A SENSOR-FUSION METHOD FOR DETECTING A SPEAKING STUDENT

S. Nishiguchi[1]*, K. Higashi*[2*]*, Y. Kameda*[3] *and M. Minoh*[3]

Graduate School of Law, Kyoto University[1]
Graduate School of Informatics, Kyoto University[2]
Academic Center for Computing and Media Studies, Kyoto University[3]
Sakyoku, Kyoto, 606-8501, Japan

ABSTRACT

In this paper, we propose a method for detecting the location of the speaker that is a target of automatic video filming in distance learning and lecture archive. It is required that a face of a speaking student is filmed in a lecture video. For this purpose, it is necessary to detect the location of a speaker. An acoustic sensor such as a microphone array is used widely to detect the location of a sound source. However, it is difficult to detect the location of a sound source precisely using only microphone array because of sound noise in a large space such as a lecture room. In this paper, we propose a method for detecting more precise location of a speaker in the lecture room using not only the microphone array but also visual sensors. The result shows that the precision ratio of detecting the location of a speaker was improved about 20% by our sensor-fusion method.

1. INTRODUCTION

Lecture archives are knowledge sources, intellectual properties of universities and material for multimedia course ware and teaching evaluation. The purpose of the lecture archives is to provide users various kinds of information about lectures without spatial and temporal restriction.

Our lecture archives consist of media information such as video, audio, slide, and event information generated as a result of interaction between a lecturer and students in a lecture room. The archives should have various useful information such as facial expression, gesture, speaking etc., so that users may retrieve a subset of the lecture archives in compliance with their requests.

In a lecture, a lecturer explains to students, and students ask the lecturer some questions. A lecture video filmed in the lecture has visual information of the lecturer and students. Users watch the video in lecture archives in order to grasp the situation of the lecture room or facial expression of the lecturer and students. The users can feel an atmosphere of the lecture room when they watch wide-angle shot, and they can see facial expression of the students when they watch zoom-up shot. Especially, filming faces of asking students is very important because the filmed video has information that can not be understood with audio only[1, 2].

In this paper, we aim to detect the location of a speaking student in the lecture room for filming his/her face. We consider the students should not be aware of the detecting system, so it is not allowed to install extra control buttons to indicate that who is the speaker. There is audio and visual based approach[3] to detect a sound source. However, the space of the experiments in that approach is much smaller than that of the lecture room, and it is assumed that there is no noise in the environment.

In this paper, we assume that there is one speaking student who asks the lecturer a question at a time. However, multiple sound sources may be detected because there is noise in real lecture room. So we use a visual based approach based on feature of students at the same time in order to select only one speaking student from the detected sound sources.

The rest of this paper is organized as follows. In Section 2, our sensor-fusion approach is described. In Section 3, a method for detecting the location of a speaking student using an acoustic and a visual based method is described in detail. Experiments are shown in Section 4. Conclusion is described in Section 5.

2. A SENSOR-FUSION APPROACH

Our goal in this paper is to detect the location of a speaking student in the lecture room. In the lecture of face-to-face style, students are sitting on the seats and listen to a lecturer while they watch slides and the white board. The students remain seated. The lecture room is divided into several rectangle regions corresponding to the seats each of which one student may occupy. R_i denotes each region, where i is the identifier of the rectangle region. There are several sound sources in the lecture room. The speaking student is one of

[*] Presently, with Osaka Gas Information System Research Institute Co.,Ltd. Osaka, 550-0023, Japan

the sound sources. Other sound sources may exist because of noise in the real lecture room. In our approach, detecting the location of a speaking student is equivalent to detecting R_i in which the speaking student exists. There should be only one speaker in the lecture room in a certain time period because we deal with a face-to-face lecture.

In order to detect the location of a sound source, we introduce the CSP (Cross-Power Spectrum Phase analysis) algorithm [4, 5, 6] using a microphone pair on the ceiling of the lecture room. A hyperboloid on the surface of which a sound source exists can be estimated by the CSP algorithm. We apply the algorithm to multiple distributed microphone pairs in order to detect the location of a sound source. Generally, multiple sound sources may be detected because there is noise in the real lecture room. Some of the detected sound sources do not represent speaking students in our assumed environment.

We need to select R_{i_0} that is occupied by the speaking student among the candidates detected by our acoustic method. In order to select R_{i_0}, we construct a student model based on his/her visual feature. Image based approaches based on the student model are applied to R_i detected by our acoustic method.

