Attention to clapping - A direct method for detecting sound source from video and audio –

Tetsushi Ikeda Hiroshi Ishiguro Minoru Asada
Dept. of Adaptive Machine Systems
Graduate School of Engineering, Osaka University
Suita, Osaka 565-0871 JAPAN
ikeda@ed.ams.eng.osaka-u.ac.jp {ishiguro,asada}@ams.eng.osaka-u.ac.jp

Abstract

The research approaches utilizing ubiquitous sensors to support human activities have become of major interest lately. One of the required features of the ubiquitous sensor system is paying its attention to our signals, such as clapping hands and uttering keywords. To detect and localize these signs, it is useful to fuse visual and audio information. The sensor fusion in previous works is performed in the task-level layer through individual representations of the sensors. Therefore, it does not provide new information by fusing sensors. This paper proposes another method that fuses sensory signals based on mutual information maximization in the signal-level layer. The fused signal provides us new information that cannot be obtained from individual sensors. As an example, this paper shows two experimental results of a sound source localization by audio-visual fusion.

1. Introduction

In proportion to extensive diffusion of many kinds of sensors in our surroundings, the concern with utilizing them to support human activities has been growing. The system needs to recognize many kinds of human actions through various sensors. Recent research activities named "intelligent room" [3], "smart room" [4], "robotic room" [6] and "perceptual information infrastructure" [7] are working on the recognition with ubiquitous sensors embedded in the environment.

Integrating various sensors, such as cameras, microphones, touch sensors, and so on, leads to many kinds of application. One of the required features of the ubiquitous sensor system is turning its attention to signs from us. When we give the sign to the system, we want it to watch us carefully for further request. Common methods of giving signs include clapping hands and uttering some words, for example. To localize the source of such signs, it is useful to integrate visual and audio information.

In this paper, we propose a novel method of sound source localization in video images with one camera and one microphone. There are two approaches in the sensor integration: one is to use each sensor in a limited situation and the other is to integrate sensory signals directly. Let us consider a task to recognize a speaking person in an environment monitored by a ubiquitous sensor system with cameras and microphones. The former approach individually uses the cameras and microphones. The system detects positions of humans by analyzing visual data taken by the cameras in the environment, and simultaneously it receives the voice with the microphones and detects the location of the sound source. Then the system identifies a speaking person by integrating voice and audition based on the representations of the positions or directions given by both of the sensors.

The latter approach solves the task more simply. Humans can recognize a speaking person by voice and lip motions. That is, a human can identify the talking person in vision and audition by directly fusing the sensory information. The latter approach is similar to such the human ability. That is the system directly finds corresponding information between different sensor signals and provides new information that directly gives the solution of the task. This paper proposes a novel method following the latter approach, which is signal-level sensor.

Several studies on signal-level sensor fusion have been reported so far. Becker and Hinton [1][2] proposed to train neural networks by using a criterion that maximizes mutual information among output from the networks. They showed that the networks extract the common information shared by inputs of the networks. Cutler and Davis [5] localized a speaker who utters a specific word by fusing visual and audio data with a TDNN (Time-Delay Neural Network). In this work, the networks need to be trained for each word, since the relation between video and audio signals is different for each word. Hershey et al. [9] and Fisher et al. [8] also localized a speaker in the image by computing mutual information between video and audio signals.

A problem of these approaches is that the statistical model for sensor signal analysis is not adaptive to dynamical changes of the scene. That is, they assume that the speaker does not move. As a method to solve the problem, this paper proposes a simple method to track the sound source prior to computation of mutual information between video and audio.

2. Measuring similarity between audio and video signals

2.1 Mutual information between sensors

Let $x(t)$ be the time sequence of sensory data from sensor X, and $y(t)$ from sensor Y, respectively. Mutual information between $x(t)$ and $y(t)$ is represented as

$$I(x;y) = H(x) + H(y) - H(x,y) \qquad (1)$$

where $H(x)$ is entropy of $x(t)$ and $H(x,y)$ is mutual entropy between $x(t)$ and $y(t)$. They are defined as:

$$H(x) = -\sum p(x(t)) \log p(x(t))$$

$$H(x,y) = -\sum_{t_x,t_y} p(x(t_x), y(t_y)) \log p(x(t_x), y(t_y))$$

Here, mutual information I is computed with a fixed-length time window whose length is T.

