Adaptive Fusion of Sensor Signals based on Mutual Information Maximization

Tetsushi Ikeda Hiroshi Ishiguro Minoru Asada
Dept. of Adaptive Machine Systems,
Graduate School of Engineering, Osaka University
ikeda@er.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. Sensor fusion is one of the fundamental issues to develop such intelligent environments. 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 also shows experimental results in an audio-visual fusion task.**

I. Introduction

As the Internet develops, the information systems supporting human daily activities is becoming more important. At the same time, several new research issues have become clear. The major problem of the current information infrastructure is in the perceptual function. In order to support human activities, the system needs to recognize them through various sensors.

Recent research activities named "intelligent room"[3],"smart room"[9], "robotic room"[8] and "perceptual information infrastructure"[7] are working on the recognition with ubiquitous sensors embedded in the environment.

One of the features of the ubiquitous sensor system is to use various sensors, such as cameras, microphones, touch sensors, and so on. Therefore, it is important to integrate/fuse such sensors and generate more robust and more task-oriented information.

There are two approaches in the sensor fusion: one is to use each sensor in a limited situation and the other is to fuse 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 fusing 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.

Before proposing our idea, let us briefly review previous works. 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 maximize 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[4] 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.[6] and Fisher et al.[5] 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 sound source does not move. In order to apply these sensor fusion methods to a dynamic environment, the system needs to deal with temporal changes of the model. 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.

II. Acquisition of Relation between Sensors based on Mutual Information

A. 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

$$I(x;y) = H(x) + H(y) - H(x,y) \tag{1}$$

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:

$$H(x) = -\sum_{t} p(x(t)) \ \log \ p(x(t))$$

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

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[6]. The mutual information can be replaced with

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

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

B. Computing mutual information in dynamically changing environments

If the relation between sensors do not change, the signals from sensors can be fused by the above method. Let us assume $x(t)$ and $y(t)$ are video and audio signal respectively. In previous approaches[5][6], $x(t)$ 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. The mutual information among sensors is acquired with a long time interval and the location of speaker is detected in the image.

In general, however, the relation among sensors change dynamically. Suppose a person walking in an environment where a sensor network observes. The target that each sensor observes often change as the person moves. Since the above assumption does not hold, localization fails in such a situation(Fig. 1).

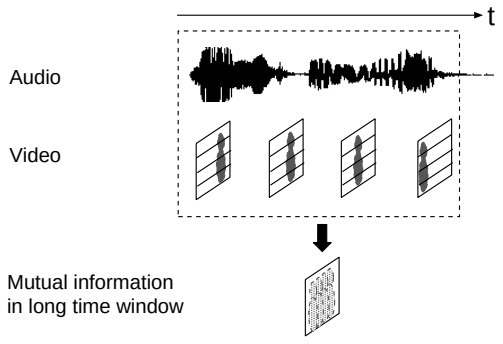


Fig. 1. Previous methods

C. Gaze control for observing moving targets

Ideally, it is required to estimate movements of all parts of targets to adapt to the changes of relation between sensors. By computing mutual information between the motion vector of each part and audio signal, the sound source is located based on this perfect tracking.

However, it is hard to estimate movements of targets precisely. 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. As shown in the next section, 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 (Fig. 2).

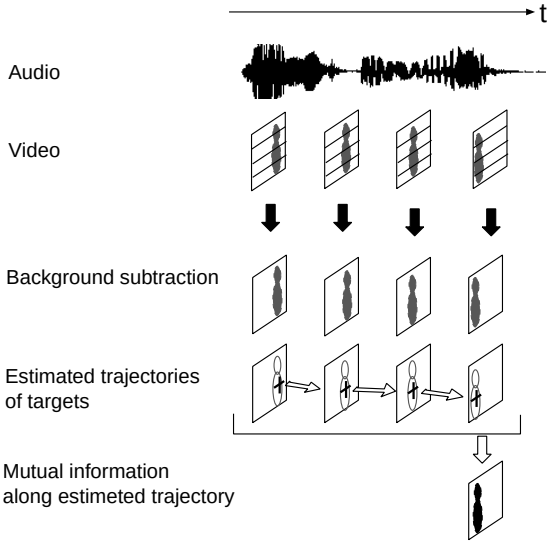


Fig. 2. Process flow of the proposed method

III. Fusion of Vision and Audition

As an example, we focus on sensor fusion between vision and audition and consider locating a walking person in video images as a task. One camera and one microphone are used. The camera observes moving legs and the microphone monitors the sound of footsteps. The relation between video and audio signals is obtained by computing mutual information between these signals.