First, we use background subtraction method in order to estimate whether there is an entity in R_i. Estimated one may represent a student, bags or coats etc. Then, students move their head, hand and upper half body when they take notes, listen to the lecturer, and watch a screen and a whiteboard in front of them. So we use an inter-frame subtraction method in order to estimate an active entity in R_i.

We describe details of this sensor-fusion method in the next section.

3. DETECTING THE LOCATION OF A SPEAKER

3.1. The CSP algorithm

In this section, the CSP algorithm is described briefly. The CSP algorithm was designed to estimate a time delay between two microphones. An amplitude of audio signal sensed at microphone k is denoted by $s_k(n), 0 \leq n < N$, where N is the number of the samples. The CSP coefficient of microphone i and j is denoted by $\mathrm{C}_{i,j}(h)$ (Eq.(1),(2)), Delay of Samples (DOS) is denoted by $\delta_{i,j}$ (Eq.(3)), and Delay of Arrivals (DOA) is denoted by $\tau_{i,j}$ (Eq.(4)). F_s indicates sampling rate.

Fourier transform is performed on the two audio signals of microphone pair by Eq.(1). Inverse Fourier transform is performed by Eq.(2). $C_{i,j}(h)$ is calculated by these equations. $\delta_{i,j}$, which has the strongest correlation between the two signals, is calculated by Eq.(3). The function argmax in Eq.(3) returns an argument h which maximizes $C_{i,j}(h)$. $\tau_{i,j}$ is obtained by Eq.(4). A 3D hyperboloid on which a sound source exists can be drawn by $\tau_{i,j}$ of one microphone pair.

$$S_{i,j}(k) = \frac{\left(\sum_{n=0}^{N-1} s_i(n) e^{-i2\pi\frac{k}{N}n}\right)\left(\sum_{n=0}^{N-1} s_j(n) e^{-i2\pi\frac{k}{N}n}\right)^*}{\left|\sum_{n=0}^{N-1} s_i(n) e^{-i2\pi\frac{k}{N}n}\right|\left|\sum_{n=0}^{N-1} s_j(n) e^{-i2\pi\frac{k}{N}n}\right|} \quad (1)$$

$$k = 0, 1, \cdots, N-1$$

$$C_{i,j}(h) = \frac{1}{N}\sum_{k=0}^{N-1} S_{i,j}(k) e^{i2\pi\frac{h}{N}k}, \qquad h = 0, 1, \cdots, N-1 \quad (2)$$

$$\delta_{i,j} = \arg\max_h(C_{i,j}(h)) \quad (3)$$

$$\tau_{i,j} = \frac{\delta_{i,j}}{F_s} \quad (4)$$

3.2. Detecting a sound source using multiple CSP

The CSP algorithm can estimate only 3D hyperboloid that contains the sound source based on the principle of Hough transform. Accordingly, at least two microphone pairs are required in order to estimate the location of a sound source. Height of the sound source from the floor is assumed to be the same of a mouth of students sitting on the seat. In order to make the estimation robust, we use multiple microphone pairs in the lecture room.

1. The CSP algorithm is applied to multiple microphone pairs, and 3D hyperboloids on which the sound source may exist are estimated for each microphone pair.
2. When the hyperboloid of a microphone pair passes within C (cm) from the center of R_i, give one vote to the R_i.
3. After all microphone pairs vote, the regions of which the number of the vote exceeds the threshold Th_{vote} are estimated as sound sources.

We call this process Multiple CSP algorithm (**M-CSP**) in this paper. In this process, multiple sound sources may be detected in the lecture room. Therefore, all regions detected by the acoustic method are not always occupied by students.

Hence we introduce visual based approach, which is not affected by sound noise in order to estimate whether there is a student or not in R_i.

3.3. Detecting students using visual method

An image based approach is used in order to check whether R_i detected in section 3.2 is occupied by a student or not.

In order to avoid occlusion caused by the students themselves, an observation camera with fish-eye lens is installed on the ceiling of the lecture room and looks down at the student area vertically. The camera captures images of the students from the ceiling. Fig.1 shows an example of fish-eye image of the student area. The size of the captured images is 640x480 pixels.

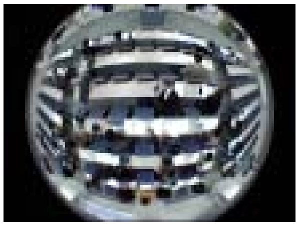

Fig. 1. Fish-eye image of students

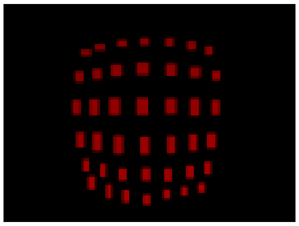

Fig. 2. Mask regions

In order to estimate the existence of a student sitting on the seat in each R_i, the mask regions are used. Fig.2 shows the masks that correspond to the seats inside R_i. $M(R_i)$ denotes a mask of R_i. $num(M(R_i))$ denotes the number of pixels included in $M(R_i)$. $I(p,t)$ denotes a pixel value in the captured image at a time t where p is the location of a pixel.