Now, let us assume that $x(t)$ and $y(t)$ are jointly Gaussian [9]. The mutual information can be replaced with

$$\frac{1}{2}\log\frac{1}{1-\rho(x,y)^2} \qquad (2)$$

where $\rho(x,y)$ is correlation function between $x(t)$ and $y(t)$.

2.2 Computing mutual information in dynamically changing environments

Let us assume $x(t)$ and $y(t)$ are video and audio signal respectively. In previous approaches [8][9], $x(t)$ is intensity of a pixel or a region in the imaging sensor. It is assumed that the target is at a fixed position in the observed image. In general, however, the relations among sensors change dynamically. Suppose a person walking in an environment where a sensor network observes. The target that each sensor observes often changes as the person moves. Since the above assumption does not hold, localization fails in such a situation (Figure 1).

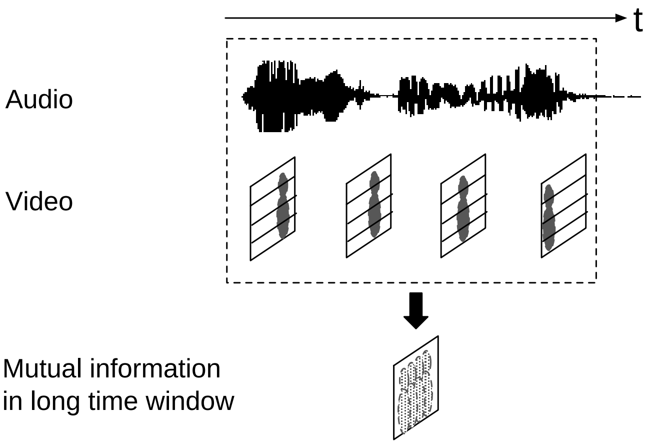


Figure 1. Previous methods

For solving this problem, our approach proposed in this paper is to track the center of the person, not all movements of parts of the body [10]. We detect the objects in the environment by background subtraction. Prior to fusing video and audio, the trajectory of each detected object is acquired. By computing mutual information between video and audio along the acquired trajectory, we obtain stable relation among sensors (Figure 2).