The video signal is sampled at 30 frames/second, the image size is 160x120, and the intensity of each pixel is referred as $x(t)$ in equation (2). The audio signal is sampled at 16 kHz, and the average energy in each video frame is computed with a Hanning window and it is referred as $y(t)$. Fig. 3 and Fig. 4 shows samples of the video and audio signals, respectively.

(a) frame 232

(b) frame 511

(c) background

Fig. 3. Example images

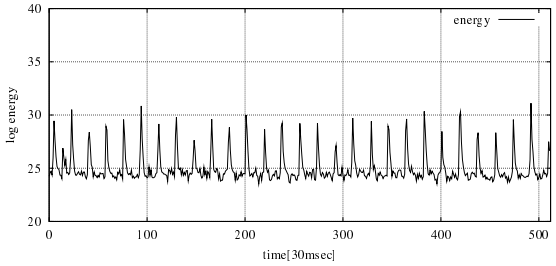


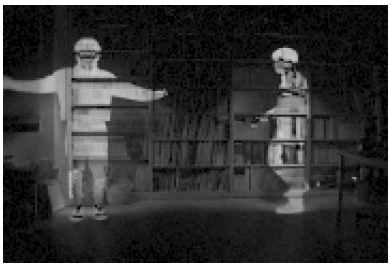
Fig. 4. Example audio signal

A. The computing process

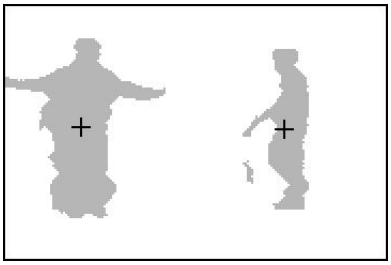
Step 1. Background subtraction and extraction of targets

First, we perform background subtraction and detect moving regions. Then perform a dilation operation two times. Each pixel in the image is substituted with the highest value in 4 nearest neighbors. Then, we perform binalization for the acquired image and extract connected regions and their centroids by labeling.

Figs. 5 shows experimental results of this step. The cross mark in Fig. 5(b) indicates the centroid of the detected region.



(a) Result of background subtraction



(b) Extracted regions and their centroids

Fig. 5. Processing results at step1

Step 2. Find correspondences between regions detected in consecutive images

In general, a few regions are extracted in step 1. These regions correspond to targets in environment that includes a sound source. To extract each sequence of the target, this step finds correspondence of the centroids of regions between consecutive images.

As a result, sequences of regions are extracted. Each sequence indicates a trajectory of a target moving in the environment.

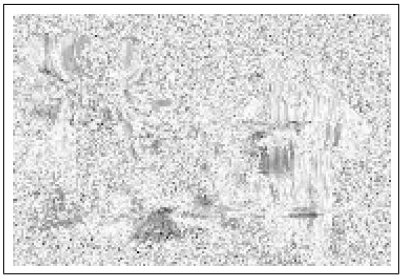
Step 3. Estimate the mutual information

Along each sequence of regions acquired in previous step, we compute mutual information along the detected trajectory in a constant time window (this window length is described as T). All detected regions in each target are aligned to overlap all of the centroid. The mutual information can be used for verifying the results of the step 2. It is possible to feed this result back to the determination of the window length. This is one of our future works.

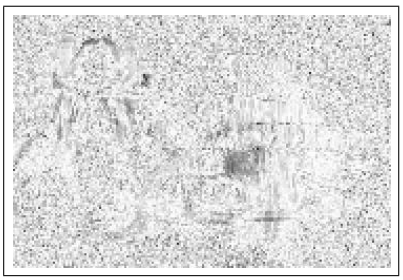
B. Results

Figs. 6 shows experimental results without tracking. Video signal is a sequence of observed intensity at a fixed position. The darker pixels indicate higher mutual information in the figure. The time window T was set to 256[frames] in all experiments. Figs. 6 (a) and (b) show the result at frame 64 and 256, respectively. The results have not included any remarkably darker regions. That means the sound source localization is failed.

