We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

18

Video Editing Based on Situation Awareness from Voice Information and Face Emotion

Tetsuya Takiguchi, Jun Adachi and Yasuo Ariki
Kobe University
Japan

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 (Ariki et al., 2006).

Audio has a key index in the digital videos that can provide useful information for video retrieval. In (Sundaram et al, 2000), audio features are used for video scene segmentation, in (Aizawa, 2005) (Amin et al, 2004), they are used for video retrieval, and in (Asano et al, 2006), multiple microphones are used for detection and separation of audio in meeting recordings. In (Rui et al, 2004), they describe an automation system to capture and broadcast lectures to online audience, 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 (Ozeke et al, 2005) (Hua et al, 2004) (Adams et al, 2005) (Wu, 2004) for home video content.

Also, there are some studies that focused on a facial direction and facial expression for a viewer's behavior analysis. (Yamamoto, et al, 2006) proposed a system for automatically estimating the time intervals during which TV viewers have a positive interest in what they are watching based on temporal patterns in facial changes using the Hidden Markov Model.

In this chapter, we are studying about home video editing based on audio and face emotion. 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 (Ariki et al, 2006).

Source: Digital Video, Book edited by: Floriano De Rango,
ISBN 978-953-7619-70-1, pp. 500, February 2010, INTECH, Croatia, downloaded from SCIYO.COM

In this chapter, our 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. Additionally, we extract facial feature points by EBGM (Elastic Bunch Graph Matching) (Wiskott et al, 1997) to estimate atmosphere class by SVM (Support Vector Machine).
One of the advantages of the digital shooting is that the 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 chapter is that home video events can be recognized using techniques based on audio and face emotion and then used as the key indices to retrieve the events and also to summarize the whole home video.
The organization of this chapter is as follows. In Section 2, the overview of the video editing system based on audio and face emotion 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 facial emotion recognition. Section 6 describes the digital camera work.

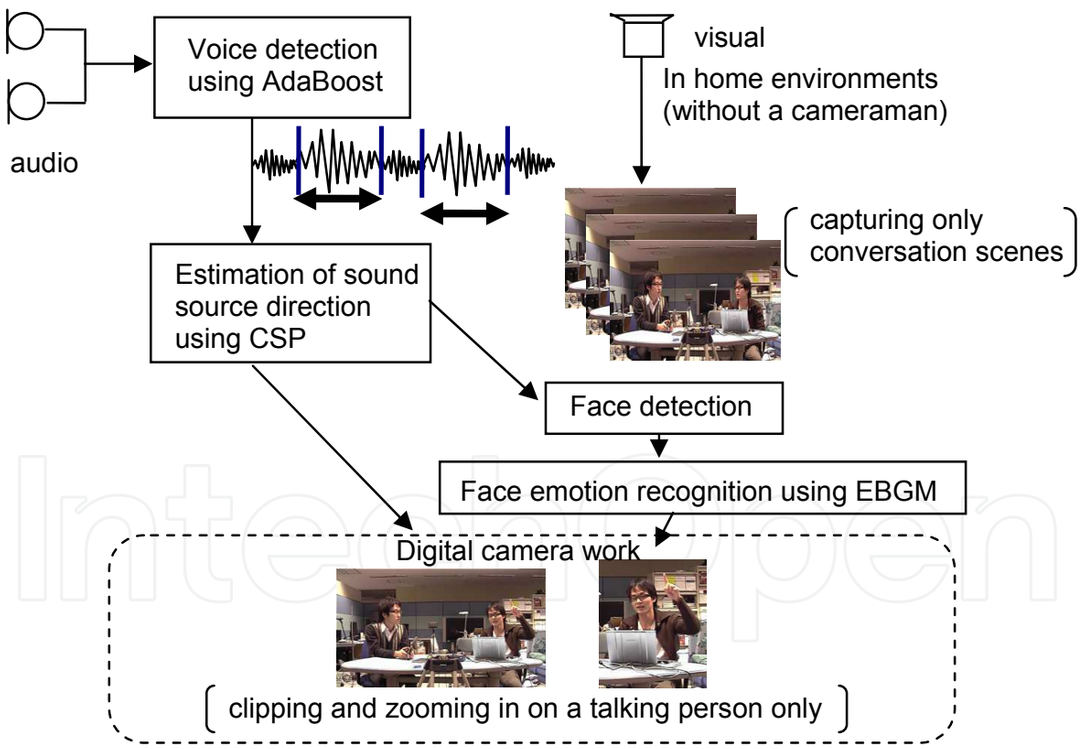


Fig. 1. Video editing system using digital camera work based on audio and face emotion

2. Overview of the system

Figure 1 shows the overview of the video editing system using digital camera work based on audio and face emotion. 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. The third step is face emotion recognition. Using the emotion result, the system can zoom out on persons who have positive expressions (happiness, laughter, etc).