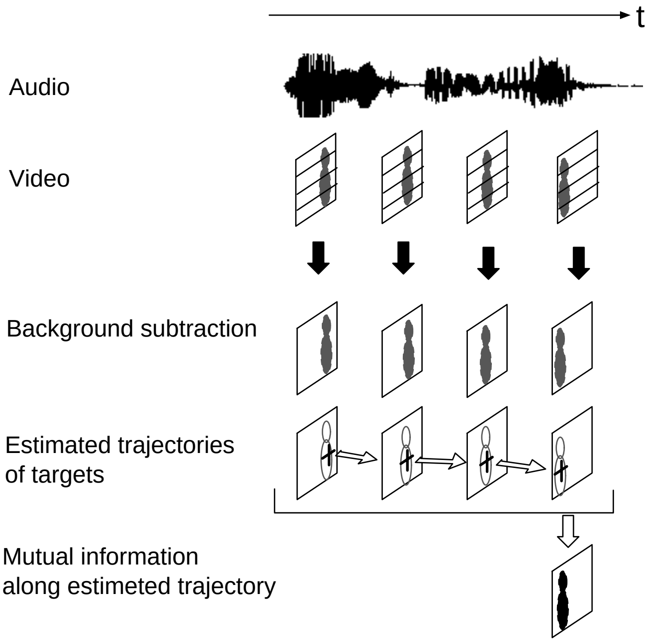


Figure 2. Process flow of the proposed method

3. Experiments

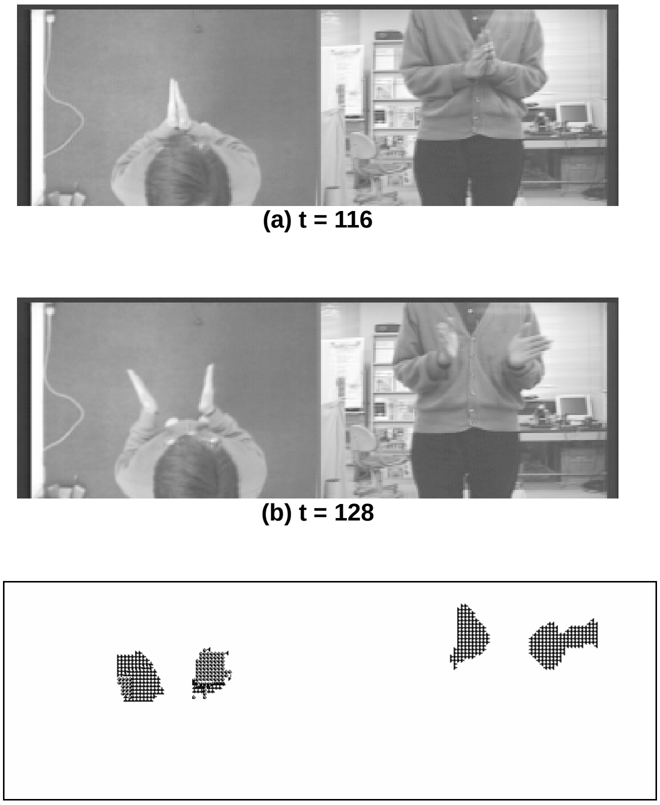
To confirm the effectiveness of the proposed method, two experiments are performed. In both examples, video signals are sampled at 30 frames/second, and audio signals are sampled at 16 kHz, and the average energy in each video frame is computed with a Hanning window and it is referred as $y(t)$.

3.1 Clapping hands

First example is detecting clapping hands. Two cameras and one microphone observe a person clapping hands. One camera observes from the top, and the other from the side. Example signals are shown in Figure 3 and Figure 4.

In this example, the person is standing at a fixed place and no tracking is needed. The mutual information between magnitude of optical flow computed from consecutive images and audio signal is computed according to equation (2). The result is shown in Figure 5. Darker pixels indicate higher mutual information.

Locations of hands are clearly located in the image.



