Sensor Fusion as Optimization: Maximizing Mutual Information between Sensory Signals

Tetsushi Ikeda Hiroshi Ishiguro Minoru Asada
Dept. of Adaptive Machine Systems
Graduate School of Engineering
Osaka University
ikeda@ed.ams.eng.osaka-u.ac.jp

Abstract

Sensor fusion is one of the fundamental issues to develop intelligent systems that recognize the scene around them precisely and robustly. Previous approaches of sensor fusion combined different kind of sensors after feature extraction and abstraction ("task-level fusion"). This paper proposes a new approach that combines sensory signals from different kind of sensors before abstraction ("signal-level fusion"). By formalizing sensory fusion as an optimization that maximizes mutual information between sensory signals, a target in a changing scene is detected by a heuristic search algorithm. As an example, experimental results of a sound source detection with one video camera and one microphone are shown.

1. Introduction

To recognize various objects precisely and robustly, much research has been performed on sensor fusion. The recognition process in sensor fusion consists of observation and integration. Each sensor creates a representation of different aspect of the objects, and an integration process reconstructs a total image of an object from each representation. One important question in sensor fusion is what kind of representation is used in the integration process. From this viewpoint, two types of sensor fusion have so far been identified. Extracting features from each sensor independently and integrating then after abstraction is one type, and the other is fusing sensors directly before abstraction. An example of the former approach is the integration of the detected location of the sound source in the video and audio. The location is detected separately by vision and audition, and then the precise position is computed by integrating the result. We call this type of sensor fusion "task-level fusion" since integration is performed based on the extracted feature after abstracting sensory signals.

In the latter approach, sensory signal from different kinds of sensors are integrated directly based on a statistical method. This approach to sensor fusion has recently been proposed. Let us review previous work. Becker and Hinton [1][2] proposed to train neural networks by using a criterion that maximizes mutual information among the output from the networks. Cutler and Davis [3] localized a speaker who utters a specific word by combining signals of one video camera and one microphone with a TDNN (Time-Delay Neural Network). Hershey et al. [5] and Fisher et al. [4] also localized a speaker in the image by computing mutual information between video and audio signals. We call this type of sensor fusion "signal-level fusion" since we can make use of the intermodal relation among sensory signals by integrating sensory signals in the early stage of processing.

However, this new approach of sensor fusion cannot be applied to the dynamic changes in the scene. By using these methods, it is difficult to detect detect a moving sound source with cameras and microphones. The reason is that the relation between sensors is assumed to be static in this approach.

In this paper, we propose to detect and track a sound source simultaneously based on the criterion of mutual information maximization. The problem of detection and tracking a sound source is formalized as an optimization problem to find the path that maximizes mutual information between video and audio signal. In section 2, the sensor fusion algorithm based on mutual information maximization is described. In section 3, we applied the algorithm to the problem of sound source localization by combining audio and visual signal. In section 4, experimental results are shown.

2. Sensor fusion based on mutual information maximization

2.1. Computing mutual information along the detected path of the target

In [5][4], the statistical relation between video and audio is assumed to be static during the computation of a statistical measure between signals. Thus, these methods cannot be applied to the environment where targets are moving.

To cope with this problem, we proposed to track the target candidates before computing mutual information [6]. This method consists of two stages. In the first stage, positions of the target candidates are computed based on background subtraction in the video. In the second stage, mutual information is computed between the audio and video signal along the detected path of each candidate.

Consider the tracking of a sound source by a camera and a microphone. Figure 2 shows the computation of the mutual information between the video signal and audio signal. In Figure 2, each ball represents an observation by one sensor and the horizontal axis represents time. Video images are shown as one-dimensional array. The line between ball in the video and audio represents signals from the same information source. Mutual information is computed between signals that are connected by arrows.

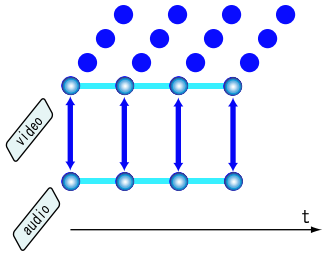

Figure 1. The relation between the video and audio signal when the target does not move.

Since previous methods compute mutual information between the audio signal and the intensity of a pixel at a fixed point, they fails to capture the relation between the signals when the target moves.

By computing mutual information along the detected path of the target, we succeeded to detect the sound source when the target is moving (Figure 3).