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 highperformance prediction rule, and AdaBoost (Freund et al, 1999) 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 uses a set of training data,

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

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

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

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

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

$$e_t = \sum_{n:h_t(X(n)) \neq Y(n)} w_t(n). \tag{3}$$

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

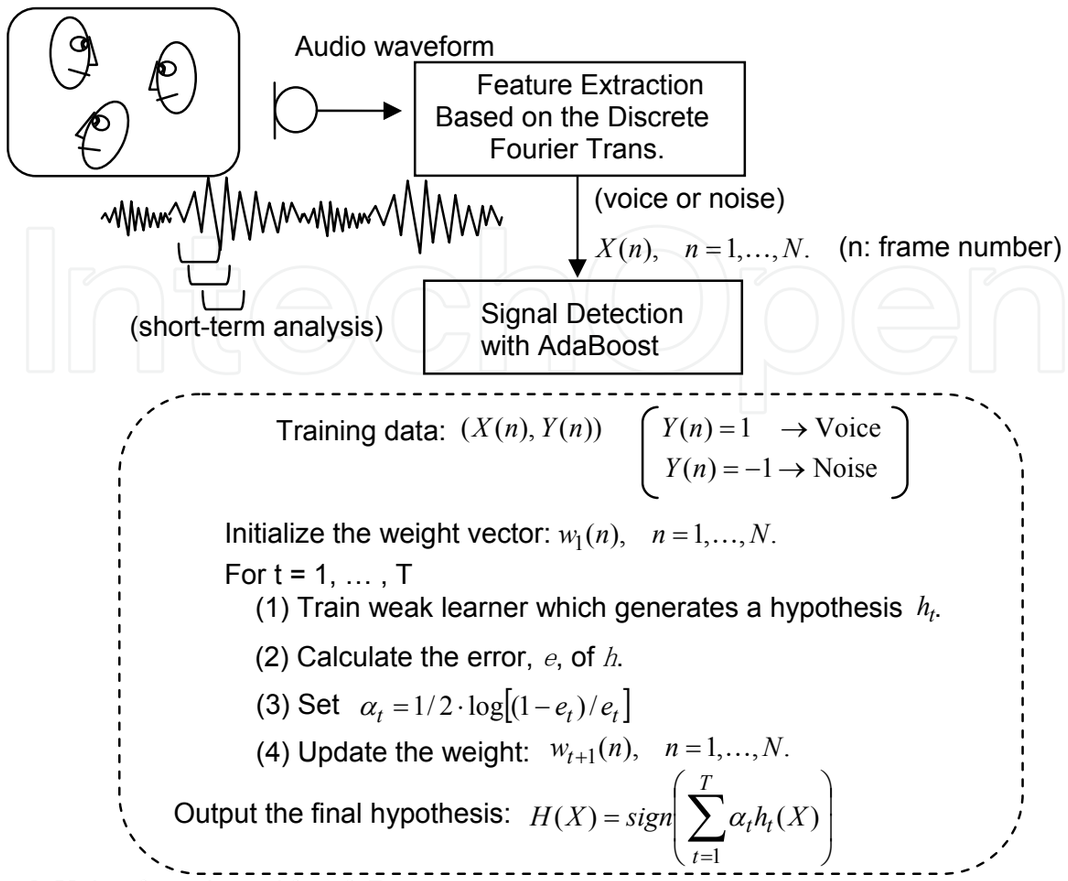


Fig. 2. Voice detection with AdaBoost

$$w_{t+1}(n) = \frac{w_t(n) \exp\{-\alpha_t \cdot Y(n) \cdot h_t(X(n))\}}{\sum_{n=1} w_t(n) \exp\{-\alpha_t \cdot Y(n) \cdot h_t(X(n))\}} \tag{4}$$

The equation (4) leads to the increase of the weight for the data misclassified by ht. 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 (Takiguchi et al, 2006), 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)

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 (Omologo et al, 1996) 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*

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

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

$$\phi(n;\omega) = \frac{X_i(n;\omega)X_j^*(n;\omega)}{|X_i(n;\omega)||X_j^*(n;\omega)|} \tag{6}$$

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.

$$C(n;l) = InverseDFT(\phi(n;\omega)) \tag{7}$$

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 chapter, the source direction has been estimated averaging angles corresponding to these delays. Therefore, a lag is given as follows:

$$\hat{l} = \arg\max_{l}\left\{\sum_{n=1}^{N} C(n;l)\right\} \tag{8}$$

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.