(a) t = 116

(b) t = 128

(c) optical flow

Figure 3. Images of clapping hands

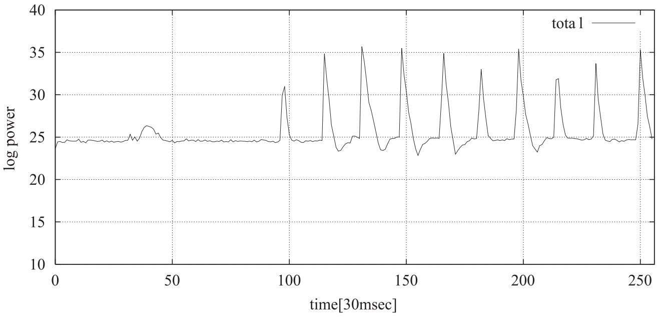


Figure 4. Audio signal of clapping hands

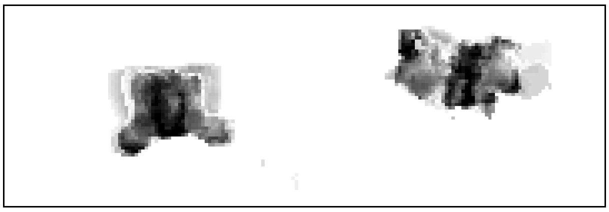


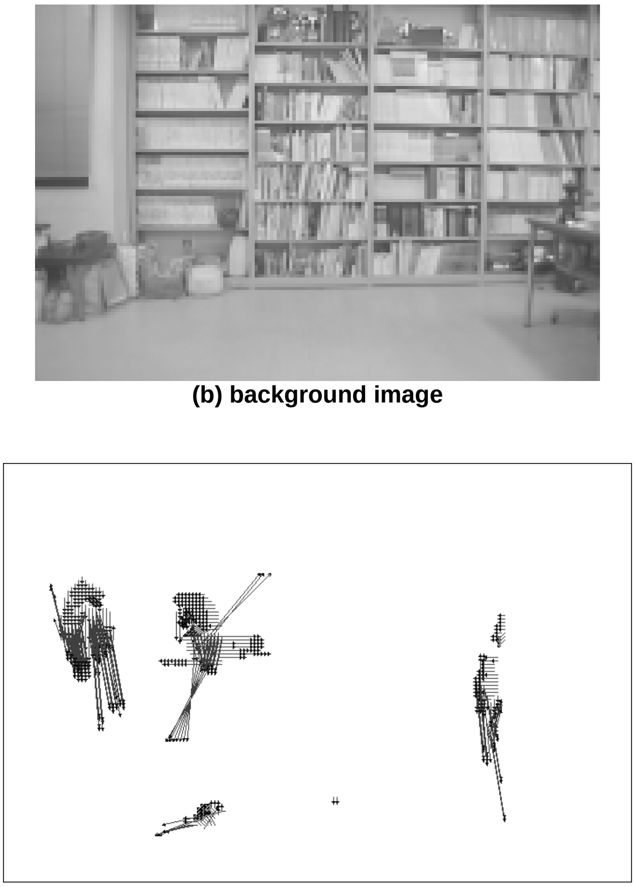
Figure 5. Result of localizing clapping hands

3.2 Walker

Second example is detecting a walking person.

One camera and one microphone observe a walking person and the other person. Example signals are shown in Figure 6 and Figure 9. The mutual information between magnitude of optical flow and audio signal is computed according to equation (2). The result is shown in Figure 7. Locations of the walker are clearly located in the image.

(a) t = 121



(b) background image

(c) optical flow

Figure 6. Images of walker

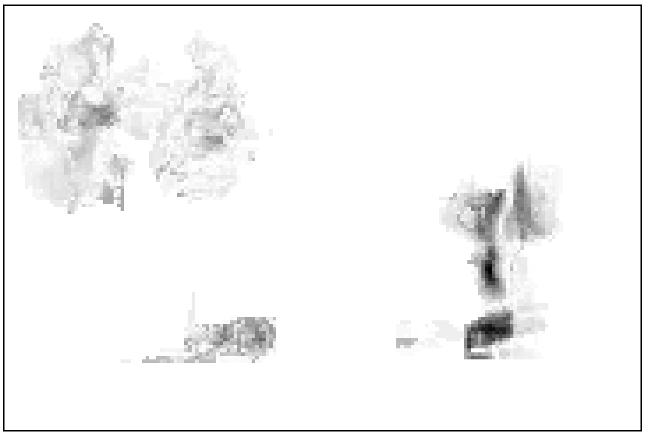
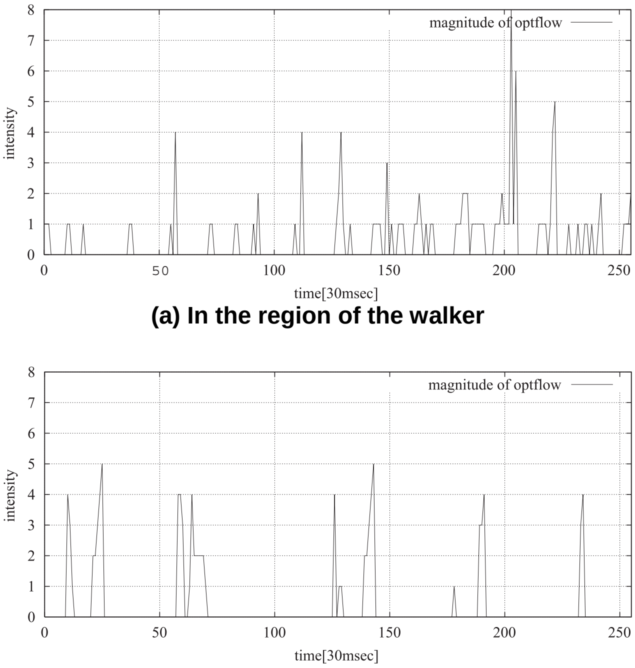