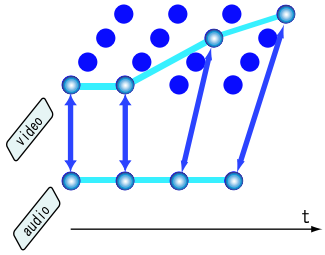

Figure 2. Computing mutual information along the path of the target.

2.2. Detection and tracking based on mutual information maximization

In this two-stage approach, the tracking process and the sensor fusion process are separated and the tracking process is performed only by image processing. In this paper, these stages are integrated into one process. We propose to detect and track the sound source simultaneously based on the criterion of mutual information maximization.

When the path of a moving sound source is unknown, the problem of detecting and tracking is regarded as the optimization problem of finding the path that maximize mutual information between the video and audio signal. Thus, a heuristic search algorithm computes the path of the sound source using mutual information as an evaluation function (Figure 4). There are many possible paths in the image sequence, and the path that maximizes mutual information is selected.

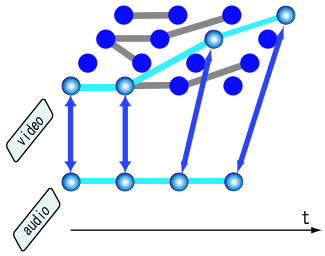

Figure 3. Detection and tracking the sound source based on mutual information maximization.

However, this search is almost breadth first at an early stage since mutual information is a poor heuristic function when the length of the input data is short. So we assume the speed of the target to be piecewise constant (Figure 5). By limiting possible paths of the target, the search process effectively finds the path that maximizes mutual information.

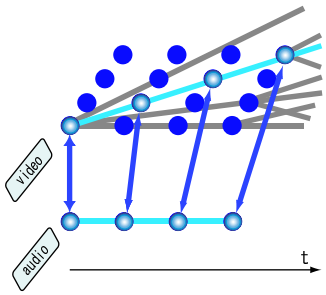

Figure 4. The speed of the target is assumed to be piecewise static.

3. Sound source localization by direct fusion of sensory signals

3.1. Computing mutual information between sensory signals

Let $A(t)$ be an audio signal, and $V(t)$ a video signal. Mutual information between $A(t)$ and $V(t)$ is represented as

$$I(A;V) = H(A) + H(V) - H(A,V) \tag{1}$$

where $H(A)$ is entropy of $A(t)$, and $H(A,V)$ is mutual entropy between $A(t)$ and $V(t)$. Here, mutual information I is computed with a fixed-length time window whose length is T.

Now, let us assume that $A(t)$ and $V(t)$ are jointly Gaussian as in [5]. The mutual information can be replaced with

$$\frac{1}{2}\log\frac{1}{1-\rho(A,V)^2} \tag{2}$$

where $\rho(A,V)$ is the correlation function between $A(t)$ and $V(t)$.

3.2. Computing mutual information along the path of the target

When the sound source moves, mutual information is computed along the path of the sound source. Now video signal V depends on time and position, and mutual information is computed according to (1) and the following formula:

$$H(V) = -\sum_t p(V(t,x(t)))\ \log\ p(V(t,x(t))), \tag{3}$$

$$H(A,V) = -\sum_t p(A(t),V(t,x(t)))\ \log\ p(A(t),V(t,x(t))), \tag{4}$$

where $x(t)$ is the position of the sound source.

3.3. Heuristic search algorithm based on mutual information maximization

To detect and track the sound source, we apply a beam search algorithm. During the search process, a list of hypothesis is updated. One hypothesis represents a possible path of the target. The number of hypotheses should be less than GLOBAL_BEAM_WIDTH.

1. Initialize hypothesis list LIST with possible positions of the target in the image.
2. For each H in LIST, create a new set of hypotheses whose positions are near to H. Substitute LIST with all of created sets of hypotheses.
3. For each H in LIST, move H by the same amount of distance as in the step 2. This step is repeated several times.
4. For each H in LIST, compute the mutual information between audio and video along the path of H.
5. Sort LIST by the computed mutual information. Select GLOBAL_BEAM_WIDTH hypotheses with high mutual information from the LIST.
6. Goto 2.

By introducing step 3, the velocity of each target is piecewise constant. All hypotheses in list are replaced in step 2 and 3. When the number of hypotheses becomes larger than a threshold (GLOBAL_BEAM_WIDTH), only hypotheses with high mutual information are selected.