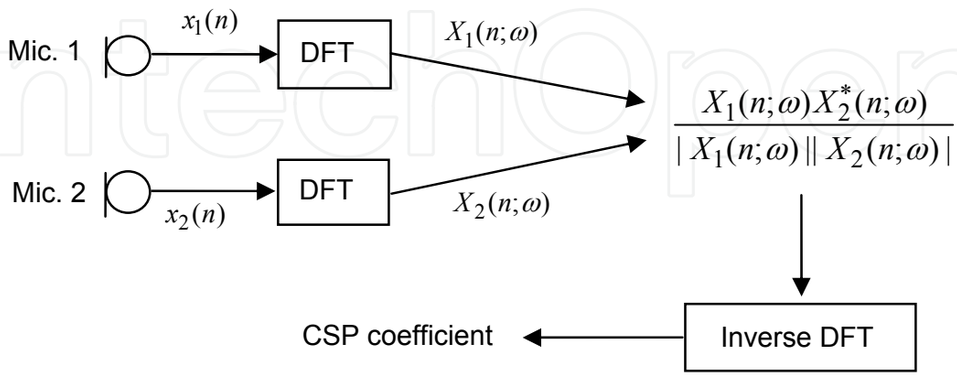
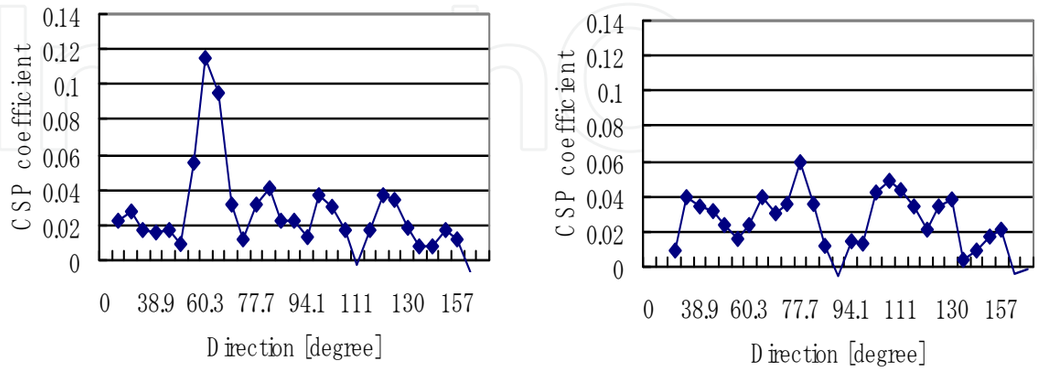


Fig. 3. Estimation of sound source direction using CSP

Figure 4 shows the CSP coefficients. The left is the result for a speaker direction of 60 degrees, and the right is that for two speakers' talking. As shown in Figure 4, the peak of the CSP coefficient (in the left 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.



(Speaker direction is about 60 deg.) (Two speakers are talking.)

Fig. 4. CSP coefficients

5. Facial feature point extraction using EBGM

To classify facial expressions correctly, facial feature points must be extracted precisely. From this viewpoint, Elastic Bunch Graph Matching (EBGM) (Wiskott et al, 1997) is employed in the system. EBGM was proposed by Laurenz Wiskott and proved to be useful in facial feature point extraction and face recognition.

5.1 Gabor wavelets

Since Gabor wavelets are fundamental to EBGM, it is described here. Gabor wavelets can extract global and local features by changing spatial frequency, and can extract features related to a wavelet's orientation.

Eq. (9) shows a Gabor Kernel used in Gabor wavelets. This function contains a Gaussian function for smoothing as well as a wave vector $\vec{k}_j$ that indicates simple wave frequencies and orientations.

$$\varphi_j(\vec{x}) = \frac{\vec{k}_j^2}{\sigma^2}\exp\left(-\frac{\vec{k}_j^2\vec{x}^2}{2\sigma^2}\right)\left[\exp(i\vec{k}_j\vec{x}) - \exp(-\frac{\sigma^2}{2})\right] \tag{9}$$

$$\vec{k}_j = \begin{pmatrix} k_{jx} \\ k_{jy} \end{pmatrix} = \begin{pmatrix} k_\nu \cos\varphi_\mu \\ k_\nu \sin\varphi_\mu \end{pmatrix} \tag{10}$$

Here, $k_\nu = 2^{-\frac{\nu+2}{2}\pi}$, $\varphi_\mu = \mu\frac{\pi}{8}$. We employ a discrete set of 5 different frequencies, index ν = 0, ..., 4, and 8 orientations, index μ = 0, ..., 7.

5.2 Jet

A jet is a set of convolution coefficients obtained by applying Gabor kernels with different frequencies and orientations to a point in an image. To estimate the positions of facial feature points in an input image, jets in an input image are compared with jets in a facial model.

A jet is composed of 40 complex coefficients (5 frequencies * 8 orientations) and expressed as follows:

$$J_j = a_j \exp(i\phi_j) \quad j = 0,...,39 \tag{11}$$

where $\vec{x} = (x, y), a_j(\vec{x})$ and ϕ_j are the facial feature point coordinate, magnitude of complex coefficient, and phase of complex coefficient, which rotates the wavelet at its center, respectively.

