Acoustic source location in noisy and reverberant environment using CSP analysis. In *Proc. ICASSP*, pages 921–924, 1996.

[9] Y. Rui, A. Gupta, J. Grudin, and L. He. Automating lecture capture and broadcast: technology and videography. In *ACM Multimedia Systems Journal*, pages 3–15, 2004.

[10] H. Sundaram and S.-F. Chang. Video scene segmentation using audio and video features. In *Proc. ICME*, pages 1145–1148, 2000.

[11] T. Takiguchi, H. Matsuda, and Y. Ariki. Speech detection using real AdaBoost in car environments. In *Fourth Joint Meeting ASA and ASJ*, page 1pSC20, 2006.

[12] P. Wu. A semi-automatic approach to detect highlights for home video annotation. In *Proc. ICASSP*, pages 957–960, 2004.