We use the background subtraction and inter-frame subtraction. Bags and coats of the students are also detected in the background subtraction method, and there is the case that students are not detected by this method when the color of clothes of the student is similar to the color of the seats. Hence, we use the inter-frame subtraction to detect the moving heads and hands of the students.

The background subtraction is performed in order to detect a visual entity from fish-eye image. This process at time t is shown by Eq.(5) and Eq.(6). A pixel value on the background image is denoted by $I_b(p)$. Eq.(5) shows that the location of a pixel has a value 1 if an absolute value of difference between $I_b(p)$ and $I(p,t)$ exceeds Th_b. Then, Eq.(6) shows that $B_i(t)$ is a ratio of pixels which have a value 1 in $M(R_i)$. $B_i(t)$ shows likelihood of existence of the entity in R_i.

$$Bin_b(p,t) = \begin{cases} 1, & \text{for } |I_b(p) - I(p,t)| \geq Th_b \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

$$B_i(t) = \frac{\int_{M(R_i)} Bin_b(p,t)dp}{num(M(R_i))} \tag{6}$$

Inter-frame subtraction of fish-eye video is performed by equations from Eq.(7) to Eq.(12). Eq.(7) shows that the location of a pixel has a value 1 if an absolute value of difference between $I(p,t-1)$ and $I(p,t)$ exceeds Th_m. Then, Eq.(8) shows that $E_m(t)$ is a ratio of pixels which have a value 1 in $M(R_i)$. $E_m(t)$ shows magnitude of motion of a moving entity in R_i at time t.

$E_m(t)$ is not used directly to estimate whether there is a moving entity or not because captured images may include visual noise. When an entity in R_i moves continuously, we consider the R_i is occupied by a student. And an entity in R_i does not move continuously, we consider the R_i is not occupied by a student but occupied by the objects like bags.

In the other cases, it is considered that there is the visual noise in R_i. In order to remove the influence of visual noise, we have introduced functions $J(t)$ and $STEP_e(t)$. In Eq.(9), Th_{E_a} is a threshold to estimate that there is a moving candidate. $STEP_e(t)$ has a positive value $+\alpha$, where $0.0 < \alpha \leq 1.0$ if an entity in R_i moves continuously. On the contrary, $STEP_e(t)$ has a negative value $-\alpha$ if an entity in R_i does not move at all. $M_i(t)$ shows an activity of a moving entity in R_i. If the value of $M_i(t)$ reaches 1.0, the value is kept for a period T_{cont}. After the period, a calculation of $M_i(t)$ begins again. $M_i(t)$ shows that an object which has moved continuously in R_i is a candidate of a student.

$$Bin_m(p,t) = \begin{cases} 1, & \text{for } |\frac{\partial}{\partial t}I(p,t)| \geq Th_m \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

$$E_m(t) = \frac{\int_{M(R_i)} Bin_m(p,t)dp}{num(M(R_i))} \tag{8}$$

$$J(t) = (E_a(t) - Th_{E_a})(E_a(t-1) - Th_{E_a}) \tag{9}$$

$$STEP_e(t) = \begin{cases} +\alpha, & \text{for } J(t) > 0, E_a(t) > Th_{E_a} \\ 0, & \text{for } J(t) \leq 0 \\ -\alpha, & \text{for } J(t) > 0, E_a(t) < Th_{E_a} \end{cases} \tag{10}$$

$$g(t) = \int_0^t STEP_e(t)dt \tag{11}$$

$$M_i(t) = \begin{cases} g(t), & \text{if } 0 \leq g(t) \leq 1 \\ 0, & \text{if } g(t) < 0 \\ 1, & \text{if } g(t) > 1 \end{cases} \tag{12}$$

Both $B_i(t)$ and $M_i(t)$ are averaged in order to estimate whether there is a student in R_i. We call this approach Active Student Detecting method (**ASD**).