5.3 Jet similarity

For the comparison of facial feature points between the facial model and the input image, the similarity is computed between jet set {*J*} and {*J'*}. Locations of two jets are represented as $\vec{x}$ and $\vec{x}'$. The difference between vector $\vec{x}$ and vector $\vec{x}'$ is given in Eq. (12).

$$\vec{d} = \vec{x} - \vec{x}' = \begin{pmatrix} dx \\ dy \end{pmatrix} \tag{12}$$

Here, let's consider the similarity of two jets in terms of the magnitude and phase of the jets as follows:

$$S_D(J, J') = \frac{\sum_j a_j a_j' \cos(\phi_j - \phi_j')}{\sqrt{\sum_j a_j^2 \sum_j a'^2_j}} \tag{13}$$

where the phase difference $(\phi_j - \phi_j')$ is qualitatively expressed as follows:

$$\phi_j - \phi_j' = \vec{k}_j \vec{x} - \vec{k}_j \vec{x}' = \vec{k}_j (\vec{x} - \vec{x}') = \vec{k}_j \vec{d} \tag{14}$$

To find the best similarity between {*J*} and {*J'*} using Eq. (13) and Eq. (14), phase difference is modified as $\phi_j - (\phi_j' + \vec{k}_j \vec{d})$ and Eq. (13) is rewritten as

$$S_D(J, J') = \frac{\sum_j a_j a_j' \cos(\phi_j - (\phi_j' + \vec{k}_j \vec{d}))}{\sqrt{\sum_j a_j^2 \sum_j a'^2_j}} \tag{15}$$

In order to find the optimal jet J' that is most similar to jet J, the best $\vec{d}$ is estimated that will maximize similarity based not only upon phase but magnitude as well.

5.4 Displacement estimation

In Eq. (15), the best $\vec{d}$ is estimated in this way. First, the similarity at zero displacement (dx = dy = 0) is estimated. Then the similarity of its North, East, South, and West neighbors is

estimated. The neighboring location with the highest similarity is chosen as the new center of the search. This process is iterated until none of the neighbors offers an improvement over the current location. The iteration is limited to 50 times at one facial feature point.

5.5 Facial feature points and face graph

In this chapter, facial feature points are defined as the 34 points shown in Fig. 5. A set of jets extracted at all facial feature points is called a face graph. Fig. 5 shows an example of a face graph.

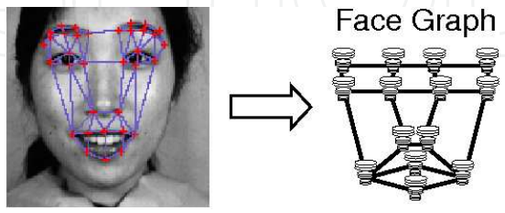


Fig. 5. Jet extracted from facial feature points

5.6 Bunch graph

A set of jets extracted from many people at one facial feature point is called a bunch. A graph constructed using bunches at all the facial feature points is called a bunch graph. In searching out the location of facial feature points, the similarity described in Eq. (15) is computed between the jets in the bunch graph and a jet at a point in an input image. The jet with the highest similarity, achieved by moving $\vec{d}$ as described in Section 5.4, is chosen as the target facial feature point in the input image. In this way, using a bunch graph, the locations of the facial feature points can be searched allowing various variations. For example, a chin bunch may include jets from non-bearded chins as well as bearded chins, to cover the local changes. Therefore, it is necessary to train data using the facial data of various people in order to construct the bunch graph. The training data required for construction of bunch graph was manually collected.

5.7 Elastic bunch graph matching

Fig. 6 shows an elastic bunch graph matching flow. First, after a facial image is input into the system, a bunch graph is pasted to the image, and then a local search for the input face commences using the method described in Section 5.4. A set of jets extracted from many people at one facial feature point is called a bunch. Finally, the face graph is extracted after all the locations of the feature points are matched.

Fig. 6. Elastic Bunch Graph Matching procedure

6. Facial expression recognition using SVM

In this study, three facial expression classes are defined; "Neutral," "Positive" and "Rejective". Table 1 shows the class types and the meanings.

Classes	Meanings
Neutral	Expressionless
Positive	Happiness, Laughter, Pleasure, etc.
Rejective	Watching other direction, Occluding part of face, Tilting the face, etc.

Table 1. Facial expression classes

The users of our system register their neutral images as well as the personal facial expression classifier in advance. The differences between the viewer's facial feature points extracted by EBGM and the viewer's neutral facial feature points are computed as a feature vector for SVM. In our experiments, Radial Basis Function (RBF) is employed as a kernel function of SVM.

7. Camera work module

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.