Figure 7. Result of locating the walker

Figure 8 shows magnitude of optical flows along trajectories. Figure 8 (a) is a video signal in the regions for the walker, and Figure 8 (b) is a video signal in the region for the other person. It is clear the signal in Figure 8 (a) is highly correlated with simultaneous audio signal (Figure 9).



(a) In the region of the walker

(b) In the region of the other person

Figure 8. Sample of the video signal

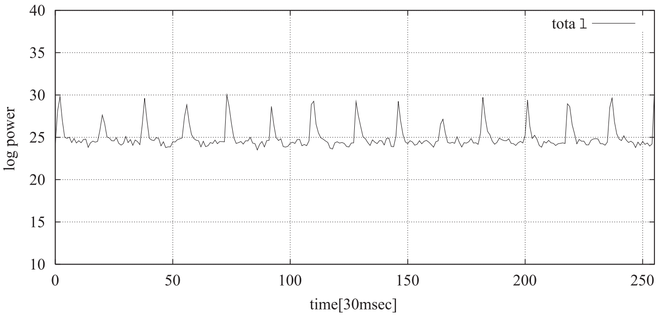


Figure 9. Audio signal

4. Conclusion

This paper has proposed a method that fuses sensory signals at the signal level based on mutual information maximization. The proposed method adapts to the dynamical changes of the relation between sensors by tracking centroids of the regions detected by background subtraction.

Two experimental results of sound source localization are carried out to confirm the effectiveness of the proposed method. In the latter experiments, the proposed method is applied when a sound source is moving. In both result, a sound source is clearly localized. The results show the effectiveness of our method.

Our next step is to deal with other sensors and to extend the method for applying to more complicated environments where many talking persons exist.

References

[1] S. Becker. Mutual Information Maximization: Models of Cortical Self-Organization. *Network: Computation in Neural Systems*, 7(1), 1996.

[2] S. Becker and G. E. Hinton. A self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355(9):161-163, 1992.

[3] R. A. Brooks. Intelligent Room Project. In *Proceedings of the Second International Cognitive Technology Conference*,1997.

[4] A. Pentland. Smart rooms. *Scientific American*, 274(4):68-76, 1996.

[5] R. Cutler and L. Davis. Look who's talking: Speaker detection using video and audio correlation. *IEEE International Conference on Multimedia and Expo (ICME'00)*, 2000.

[6] T. Mori and T. Sato. Robotic Room: Its concept and realization. *Robotics and Autonomous Systems*, 28(9):141-148, 1999.

[7] H. Ishiguro. Distributed vision system: A perceptual information infrastructure for robot navigation. *Proceedings of International Joint Conference on Artificial Intelligence*, 36-41, 1997.

[8] J. W. Fisher III, T. Darrell, W. T. Freeman, and P. Viola. Learning Joint Statistical Models for Audio-Visual Fusion and Segregation. *Advances in Neural Information Processing Systems*, 2000.

[9] J. Hershey, H. Ishiguro, and J. R. Movellan. Audio Vision: Using Audio-Visual Synchrony to Locate Sounds. *Proceedings of Neural Information Processing Systems (NIPS'99)*, 1999

[10] T. Ikeda, H. Ishiguro, and M. Asada. Adaptive Fusion of Sensor Signals based on Mutual Information Maximization. *IEEE International Conference on Robotics and Automation (accepted)*, 2003