4. Experiment

To confirm the effectiveness of the proposed method, we apply the method to a sound source localization problem with one microphone and one video camera. The video signal is sampled at 30 frames/second, and the image size is 160x120. $V(t,x)$ in (3), (4) is the intensity of the pixel at time t and position x. The audio signal is sampled at 16 kHz, and the average energy in each video frame is computed with a Hanning window. $A(t)$ is the average energy of the audio signal. Figure 5 shows examples of the video and audio signals, respectively.

Figures 6 and 7 show the process of the search algorithm when the initial hypothesis is the position of each person. Graphs in Figures 6,7 show the paths of the five best hypotheses at frame 64, 128, 192, 256, respectively. The computed intensity of the mutual information and average mutual information of the best hypothesis are also shown on the right of each graph. Mutual information is computed from the first frame to the last frame in each graph. The moving path of the sound source is correctly tracked by the criterion of the mutual information maximization.

5. Conclusion

This paper has proposed a novel sensor fusion method at the signal level. By fusing different kinds of sensors at the

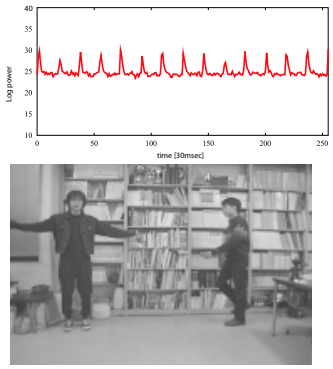

Figure 5. Example of audio and video signal.

signal level, multimodal information would have been lost in a abstraction process can be effectively used. We applied this signal level sensor fusion method to detect and track a sound source.

The problem of detection and tracking is regarded as an optimization problem to find the path that maximizes mutual information between the video and audio signal. This paper has proposed to solve this problem by a heuristic search algorithm and mutual information is used as the heuristic function.

In the experiment, the search process finds the path that maximizes mutual information between the audio and video signal. That path corresponds to the true path of the sound source.

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

[4] 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.

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

[6] T. Ikeda, H. Ishiguro, and M. Asada. Attention to clapping - a direct method for detecting sound source from video and audio -. In *Proceedings of IEEE Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI2003)*, pages 264–268, Jul. 2003.

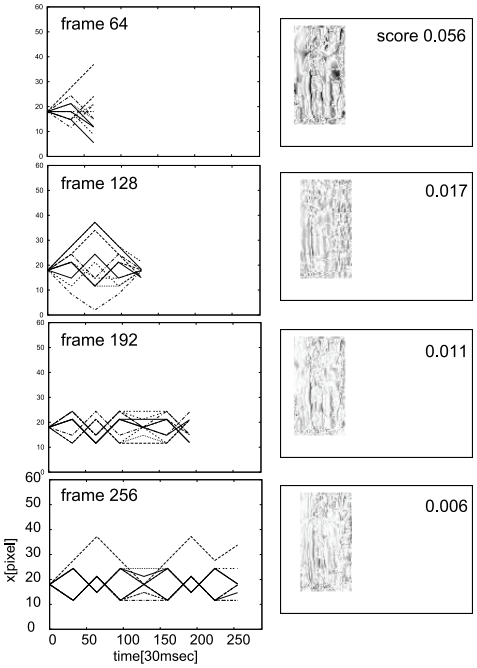

Figure 6. Process of the search when the initial region is set to the left person.

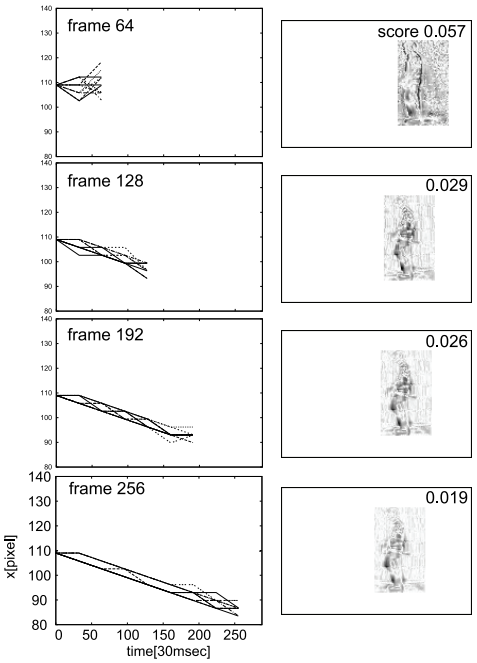

Figure 7. Process of the search when the intial region is set to the right person (correct signal source).