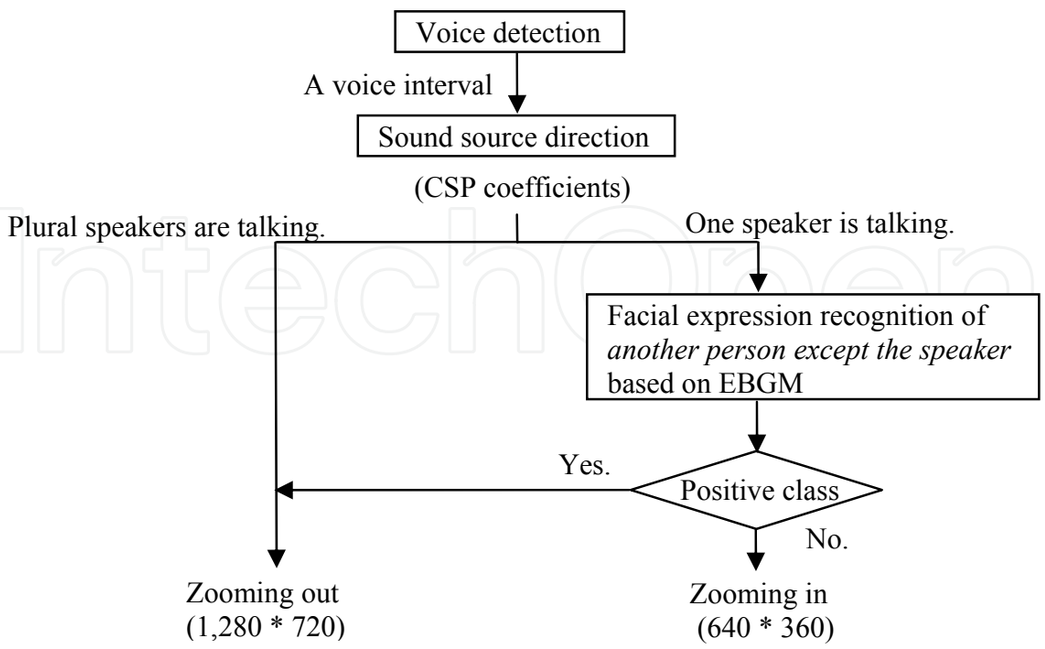


Fig. 7. Processing flow of digital zooming in and out

7.1 Zooming

Figure 7 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 system estimates the atmosphere of another person. When the atmosphere of another person is "positive" class, a wide shot (1280*720) is taken. When the atmosphere of another person except the speaker is not "positive" class, the digital camera work zooms in the speaker. In our experiments, the size of the clipping window (zooming in) is fixed to 640*360.

7.2 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.

The face regions are detected within the 200 pixels of the sound source direction in a voice interval as shown in Figure 8. Then the average centroid is calculated in order to decide that of the clipping window.

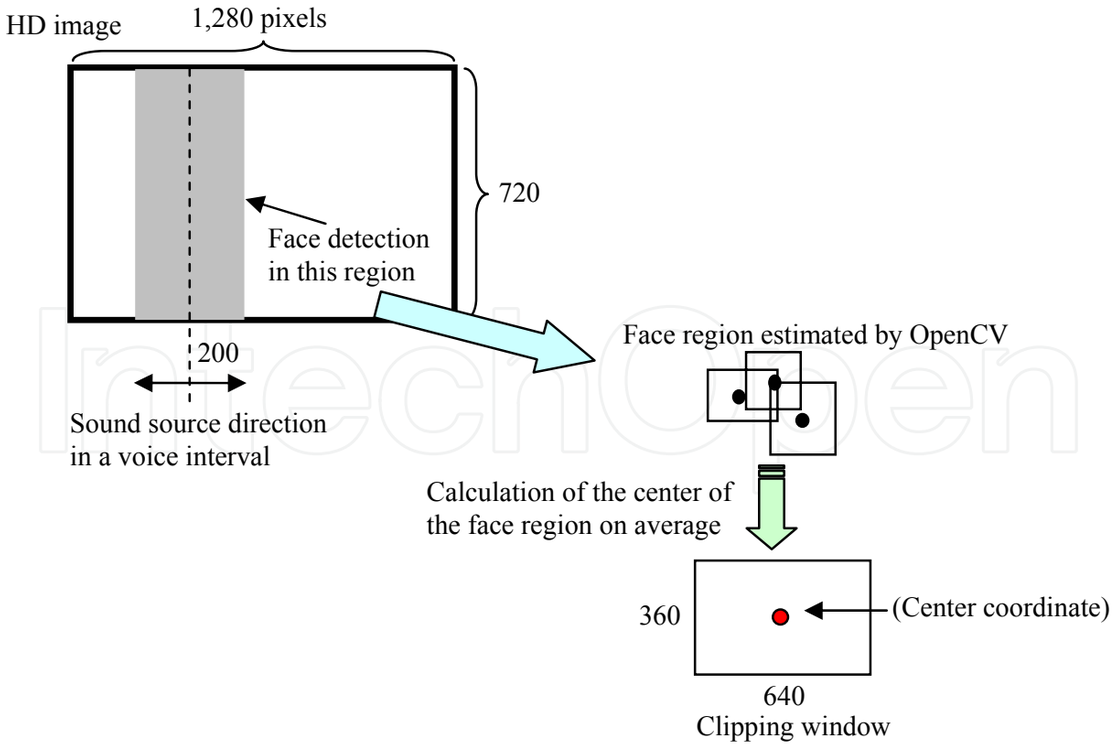


Fig. 8. Clipping window for zooming in

8. Experiments

8.1 Voice detection and sound source direction

Preliminary experiments were performed to test the voice detection algorithm and the CSP method in a room. Figure 9 shows the room used for the experiments, where a two-person conversation is recorded. The total recording time is about 303 seconds.