Figs. 7 shows results by proposed method. Video signal is a sequence of observed intensity in a moving region. By computing mutual information along the extracted trajectory, regions that correspond to walker's leg have not been clearly detected.



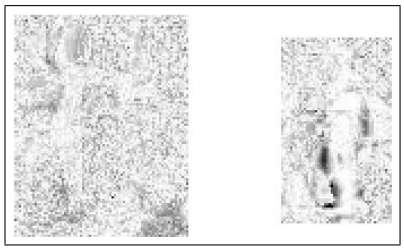
(a) at frame 64



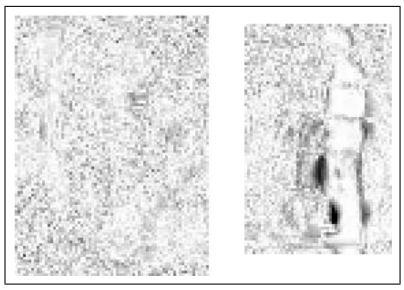
(b) frame 256

Fig. 6. Result (without tracking)

Fig. 8 shows samples of the video signals along trajectories. Fig. 8(a) is a video signal in the regions for the walker, and Fig. 8(b) is a video signal in the region for the other person. It is clear the signal in Fig. 8(a) is highly correlated with audio signal (Fig. 9).



(a) at frame 64



(b) frame 256

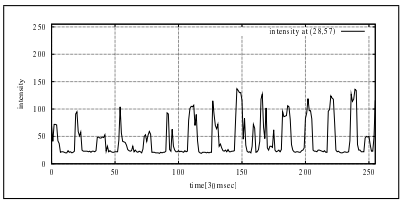
Fig. 7. Result (with tracking)

IV. Conclusion

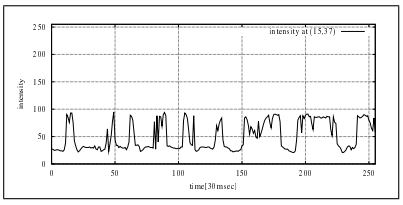
This paper has proposed a method that fuses sensory signals at the signal level based on mutual information maximization. The proposed method adapts to the dynamical changes of the relation between sensors by tracking centroids of the regions detected by background subtraction.

To confirm the effectiveness of the proposed method, we have applied it to track a walking person. The experimental result has shown that the proposed method can fuse audio signal and video signals and localize the signal source under a condition where the relation between sensors is dynamically changing.

In this paper, we have focused on audio and vision. But, it is possible to apply this method to other sensors used in a perceptual information infrastructure. 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.



(a) In the region of the walker



(b) In the region of the other person

Fig. 8. Sample of the video signal

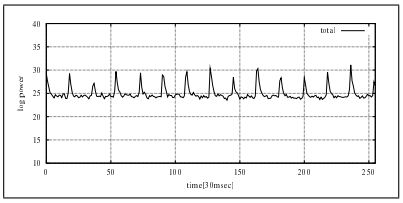


Fig. 9. Audio signal

V. REFERENCES

[1] S. Becker. Mutual information maxmization: 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 *Proc. of the Second International Cognitive Technology Conference*, 1997.

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

[5] J. W. Fisher III, T. Darrell, W. T. Freeman, and P. Viola. Learning joint statistical models for audio-visual fusion and segregation. In *Advances in Neural Information Processing Systems*, 2000.

[6] J. Hershey, H. Ishiguro, and J. R. Movellan. Audio vision: Using audio-visual synchrony to locate sounds. In *Proc. of Neural Information Processing Systems (NIPS'99)*, 1999.

[7] H. Ishiguro. Distributed vision system: A perceptual information infrastr ucture for robot navigation. In *Proc. Int. Joint Conf. Artificial Intelligence*, pp. 36–41, 1997.

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

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