3.4. Estimating the location of a speaker

We estimate the location of a speaker by following sensor-fusion based approach. The number of detected regions by **M-CSP** is denoted by N_{CSP}. When sound volume detected by a microphone is small or N_{CSP} is equal to 0, it is considered that there is no speaker. When N_{CSP} is equal to 1, it is considered that the region is occupied by the speaker. When N_{CSP} is more than 1, R_i that meets $\frac{B_i(t)+M_i(t)}{2} \geq Th_{exist}$ is selected. If several R_i meet the equation, we select one R_i that has the highest value of $M_i(t)$.

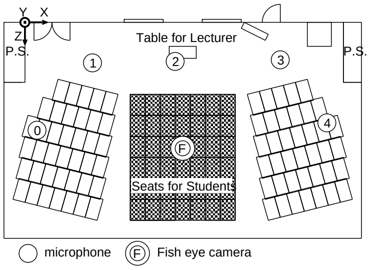

Fig. 3. Lecture Room

Table 1. Result of detecting students

	correct		incorrect	
Students	present	absent	present	absent
ASD output	detected	not detected	not detected	detected
Average	14.1/42 (33.7%)	17.3/42 (41.2%)	5.9/42 (14.0%)	4.7/42 (11.2%)
Total	74.9%		25.1%	

4. EXPERIMENT

4.1. Lecture room

We have implemented our method in the lecture room shown in Fig.3. The size of the room is 15.5(m) × 9.6(m), and the height is 3.2(m). Five microphone pairs are used in M-CSP. They are $\{(0,1),(1,2),(2,3),(3,4),(1,3)\}$ in Fig.3. Students are in the center area (student area) of the lecture room in this experiment (gray area in Fig.3). The area has $6 \times 7 = 42$ seats. Hence the total number of R_i is 42.

4.2. Result of Detecting students by ASD

Table1 shows the result of detecting students by the ASD method described in 3.3. 20 students exist during the lecture in the student area and we take a video for 5 minutes using a fish-eye camera. The ASD method was applied several times in a second. 74.9% accuracy about the presence of students has been achieved on average.

4.3. Result of Detecting the location of a speaker

We have compared the result of M-CSP with the result of M-CSP + ASD for two short lectures. Recall ratios when using M-CSP + ASD are 57.1% at sample lecture 1 and 65.6% at sample lecture 2. The recall ratios when using only M-CSP are equal to ones of M-CSP + ASD because ASD is applied only to the regions that M-CSP detects, and ASD always select only a sound source as the speaker. Table2 shows precision ratios of these two methods. Denominator of fractions in this table is the number of times of the student speaking. This table shows that precision ratios are improved to 100.0(%) by M-CSP + ASD in our short lectures. The precision ratio was improved about 20%. The result indicates that our method may miss the speaker, while it never films an object which is not a student.

Table 2. Precision ratio of detecting location of speaker

Method	Sample Lecture 1	Sample Lecture 2
M-CSP only	80.0% (8/10)	87.0% (20/23)
M-CSP + ASD	100.0% (8/8)	100.0% (20/20)

5. CONCLUSION

In this paper, we proposed the method for detecting the location of a speaker in the lecture room. In order to estimate the location of the speaker, we developed M-CSP method to detect sound sources and ASD method which utilizes visual processing approach in order to improve the precision rate. The result of our preliminary experiment shows that it has high performance especially for the precision ratio. In the future, it is necessary to improve the recall ratio so that we can construct more reliable filming system.

6. REFERENCES

[1] Yoshinari KAMEDA, Kentaro ISHIZUKA, and Michihiko MINOH, "A live video imaging method for capturing presentation information in distance learning," in *IEEE International Conference on Multimedia and Expo*, 2000, vol. 3, pp. 1237–1240.

[2] Keisuke Yagi, Yoshinari Kameda, Motonori Nakamura, Michihiko Minoh, and Maha Ashour-Abdalla, "A novel distance learning system for the tide project," in *Proceedings of ICCE/ICCAI 2000*, 2000, vol. 2, pp. 1166–1169.

[3] Harsh Nanda Dmitry Zotkin, Ramani Duraiswami and Larry S.Davis, "Multimodal tracking for smart videoconferencing," in *Proceedings of ICME 2001*, 2001, pp. 37–40.

[4] M.Brandstein and H.Silverman J.Adcock, "A closed-form method for finding source locations from microphone-array time-delay estimates," in *Proc. ICASSP95*, 1995, pp. 3019 – 3022.

[5] P.Svaizer M.Omologo, "Acoustic source locationin niosy and reverberant environment using csp analysis," in *Proc. ICASSP96*, 1996, pp. 921 – 924.

[6] M.Omologo and P.Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," *IEEE Trans. on Speech and Audio Processing*, pp. 288 – 292, 1997.