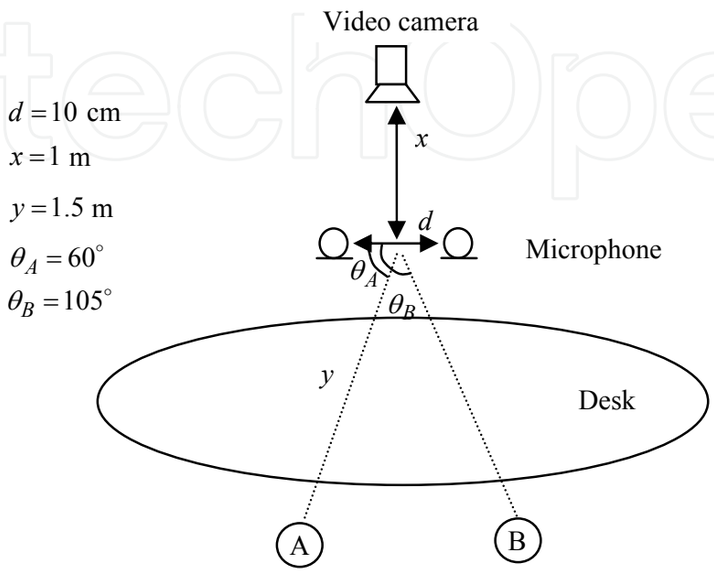


Fig. 9. Room used for the experiments. A two-person conversation is recorded.

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.)

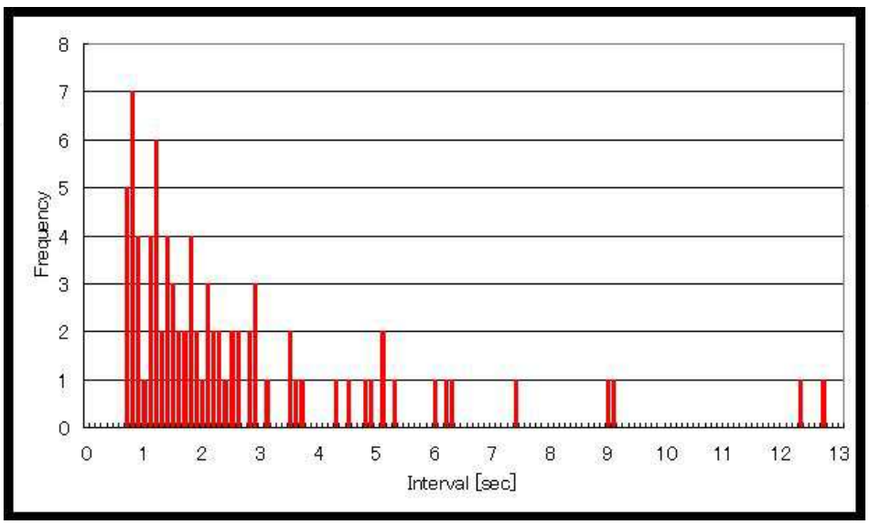


Fig. 10. Interval of conversation scene that was estimated by AdaBoost

Figure 10 shows the interval of the conversation scene that was estimated by AdaBoost. The max interval is 12.77 sec., and the minimum is 0.71 sec. The total number of conversation scenes detected by AdaBoost is 84 (223.9 sec), and the detection accuracy is 98.4%.
After capturing conversation scenes only, the sound source direction is estimated by CSP in order to zoom in on the talking person by clipping frames from videos. The clipping accuracy is 72.6% 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 11 shows an example of time sequence for zooming in and out.

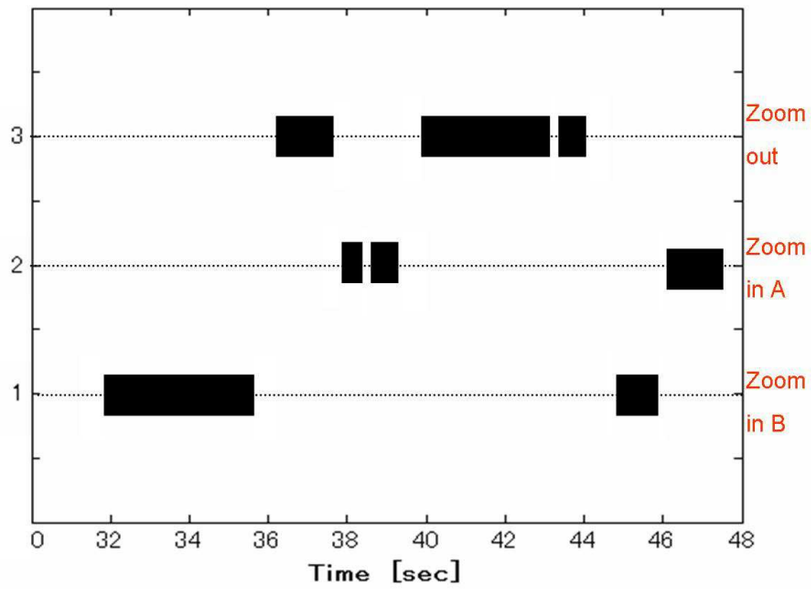


Fig. 11. Example of time sequence for zooming in and out

8.2 Facial expression recognition

	Neutral	Positive	Rejective	Total
Subject A	3,823	2,517	153	6,493
Subject B	3,588	2,637	268	6,493

Table 2. Tagged results (frames)

Next, we tagged the video with three labels, “Neutral,” “Positive,” and “Rejective.” Tagged results for all conversational frames in the experimental videos are shown in Table 2.

Facial regions were extracted using AdaBoost based on Haar-like features (Viola et al, 2001) in all frames of conversation scenes except Reject frames. Extracted frames were checked manually to confirm whether they were false regions or not. The experimental results are shown in Table 3 and Table 4. The extraction rate of the facial region for subject B was not good, compared with that for subject A. The reason for the worse false extraction rate for subject B is attributed to his face in profile, where he often looks toward subject A.

	Neutral	Positive
False extraction	13	16
Total frames	3,823	2,517
Error rate [%]	0.34	0.64

Table 3. Experimental results of facial region extraction for subject A

	Neutral	Positive
False extraction	1,410	1,270
Total frames	3,588	3,719
Error rate [%]	39.3	48.2

Table 4. Experimental results of facial region extraction for subject B

For every frame in the conversation scene, facial expression was recognized by Support Vector Machines. The 100 frames for each subject were used for training data, and the rest for testing data. The experiment results were shown in Table 5 and Table 6.

	Neutral	Positive	Rejective	Sum.	Recall [%]
Neutral	3,028	431	364	3,823	79.2
Positive	230	2,023	264	2,517	80.4
Rejective	32	10	121	153	79.1
Sum.	3,280	2,464	749	6,493	-
Precision [%]	92.3	82.1	16.2	-	-

Table 5. Confusion matrix of facial expression recognition for subject A

	Neutral	Positive	Rejective	Sum.	Recall [%]
Neutral	1,543	214	1,831	3,588	43.0
Positive	194	1,040	1,403	2,637	39.4
Rejective	34	24	210	264	78.4
Sum.	1,771	1,278	3,444	6,493	-
Precision [%]	87.1	81.4	6.1	-	-

Table 6. Confusion matrix of facial expression recognition for subject B

The averaged recall rates for subject A and B were 79.57% and 53.6%, respectively, and the averaged precision rates for subject A and B were 63.53% and 58.2%, respectively. The result of the facial expression for subject B was lower than that for subject A because of the worse

false extraction rate of the facial region. Moreover, when the subjects had an intermediate facial expression, the system often made a mistake because one expression class was only assumed in a frame.

Figure 12 and Figure 13 show an example of the digital shooting (zooming in) and an example of zoom out, respectively. 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, and subjective evaluation of video production will be described.

9. Conclusion

In this chapter, we investigated about home video editing based on audio with a two-channel (stereo) microphone and facial expression, 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. Also, we extract facial feature points by EBGM (Elastic Bunch Graph Matching) to estimate atmosphere class by SVM (Support Vector Machine). When the atmosphere of the other person except the speaker is not "positive" class, the digital camera work zooms in only the speaker. When the atmosphere of the other person is "positive" class, a wide shot is taken. 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.

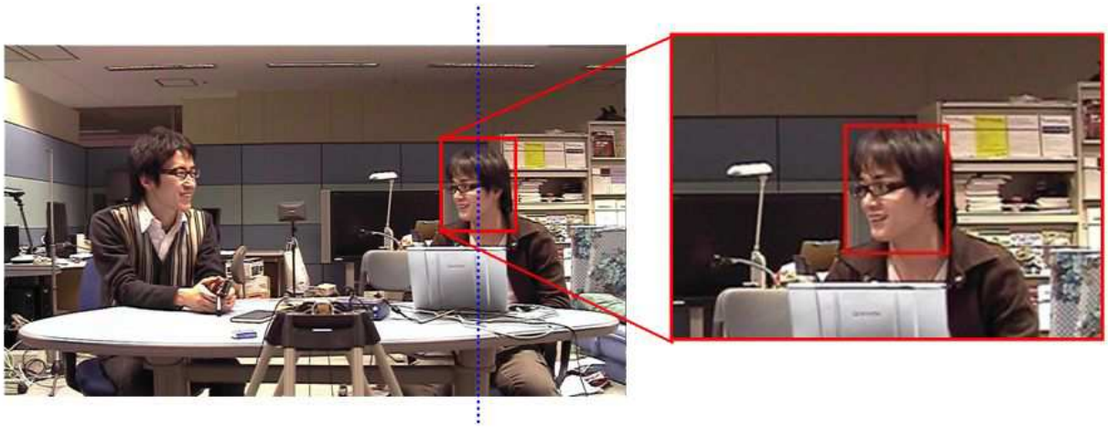


Fig. 12. Example of camera work (zooming in)

Fig. 13. Example of camera work (zoom out)

10. References

Y. Ariki, S. Kubota, & M. Kumano (2006). Automatic production system of soccer sports video by digital camera work based on situation recognition, *Eight IEEE International Symposium on Multimedia (ISM)*, pp. 851-858, 2006.

H. Sundaram & S.-F. Chang (2000). Video scene segmentation using audio and video features, *Proc. ICME*, pp. 1145-1148, 2000.

K. Aizawa. Digitizing personal experiences: Capture and retrieval of life log, *Proc. Multimedia Modelling Conf.*, pp. 10-15, 2005.

T. Amin, M. Zeytinoglu, L. Guan, & Q. Zhang. Interactive video retrieval using embedded audio content, *Proc. ICASSP*, pp. 449-452, 2004..

F. Asano & J. Ogata. Detection and separation of speech events in meeting recordings, *Proc. Interspeech*, pp. 2586-2589, 2006.

Y. Freund & R. E. Schapire. A short introduction to boosting, *Journal of Japanese Society for Artificial Intelligence*, 14(5), pp. 771-780, 1999.

Y. Rui, A. Gupta, J. Grudin, & L. He. Automating lecture capture and broadcast: technology and videography, *ACM Multimedia Systems Journal*, pp. 3-15, 2004.

M. Ozeke, Y. Nakamura, & Y. Ohta. Automated camerawork for capturing desktop presentations, *IEEProc.-Vis. Image Signal Process.*, 152(4), pp. 437-447, 2005.

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

B. Adams & S. Venkatesh. Dynamic shot suggestion filtering for home video based on user performance, *ACM Int. Conf. on Multimedia*, pp. 363-366, 2005.

P. Wu. A semi-automatic approach to detect highlights for home video annotation, *Proc. ICASSP*, pp. 957-960, 2004.

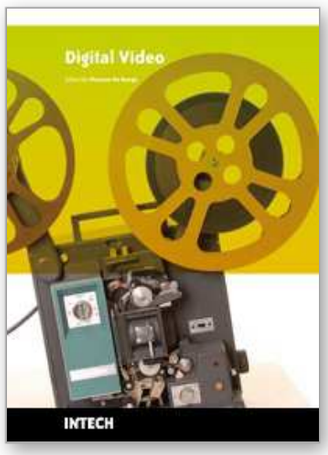
M. Yamamoto, N. Nitta, N. Babaguchi, Estimating Intervals of Interest During TV Viewing for Automatic Personal Preference Acquisition. *Proceedings of The 7th IEEE Pacific-Rim Conference on Multimedia*, pp. 615-623, 2006.

L. Wiskott, J.-M. Fellous, N. Kruger, & C. von der Malsburg, Face Recognition by Elastic Bunch Graph Matching, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19(7), pp. 775-779, 1997.

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

M. Omologo & P. Svaizer. Acoustic source location in noisy and reverberant environment using CSP analysis, *Proc. ICASSP*, pp. 921-924, 1996.

P. Viola & M. Jones, Rapid object detection using a boosted cascade of simple features, *Proc. IEEE conf. on Computer Vision and Pattern Recognition*, pp. 1-9, 2001.



Digital Video

Edited by Floriano De Rango

ISBN 978-953-7619-70-1

Hard cover, 500 pages

Publisher InTech

Published online 01, February, 2010

Published in print edition February, 2010

This book tries to address different aspects and issues related to video and multimedia distribution over the heterogeneous environment considering broadband satellite networks and general wireless systems where wireless communications and conditions can pose serious problems to the efficient and reliable delivery of content. Specific chapters of the book relate to different research topics covering the architectural aspects of the most famous DVB standard (DVB-T, DVB-S/S2, DVB-H etc.), the protocol aspects and the transmission techniques making use of MIMO, hierarchical modulation and lossy compression. In addition, research issues related to the application layer and to the content semantic, organization and research on the web have also been addressed in order to give a complete view of the problems. The network technologies used in the book are mainly broadband wireless and satellite networks. The book can be read by intermediate students, researchers, engineers or people with some knowledge or specialization in network topics.

InTech Europe

University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China

Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen