Doctoral Dissertation

Studies on Sensor Integration Based on Signal Level Correlation

Tetsushi Ikeda

January 2014

Graduate School of Engineering,
Osaka University

Abstract

For artificial systems that behave in response to the external world, we must use sensors that observe its state. Many kinds of sensors have been developed to describe and understand outside scenes and objects, and much work has been conducted on signal processing and pattern recognition. Integrating many kinds of sensors and generating descriptions of scenes and objects are fundamental problems in this research area.

The objective of this research is to propose a new sensor integration method based on signal correlation, which appears in multimodal observations using different kinds of sensors. Since previous methods associate observations in a common position coordinate, applying them to sensors that do not directly measure positions is difficult. To associate the observations in different kinds of sensors, we focus on signal correlation, which is a signal structure that appears when localized multiple sensors observe a common scene. Although the observed physical quantity is different, the changes in the scene result in correlated changes in the observed signals. By evaluating the signal correlation among multimodal observations, our method can associate observations without measuring a representation in a common coordinate and be applied to integrate various sensors. Since our method focuses on signals at the lower level of abstraction before computing the higher level features and performing pattern recognition, we call it *signal level integration*.

The relationship among sensory signals is also important in the area of media conversion that converts signals from one modality to another. In situations when we canft use specific modality for communication and presentation, it is effective to use a different modality in a complementary manner by media conversion. In previous media conversion methods, signals in different media are associated and converted in a common symbolic coordinate such as recognized patterns and words that describe the impression of signals. However, much information is lost when we represent signals in one medium in symbols. By converting signals in one medium into another by keeping their signal correlation, many features in the original medium are converted into another medium.

However, simply computing the correlation function does not extract clear and stable relationships among multimodal observations. When objects move in a scene, it is difficult to keep their multimodal correlations since each sensor only observes its local area. When integrating observations in binary representations, we need to design a suitable method to

compute the correlation. Since the observations are not always stable, we must consider the instability in computing the correlations.

To achieve this goal, we expand the previous signal level integration method in three points. First, we expand it so that it associates observations when the observed target moves and the correlation among sensory signals is not stable by proposing a method that estimates the target positions and simultaneously associates observations based on the maximization of the correlation among the sensory signals. Second, we propose a new signal association method for binary observations based on a statistical test. Third, when the observation confidence changes based on the situation and affects the signal correlation, it is difficult to stably associate observations. We propose an association method that evaluates observation confidence and apply it to associate the leg motion of pedestrians and wearable accelerometers to estimate stable signal correlation. Finally, we propose a new media conversion method that converts omni-directional video to sound that keeps the impressions to signals in the original media by considering the signal correlation.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

For artificial systems that behave in response to the external world, it is necessary to use sensors to observe its state. To describe and understand outside scenes and objects, many kinds of sensors have been developed, and much work on signal processing and pattern recognition has been conducted. In recent years, progress in communication networks and sensors has greatly increased the amount of observation data from sensors. Toward recognizing human behavior, the big data statistical analysis approach, which uses many sensors in public locations and wearable devices, has drawn increasing attention. The integration of many sensors is one fundamental problem in this research area.

Just as humans combine various kinds of sensory organs to observe scenes and objects, it is also effective for artificial perceptual systems to combine many kinds of sensors. There are two main advantages of integrating multiple sensors.[1)]

- *Redundant observation*: By observing the same features by multiple sensors and integrating the observations, measurement results are obtained with higher accuracy. For example, by integrating a camera and a distance sensor, both of which are observing the same object, more accurate position estimations are obtained. By integrating the same physical quantity observed by multiple sensors, accurate representations of targets are obtained.
- *Complementary observation*: By integrating sensors that cover different observation

areas, the total amount of measurement results increases. By integrating sensors that observe different physical quantities, the number of features that describes targets increases. For example, by integrating a camera and a microphone, we obtain visual and audio representations of targets. By integrating different physical quantities observed using different kinds of sensors, we obtain multifaceted representations of targets.

To exploit these advantages, many studies have been conducted in the area of sensor integration[2].[3] Focusing on redundant observations by using multiple sensors, integration methods based on the Kalman filter[4] are widely used, and other methods based on the particle filters have been proposed[5] to compute Bayes filter-based integration. In the research area of ambient intelligence, many kinds of sensors are installed in office and home environments to observe and identify daily behavior in projects like Smart House,[6] Intelligent Room,[7] Aware Home,[8] and Ubiquitous Experience Media.[9] Progress in smaller and lower energy-saving sensors has fueled the development of wearable sensors. Studies on position estimation by integrating many wearable sensors and integrating wearable and environmental sensors have been conducted[10][11][12].[13] Also in the research area of robotics that focuses on human-robot communication, the effectiveness of integrating many kinds of sensors has been shown, including audio-visual perception,[14] multimodal interaction, and multimodal robot learning.[15]

Focusing on redundant and complementary observations using multiple sensors, increased attention has been drawn to media conversion studies that propose algorithms to convert signals in one modality to another. For example, in situations when we can ’ t use a specific modality for communication and presentation, it is effective to use different modalities in a complementary manner by media conversion. Another application of media conversion expects a synergistic effect from using multiple modalities by adding another modality to the original contents. By focusing on the possibilities of media conversion, much work has been proposed to transfer the descriptions and the impressions of the original scene to another medium, especially in the area of *kansei information processing*.

However, the previous studies on the integration of different kinds of sensors suffer from the following problems:

- Integration is performed on the representation in a common coordinate. In most studies, since integration is in a common position coordinate or a symbol system, it is difficult to apply to sensors that cannot measure the representation in a given coordinate alone.

- In previous studies, the abstract level of the representation is high. The correlated multimodal information from different kinds of sensors is lost during the independent conversion of each signal to a given representation. We cannot use the lost information in the integration process.

In this thesis, we present solutions for the above problems for a new sensor integration framework.

1.2 Limitation of Previous Sensor Integration Methods

1.2.1 Perceptual Binding Problem and Position-based Association of Multisensory Signals

Consider the observation of a scene with different kinds of sensors. Before integrating the observations, each sensor observes the scene independently and its observations are represented at each coordinate system. At this stage, since the relationship between these representations is unknown, we must associate the observations to integrate them and construct an integrated multimodal representation of the objects.

This problem is related to the binding problem in human perception. When we observe objects using many features, a mechanism is needed to bind the information that is related to each object and to distinguish it from others. The mechanism has not been identified until now and called *perceptual binding problem*. We call the problem in artificial sensor integration as the *perceptual binding problem in engineering*.

For example, imagine observing a few objects in a scene with cameras and microphones. The objects in it are detected and represented in each sensor coordinate. Cameras represent them by shapes and colors; microphones represent them by sound intensity. Generally, when multiple objects are detected, the observations must be associated from both sensors that represent the same object to construct a multi-sensor representation of each object.

To solve this association problem, conventional methods assume that each sensor generates representation in a common coordinate, where observations are associated. In the previous example of observing with cameras and microphones, the conventional method assumes that both cameras and microphones detect the positions of objects and associates those that are close to each other in the common position coordinate based on the knowledge of the relationship between the camera and microphone coordinates. However, we can only apply the position-based association method to integrate sensors that can measure the positions of the objects in the scene. It is difficult to apply it to other kinds of sensors that do not directly observe positions.

1.2.2 Representation in a Common Coordinate System at Higher Level of Abstractions

Previous integration methods associate observations from different kinds of sensors on the representation in a common coordinate. Usually the positions on the common coordinate are estimated using each sensor and then close observations are associated.

However, since the representation in the coordinate is sometimes too abstracted, correlated multimodal information is lost during the feature extraction and recognition processes to obtain representation. We cannot use the lost information in the integration process. The representation in a common coordinate sometimes lack important cues for integration.

Also, a common method toward media conversion is the pattern recognition-based method, which defines the conversion rules between the patterns in two kinds of media. When a pattern is detected in one medium, the associated pattern is presented in another. Another approach to media conversion is based on words that represent impressions. This approach associates patterns in multiple media by evaluating the relationship between the adjectives and patterns in both media. These approaches might be limited since they associate media by symbolic representations like pre-defined patterns and natural language. The following are the restrictions of pre-defined symbol-based association: 1) patterns that are difficult to represent in symbols are discarded in the recognition and abstraction processes, 2) the association methods only detect and convert pre-defined patterns and unknown patterns are discarded.

Therefore, in previous multi-sensor integration methods, the sensing system designers define a common coordinate where the observed information is integrated and converted. An advantage of the previous methods is that we can rely on developed techniques to detect the representation on the coordinate since the meaning of the coordinate is easy to understand. However, one problem is that important information is lost during the abstraction process to compute the representation.

1.3 Sensor Integration Based on Signal Correlation

To cope with the problems, we provide a new sensor integration framework. We propose an integration method that focuses on signal correlation in the observation of different types of sensors. Our concept is called *signal level integration*, where observations are integrated and converted before the abstraction and recognition processes to obtain representations in common coordinates. Since we do not have to independently compute the representations in a given common coordinate using each sensor, our method can be applied to integrate many kinds of sensors, especially those that do not directly measure positions.

1.3.1 Signal Correlation in Human Perception

In general, we obtain redundant observations when we use multiple perceptual organs to observe a scene. For example, when we observe with a camera and a microphone a person clapping his hands, and the motion of his hands generates a sound, the motion observed in the video and the sound observed by the microphone are expected to correlate. Although these observations are independently performed, the observed signals share correlated components. Signals that represent the same target in different modalities superficially look differently, but they share similarity and correlation.

This correlation among observations is a signal structure that appears when localized multiple sensors observe a common scene. Humans always use multiple perceptual organs to observe scenes and are exposed to correlated signals among these observations. It is natural to assume that our perceptual systems are highly developed to perceive signal correlations. For example, humans tend to associate correlated observations perceived by different organs since they are likely to observe the same object.

Signal correlation plays an important role when we use our perceptual system in a complementary manner. Suppose we usually observe a scene using multiple perceptual organs. When we can observe the same scene by only one perceptual organ, the observation with the other organ is recalled and our perceptual experience is complemented. Humans resemble information processing machines that watch not only the perceptual observations themselves but always the signal correlations among observations as well.

In previous studies on artificial perception, much work has focused on the information processing of each individual type of sensor. Although humans always exploit the relationship among correlated signals, little attention has been given to artificial perception techniques that evaluate and utilize signal correlation. We focus on the correlation among observed signals and propose signal level integration methods of different kinds of sensors.

1.3.2 Integration at a Signal Level That Keeps Correlation Among Multimodal Signals

We associate multimodal observations and convert the observed signals in one modality into another based on the evaluation of the signal correlation among observations. Our method integrates observed signals at an earlier level of abstraction. Compared to higher level feature extraction and pattern recognition results, our method focuses on signals at earlier levels of abstraction where the signal correlation among multimodal observations is preserved. We call multisensory integration at this level *signal level integration*.

In our proposed signal level association method, we do not have to independently compute the representation in the common coordinates from each sensor. Our method associates observations from different types of sensors by evaluating their signal correlation. Therefore, we can associate the observations of sensors even if we cannot measure the position just by sensors. For example, suppose we observe a person walking with cameras and wearable accelerometers (Figure 1.1). From both sensors we obtain the signal components of vibrations due to the walking. Our proposed method associates various types of sensors that are difficult to associate in previous integration methods.

In our proposed signal level media conversion method, we do not rely on such symbolic representation as pre-defined patterns and natural language. By getting correlated signals among multimodal observations, observing targets in only one modality will recall the miss-

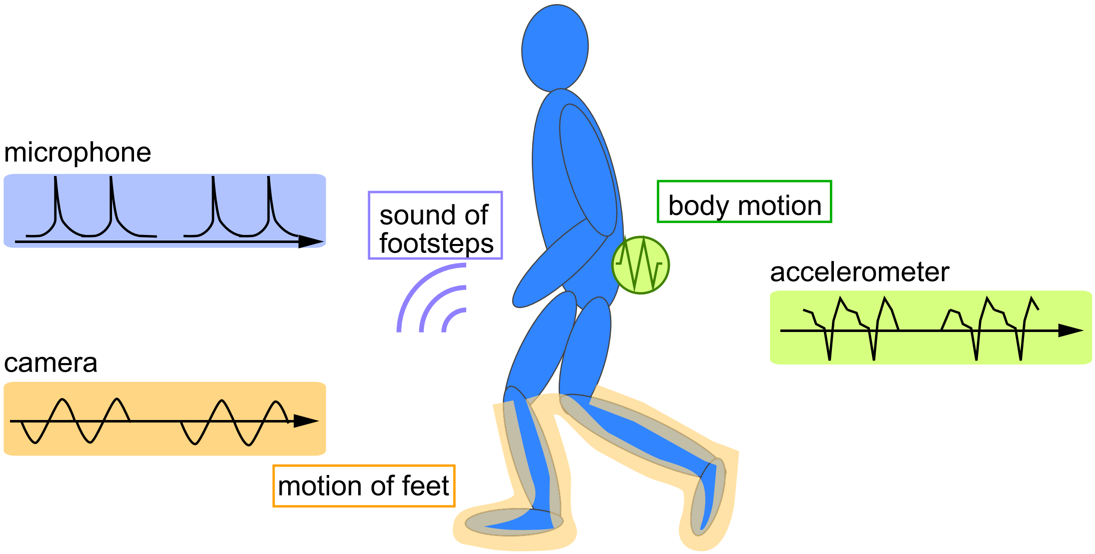


Figure 1.1 Signal level correlation in observations

ing observations in another modality. Based on this complementary action of our sensory systems, we propose a new media conversion method. For example, we can generate the converted signals in the target modality so that humans can recall the signals in the original modality by keeping the signal correlation between signals. This new media conversion is based on the similarity among observations in different media at the signal level.

Figure 1.2 shows the process of observation and feature extraction from the observation and the association among the observations. The vertical axis represents the abstraction level of the observations, and the horizontal axis represents the types of sensors. Association methods based on positions link the detected positions. In contrast, our proposed association method based on signal level correlation associates at an earlier stage of abstraction: at the *signal level*. In this paper, we propose methods of signal level association to apply to many types of sensors and in many situations.

Recently, a new sensor integration method associates different kinds of sensors at an earlier stage of perception.[16)] They focused on the fact that when two sensors observe a target there is signal correlation among the signals and associated different physical quantities. They observed two people with a camera and a microphone and showed that when they speak in turn there are signal correlations between the intensity of the pixels close to the mouth and the sound intensity. By focusing on the redundant correlation among signals, their method

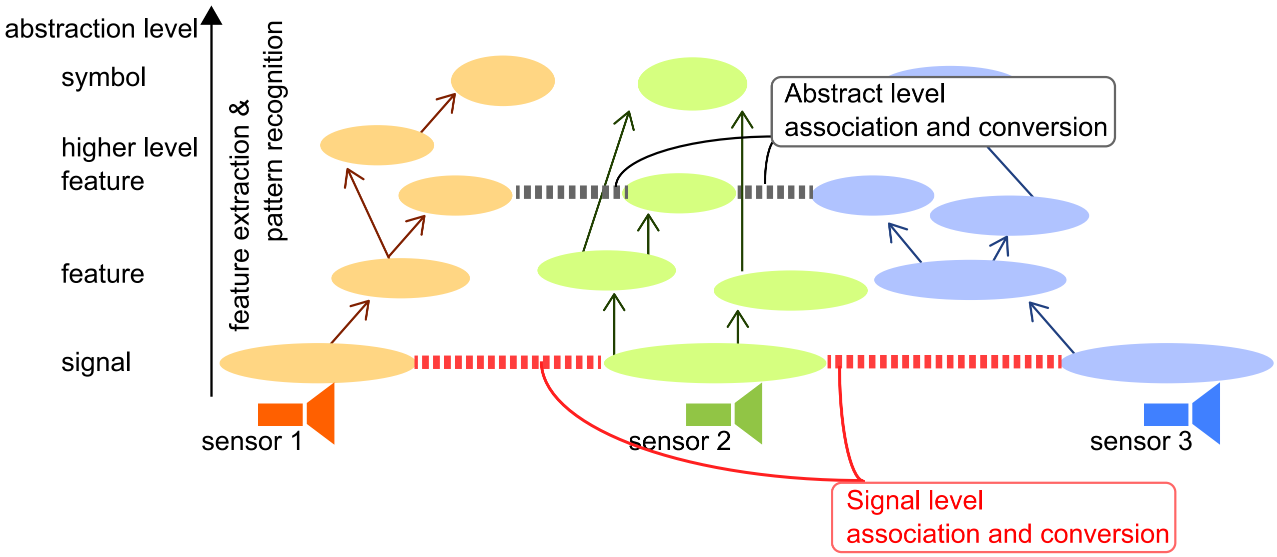

abstraction level
symbol
higher level feature
feature
signal
feature extraction & pattern recognition
Abstract level association and conversion
sensor 1
sensor 2
sensor 3
Signal level association and conversion

Figure 1.2 Integrating different kinds of sensors.

associates signals at an earlier stage of perception before the signal correlation is abstracted away in the recognition process. This approach has been expanded and generalized by many researchers[17)18)19)20)21)22)].[23)]

However, since studies have only been conducted for integrating a microphone and a camera, this approach must be expanded to apply it to integrate other kinds of sensors. This thesis expands this method based on many kinds of sensors and increases the possibilities of an association method based on signal correlation.

1.4 Resaerch Issues and Approaches

To apply signal level association methods in many kinds of sensors and in many situations, following research issues are significant.

Signal level association in various situations

We expand methods to associate moving targets in array sensors.

- **Association of a moving target in video images**: When the signals source moves, previous method[16)] cannot associate observations. To cope with the problem, we propose a method that estimates the positions of targets and associates observations simultaneously based on maximization of correlation among sensory signals.

- **Association of moving targets in binary touch sensors**: Observations of sensors like touch sensor and event detection sensor are binary. Previous method assumed continuous observation signals. We propose a new signal association method based on a statistical test.

Signal level association based on unreliable observations

We propose methods that reliably extract signal level correlations.

- **Association of unreliable observations**: When observation confidence changes according to situation and it affects to signal correlation, it is difficult to associate observations in stable manner. We propose an association method that evaluates observation confidence that is specific to types of sensors.
- **Construct time-dependent correlation model**: When observations are limited, computed correlation among sensors is sometimes not reliable. We propose to model time-dependent correlation model to associate observations.

Applications of signal level relationship

We use the signal level relationship in the area of tracking and indentification and media conversion.

- **Application to people tracking with identification**: Recently people behavior in public locations is observed and statistically analyzed and the results are attracting attention in the area of environment design and marketing. We propose a method that not only estimates positions but also identifies each person who carries wearable sensors.
- **Application to convert signals between different media**: Based on signal level relationship between video and music, we propose a method that convert omni-directional video to music that conveys similar impression. By listening generated music, the user can imagine impression of the original scene.

1.5 Thesis Outline

In chapter 2, we simultaneously detect and track a sound source based on the criteria of mutual information maximization. The problem of detecting and tracking a sound source is solved as an optimization problem to find the path that maximizes the mutual information between video and audio signals. We describe a sensor fusion algorithm based on mutual information maximization and apply it to the problem of sound source localization by combining audio and visual signals.

In chapter 3, we propose a method that associates binary signals based on signal correlation and explain an integration method of wearable and floor sensors that detects the positions of people. Floor sensors consist of small unit sensors, each of which returns ‘1’ when someone is standing on one of them and ‘0’ otherwise. To integrate these binary and acceleration signals from wearable sensors, we propose an integration method that evaluates signal correlation based on a statistical test.

In chapter 4, we propose an association method from among different kinds of sensors that considers confidence in observation. Different kinds of sensors have different reliability characteristics depending on the situation. We focus on the association of laser range finders (LRFs) and wearable gyroscopes to track and identify each person and propose an association method that considers the reliability of LRF observations.

In chapter 5, we propose a method that associates the leg motion of pedestrians and wearable accelerometers. LRFs observe pedestrians at the height of their feet and extract features from a bipedal walking pattern. Wearable accelerometers also observe walking patterns. Since walking rhythms differ from person to person, our proposed method can distinguish pedestrians walking in a line. Another characteristic is that it only uses an accelerometer in the wearable devices.

In chapter 6, we propose a new feature level media conversion method that generates comfortable sounds to listen to the transfer impressions of visual scenes. We define a set of low-level visual and musical features and conversion rules between them. Since the method does not assume pre-defined entities in visual scenes, it transfers the impressions of unseen visual scenes with unknown entities. By introducing music constraints in the generated sound, listening becomes more comfortable.

In chapter 7, we summarize and conclude our thesis.

Chapter 2

Signal Level Associatition of Moving Targets Based on Mutual Information Maximization

The signal association method proposed by Hershey et al.[16)] associates different kinds of sensors based on signal correlation. Their method is expanded and many audio-visual signal association methods have been proposed. However, when the signal source moves in images and the signal correlation is not stable, it is difficult to provide stable association. In this chapter, expand the signal association method and solve the both the signal association problem and the position estimation problem by maximizing signal correlation between observed signals. Experimental results show the effectiveness of the proposed method in sound source localization problems for moving targets.

In this chapter, we propose to detect and track a sound source simultaneously based on the criteria of mutual information maximization. The problem of detection and tracking a sound source is solved as optimization problem to find the path that maximizes mutual information between video and audio signal. In section 2.1, the sensor fusion algorithm based on mutual information maximization is described. In section 2.2, we applied the algorithm to the problem of sound source localization by combining audio and visual signal. In section 2.3, experimental results are shown. In section 2.4, we conclude this chapter.

2.1 Previous Work on Integration of Different Kinds of Sensors at Earlier Stages

Recently, to associate visual and audio signal at the earlier stage of integration, Hershey et al.[16] proposed a direct integration method based on computing mutual information between observed signals. They focused on signal correlation between the pixel intensities of the speaker's mouth and sound intensity. By extracting the redundancy of signals that observed same target, they associate different kinds of sensors. Fisher III et al.[17] expanded this method and computed mutual information without assuming any model of signals. In these works, different kinds of sensors are integrated at early stages based on statistical methods. However, it is assumed that a few sensors continue to observe a signal source and that the relation between sensory signals is stable for a period of time. So it is difficult to apply these methods to a case where a signal source moves in the environment.

To cope with moving signal sources, object detection is applied in several studies. Slaney and Covell[18] applied face detection to detect pixels related to the speaker based on the canonical correlation analysis. Li et al.[20] computed the mutual information between sensory signals in the projection space where sensory signals are similar to each other. Ikeda et al.[24] extracted objects based on background subtraction and computed the mutual information between sensory signals. Fisher III and Darrel,[22] Nock et al.[23] also applied face detection and computed the relation between sensory signals based on their previous methods[17].[19] However, targets are limited since these methods require models to detect targets. Furthermore, these methods are not robust since a segmentation process is performed before the integration process.

In this chapter, we propose to detect and track a signal source simultaneously based on maximizing mutual information with a jointly Gaussian assumption. The problem of detecting and tracking a signal source is solved as an optimization problem to find the path that maximizes the mutual information between the video and audio signals.

2.2 Signal Level Association Based on Mutual Information Maximization

2.2.1 Common Signal Source Detection Using Mutual Information

Suppose we observe an information source with different kinds of sensors. Though the observed signals are in different representations, these signals are correlated and share common components. By detecting and evaluating correlated components from the observed signals, the relations between sensors are measured and signals are integrated before the abstraction process. When a pair of signals is correlated with each other, the knowledge of one signal enables us to predict another signal. In the information theory, the predictability between signals is defined as mutual information. So it is natural to use mutual information to estimate the correlation between sensory signals. Hershey et al.[16] used mutual information to measure the correlation between audio and video signals.

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

$$I(A; V) = H(A) + H(V) - H(A, V) \tag{2.1}$$

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

$$H(A) = -\sum_t p(A(t)) \ \log \ p(A(t)), \tag{2.2}$$

$$H(V) = -\sum_t p(V(t)) \ \log \ p(V(t)), \tag{2.3}$$

$$H(A, V) = -\sum_t p(A(t), V(t)) \ \log \ p(A(t), V(t)). \tag{2.4}$$

.

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.[16] The mutual information can be replaced with

$$\frac{1}{2}\log\frac{1}{1-\rho(A,V)^2} \tag{2.5}$$

where $\rho(x, y)$ is correlation function between $A(t)$ and $V(t)$.[25)]

A more general estimation method without assuming any distribution model is proposed in.[17)] A few criteria are compared to estimate the correlation between the audio signal and the video signal[19)] and they reported that the mutual information with a jointly Gaussian assumption is the best in their experiments.

2.2.2 Detection of a Moving Signal Source

In,[16)] 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 a case that where signal source is moving in the environment. Suppose we are tracking a sound source by using a camera and a microphone. Fig. 2.2 shows the computation of the mutual information between the video signal and the audio signal. Since previous methods computed the mutual information between the audio signal and the brightness of each fixed pixel, they failed to capture the relation of the signals when the signal source moved.

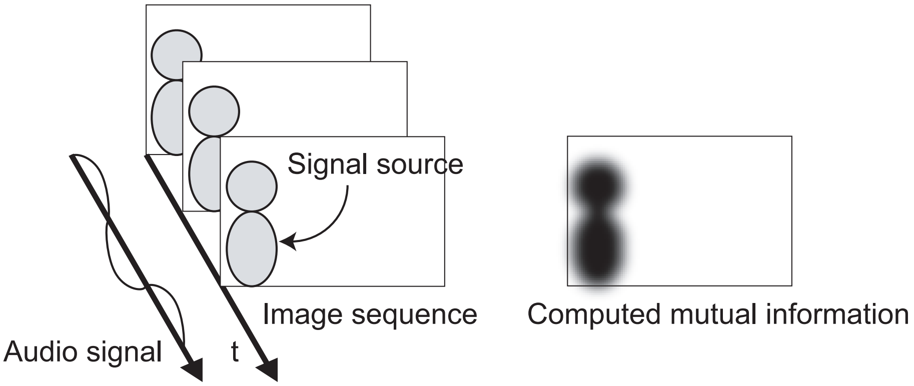


Figure 2.1 Computing mutual information between sensory signals when the signal source does not move.

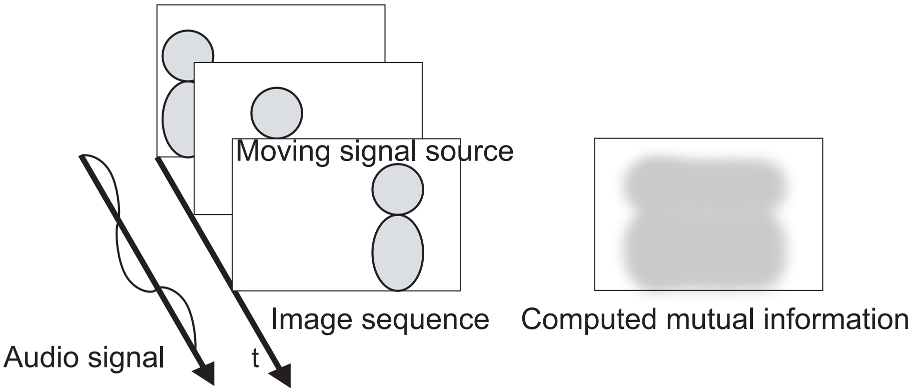


Figure 2.2 Computing mutual information between sensory signals when the signal source moves.

2.2.3 Computing Mutual Information Along an Estimated Trajectory

To cope with this problem, an object is detected before the mutual information is computed[24)23)20)].[26)] These methods consist of two stages. In the first stage, the positions of the target candidates are computed by image processing. In the second stage, mutual information is computed between the audio signal and the video signal along the detected trajectory of each candidate. By computing the mutual information along the detected path of the target, it is possible to detect the sound source when the target moves (Figure 2.3).

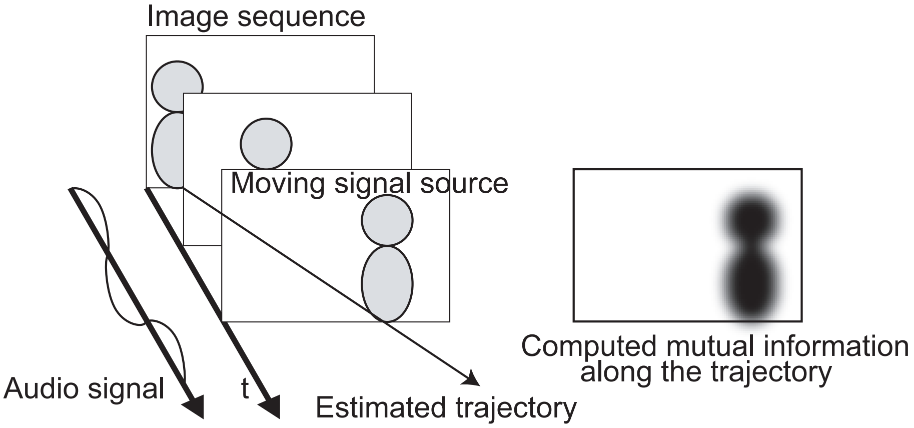


Figure 2.3 Computing mutual information along the trajectory of the signal source.

When the sound source moves, mutual information is computed along its path. Now video signal $V(t, x)$ depends on time and position, and mutual information is computed according to Equation 2.1 and the following formula:

$$H(V) = -\sum_{t} p(V(t, x(t)))) \ \log \ p(V(t, x(t)))), \tag{2.6}$$

$$H(A, V) = -\sum_{t} p(A(t), V(t, x(t))) \ \log \ p(A(t), V(t, x(t))). \tag{2.7}$$

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

2.2.4 Detecting and Tracking a Moving Signal Source Based on Mutual Information Maximization

In the two-stage approach, the tracking process and the sensor integration process are separated and it is difficult to recover tracking errors in the integration process. In this paper, these stages are integrated into one process. We propose to detect and track the sound source simultaneously based on the criteria of mutual information maximization. Since the detection and tracking process are performed according to a unique criterion, this method does not suffer from the segmentation error of the detection process prior to the integration process.

When the trajectory of a moving signal source is unknown, the problem of detecting and tracking is regarded as an optimization problem to find the trajectory that maximizes the mutual information between video and audio signals. We propose to find the trajectory of the sound source by performing a heuristic search using mutual information as a heuristic evaluation function (Figure 2.4). There are many possible trajectories in the image sequence, and the trajectory that maximizes the mutual information is selected.

2.2.5 Introducing Heuristics for Robust Estimation

Computing mutual information in moving regions

The estimated trajectory of a single pixel based on computing mutual information has significant noise. Computed mutual information in a specific trajectory in the background

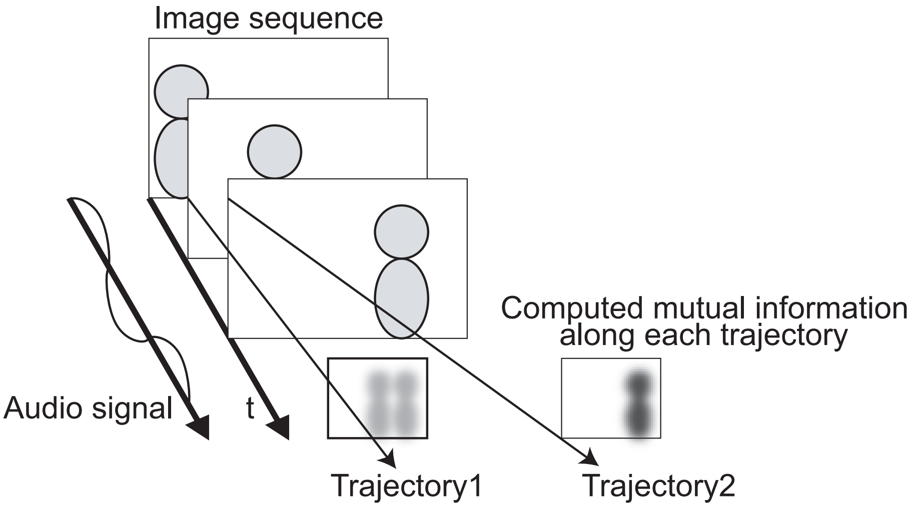


Figure 2.4 Detecting and tracking a target based on mutual information maximization between sensory signals

region sometimes has a large value. In this paper, we propose to compute the mutual information along trajectories of regions. Each pixel in the region is supposed to move in parallel and mutual information is computed along the trajectory. The correlation between the audio signal and the trajectory of the region is evaluated with the average of the mutual information of pixels in the region.

Introducing Motion model

The search process that maximizes the mutual information between sensory signals is almost breadth first search at an early stage since mutual information is a poor heuristic function when the length of the signals is short. So we introduce a motion model of the target. Then the search process effectively finds the path that maximizes mutual information.

Applying smoothing filter on computed mutual information

To make search process stable, we propose to apply a smoothing filter on the array of computed mutual information. We applied a spatial averaging filter on the computed mutual information at each frame.

2.2.6 Heuristic Search Algorithm Based on Mutual Information Maximization

To detect and track a sound source, we apply a heuristic beam search algorithm (Figure 2.5 (a)). In the algorithm, a hypothesis represents a trajectory of the sound source in the images. During the search process, a list of hypotheses is updated. Where LIST is the hypothesis list, H is a hypothesis in the list, and BEAM is a threshold of the number of hypotheses in the list. In step 3, the velocity of each target is piecewise constant (Figure 2.5 (b)). All hypotheses in list are replaced in step 2 and step3.

2.3 Experiments

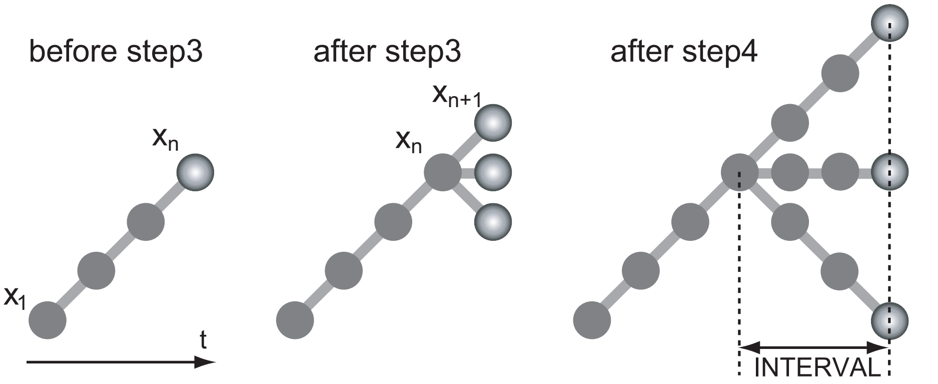
To confirm the effectiveness of the proposed method, we apply the method to a sound source location problem using one microphone and one video camera. The video signal is sampled at 30 frames/second, and the image size is 160x120. $V(t, x)$ in Equation 2.6 and 2.7 is the brightness of the pixel at time 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 audio signal. Fig. 2.6 shows samples of the video and audio signals, respectively.

Search	
Maximum number of hypotheses	1000
Region size [pixel]	100 x 100
Length of signals [frame]	256
Motion model	
Update interval [frame]	32
Maximum acceleration [pixel/frame2]	1.0
Maximum velocity [pixel/frame]	2.0

Table 2.1 Parameters used in the experiment

1. Initialize hypothesis list LIST with possible positions of the target in the image.
2. If the length of hypotheses in LIST is T, output the hypothesis with maximum mutual information from LIST and exit.
3. For each hypothesis H = $x_1, \ldots, x_n$, create new hypothesis H' by adding next positions x_{n+1} that are predicted according to the motion prediction model. Remove H from LIST and H'. Note that multiple positions may be generated according to the prediction model.
4. For each hypothesis H, create new hypothesis H' by adding next positions x_{n+1} that are predicted according to the piecewise constant velocity model motion. Remove H from LIST and H'. Repeat this step INTERVAL-1 times.
5. For each hypothesis H, compute mutual information along the trajectory between audio and video.
6. Sort LIST in descending order based on the computed mutual information and select highest BEAM hypotheses and discard others.
7. Goto 2.

a) Algorithm



b) Expansion of hypotheses. Each hypothesis in the hypothesis stack corresponds to a leaf of the search tree.

Figure 2.5 Heuristic beam search algorithm that finds the trajectory that maximizes mutual information computed along the trajectory

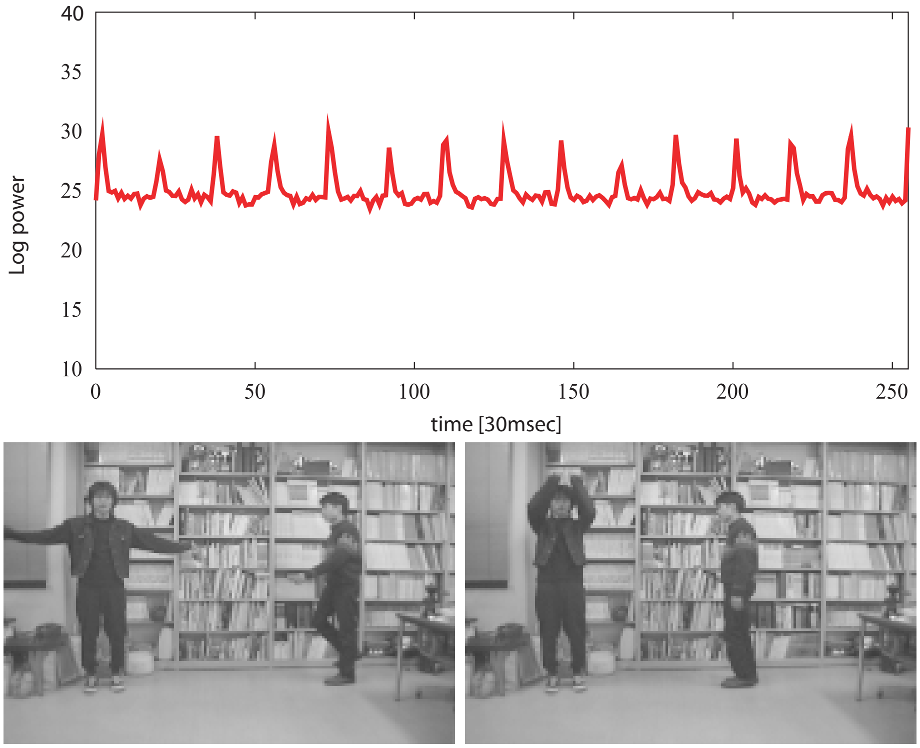


Figure 2.6 The average energy of the audio signal (upper), example images in the video signal (lower).

The parameters of the experiment are shown in Table 2.1. In this experiment, we introduced a motion model of the target that assumes the acceleration of the target is constant for an interval. The update interval of the acceleration is set to 32 frames (about one second), and the range of the acceleration and velocity is limited. This model assumes that the person don't walk fast and don't change the direction frequently. The search is performed in the horizontal direction in the images. The maximum number of hypotheses should be set to any large value, and larger value will result in more precise search and longer time of the execution. The region size is fixed to 100x100 that is about the size of the people in images in the experiment. The length of signals used to estimate mutual information is 256 frames. The length should be large enough to detect correlation between signals.

2.3.1 Results of a Sound Source Detection

Figure 2.7 shows the result based on previous signal level fusion methods which assume that the sound source does not move. The color of pixels in Figure 2.7 indicates the intensity of the mutual information between the brightness of the pixel in the images and the energy of the audio, where a darker color indicates a higher intensity of mutual information. The results do not include any remarkably darker regions. This means the sound source localization has failed.

2.3.2 Results of Searching Trajectory that Maximizes Mutual Information

Figure 2.82.9 shows the process of the search based on the proposed method. The graphs in the figure show the trajectories of the best twenty hypotheses at frame = 32, 64, 128, and 256, respectively, where the horizontal axis is time and the vertical axis is the position of the target. The computed intensity of mutual information of the best hypothesis at each frame is also shown on the right of each graph. The mutual information is computed from the first frame to the last frame in the left graph.

Figure 2.8 shows the results of search process of a hypothesis that are initially located in the area of a walking person. The trajectory of the signal source is estimated by a heuristic search that maximizes mutual information between the video signal and the audio signal. The trajectories in hypotheses have converged and the intensity of mutual information is remarkably high and location of the sound source is determined.

In contrast Figure 2.9 shows the result when the hypothesis is initially located at the area of another person. The trajectories in hypotheses do not converged and intensity of mutual information along any path is not high.

2.3.3 Detected Correlation of the Signals

Figure 2.10 shows the changes of the intensity of a pixel in the region of the left person's hand and the audio signal. There is no correlation between the signals. Figure 2.11 shows the intensity of a pixel in the region of the right person's knee on the detected trajectory.

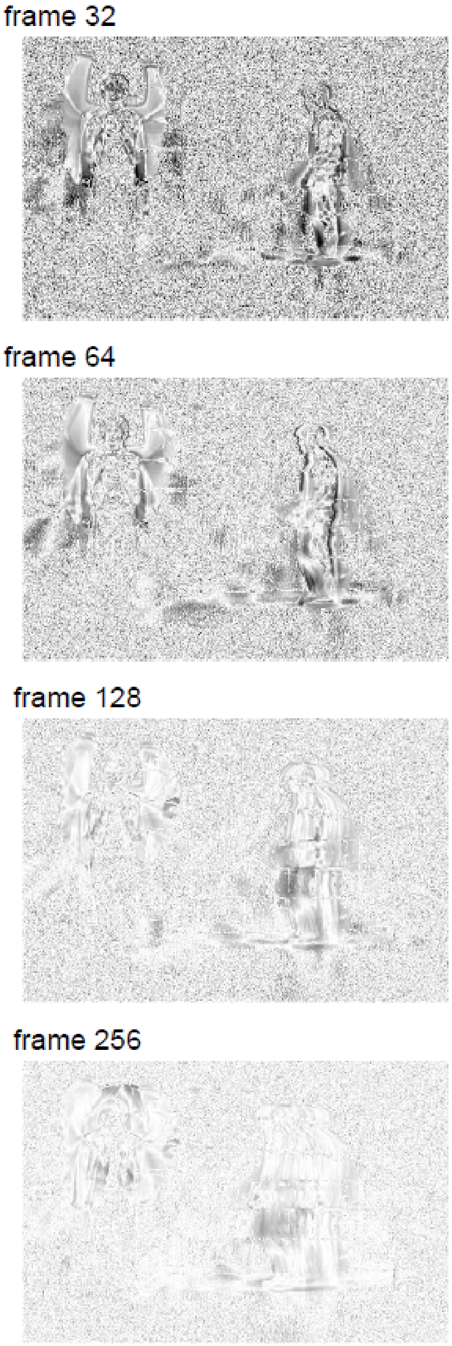


Figure 2.7 The changes of the intensity of the mutual information based on previous direct fusion method. Computed mutual information is dissipated in the image.

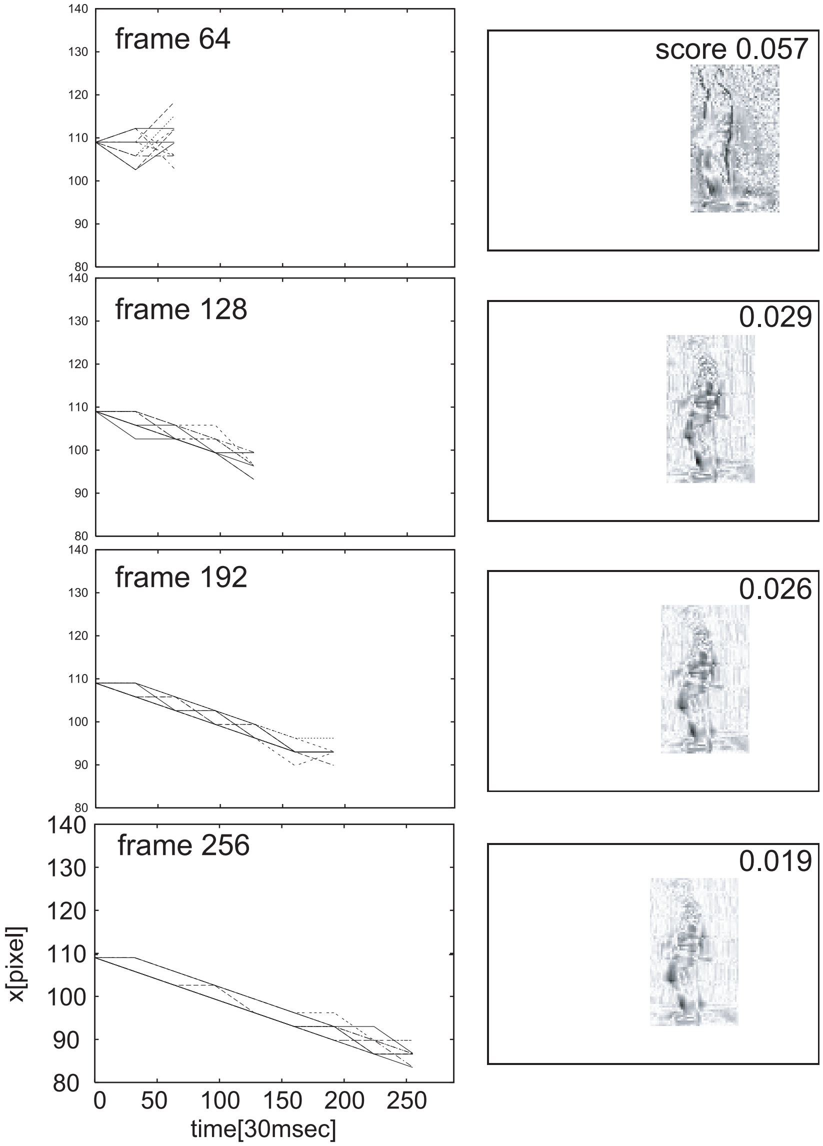


Figure 2.8 The search process of the proposed direct fusion method (1). In the left graphs, the best hypothesis of the trajectory (solid line) and best twenty hypotheses (dashed line) at frame = 32, 64, 128, 256 (t=1,2,4,8 [sec], respectively) are shown. Right figures show computed mutual information in the region along the best hypothesis.

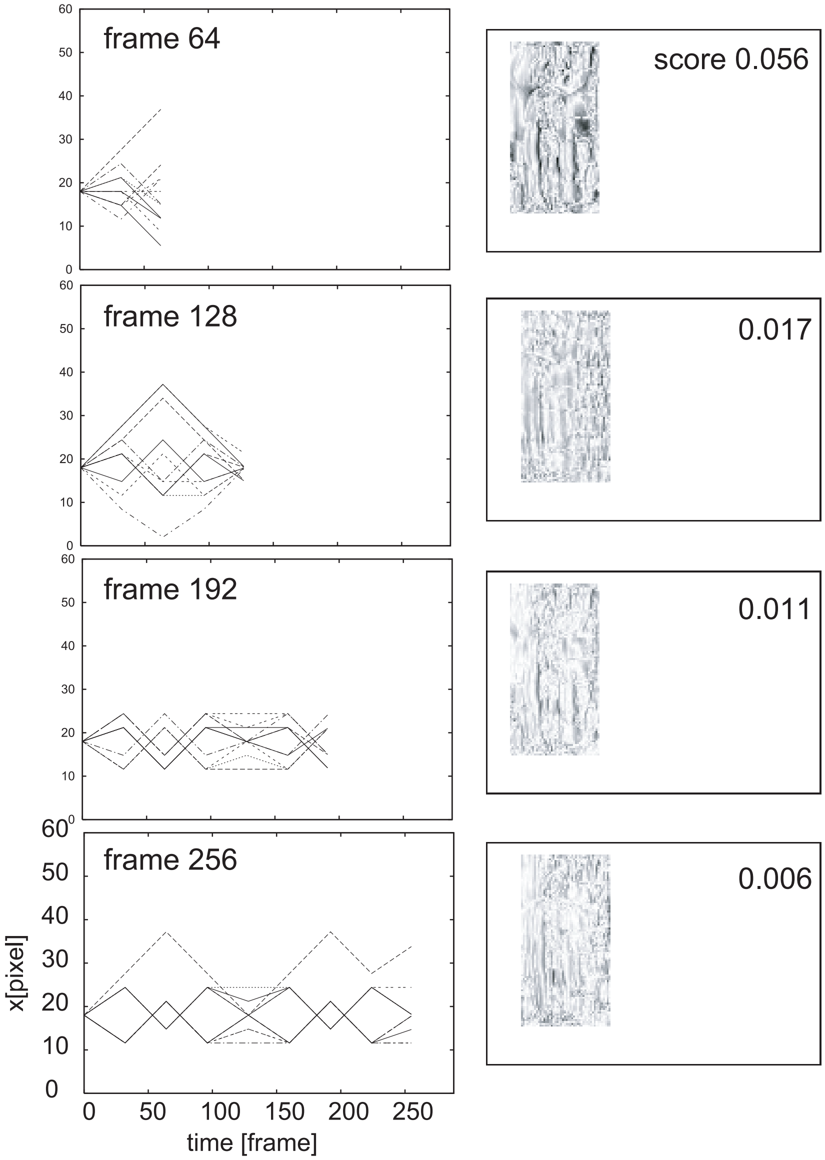


Figure 2.9 Search process of the proposed direct fusion method around the area of the people who shake his hands.

They show strong correlation and they share many peaks in the graph. The pixel is in the dark region in the Figure FS:fig:search-walker.

Figs. 2.8 show the process of the search algorithm when initial hypothesis is the correct position of the sound source. (a)(b)(c)(d) show the hypotheses with high mutual information at frame 64,128,192,256, respectively.

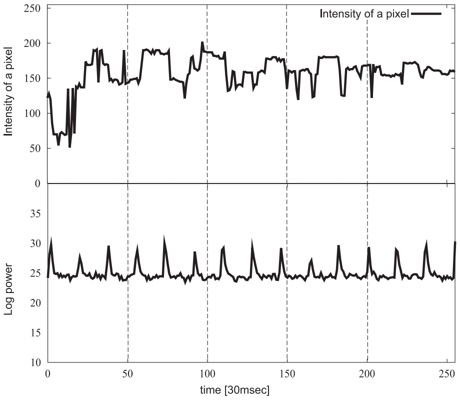


Figure 2.10 The brightness of a pixel in the region of the left person who shakes his hand (upper) and sequence of average power of sound signal (lower).

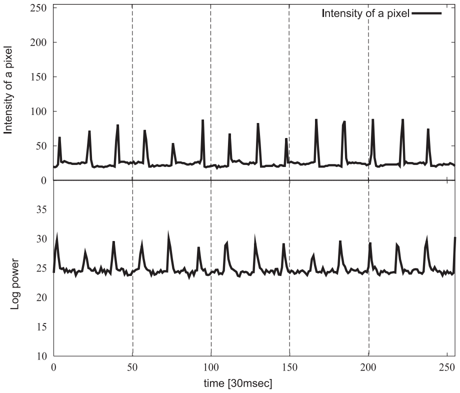


Figure 2.11 The brightness of a pixel in the region of the right person who walks with sounds of footsteps (lower) and sequence of average power of sound signal (upper).

The moving path of the sound source is correctly tracked by the criterion of the mutual information maximization.

2.4 Conclusion

In this paper we have presented a novel sensor fusion method at an early stage of processing. By fusing different kinds of sensors at the signal level, correlated multimodal information that is lost in abstraction process can be effectively used. To cope with moving signal sources, we have proposed a method for finding the trajectory of the signal source by evaluating the correlation between sensory signals. In our framework, the problem of detecting and tracking of a signal source is regarded as an optimization problem to find the trajectory that maximizes the signal correlation between sensory signals. We have proposed solving this problem by a heuristic search algorithm, using mutual information with a jointly Gaussian assumption as a heuristic function, and introduced a target motion model and a region-based evaluation method for trajectories to effectively search for the trajectory. Compared to the previous signal level sensor fusion method with object recognition, the proposed method does not suffer from segmentation error in the detection process. We applied the proposed sensor fusion method to detect and track a sound source using a camera and a microphone. In the experiment, two people in motion are observed, and a walking person with audible footsteps is detected and tracked in the images. We introduce a simple motion model of the signal source. When the motion of the target is more complex, a more detailed model of the target will be required. We plan to adapt the region size and the length to compute correlation in search process. By preparing various initial hypotheses with different sizes of regions, the hypothesis with the best size will be selected as the search process proceeds. By changing the length to compute correlation according to the intensity of mutual information, the proposed method will be expanded to cover wider applications with various time constants. In the future, we plan to apply the proposed method to other kinds of sensors, and to investigate a robust method for estimating the mutual information of signal sources in which the jointly Gaussian assumption is not appropriate.

2.5 Associating Signals in Different Dimension

Dimension of observations of different kinds of sensors are sometimes different. To associate and integrate many kinds of sensors, it is important to associate observations in different dimensions. In previous methods, scalar observations are associated based on normalized correlation and mutual information. In this chapter, we propose to associate three-dimensional accelerometer and a pixel in video images.

Figure 2.12 shows the typical acceleration sensor signal when two people walks. Each person has an acceleration sensor on the right hand. The acceleration signal is averaged in each video frame. We consider each pixel as an independent sensor. The figures show

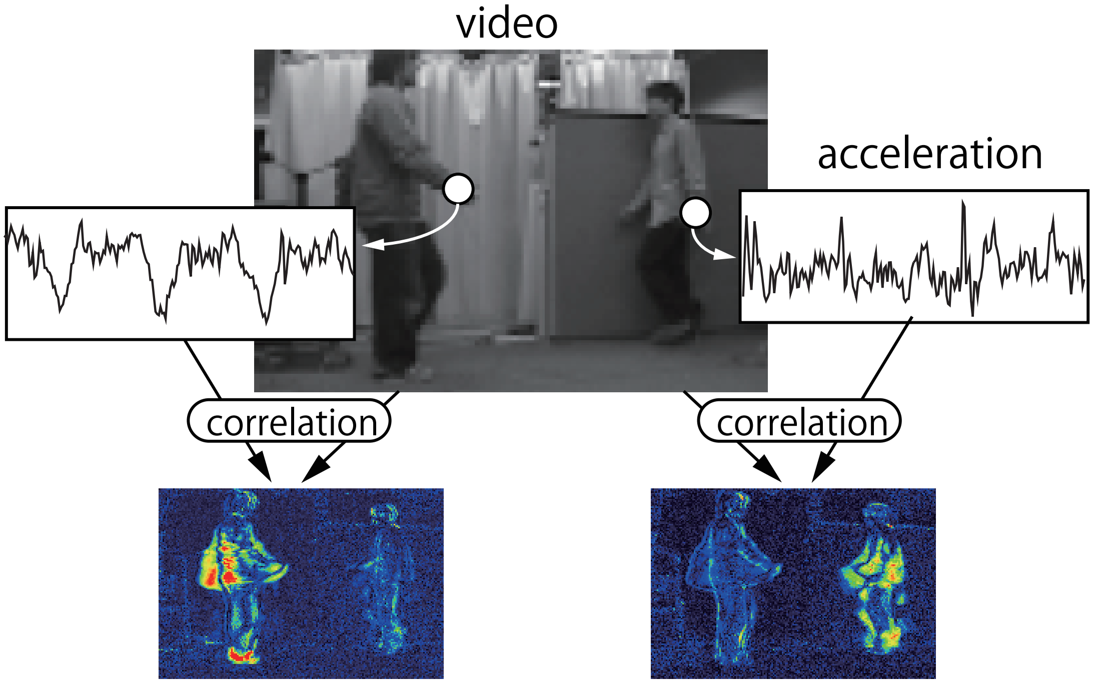


Figure 2.12 Association of wearable sensor and video. Computed absolute value of correlation function between acceleration from wearable accelerometer and change of inensity of each pixel in video.

the absolute normalized correlation between acceleration signal of a direction and intensity of pixels. The figure shows clear correlation between these sensors. However, we need to determine the direction of the acceleration sensor with highest correlation.

In general, when we observe motion of a person by using sensors in multiple dimensions, there is a component that shows clear correlation among other sensors. We applied the canonical correlation analysis (CCA) to estimate the direction that maximizes correlation between signals. CCA finds the linear mapping that maximizes correlation between two input signals.

2.6 Computing Canonical Correlation Between Acceleration Signals and Video Signals

Suppose we have N observations from two sensors and the dimension of observations are p and q respectively. A data matrix is defined as

$$X = [\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_p]^T$$

$$Y = [\mathbf{y}_1, \mathbf{y}_2, \ldots \mathbf{y}_q]^T$$

where the row vectors $\mathbf{x}_i^T (1 \leq i \leq p), \mathbf{y}_j^T (1 \leq j \leq q)$ in X and Y are vectors that represent N data samples. Let $\tilde{X}, \tilde{Y}$ represents the results of subtracting average of each line. Our purpose is to find a linear transformation $\mathbf{a}, \mathbf{b}$ to $\tilde{X}, \tilde{Y}$ that maximizes correlation between $\mathbf{a}^T \tilde{X}$ and $\mathbf{b}^T \tilde{Y}$. $\mathbf{a}, \mathbf{b}$ are computed based on canonical correlation analysis:

$$r(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a}^T S_{XY} \mathbf{b}}{\sqrt{\mathbf{a}^T S_{XX} \mathbf{a}} \sqrt{\mathbf{b}^T S_{YY} \mathbf{b}}} \tag{2.8}$$

where $S_{XX}, S_{YY}, and S_{XY}$ are variance covariance matrices:

$$S_{XX} = \frac{1}{N} \tilde{X} \tilde{X^T} \cdots\cdots (2.9)$$

$$S_{YY} = \frac{1}{N} \tilde{Y} \tilde{Y^T}$$

$$S_{XY} = \frac{1}{N} \tilde{X} \tilde{Y^T}$$

In order to have only a single solution, $\mathbf{a}, \mathbf{b}$ we impose following constraints.

$$\mathbf{a}_i^T S_{XX} \mathbf{a}_i = 1 \cdots\cdots (2.10)$$

$$\mathbf{b}_i^T S_{YY} \mathbf{b}_i = 1$$

The $\mathbf{a}, \mathbf{b}$ are obtained by applying singular value decomposition.

$$S_{XX}^{-1} S_{XY} S_{YY}^{-1} = \mathbf{A \Lambda B^T} \tag{2.11}$$

Matrices $\mathbf{A}, \mathbf{B}$ are othogonal matrices and $\mathbf{\Lambda}$ is a diagonal matrix. The first column vector of $\mathbf{A}, \mathbf{B}$ is the canonical correlation vectors $\mathbf{a}, \mathbf{b}$, respectively.

By applying cannonical correlation to each pixel in video images and three dimensional accelerometer, the most correlated direction of the accelerometer is derived. Let X represents acceleration signals in N video frames.

$$X = [\mathbf{a}_x, \mathbf{a}_y, \mathbf{a}_z]^T$$

where $\mathbf{a}_x$ represents x components of acceleration sensors at $1 \cdots N$.

$$\mathbf{a}_x = [a_x(1), \cdots, a_x(N)]^T$$

$(a_x(t), a_y(t), a_z(t))$ represents acceleration at time t. A sequence of pixel intensity is represented as

$$Y = [I_{x,y}(1), \cdots, I_{x,y}(N)]$$

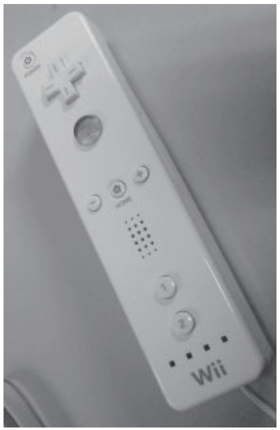
where $I_{x,y}(t)$ is pixel intensity at time t and (x,y) in image coordinate. Signal correlation between three dimensional accelerometer and a pixel are derived by estimating transformation. We set N to number of frames in a few second. By repeating this estimation, a sequence of direction of acceleration that maximizes signal coorelations are derived. Each pixels are associated to one of acceleration sensors based on the cannonical correlation coefficient (Equation 2.11).

Then the absolute value of canonical correlation coefficients are computed and the value is set to zero if it is under a threshold. Further erosion and dilation are performed to remove isolated noisy pixels. Image regions are associated to one of accelerometers that maximizes averaged absolute thresholded correlation.

2.7 Experiments

2.7.1 Associate Image Regions and Acceleration Sensors

To confirm the effectiveness of the proposed method, we apply the method to detect and track people in a room. We observe two people by wearable accelerometers and two cameras. The acceleration sensor used in the experiments is ADXL330 (Analog Devices, Inc.). All people in the environment are assumed to have the sensor with the right hand. The motion of the Wiimote is sensed by a 3-axis linear accelerometer located slightly left of the center of



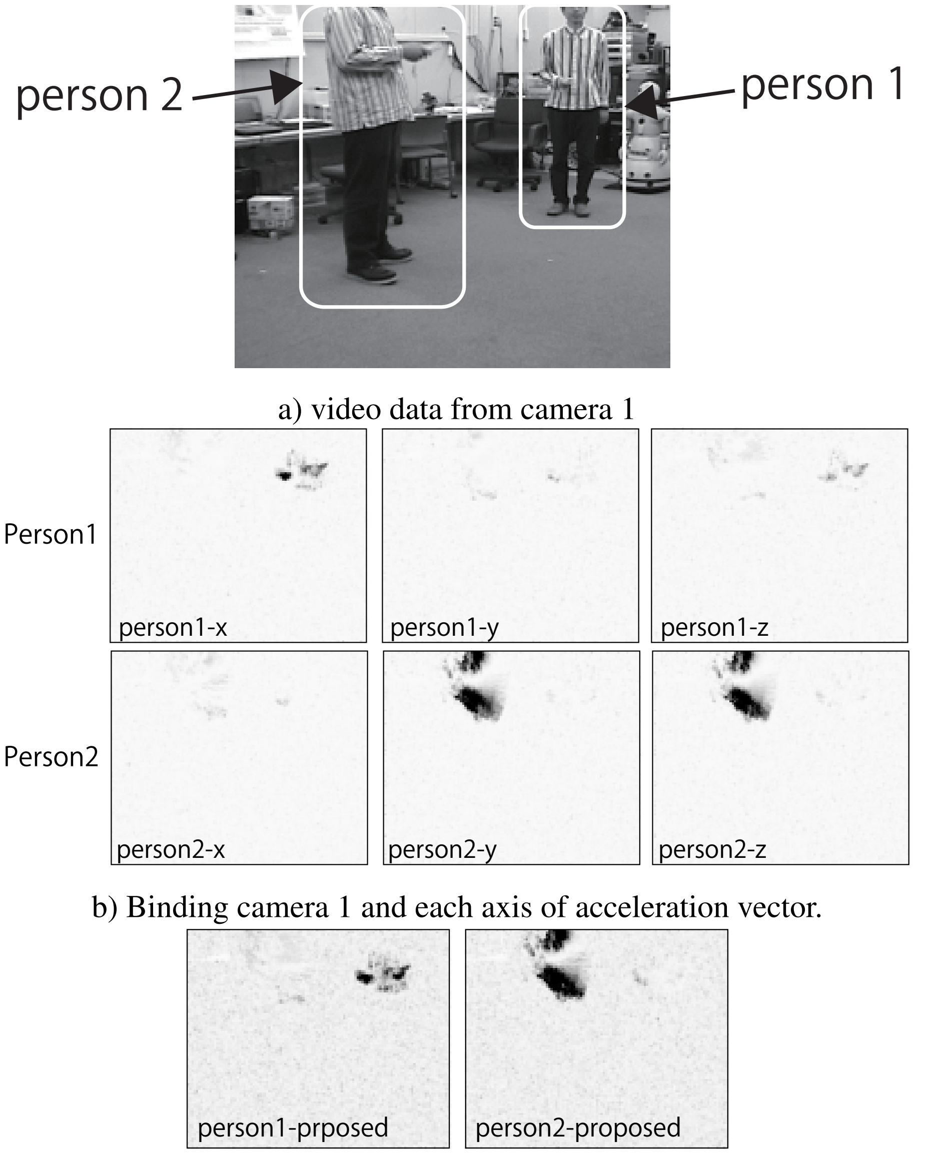
Specifications	
Output signal	8 bit integer
Sampling frequency	70 Hz
Inteface	Bluetooth

Figure 2.13 Acceleration sensor used in the experiments and specifications.

the controller (Figure 2.13). The signal is sampled at 70 kHz, and the signals are transmitted via Bluetooth. The average of the signal is computed in each video frame. The video signal is sampled at 30 frames/second, and the image size is 360x240. We compute inter frame difference of intensity in each pixel. Typical signals of pixel intensity are shown in Figure 2.12.

First we recorded two people that are shaking their arms. Figures 2.14,2.15 shows the computed correlation function between each pixel in the images and each acceleration sensor.

Next we recorded two people that walk across. Figure 2.16 shows the detection and tracking results when two people go across. The region with the highest average correlation is detected and tracked in images for each acceleration sensor signal. The right of the figure shows the trajectory of the region that maximizes correlation between sensory signals.

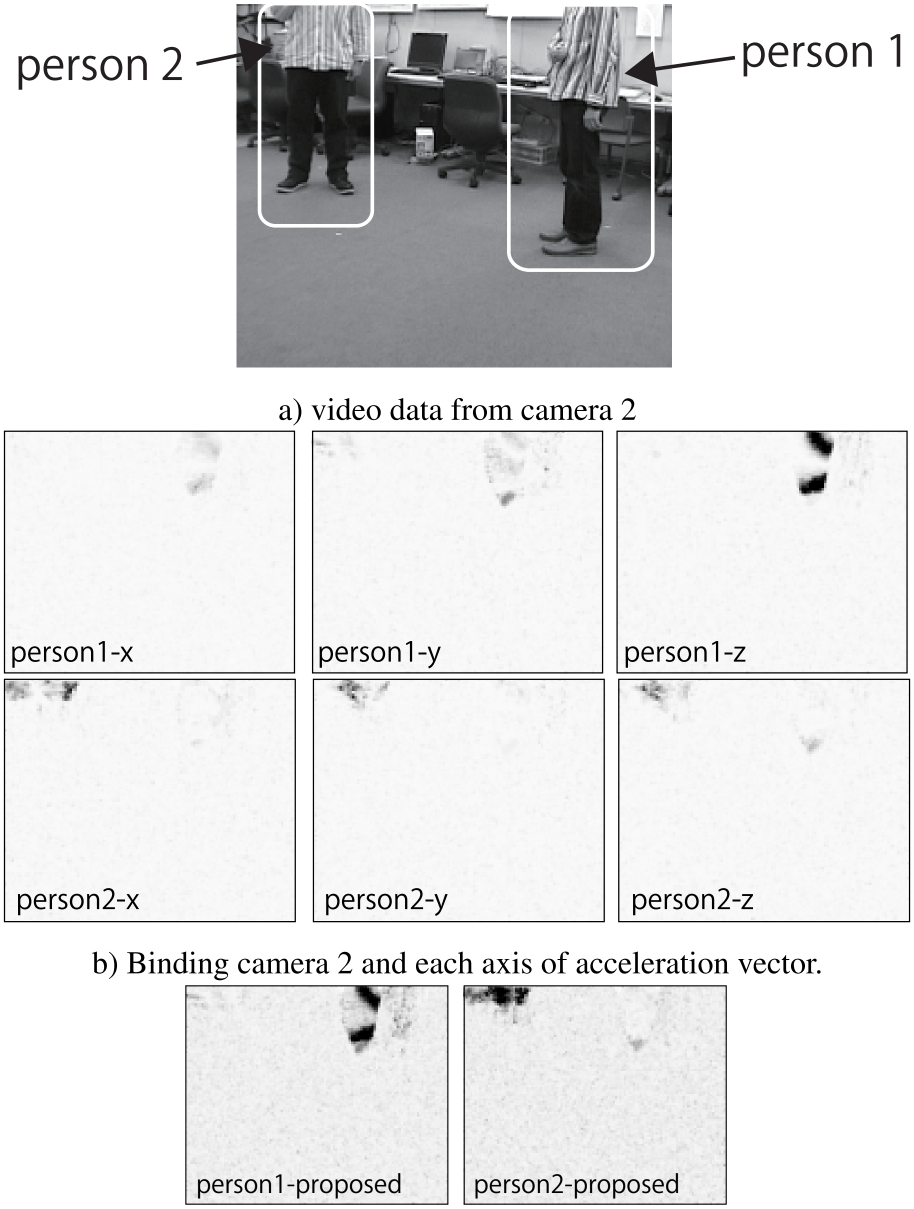


a) video data from camera 1

b) Binding camera 1 and each axis of acceleration vector.

c) Binding camera 1 and acceleration vector based on the proposed method.

Figure 2.14 Binding wearable accelerometer of each person and pixels in video (camera 1).



a) video data from camera 2

b) Binding camera 2 and each axis of acceleration vector.

c) Binding camera 2 and acceleration vector based on the proposed method.

Figure 2.15 Binding wearable accelerometer of each person and pixels in video (camera 2).

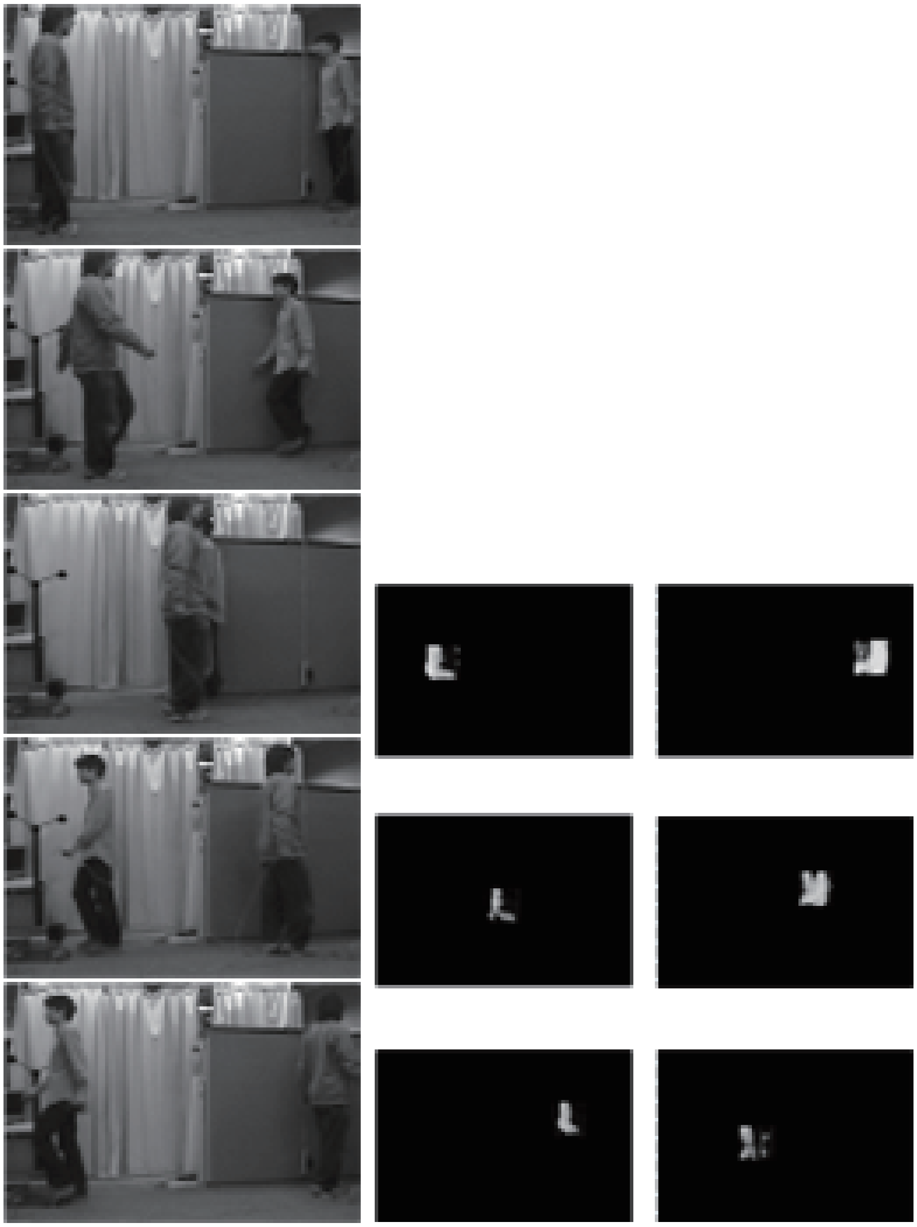


Figure 2.16 Two people walk across. The left figures show the original images and the right show computed correlation function between intensity of each pixel and the signal from the acceleration sensor on the left person.

Chapter 3

Signal Level Association of Binary Signals

Binary signals are common observation signals from simple sensors like switches. In this section, we propose a method that associates binary signals based on signal correlation. We explain an integration method of wearable sensors and floor sensors that detect positions of people. Floor sensors consist of small unit sensors and each sensor returns ’1’ when someone is on a sensor and ’0’ otherwise. In order to integrate these binary signals and acceleration signals from wearable sensor, we propose an integration method that evaluates signal correlation based on statistical test.

3.1 Introduction

In order to realize intelligent environment that supports human activities, estimation of positions and IDs of people is one of important issues. Floor sensors that consist of small touch sensors on the floor are one of promising sensors that reliably detect our locations. However, since floor sensors observe discrete footsteps, association ambiguity arises when two people go across. A typical problem to track people with floor sensors is shown in Figure 3.1. Two candidates of associations from observations to person are shown. It is difficult to distinguish these two associations only by floor sensor observations. To solve the problem, we propose to combine wearable acceleration sensors. Wearable devices can keep ID infor-

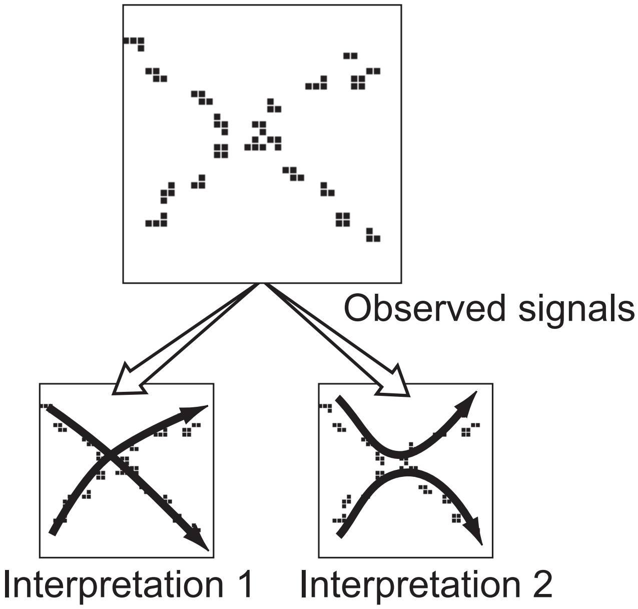


Figure 3.1 The ambiguity of the association of floor sensor signals when two people get close. Observations are superimposed in a certain period of time. Two candidates of associations are shown on the right.

mation of the user, but it is difficult to locate the person by using wearable devices alone. By combining observed positions by using floor sensors and IDs from wearable devices, the system will be able to provide trajectories of all people with IDs. However, there are ambiguity in associations between trajectories and IDs. In this chapter, we propose to associate these observations and estimate consistent trajectories based on signal correlation. Since the signals from floor sensors and wearable accelerometer synchronize when they observe same walking person, these two signals are not independent. The synchrony between the signals is evaluated based on statistical test to find correct association. People tracking examples are shown to confirm the effectiveness of the proposed method. Significant improvement in correct association rate is achieved compared to the results only by floor sensors.

3.2 Related Works

Detecting positions and IDs of people is one of important functions for intelligent environments. So far many works have been proposed to observe people and provide services by integrating sensors in the environment and wearable devices[6)7)].[9)] The sensors used in previous works are classified into three main groups.

Vision sensors Vision sensors are widely used to understand the scene in the environment, and much works have been done to recognize human behavior using vision sensors[27].[28] Vision sensors provide much information about people in the environment, not only their positions but shape, color, and gestures. A problem with cameras is that they suffer from changes in the lighting conditions in the environment. Also, using cameras in public spaces for identification purpose sometimes causes privacy issue.

Floor sensors By spreading touch sensor or pressure sensor network on the floor, the positions of people are accurately detected. Recently, floor sensors have received increasing attention and several studies have been proposed to recognize human behavior using floor sensors[29][30][31][32][33].[34]

Addlesee et al.[29] and Orr et al.[30] identified people based on changes in pressure at the time of the landing by using a load cell. They focused on a footstep and did not track people. Liau et al.[35] put load cells on the floor, Murakita et al.[32] put touch sensors on the floor and they track people who walked on the sensors. Our method used similar people tracking method by using floor sensors and also identify people by integrating other kinds of sensors. Yamanishi et al.[36] put pressure sensors on the floor and tracked and identified people by shapes of footsteps. Since their method need high resolution pressure sensors, the amount of data become too large to cover large area. Whereas floor pressure sensors are used in previous studies[29][30][31],[36] we used touch sensors. Good points of touch sensors is the simple structure of the sensor and the small amount of data to transmit, which enable to cover large area in low cost.

Wearable devices

In ubiquitous computing, wearable devices have been used to locate people.[10] Devices that have been studied include IR tags,[37] ultrasonic wave tags[27] RFID tags[38],[39] Wi-Fi,[40] and UWB.[41] Kourogi et al.[11] integrates many kinds of wearable devices, such as accelerometers, gyroscopes, geomagnetic sensors, and cameras, and they estimated positions by only using wearable sensors. If the device ID is registered with the system, the person carrying that specific device can be located and identified. However, tag-based methods require the placement of many reader devices in order to locate people accurately, so the cost of installing reader devices is problematic in large public places. Wi-Fi- and UWB-based methods do not provide enough resolution to distinguish one person in a crowd. Furthermore, if users of the system have to carry additional devices just to use the location service, the cost and inconvenience should also be considered.

Wearable inertial sensors have also been used to locate a person by integrating observations[42)][12)].[10)] Since integral drift has been problematic, it is important to combine observations with those of other sensors. Recently, many types of cellular phones have started to incorporate accelerometers, and some people are carrying them in their daily lives. Therefore, the approaches using acceleration sensors for locating people can effectively use the infrastructure.

Another important point is to carry devices natural manner. Carrying additional devices in daily life is not very comfortable.

Integrating of Environmental Sensors and Wearable Devices

Kourogi et al.[11)] integrated wearable inertial sensors, a GPS function, and an RFID tag system. Woodman and Harle[43)] also integrated wearable inertial sensors and map information. Schulz et al.[13)] used LRFs and ID tags to locate people in a laboratory, and they proposed a method that integrates positions detected using LRFs and identifies people by using sparse ID-tag readers in the environment. Mori et al.[44)] installed floor sensors and RFID tag readers in a room and tracked and identified people who carry ID tags. They associated anonymous trajectories and IDs from tags when a trajectory and an ID tags are observed close location. Difficulties in their method are ambiguity remains when more than one ID tags are detected from a reader device, and association is possible only when a person is close to a reader device. More dense installation of reader devices will increase the spatial resolution to some extent, but conflicts between reader devices are inevitable.

Integration based on Signal Correlation

In this section, we propose to integrate floor sensors and wearable acceleration sensors and propose a method that associates observed trajectories and IDs by evaluating their correlation

Table 3.1 Methods for people tracking and ID detection

method	Accuracy of	Accucacy of	Privacy	Ease of
	position estimation	personal identification	issue	carry
Environmental sensors				
cameras	★★★	★★	★★	★★★
floor sensors	★★★	★	★★★	★★★
Wearable devices				
ID tags	★	★★★	★★★	★
ID tags with accelerometer	★★	★★★	★★★	★★
Integration				
floor sensors + ID tags	★★★	★★★	★★★	★
floor sensros + ID tags with accelerometer	★★★	★★★	★★★	★★

In privacy issue, we evaluate if using the contents of the sensors always are socially acceptable, especially using face and body images. In ease of carry, we evaluate convenience of having to carry around the wearable device. The accuracy of position estimation by using ID tags depends on the spatial density of reader devices.

statistically. Floor sensors are very reliable, but it provide ambiguous tracking results and does not provide any information to distinguish each person. This characteristic causes ambiguity of association between observations and people. By combining reliable observations of positions by floor sensors and wearable devices that users carry, the system can estimate both accurate positions and IDs from wearable devices.

So far, sensor integration based on signal correlation has been studied mainly in the area of audio-visual integration[45][22].[46] However, the floor sensors generate binary position information and new integration method is required. The proposed method in this section is general method to integrate floor sensors and wearable devices.

Table 3.1 summarizes methods to estimate positions and IDs of people. There are many possible combinations of integrating sensors and integration of floor sensors and wearable accelerometers is one of promising approach. There are many ways to carry accelerometers: attach to the foot, hand, and waist. We put accelerometer on the waist, since it is one of common ways to carry cellular phones. Recently cellular phones are equipped with accelerometer and wireless LANs, so signals from cellular phones are available without burden on the user and practical applications of using these wearable sensors are expected[11].[47]

3.3 Integration of Different Kinds of Sensors Based on Statistical Test

In this section, we propose a integration method of following two types of sensors:

position sensors Sensors that observes positions of targets. The observations are not labeled with target IDs.

wearable sensors Sensors that is attached to each target. It observes the motion of the target.

3.3.1 Multiple Target Tracking Using Position Sensors

Since the observations from the position sensor contain only position information and do not contain ID information, ambiguities may arise in estimating trajectories of multiple targets. Tracking multiple targets using the position sensors requires solution to both data association and state estimation problems[48][49].[50] The most successful algorithm is the multiple hypothesis tracker (MHT).[48] MHT generates and maintains a set of hypotheses, where each hypothesis associates past observations with targets in a different way. Each hypothesis is evaluated by its posterior probability and the final result is the hypothesis with the highest probability. However, since MHT postpone decisions and examine all possible combination association as a new set of observation arrives, the number of hypotheses grows exponentially. The growth of the hypotheses is shown in Figure 3.2, where each branch denoting a different assignment of an observation to a target. Though several heuristics are proposed to cope with this problem, it is essentially difficult to select correct association only from position sensors. A promising approach is to combine different kinds of sensors.

3.3.2 Evaluating Association Hypotheses by Integrating Different Kinds of Sensors

In this paper, we propose to disambiguate the association by integrating wearable sensors. For example, the signals from floor sensors and acceleration sensors on the body change in

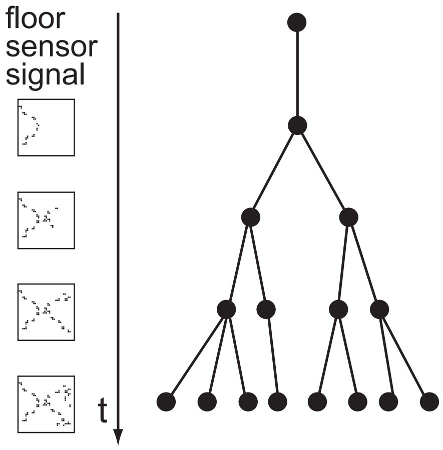


Figure 3.2 Exponential growth of the number of association hypotheses in the Multiple Hypothesis Tracker. A path from the root node to a leaf node represents an association hypothesis.

correlated manner if they observe same person. By evaluating the correlation between these sensors, probable association hypotheses are selected and the number of hypotheses becomes tractable (Figure. 3.3). The problem is how to evaluate correlation between position sensors and wearable sensors. Since the signals are in the different representation, a method that computes correlation between different kinds of sensors is required.

Position sensors and wearable sensors display synchrony and the signals are not independent if they observe same information source. Recently, several studies of sensor integration have been proposed by extracting synchrony between the signals from different kinds of sensors based on statistical methods. Hershey et al.[45] observed people speaking alternately with a camera and a microphone. They extracted synchrony between the audio signal and the brightness of the pixel around the speaker's mouth. They localized the speaker in the image by computing mutual information between the signals. This method has extended and has been applied to especially sound source localization problem[22].[23] A limitation of the method is the assumption that the target does not move in the images. In coping with a moving target, object detection is applied[18].[24]

However, previous statistical sensor integration methods cannot be applied to the case that multiple signal source overlap in the array sensor signals like video cameras and floor sensors. Furthermore, since floor sensors are binary sensors that report the position and the

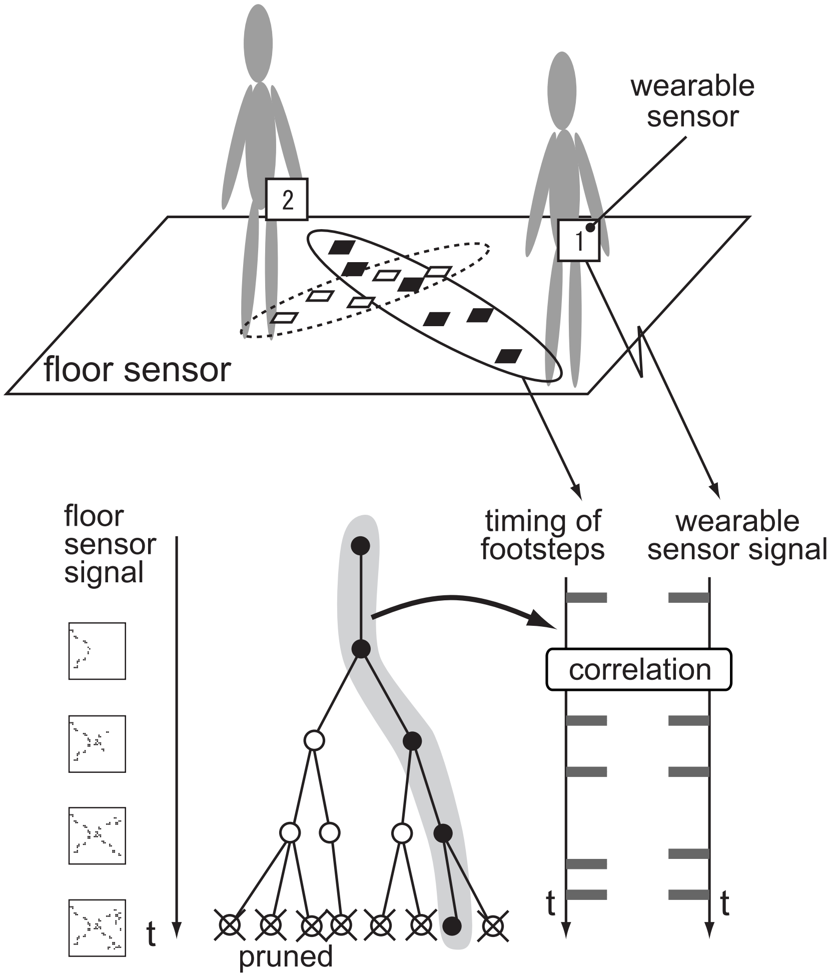


Figure 3.3 Evaluation and selection of association hypotheses using correlation between floor sensors and wearable sensors.

time and acceleration sensors are continuous sensor, another difficulty arises to use previous method to evaluate synchrony.

3.3.3 Evaluate Synchrony Based on the Chi-square Test of Goodness-of-fit

In order to evaluate synchrony between floor sensors and wearable accelerometer, we focused on the timing of footsteps. We propose to convert both signals into binary representation and apply a statistical test to evaluate correlation between binary signals. We generates

Table 3.2 Two-way contingency table

	y=1	y=0	計
x=1	z_{11}	z_{10}	$z_{1.}$
x=0	z_{01}	z_{00}	$z_{0.}$
total	$z_{.1}$	$z_{.0}$	$z_{..}$

The dot . represents that the sum is taken for all possible value of the index at the dot.

$$z_{.j} = \textstyle\sum_i z_{ij},\ z_{i.} = \sum_j z_{ij},\ z_{..} = \sum_i \sum_j z_{ij}.$$

multiple hypotheses of the trajectories from the position sensor array based on MHT and select hypotheses by evaluating correlation between position sensors and wearable sensors. We evaluate if the acceleration sensor signals on a target are independent from the sequence of floor sensor signals in a MHT hypothesis. The association problem between signals from different kinds of sensors is described as a hypothesis test.[51)] Whether time series $x(t), y(t)$ observes same information source is decided by a hypothesis test:

$$\begin{array}{rcc} H_0 : x(t), y(t) & \sim & p(x)p(y) \\ H_1 : x(t), y(t) & \sim & p(x, y) \end{array} \tag{3.1}$$

where H_0 states that the observations are statistically independent and H_1 states dependent. In the case of testing dependency between discrete signals, the chi-square test of goodness-of-fit is applied.[51)] When both signals are in binary representation, a two-way contingency table is created (Table 3.2).

When the null hypothesis H_0 is presumed true (signals are independent), theoretical frequency $\hat{z}_{ij}$ are estimated as follows given peripheral frequencies ($z_{1.},\ z_{0.},\ z_{.1},\ z_{.0}$) are fixed:

$$\hat{z}_{ij} = \frac{z_{i.}\ z_{.j}}{z_{..}} \tag{3.2}$$

where the dot . represents that the sum is taken for all possible value of the index at the dot.

Then χ^2 value is computed after observing frequency z_{ij} when the null hypothesis H_0 is presumed true:

$$\chi^2 = \Sigma_i \Sigma_j \frac{(z_{ij} - \hat{z}_{ij})^2}{\hat{z}_{ij}} \tag{3.3}$$

Then χ^2 value is distributed as chi-square with $\nu = (n_x - 1)(n_y - 1)$ degree of freedom. Where n_x, n_y are possible kinds of values of x, y. When both values are binary $\nu = 1$, in the case of

Table 3.2. When there are correlation between two signals, the H_0 will be rejected since they are not independent. From a set of hypotheses in MHT 3.2, we remove hypotheses that H_0 are not rejected and considered independent (Figure 3.3).

Proposed algorithm is summarizes as Table 3.3.

1. As new observations are obtained, generate hypotheses of target trajectories based on the observation of the position sensors.
2. For all generated hypotheses, convert the floor sensor signals associated to each target to binary signals. Convert signals from acceleration sensors to binary signals.
3. For all hypotheses, perform the chi-square test of goodness-of-fit to the binary signals computed in 2.
4. If the pair of signals is considered independent in 3, remove the hypothesis.
5. Go to 1.

Table 3.3 Multiple Hypotheses Tracking Algorithm by Integrating Floor Sensors and Acceleration Sensors.

3.4 Algorithm to Integrate a Floor Sensor and Acceleration Sensors

In this section, the details of the algorithms to integrate floor sensors and acceleration sensors based on the proposed association method are described.

3.4.1 Multiple Target Tracking Using the Position Sensors

An example of the observations of the floor sensor is shown in Figure 3.4. In the figure, the observations in a period are overlapped. The problem of estimating trajectories of multiple people based on floor sensors are:

1. Exponential growth of the number of hypotheses. Keeping all hypotheses is not realistic.

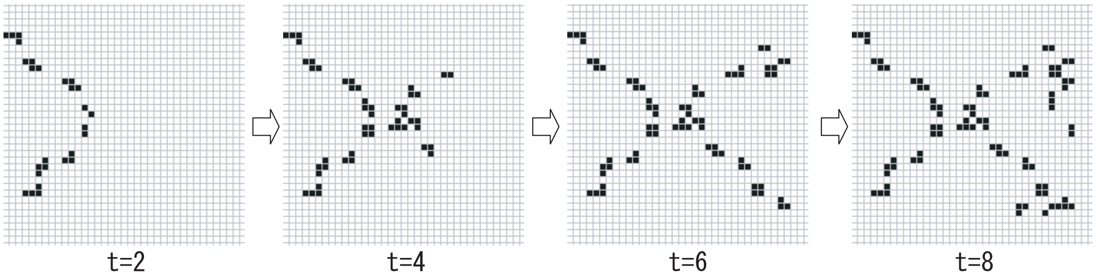


Figure 3.4 Examples of floor sensor signals

2. Discrete changes of the observation. Since the floor sensors observes the feet of people, only discrete positions of the feet are observed. And the observation of both feet sometimes disappears.

We apply MHT to estimate possible trajectories of the targets. It is difficult to handle huge number of hypotheses, but in the next step it is better to evaluate longer observation to decide dependency between signals. We introduce following models and assumptions.

Modeling trajectories based on the Kalman Filter When body of each person moves smoothly, floor sensors observe discrete footsteps. To estimate and predict trajectories from observed footsteps, Kalman filter based state estimation is applied for each person. The state vector of the target i at t consist of positions and velocities in two-dimensional coordinates.

$$X_{i,t} = [x, y, \dot{x}, \dot{y}]' \tag{3.4}$$

State model is based on the assumption that people do not change moving direction suddenly.

$$X_{i,t} = FX_{i,t-1} + w_t \tag{3.5}$$

where w_t is noise vector with average 0 and

$$F = \begin{bmatrix} 1 & 0 & \Delta t & 0 \\ 0 & 1 & 0 & \Delta t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{3.6}$$

where Δt is sampling period. The observation vector $Z_{i,t}$ is computed using following observation model:

$$Z_{i,t} = HX_{i,t} + v_t \tag{3.7}$$

where

$$H = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \tag{3.8}$$

where v_t is noise vector with average 0. Positive values are set to diagonal elements in w_t, v_t. A standard Kalman filter update rule is applied. To cope with intermittent observation, only the prediction step of the Kalman filter is executed when there are no observations that are associated to the target.

association constraints

To reduce number of the association between the observation and the target, we introduce following constraint.

1. Associate all observations that are adjacent to each other to the same target, and associate the succeeding observation at same position to the same target.
2. Limit the observations that are associated to the target to positions whose Mahalanobis Distance between the observation and the estimated position by the Kalman filter is less than a threshold (f_md_threshold).[50)]

In the experiments, the constraint 1 was not broken even once when two subjects got very closer. It depends on the size of unit floor sensor (10cm in experiments). When the size is smaller, the constraint is always satisfied. Otherwise, the number of ambiguous association becomes smaller and there are no need to introduce this constraint. Constraint 2 prunes the observation that does not match the linear prediction model of Kalman filter. It is based on the assumption that trajectories are smooth and do not jump.

3.4.2 Extracting the Binary Signal That Represents the Time of Contact

From both observed signals, binary signals are extracted that represents timings of footsteps.

Preprocess floor sensor signals

For each hypothesis, a binary signal $f(t)$ that represents the time of contact is computed.

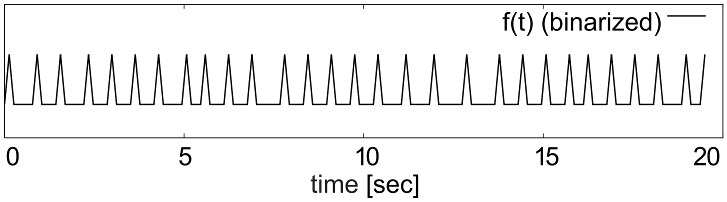


Figure 3.5 Extraction of the contact timings from floor sensor signals.

This conversion process is similar to differential processing in one dimensional signal.

$$f(t) = \begin{cases} 1 & |x(t) - x(t')| > \text{f\_step\_threshold}, \\ 0 & \text{otherwise} \end{cases} \tag{3.9}$$

where

$x(t)$ center of the current observations at t

$x(t')$ center of the observations at t', t' is the latest time s.t. $f(t') = 1$

and f_step_threshold is a threshold

We introduce the threshold since the center of the observation sometimes moves a little after the contact of a foot to floor sensors. The threshold should be larger than the length of the move, and we set the value to about half length of the sole (20cm in experiments). Then $f(t)$ is a binary signal that is 1 when there are new observation at a distance from previous 1s, otherwise 0. An example of the extracted binary signal from a person's walk is shown in Figure 3.5.

Preprocess acceleration signals

Figure 3.6 (1) shows an acceleration signal from a person who put an accelerometer on the waist. Signals from acceleration signal show almost binary property that is correlated to the walking motion. We convert acceleration signals to binary signals $a(t)$ that represent the time of contact (Figure 3.6) as follows.

1. An original observed signal.
2. Extract maximum difference of signals in a_diff_wlen frames. The large change of the value is detected.
3. Smooth the signal computed in 2 in a small window (Hanning window, window length = a_smooth_wlen). Compute local average signal in a large window (window length = a_localavg_wlen)

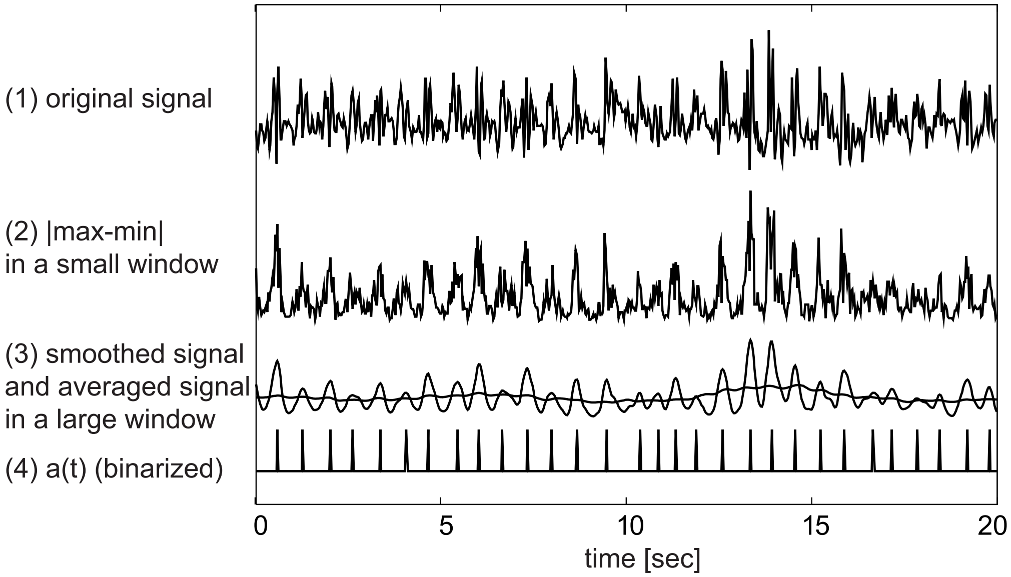


Figure 3.6 Extraction of the contact timings from acceleration sensor signals.

4. Extract peak of the signal computed in 2 if the value is larger than local average. Set binary signal a(t) as follows:

$$a(t) = \begin{cases} 1 & \text{if threre is a peak at } t \text{ and} \\ & \text{greater than the local average,} \\ 0 & \text{otherwise} \end{cases} \tag{3.10}$$

3.4.3 Evaluating Synchrony Between Sensory Signals Based on Chi-square Test of Goodness-of-fit

For each association hypotheses, the chi-square test of goodness-of-fit is performed to test if the binary signals $a(t), f(t)$ are independent. In general, since sampling period is different for each sensor, signals are averaged so that both binary signals has same sampling period. In this case of associating floor sensors and accelerometers, floor sensors has longer sampling period. So we generate $a(t)$ in longer sampling period by assigning '1' if original binary acceleration signal is 1 at least once in the period of one floor sensor signal, and '0' otherwise.

After sampling period is aligned, we compute a two-way contingency table based on last correlation_length frames $(a(t), f(t))$ (Table 3.2). (Table 3.2). z_{ij} in the table represents frequency of $(a, f) = (i, j)$ in the observation sequence. If $a(t), f(t)$ are observation of same person, the diagonal elements in the table will becomes larger. Then the test statistic is computed according to Equation 3.3 and test following hypotheses.

H_0: (a, f) are independent
H_1: (a, f) are not independent

Since our purpose is selecting correlated and not independent associations, we select associations if H_0 is rejected.

3.5 Experiments

We applied the proposed method to track two people in the room. Since difficulty in tracking by using floor sensors arises when people get close, we focused on the situation that two people go across. To apply our method in more crowded situations, resolving ambiguity when two people go across is one of fundamental problems. This experiment confirms the basic function of the proposed method.

3.5.1 Experimental Setup

Floor sensors

The floor sensor used in the experiments is VS-SS-SF55 (Vstone[52]) shown in Figure 3.7. A square region in Figure 3.7 is a unit sensor. A carpet is laid on floor sensor in use.

Accelerometers

The acceleration sensor used in the experiment is ADXL202 (Analog Devices, Inc.). The sensor is attached to the waist of the body (Figure 3.8). The sensor measures acceleration of two axes, and the signals of an axis that are close to horizontal plane are used in the experiments as shown in Figure 3.8. The observed signals are sent to a host PC via Bluetooth.

Parameters The parameters used in the experiment are shown in Table 3.4. The parame-

Specifications	
Size of the detection unit	100mm × 100mm
Number of the detection unit	1400
Gross area	14.0 m^2
Data format	binary (on/off)
Sampling frequency	8 Hz (*)
Inteface	RS-232C

(*) depends on the number of detection units.

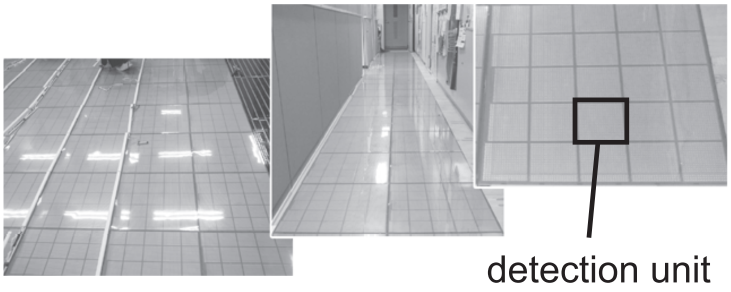


Figure 3.7 The floor sensor network used in the experiment.

Specifications	
Data format	16bit integer
Sampling frequency	36Hz (*)
Inteface	RS-232C

(*) an actual measured value in the experiments

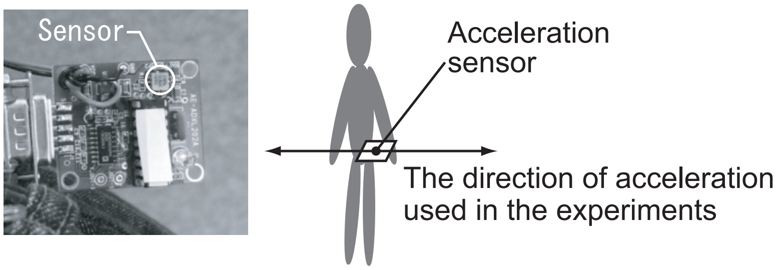


Figure 3.8 The acceleration sensor

Table 3.4 Parameters in the experiments.

Parameters used in processing floor sensor signals	
Variance of w (position) (eq. (3.5))	$(0.1)^2$
Variance of w (velocity) (eq. (3.5))	$(0.1)^2$
Variance of v (eq. (3.7))	$(0.5)^2$
f_md_threshold	1.2
f_step_threshold	20

Parameters used in processing acceleration sensor signals	
a_diff_wlen	3
a_smooth_wlen	16
a_localavg_wlen	48

Parameters used in the statistical test	
level of significance (α)	0.05
correlation_length	40

ters related to floor sensors are determined empirically so that our tracker correctly tracks one person. Window length parameter in preprocessing accelerometers (a_smooth_len) are determined to compute average of one step walk. Windows length to compute correlation (correlation_length) is determined to enough length to compute synchronization in this experiments.

Data acquisition

We captured two types of trajectories shown in Figure 3.9. The trajectories of two people cross in type 1 (”cross”) and do not cross in type 2 (”pass”). Seven type1 data and six type2 data are used in the experiment. There is no false alarm in floor sensor data. The scenes are recorded in video and correct associations are obtained for evaluation. We assumed the number of people is known and all people in the environment have an acceleration sensor.

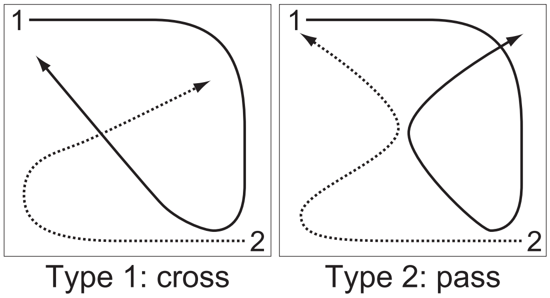


Figure 3.9 Two types of trajectories in the experiments.

Table 3.5 Experimental results. These figures show the times that the alogorithm correctly associate floor senosor data to each person.

Expermental data		Number of correct associations	
Type	Number of data	By floor sensor	By integration
cross	7	7	7
pass	6	2	6

3.5.2 Results

We performed two people tracking experiments based on the proposed method and computed correct association rate between observation of floor sensors and target tracks. For comparison, the results by only floor sensors (section 3.4) are computed.

The results are shown in Table 3.5. For ”cross” data, the baseline method correctly tracks all case until the end. By introducing velocity term in the state of Kalman Filter, the method succeeded to track straight trajectories. For ”pass” data, the baseline method often failed to track people until end and failed to associate IDs to observed trajectories. In contrast, the proposed method tracks people and associate correct IDs for all case by integrating acceleration sensors.

The results are shown in Table 3.5. The figures in the table show the times each method correctly associated floor sensor data to each person.

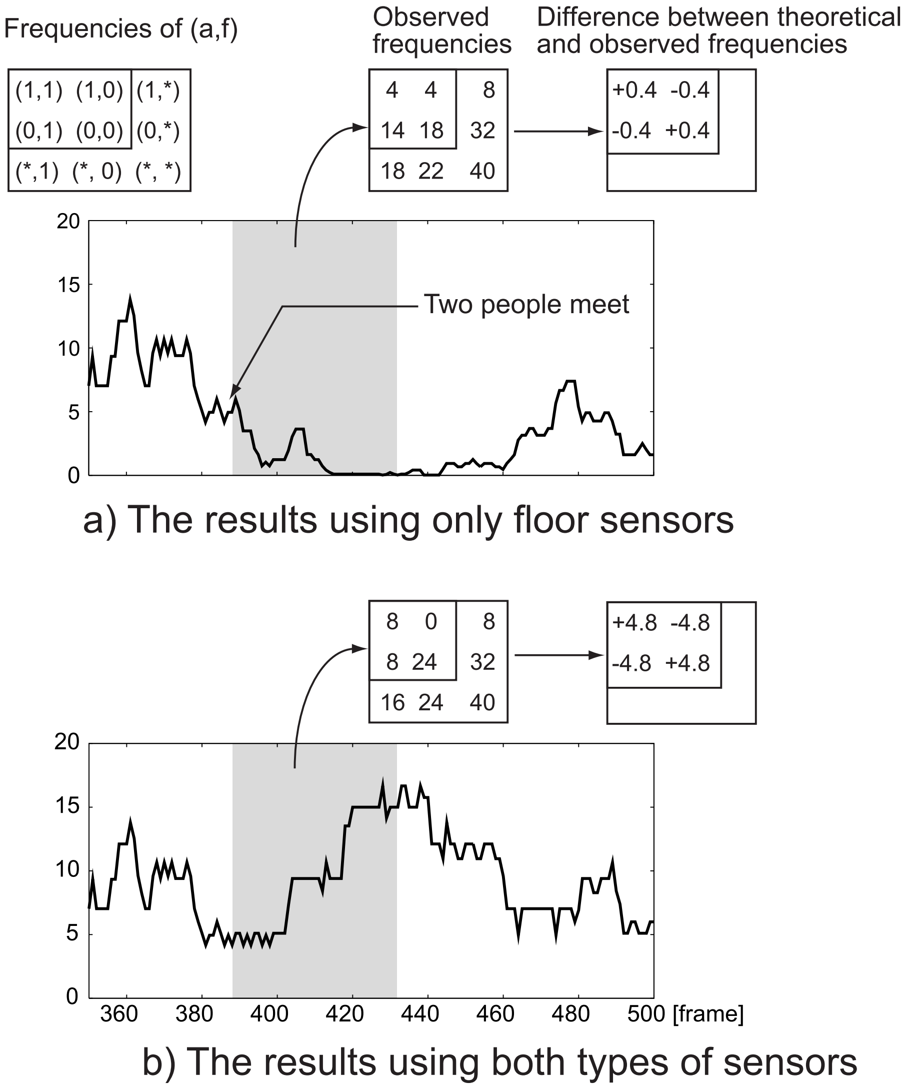
Figure 3.10 Experimental envirnment.

3.5.3 Change in the Statistics

Here we analyze the change in the statistics in one typical result that the propose method correctly associated and the baseline method failed. Figure 3.11 (a) shows the change in the statistics (Equation 3.3) by using the baseline method and (b) shows the proposed method. The arrow in figure shows the time two people approached. After the time, the statistics decreased in (a) since the tracking failure by using floor sensors. In contrast, the statistics increased in (b) by associating the hypothesis that maximizes correlation between the floor sensors and the acceleration sensor. Above each graph in Figure 3.11, a two-way contingency table and the difference between the theoretical frequency in shaded period are shown. In Figure 3.11 (b) there are large difference between the theoretical frequency that shows signals are not independent.

3.6 Discussion

Property of floor sensors Major specifications of floor sensors are spatial and time resolution, which are determined by the size of unit sensor (10 cm in the experiments) and



a) The results using only floor sensors

b) The results using both types of sensors

Figure 3.11 The changes of the statistics in Eq.(3). The two-way contingency tables after two people meet are also shown. The upper figure shows the value computed for the tracking result only by floor sensors. The lower figure shows the value when the acceleration sensor is integrated. In the integrated case, the hypothesis that signals are not independent is selected and the difference between the theoretical and the observed frequency is larger.

sampling period (0.125 seconds). They affect the performance of the proposed method With regard to the time resolution, the walking frequency of average adults is 1.75 Hz[53] (sampling period is about 0.57 seconds). To distinguish timings of footsteps of one person, sampling period should be no more than 0.5 second. Next, we discuss required sampling period to

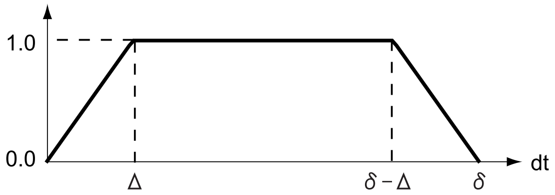


Figure 3.12 The probability that steps of two people are separately observed when the time difference of the steps is dt. Δ represents the sampling period, and δ represents the period of walking steps. This figure shows the case $\Delta < \delta/2$.

distinguish timings of footsteps of two people. Suppose sampling period of floor sensors is Δ. For simplicity, we assume period of footstep δ is same for both people. When the time difference of landing between two people is dt, floor sensors always distinguish two steps if $\Delta \leq dt$, distinguish probabilistically if $\Delta > dt$. Figure 3.12 shows the probability with respect to the difference of landing dt. As the Figure 3.12 shows, the average probability that floor sensors distinguish two steps is $1 - \Delta/\delta$. In the experiments the probability is 0.78 given $\Delta = 0.125$ and $\delta = 0.57$. However it is ideal case and the probability will decrease when there are observation error in sensors and observation delays as discussed in next topic.

In the observation in this experiment, floor sensor failed to observe landings in synchronized to the corresponding accelerometer once a few seconds. It decreases the probability to distinguish two steps from different people. On the other hand, the probability increases by observing more footsteps. Therefore, by observing footsteps in appropriate periods, the proposed method correctly distinguishes two people. However, longer observation will result in exponential increase of MHT hypotheses and computational resource. In this experiment, we determine based on the observation in five seconds.

With regard to the spatial resolution, in order to distinguish two close footsteps, smaller unit sensor is preferable. To distinguish two lined feet of 10cm, the size of unit sensor should be smaller than 10cm. The size in this experiment is 10cm. In observed data, two feet rarely become very close and no two feet contact. Assuming that two feet may approach to the closeness of 10cm, the size of unit sensor should be smaller than 20cm. However, smaller size of unit sensor will result in larger amount of data to transmit and longer sampling period. By improving response of floor sensors and introducing parallel sensing, we can realize shorter sampling period. Testing the proposed method by using floor sensors in different specifications is our future plan.

Signal delay Depending on the size of the environment, the delay in transmission of acceleration signals and it becomes difficult to synchronized observation. To cope with the problem, 1. estimate transmission delay in advance and correct delay when the delay is constant, 2. assign timestamps before transmission when the delay changes. In this experiment, since we used the time when the signal arrives at the host PC, the delay affect performance of the proposed method. In fact, acceleration signals sometimes do not synchronize to the footsteps of the same person observed by using floor sensors. To evaluate synchrony in stable manner, we currently observe signals enough time. Improving the accuracy of the timestamps will result in the shorter observation time to evaluate synchrony. It is desirable to use timestamps in the wearable device and it is one of our plans. In addition, more flexible evaluation method including modeling delay is promising approach.

Computational complexity The computational cost of the proposed method is proportional to the number of MHT hypotheses. When the positions of people are not close, association is not ambiguous and number of hypotheses does not increase. When people get close, the number of hypotheses increases during the approach of two people. In this experiment, it is difficult to compute online when the number of hypotheses becomes large. When the number of people increases, the computational cost grows accordingly. Furthermore, it becomes more difficult to distinguish different footsteps when number of people becomes larger, which result in increased number of hypotheses. To cope with the problem, it is effective to use floor sensors that observe smaller sampling period to measure timing of footsteps with high precision. Shorter sampling period realizes accurate decision of synchrony and shorter time to accumulate enough statistics and finally smaller computational cost.

Recovery from false association Since the proposed method associates signals by evaluating synchrony in enough length of observations, the method find correct association. However once it associates wrongly after people go across, it is necessary to recover from false association. For example, when a person changes his direction suddenly and the motion model does not cover the strange motion, floor sensors do not generate correct hypothesis of trajectory and the method does not generate correct results. To cope with the problem, it is possible to recover correct association by re-estimating association after the distance between people becomes larger. Now floor sensors correctly estimate each correct trajectory and association between signals will work perfectly. Since our system repeatedly re-estimate association based on the latest observations, it recovers correct association though it wrongly associate once.

Using many accelerometers In this experiment, we assume each person carries one accelerometer. It is promising to put another accelerometer to shoes, which enable to observe walking motion in more details. More accelerometer will enable us to understand various daily behaviors not only walking. When we increase number of wearable sensors, it is important that we easily carry the sensors. Future plan includes proposing a general framework of adding sensors in a flexible manner depending on the purpose.

3.7 Conclusion

In this section, we proposed a method that associates binary signals based on signal correlation. To compute correlation between floor sensors that each unit sensor generates '0' or '1' and accelerometers that observe continuous acceleration, we extract binary signals that represent footsteps from each sensor and associate them based on statistical test. We applied the proposed association method to track multiple people by associating floor sensors and acceleration sensors that are attached to the human body. By associating trajectories observed by using floor sensors and IDs from wearable accelerometer, our method estimate both positions and IDs of people. Since many cellular phones have an acceleration sensor, the proposed approach is realistic for this application. By using only floor sensors, it is difficult to estimate correct associations between observations and people since floor sensors does not provide any ID information and trajectories are ambiguous. By selecting association hypotheses that maximizes correlation, the correct association hypothesis is estimated. In experiments, significant improvement in correct association rate is achieved.

Chapter 4

Tracking and Identifying People

Different kinds of sensors have different characteristics of reliability depending on situation. To associate signals of different kinds of sensors in stable manner, it is important to consider the reliability of observation of each sensor. In this chapter, we focus on association of LRFs (laser range finders) and wearable gyroscopes to track and identify each person, and propose an association method that consider reliability of observations of LRFs.

4.1 Related Works

4.1.1 Locating Pedestrians Using Environmental Sensors

LRFs have recently attracted increasing attention for locating people in public places. As they have become smaller, it becomes easier to install them in environments. Since LRFs observes only the positions of people, installation of LRFs does not raise privacy issue. Cui et al.[54)] succeeded in tracking a large number of people by observing feet of pedestrians. Zhao and Shibasaki[55)] also track people by using a simple walking model of pedestrians. Glas et al.[56)] placed LRFs in a shopping mall to predict the trajectories of people by observing customers at the height of waist. In summary, LRFs placed in the environment are good at locating people precisely. However, it is difficult to use them to identify pedestrians when they are walking in a crowded environment.

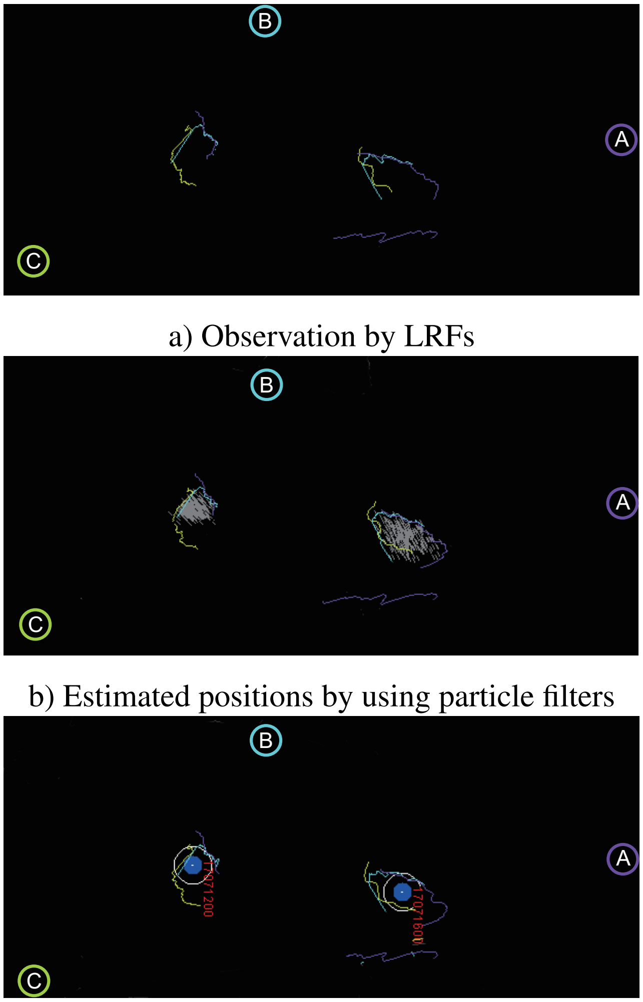
4.1.2 Locating People by Using LRFs

In this chapter, we use a network of LRFs to estimate positions of people. LRFs are one of promising devices that provide positions of people since they estimates positions accurately in large crowded locations day or night, outdoors and indoors. As they have become smaller, it becomes easier to install them in environments. Since LRFs observes only the positions of people, installation of LRFs does not raise privacy issue. Cui et al.[54] succeeded in tracking a large number of people by observing feet of pedestrians. Zhao and Shibasaki[55] also track people by using a simple walking model of pedestrians. Glas et al.[56] placed LRFs in a shopping mall to predict the trajectories of people by observing customers at the height of waist. Our method is based on Glas et al.,[56] and we expands the methods according to each experiment.

Figure 4.1 (a) shows raw observations by using LRFs. An LRF measures distance from sensor to close targets. The sensor we mainly uses in experiments is UTM-30LX (Hokuyo Automatic), which observes the distance of 30m and the angular range of 270 degrees in specification (Table 4.1). A background model is first computed for each sensor by analyzing hundreds of scan frames to filter out noise and moving objects. Points detected in front of this background scan are grouped into segments within a certain size range and ones that persist over several scans are registered as human detections. Each person is then tracked by the particle filter using a linear motion model (Figure 4.1 (b)). Likelihood is evaluated on the basis of the potential occupancy of each particle's position. For example, humans cannot occupy spaces that have been observed to be empty. Figure 4.1 (c) shows observed positions of people in the environment with anonymous IDs. This tracking technique provides quite stable and reliable position data, with a position error 6 cm. Further details on this algorithm are presented in.[56]

4.2 Associating LRFs and Werable Accelerometers

To associate signals from different kinds of sensors, it is important to evaluate confidence of observations since each sensor has different characteristics of observations. In this section, we associate LRFs installed in the environment and wearable accelerometers and locate



a) Observation by LRFs

b) Estimated positions by using particle filters

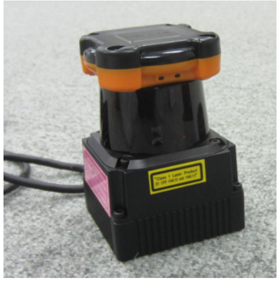
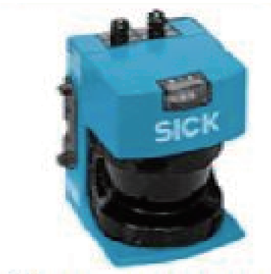
c) Estimated positions with anonymous IDs

Figure 4.1 Person position estimation using LRFs

positions of people who carry accelerometers. We focus on angular velocity signals around the vertical axis that are observed from environmental and wearable sensors. After angular velocity is observed from two types of sensors, signals are compared to determine whether two signals come from same person. In this framework, the problem of locating the person with a wearable sensor is to compare the signal from the wearable sensor to all signals from the people detected by environmental sensors and selects the person with the most similar signal (Figure 4.2).

However, observed angular velocity by using LRFs is not reliable depending on the be-

Table 4.1 Specification of LRFs used in experiments

Product name	Hokuyo Automatic UTM-30LX	SICK LMS200
Scanning angle [degree]	270	180
Angular resolution [degree/sec]	0.25	0.25,0.5,1.0
Sampling frequency [Hz]	40	75.0, 37.5, 18.75
Range [m]	30	80 (max)
Measurement accuracy [mm]	± 30	± 15 - 40
		

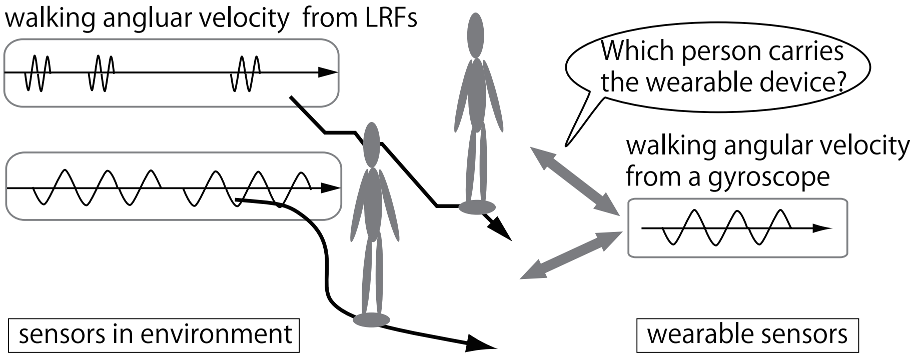

walking angluar velocity from LRFs
Which person carries the wearable device?
walking angular velocity from a gyroscope
sensors in environment
wearable sensors

Figure 4.2 Locate a person carrying a specific gyroscope by computing correlation between wearable and environmental sensors.

havior of people. To associate LRFs and wearable accelerometer accurately, we propose to evaluate confidence of observation of LRFs and introduce association method based on the evaluation.

4.2.1 Estimating Angular Velocities by Using Environmental Sensors

Our method expands the tracking algorithm explained in section 4.1. For each observed trajectory, angular velocity is computed as:

$$\begin{aligned} \mathbf{v}(t) &= (\tilde{\mathbf{x}}(t) - \tilde{\mathbf{x}}(t-1))/\Delta \\ \theta(t) &= \arg(\mathbf{v}(t)) \qquad (4.1) \\ \omega_L(t) &= (\theta(t) - \theta(t-1))/\Delta \end{aligned}$$

where $\tilde{\mathbf{x}}(t)$ is smoothed position vector, $\mathbf{v}(t)$ is velocity vector, Δ is sampling period, $\arg(\mathbf{v})$ represents a function that returns direction of $\mathbf{v}$, $\theta(t), \omega_L(t)$ are moving direction and angular velocity, respectively. In general, the position and angular velocity of a person can change independently. However, when people walk in daily lives, changes in angular velocities could be mainly caused by changes of the walking directions. In fact, we found angular velocity estimated by using LRFs are similar to that observed by using wearable gyroscopes (Figure 4.3).

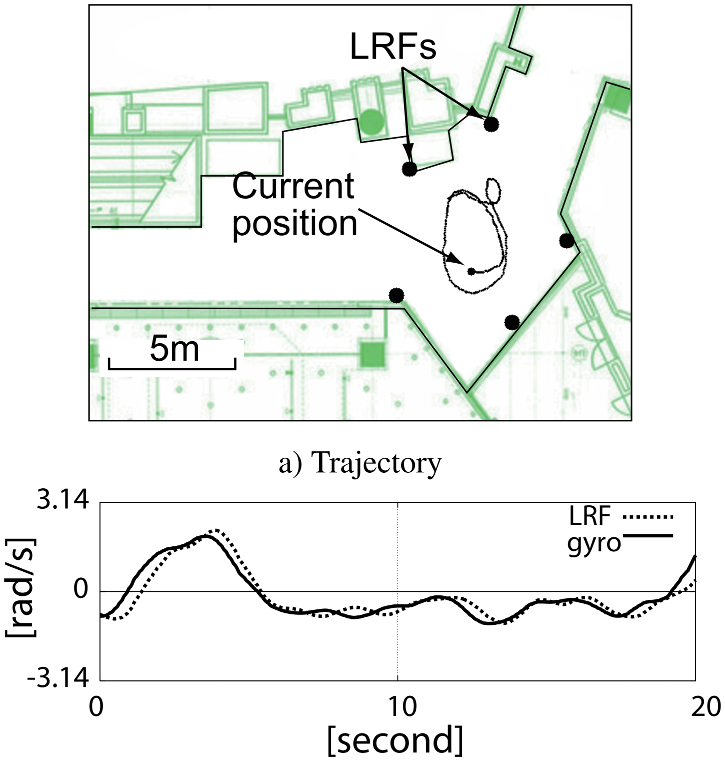
4.2.2 Estimating Angular Velocities by Using Wearable Inertial Sensors

Next we estimate angular velocities around the vertical axis of people who carry wearable accelerometers. Angular velocities around three axes are observed for each person by using body-mounted 3-axis gyroscopes. To estimate component around vertical axis of angular velocities ω_G,

$$\omega_G = \mathbf{\Omega} \cdot \mathbf{e_z} \qquad (4.2)$$

where $\mathbf{\Omega}$ is observed angular velocity vector, and $\mathbf{e}_z$ is the unit vector of the vertical axis. The suffix G of ω_G represents that it is estimation by using gyroscopes. In principle, $\mathbf{e}_z$ is estimated by integrating the angular velocity signals,[57] but we need initial posture of the sensor and it is difficult to estimate accurate $\mathbf{e}_z$ since drift error grows with time. Therefore, we use accelerometers and compute the short-time average of the observed accelerometers to estimate $\mathbf{e}_z$:

$$\hat{\mathbf{e}}_z(t) = -\frac{1}{Lg} \sum_{\tau=t-L+1}^{t} \mathbf{a}(\tau) \qquad (4.3)$$



a) Trajectory

b) Angular velocity

Figure 4.3 An example signals from LRFs and a gyroscope in 20 seconds. a) Estimated trajectory using LRFs. b) Estimated angular velocities. The vertical axis is the angular velocity. Two signals are quite similar.

where **a** is the acceleration vector and g is the gravitational constant. In the experiments, we set the length L to the number of samples for eight seconds. Though this estimation is incorrect when people are walking, it does not suffer from drift error. In preparatory experiments, we confirmed that this simple averaging can be used to estimate $\mathbf{e}_z$ for our purpose. When body motion is measured using inertial sensors, the sensor's attachment position is important. In preparatory experiments, we tested three different attachment positions: on the head, chest, and waist. We found that the results for the head-mounted sensor were noisy, while the results for the other positions were adequate and almost the same. In the following experiments, the inertial sensor was placed on the person's chest.

4.2.3 Integration by Evaluating Correlation Between Angular Velocity Sequences

Suppose many people carry wearable gyroscopes, we observe angular velocity $\omega_G^{(j)}$ from sensor j. At the same time, by using LRFs in the environment, we estimate angular velocity $\omega_L^{(i)}$ from observed trajectory of a person i. By associating two sensors (i, j) that observes same person, we can estimate the position of a person who carries specific sensor j.

Before evaluating correlation between two sensors, we align sampling period of signals. Since generally sampling periods of sensors are different, we average signals from sensors with shorter sampling period and generate signals with same sampling period as another sensors. Here we average signals from gyroscopes and generate averaged angular velocity $\omega_G^{(j)}(t)$ with same sampling period as $\omega_L^{(i)}(t)$ from LRFs. A simple method to evaluate correlation between $\omega_G^{(j)}(t)$ and $\omega_L^{(i)}(t)$ is:

$$f_1(i, j) = \frac{1}{T} \sum_{t=1}^{\mathrm{T}} |\omega_G^{(j)}(t) - \omega_L^{(i)}(t)| \tag{4.4}$$

where Equation （4.4） is cost function and smaller cost means higher correlation. T is time length to compute correlation (five seconds in experiments). We can estimate positions of the person who carries sensors j by estimating person i that minimizes cost function Equation （4.4） .

$$i^* = \underset{i}{\operatorname{argmin}}(f_1(i, j)) \tag{4.5}$$

when no observation is available from either types of sensors, a small constant is used instead of absolute difference of angular velocities in Equation 4.4. For each wearable device j, a trajectory i is associated and our system can estimate the location of the person with sensor j.

Smoothed angular velocity signals for 20 s from LRFs and from a gyroscope are shown in Figure 4.3. Though these signals were observed from different viewpoints, they are quite similar. Since observing a body's angular velocity by using a gyroscope is straightforward and free from drift error and since positions are estimated precisely using environmental

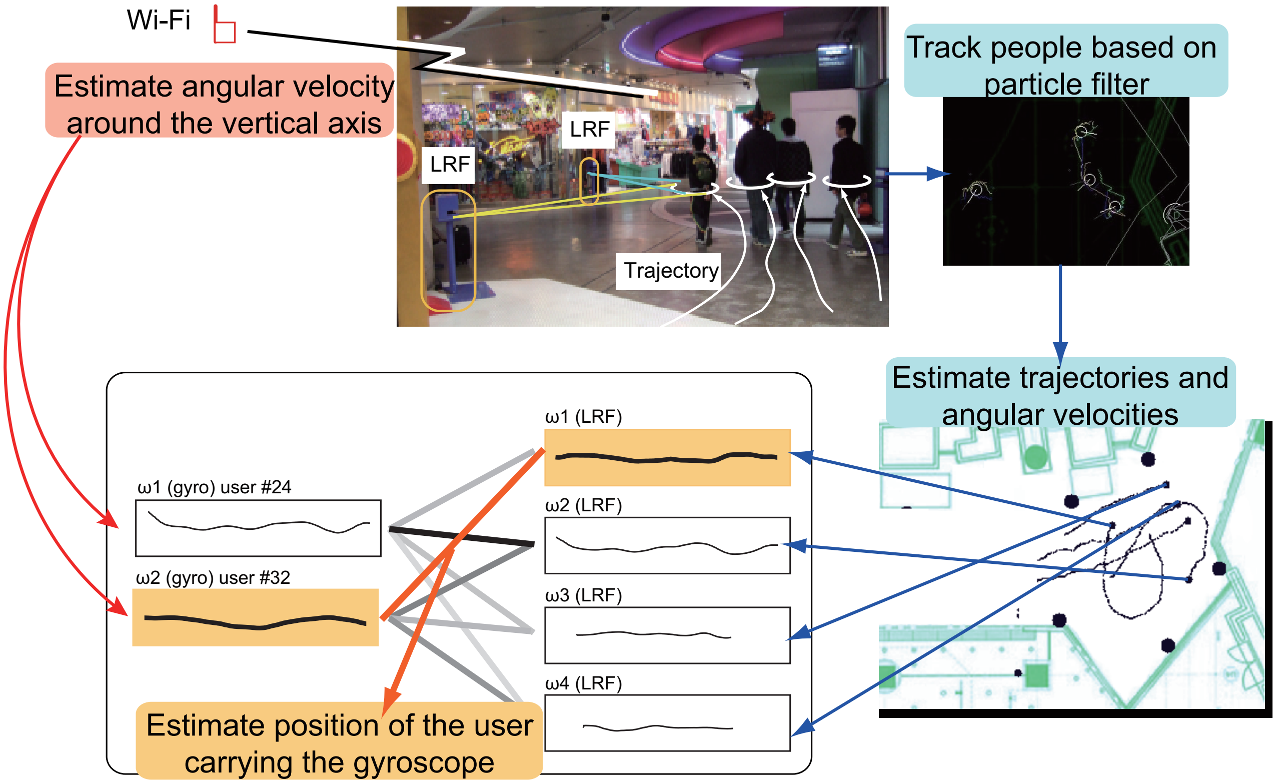

Wi-Fi
Estimate angular velocity around the vertical axis
LRF
LRF
Trajectory
Track people based on particle filter
Estimate trajectories and angular velocities
ω1 (LRF)
ω2 (LRF)
ω3 (LRF)
ω4 (LRF)
ω1 (gyro) user #24
ω2 (gyro) user #32
Estimate position of the user carrying the gyroscope

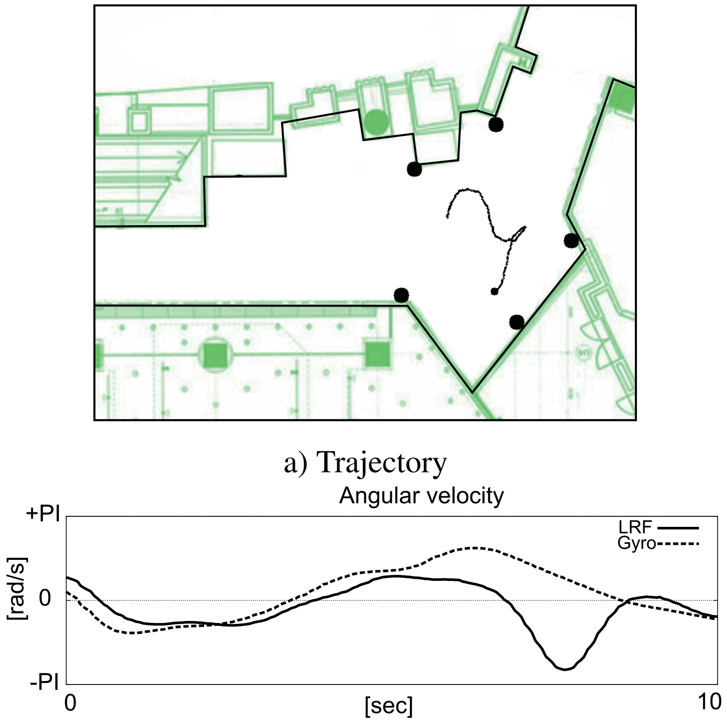
Figure 4.4 System overview

sensors, our method enables a robust and precise location system. Figure 4.4 summarizes the proposed system.

4.3 Sensor Integration Considering Confidence in Observation

4.3.1 A problem of Estimating Angular Velocity Using Position Sensors

The simple method of comparing angular velocities (Equation 4.4) does not always provide reliable results. Angular velocities estimated using LRFs and gyroscopes are shown in Figure 4.5. In the data for 20 s, the estimated angular velocities differed significantly when the person stopped and changed direction. This difference arises because the error in the angular velocity estimated using LRFs is larger when the velocity is low. In general, when a target's angular velocity is estimated using position-observing sensors, the confidence in the



a) Trajectory

b) Angular velocity

Figure 4.5 Example of signals produced for a low walking speed. a) Trajectory. The person stopped once and changed direction. b) Estimated angular velocity signals differed significantly when the person stopped.

estimated value depends on the target's velocity.

Typical changes in confidence while estimating direction are shown in Figure 4.6. When the position is observed with a certain precision, the direction is estimated from the difference between the subsequent positions. The estimated direction is limited to a certain distribution according to the target's velocity (Figure 4.6 (a)). However, if the velocity is low, the distribution is broad and the confidence is low (Figure 4.6 (b)). Since angular velocity is estimated from the difference in directions, the confidence in the angular velocity also depends on the target's velocity. When the velocity is close to zero, it is difficult to estimate angular velocity. This causes a problem when we locate people on the basis of Equation 4.4.

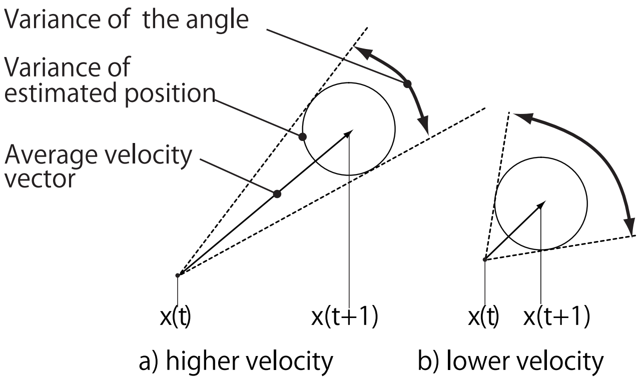


Figure 4.6 Relationship between the target's velocity and the variance of the estimated angle.

Evaluating Confidence in Observed Angular Velocity

When we observe positions by using sensors in the environment and estimate angular velocities, the confidence in the estimated angular velocity depends on the person's velocity. Since we cannot trust the estimated angular velocity when the velocity is low, a simple matching method using Equation 4.4 will fail to locate the person carrying a wearable sensor.

One approach for dealing with this problem is to consider the confidence in estimated angular velocity when matching angular velocities. To confirm the effectiveness of this approach, we introduce a simple cost function based on target's velocity. The cost function uses a simple heuristic but is a robust method of evaluating observation confidence. A weight term that depends on the target's velocity is added to Equation 4.4:

$$f_2(i, j) = \frac{1}{T} \sum_{t=1}^{\mathrm{T}} \frac{1}{\mathrm{w}(v)} |\omega_G^{(i)}(t) - \omega_L^{(j)}(t)| \tag{4.6}$$

where v represents person's velocity estimated by using LRFs, the term $\mathrm{w}(v)$ represents the uncertainty of the estimated angular velocity, which depends on the target's velocity. Larger w results in lager uncertainty. In the following experiments, we approximated the uncertainty of the angular velocity by using following formula:

$$\mathrm{w}(v) = \sin^{-1} \frac{\sigma_L}{v} \tag{4.7}$$

where σ_L is the fixed standard deviation of the position sensor's estimation error. As shown in Figure 4.7, Equation 4.7 is based on simple geometrical estimation.

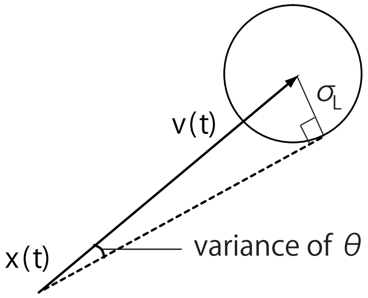


Figure 4.7 Estimation of variance of direction based on the observed position

4.4 Experiments

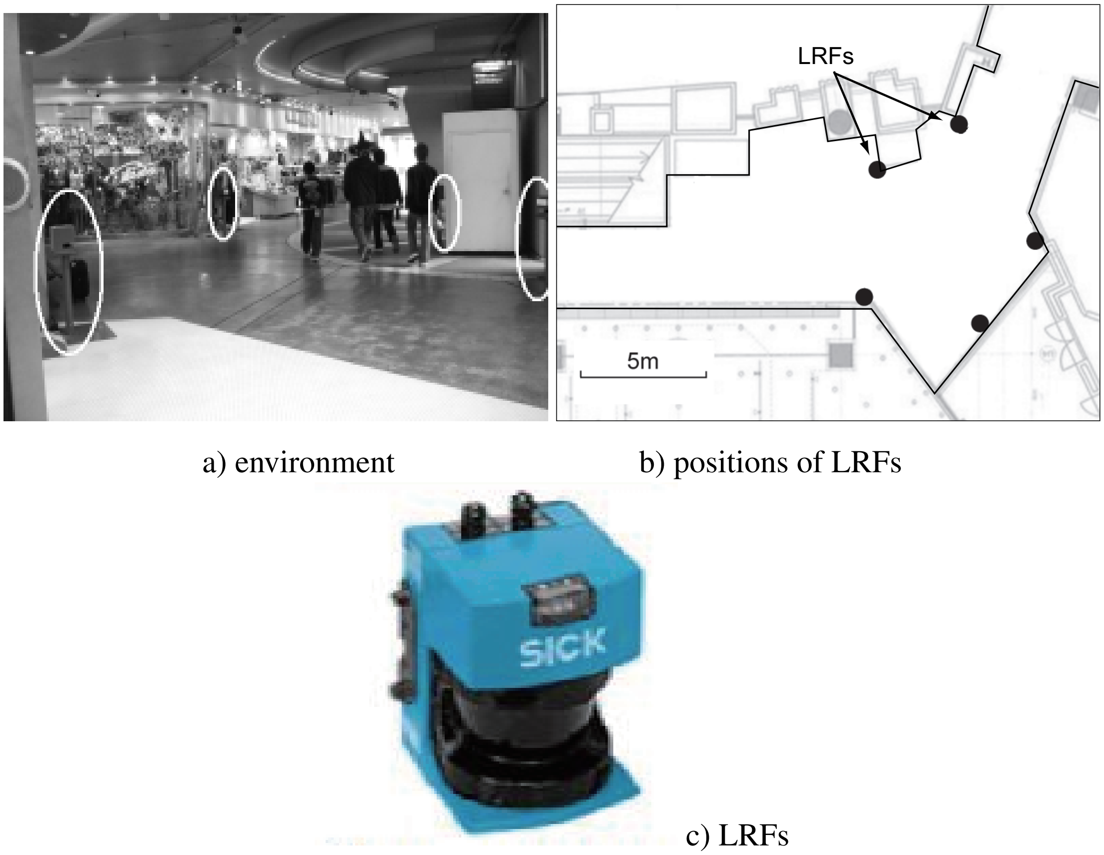
4.4.1 Experimental Setup

We conducted experiments in an entertainment/shopping arcade located near the entrance to Universal Studios Japan, a major theme park. We located people in a 20-m-radius area of the arcade containing shops selling clothing and accessories on one side and an open balcony on the other side. People in this area were monitored via a sensor network of consisting of five SICK LMS-200 LRFs mounted at a height of 85 cm (Figure 4.8 (c)). We expanded the system in[56)] by integrating wearable sensors to locate and identify people.

Each person in the environment was detected and tracked with a particle filter. By computing the expectation of the particles, we estimated the position and velocity 25 times per second. This tracking algorithm ran very stably and reliably with a measured position accuracy of less than 6 cm for our environment.[56)] Two people in the environment each carried one wearable sensor (WAA-006, ATR Promotions) with a three-axis gyroscope and a three-axis accelerometer (Figure 4.9). In the experiments, the observed angular velocity and acceleration signals were timestamped and sent to a host PC via Bluetooth.

Since our method locates people by comparing angular velocity time sequences, it is important to adjust the clocks of the LRFs and wearable sensors. In the following experiments, the wearable sensor clocks were synchronized with the host PC when they initially established a Bluetooth connection.

Another problem is the delay in the transmission from the wearable sensors to the host PC. In the following experiments, signals were sent with timestamps added by the wearable



a) environment b) positions of LRFs

c) LRFs

Figure 4.8 Experimental environment. The circles in the photograph show the locations of LRFs.

sensors. If the timestamp were set after the signals had been sent (e.g., by the host PC), the results would be affected by sudden transmission delays.

4.4.2 Results

Estimated angular velocities of a walking person

Figure 4.10 a) shows the estimated angular velocity of a person who walked around in the environment while carrying an inertial sensor. The angular velocity was estimated by two different methods: using LRFs and using a gyroscope. The two estimates were similar and changed in a correlated manner except for a few times (e.g. t =60 to 70, 100 to 110). Figure 4.10 b) shows the person's estimated walking speed. It is clear that significant differences

Product name	WAA-006 (ATR Promotions)
Weight	20 [g]
Sampling frequency	500 [Hz]
Range (Gyroscope)	± 300 [deg/s]
Range (Acceleration)	± 2G
Interface	Bluetooth

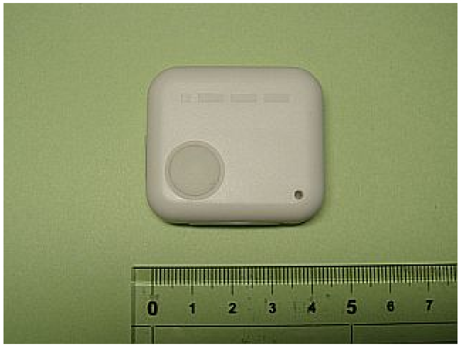
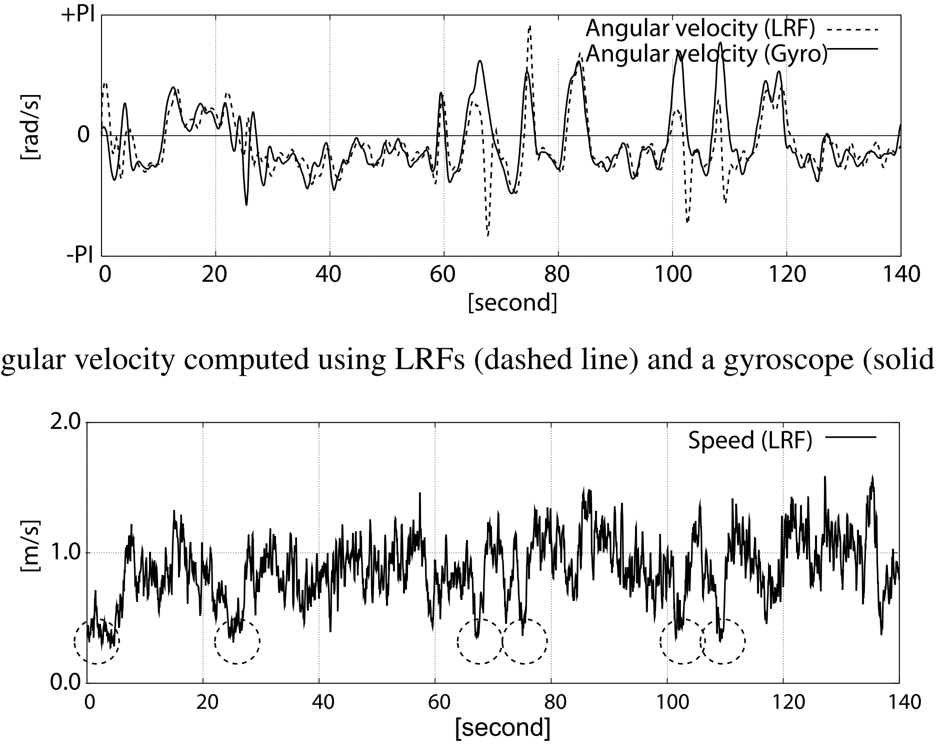


Figure 4.9 Wearable sensor device used in experiments.

between angular velocity estimates in 4.10 a) occurred only when the walking speed was very low (dashed circles in the lower graph).

Identification of target people based on cost function

Figure 4.11 shows the computed cost function between the sensor-equipped person and all the other people in the environment during a 20-s period based on Equation 4.6. The number of lines represents the number of people and smaller values represent higher correlation between signals. In Figure 4.11, the cost function of the target person (solid line) is clearly lower than those of the other people (dashed lines). This means that the cost function was lowest for the target person, who could be located very precisely using the tracking system using LRFs. These results show clearly how our algorithm distinguished the person carrying an inertial sensor when there were many people in the environment. Figure 4.11 c) shows the estimated angular velocities of one person by using two kinds of sensors, which show similar estimation results.



a) Angular velocity computed using LRFs (dashed line) and a gyroscope (solid line).

b) Walking speed using LRFs. When the person walks slowly (dashed circles), the angular velocity estimates differed significantly.

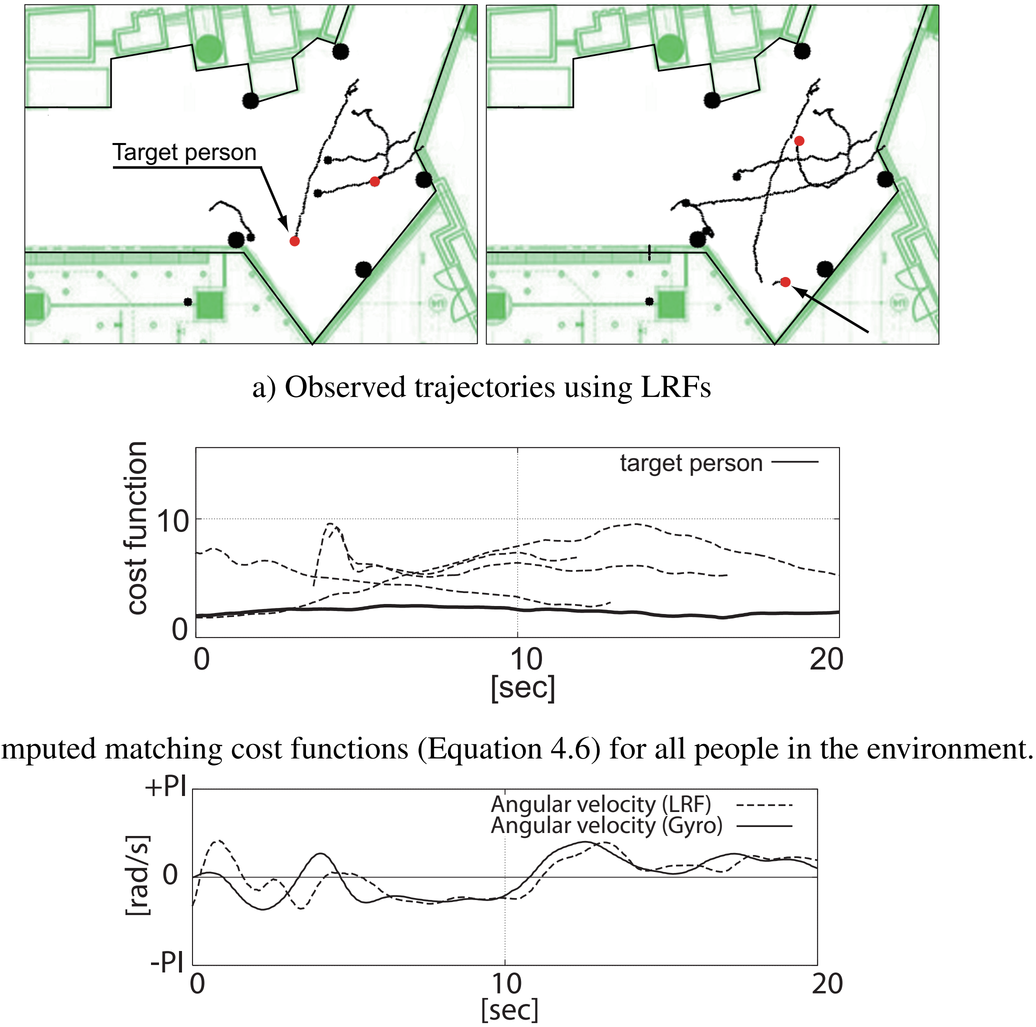
Figure 4.10 Estimated angular velocities and velocities of a user walking in the environment.

The effect of introducing the weight to the cost function

Figure 4.13 shows the effect of evaluating observation confidence by introducing our weight function. Figure 4.13 a) shows the cost function computed with a fixed averaged weight in Equation 4.6, Figure 4.13 b) with the proposed weight function. The solid line shows the cost function with the correct association. In the upper graph, the two lines sometimes touch and this could be the cause of failures. The lower graph enables the person to be distinguished from other people much more clearly.

Effect of the length of observations

Comparative results for various lengths of the computing cost function (parameter T in Equation 4.6) are shown in Figure 4.14. It was difficult to locate the person from only instant observation. When T was set to at least 64 frames (about 3 s), the person was located almost



a) Observed trajectories using LRFs

b) Computed matching cost functions (Equation 4.6) for all people in the environment.

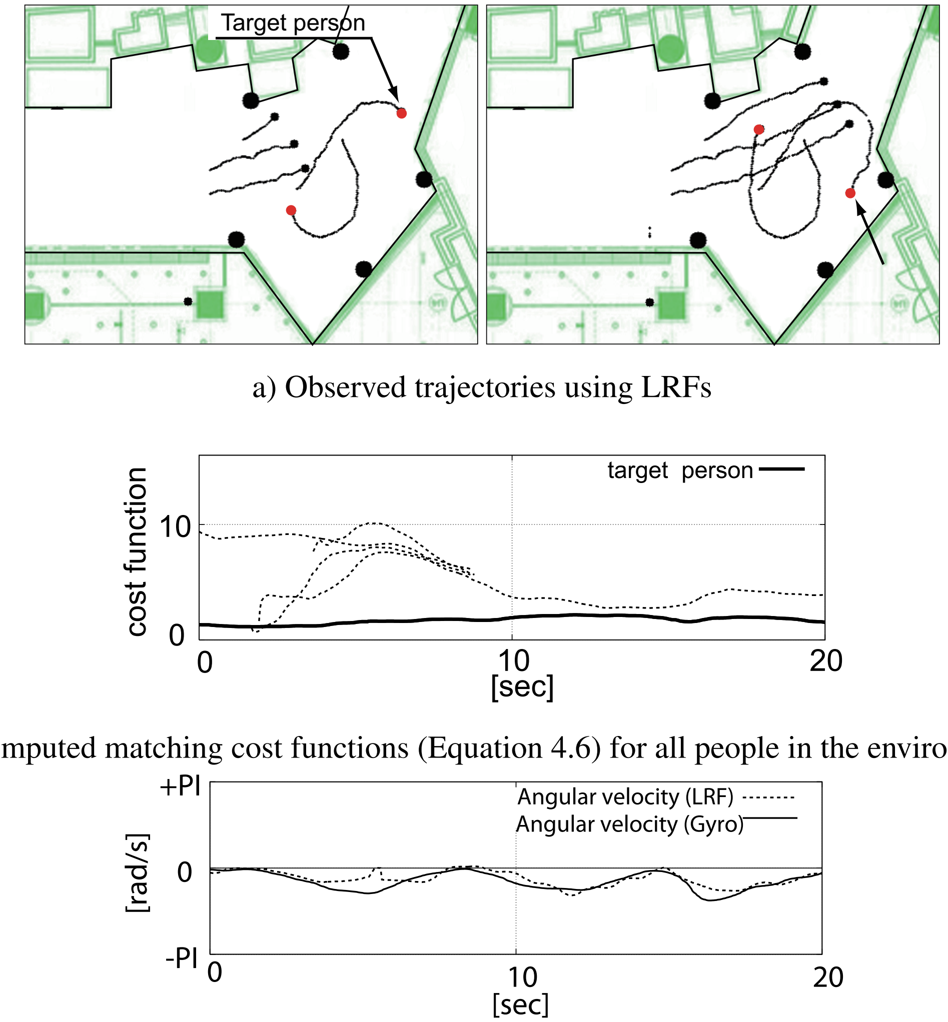
c) Observed angular velocities of a person using both LRFs and a gyroscope

Figure 4.11 Results for locating a person carrying a wearable sensor in an environment containing several people. The cost function for the person carrying the sensor was the lowest and this person was clearly located.

correctly. In the bottom graph in Figure 13, for which T was set to 192 frames (about 8s), the result is very clear.

4.4.3 Effect of Calibration Errors and Estimation Errors

Effect of error in time synchronization



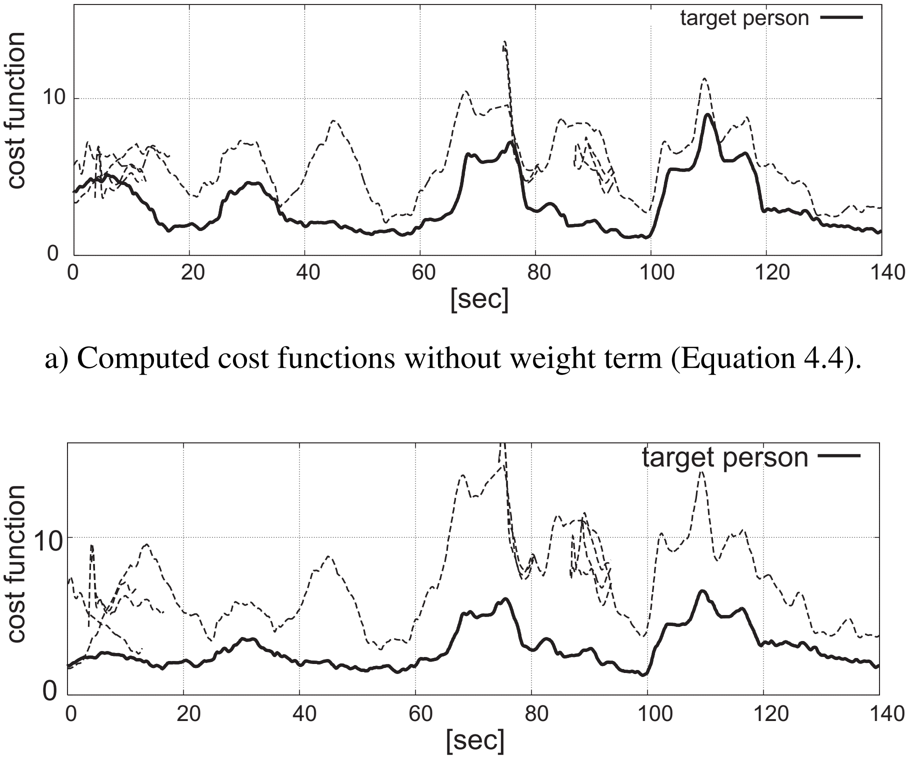
a) Observed trajectories using LRFs

b) Computed matching cost functions (Equation 4.6) for all people in the environment.

c) Angular velocity estimates for the person: they are very similar.

Figure 4.12 Results for locating another person carrying a wearable sensor in the same environment as in Figure 4.11.

The proposed method assumes time synchronization between wearable devices and sensors in the environment. In experiments, we synchronized time clocks of wearable devices to a host PC. However, in real system time synchronization is not perfect and investigation the change of accuracy in association due to the time synchronization error is important issue. We added error artificially to the time clock of a wearable device and evaluate error in association. Figure 4.15 shows the correct association ratio with respect to the error (based on Equation 4.6). When the error is positive, the clock of the wearable device is fast. The graph



a) Computed cost functions without weight term (Equation 4.4).

b) Computed cost functions with proposed weight term (Equation 4.6).

Figure 4.13 Effect of introducing the weight term into the cost function. When the cost function was computed using the weight term (b), the person could be distinguished from other people very clearly.

shows robustness of our method to the synchronization error. Figure 4.16 shows the result based on Equation 4.4 when we do not use proposed weight function. The graph shows a decrease in association ratio when the error is negative, and this results shows effectiveness of the proposed method.

Effect of calibration error of gyroscopes

In experiments, zero point offsets of gyroscopes are measured and calibrated. Since in real situation it is not realistic to calibrate scale factor errors that depends on temperature, we did not calibrate scale factor error. To confirm the effect of scale factor error, we introduced error artificially and investigate association accuracy. Figure 4.17 shows the correct association ratio with respect to the error. When the scale factor error is larger than 1.0, observed angular

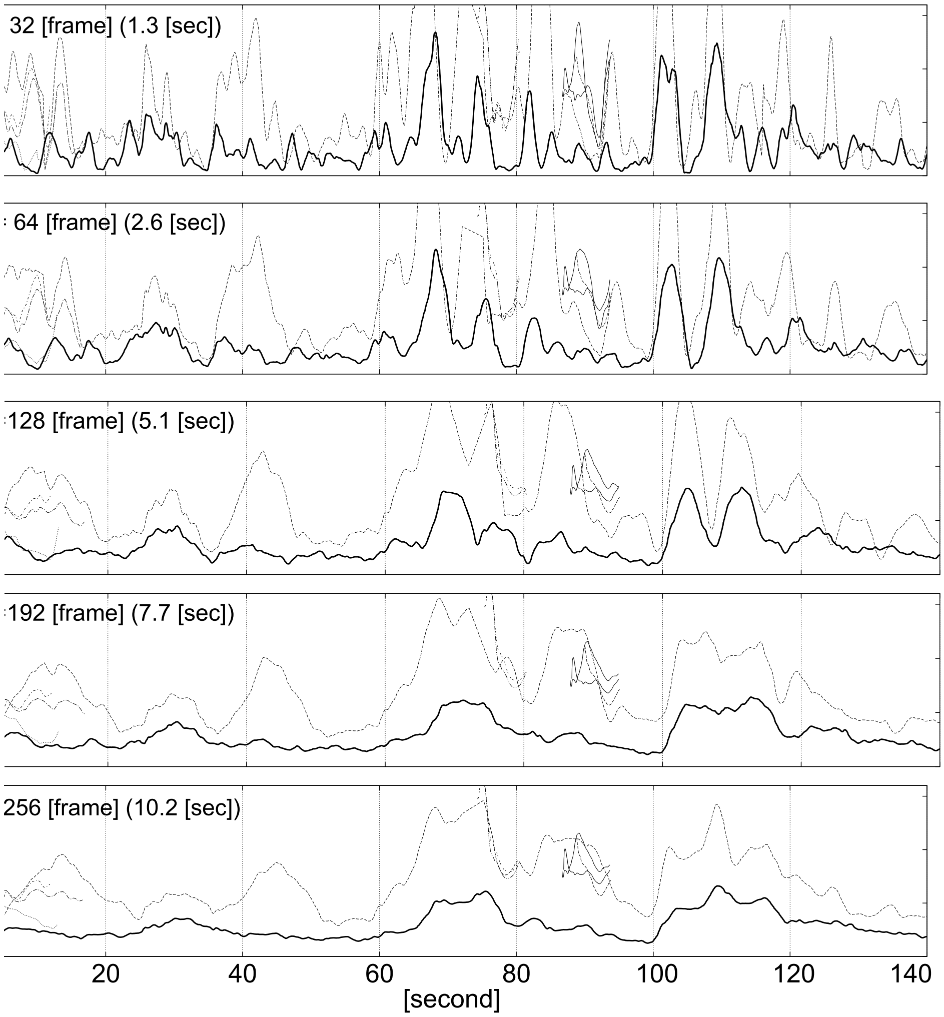


Figure 4.14 The cost function computed for different time period T. The costs are computed for all people detected using LRFs. The ID of gyroscope is associated to the trajectory with the lowest cost. These graphs represents results for T = 0.04, 1.3, 2.6, 5.1,7.7 [s] from the top to the bottom.

velocity is larger than the true value. The classification accuracy does not change when we introduced error in all axes. Figure 4.18 shows the result based on Equation 4.4 when we do not use proposed weight function. Still the accuracy is not very different and these results show the robustness of this application to the scale factor error.

Effect of estimation error of vertical direction

To estimate angular velocities around the vertical axis, we approximate the vertical di-

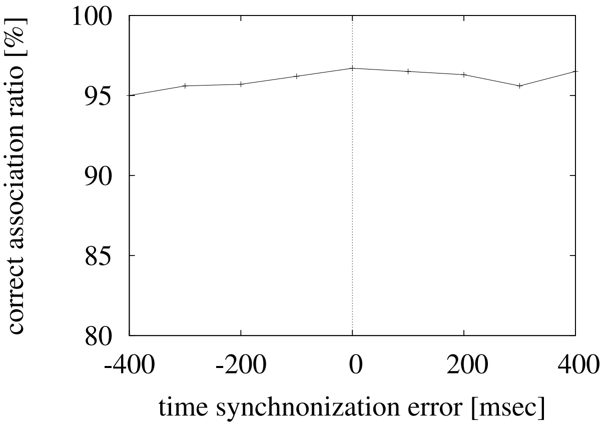


Figure 4.15 Effect of the time synchronization error in werable devices (proposed method)

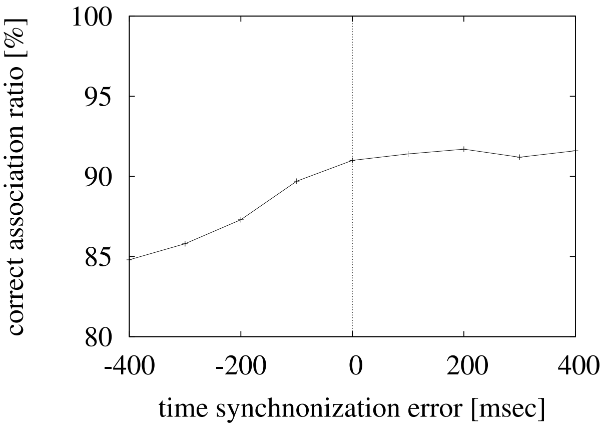


Figure 4.16 Effect of the time synchronization error in werable devices (fixed window)

rection based on Equation 4.3. To confirm the effect of the error in the estimated vertical direction, we added error to the estimated direction artificially.

Figure 4.19 shows the correct association ratio with respect to the error. The classification accuracy does not change when we introduced error in both x and y axes. Figure 4.20 shows the result based on Equation 4.4 when we do not use proposed weight function. Still the accuracy is not very different and these results show the robustness of this application to the scale factor error.

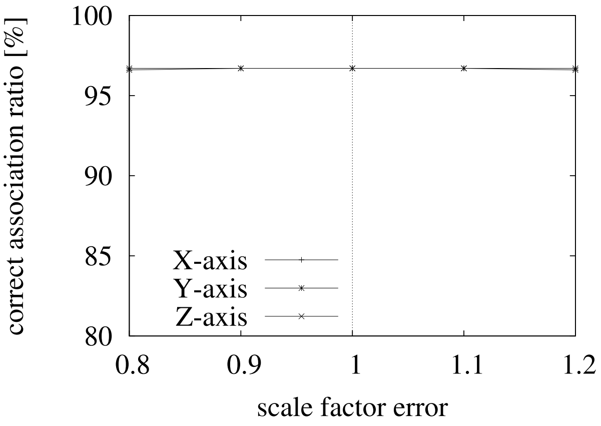


Figure 4.17 Effect of the scale factor error in gyroscopes (proposed method)

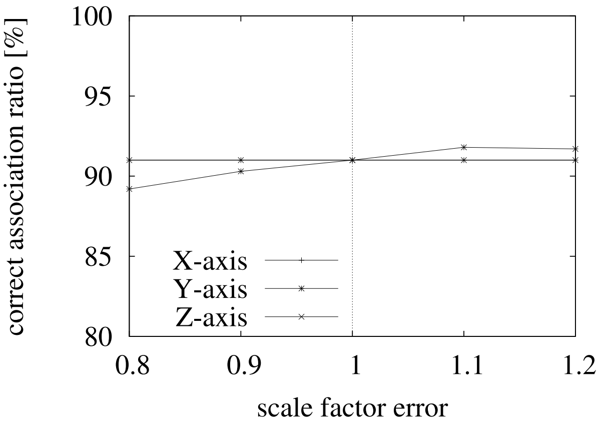


Figure 4.18 Effect of the scale factor error in gyroscopes (fixed window)

4.5 Conclusion

In this chapter, we propose a method to evaluate confidence of observations to associate different kinds of sensors. Since different kinds of sensors have different characteristics, it is important to consider observation error and confidence of observations to compute correlation between observed signals. We focused on association of LRFs and wearable accelerometers and proposed a weight function to evaluate confidence of observations by using LRFs.

LRFs are one of promising devices to estimate positions of people in crowded public lo-

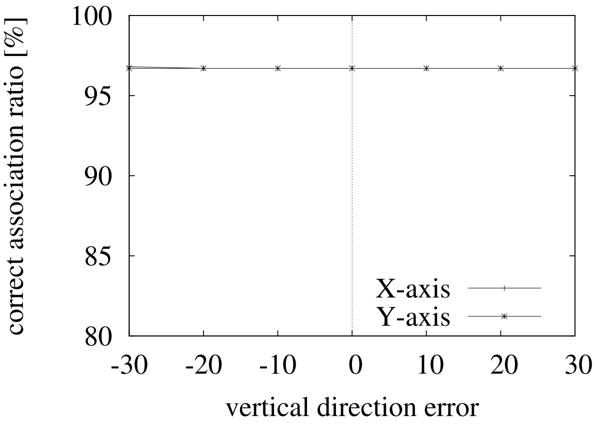


Figure 4.19 Effect of the estimaed vertical direction error (proposed method)

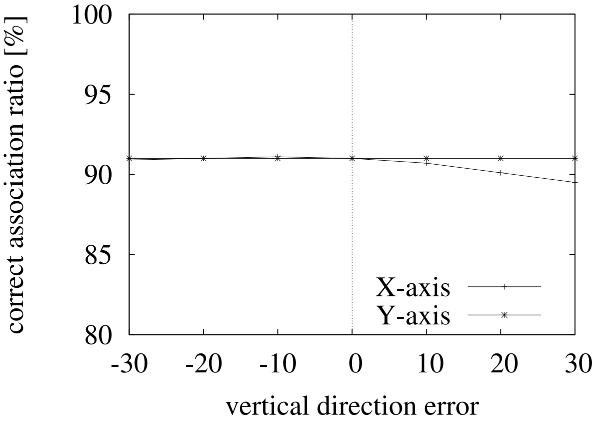


Figure 4.20 Effect of the error in the estimated direction (fixed window)

cations. A network of LRFs observes positions of people accurately in large area in both outdoors and indoors. However, it is difficult for LRFs to identify each person. By associating mobile devices to observations by using LRFs, the system can provide both accurate positions and IDs, which enable services that depend on positions in public locations. Proposed method can provide a fundamental infrastructure for such applications.

Chapter 5

Integration of Foot Motion and Trajectories Based on Phase Depend Features

5.1 Introduction

I nformation infrastructure that provides personal and location-dependent services in public spaces like a shopping mall permits a wide variety of applications. Such a system will provide the positions of friends who are currently shopping in the mall. When they have many bags, users will call a porter robot, which can reach them by using the location system. To enable location-dependent and personal services, we propose a system that locates and identifies a pedestrian, who carries a mobile information terminal, anywhere in a crowded environment.

Many kinds of location systems have been studied that provide the positions of pedestrians by using sensors installed in the environment. For example, location systems using cameras and laser range finders (LRFs) can track people in the environment very precisely. However, it is difficult to identify each pedestrian or a person carrying a specific wearable device by using only sensors in the environment.

On the other hand, in ubiquitous computing, many kinds of wearable devices have been used to locate people. Since a location system using ID tags requires the installation of many reader devices in the environment for precise localization, it is not a realistic solution in large public spaces. Wearable inertial sensors are also used to locate people, but the cumulative estimation error is often problematic. For a precise location system, it is important to integrate other sources of information.

In order to locate a pedestrian carrying a specific mobile device anywhere in an environment, a promising approach is to integrate environmental sensors that observe people from the environment and wearable sensors that locate the person carrying them. In this paper, we propose a novel method integrating LRFs in the environment and wearable accelerometers to locate people precisely and continuously. Since location systems using LRFs have been successfully applied for tracking people in large public spaces like train stations and the sizes of LRF units are becoming smaller, LRFs are highly suitable for installation in public spaces. Since many cellular phones are equipped with an accelerometer for a variety of applications, users who have a cellular phone do not have to carry any additional device. In chapter 4, LRFs and wearable gyroscopes are integrated based on body rotation around the vertical axis from both types of sensors. However, it was difficult to distinguish pedestrians who move in a line when the trajectories are similar. Another problem in chapter 4is the method's use of gyroscopes, even though cellular phones equipped with gyroscopes are not yet so common.

In this chapter, to cope with these problems, we propose a new method that extracts features from a bipedal walking pattern. LRFs observe pedestrians at the height of feet and estimate the positions of people and walking rhythms. The wearable accelerometer also observes walking rhythms. Since walking rhythms differ from person to person, the proposed method can distinguish pedestrians walking in a line, and it uses only an accelerometer in the wearable devices.

The rest of this chapter is organized as follows. First, we review previous studies. Then, we discuss a method of integrating LRFs and accelerometers and how it can provide reliable estimation. Finally, we discuss the application of our method to a practical system and present the results of an experimental evaluation.

5.2 People Tracking and Identification Using LRFs and Wearable Accelerometers

5.2.1 Associating Signals from Environmental and Wearable Sensors

To locate each person carrying a wearable sensor, we focus on correlation of signals that are observed from environmental and wearable sensors. After features of the motion are observed using two types of sensors, signals are compared to determine whether the two signals come from the same person.

In this framework, the problem of locating the person with a wearable sensor is reduced to comparing the signal from the wearable sensor to all signals from the people detected by the environmental sensors and then selecting the person with the most similar signal (Figure 5.1).

Suppose feet of pedestrians are tracked by using LRFs in the environment, and the motions of both feet are estimated (Figure 5.2). Simultaneously, the timings of footsteps are observed by using wearable accelerometers. If the signals from both kinds of sensors are from the same pedestrian, we can assume that the two signals are highly correlated, since they were originally generated from a common walking rhythm. We found the acceleration signal from wearable sensors and acceleration of both feet that are estimated from tracking results are highly correlated. In this paper we focus on walking behavior and propose an association method of signals from both kinds of sensors based on signal correlation.

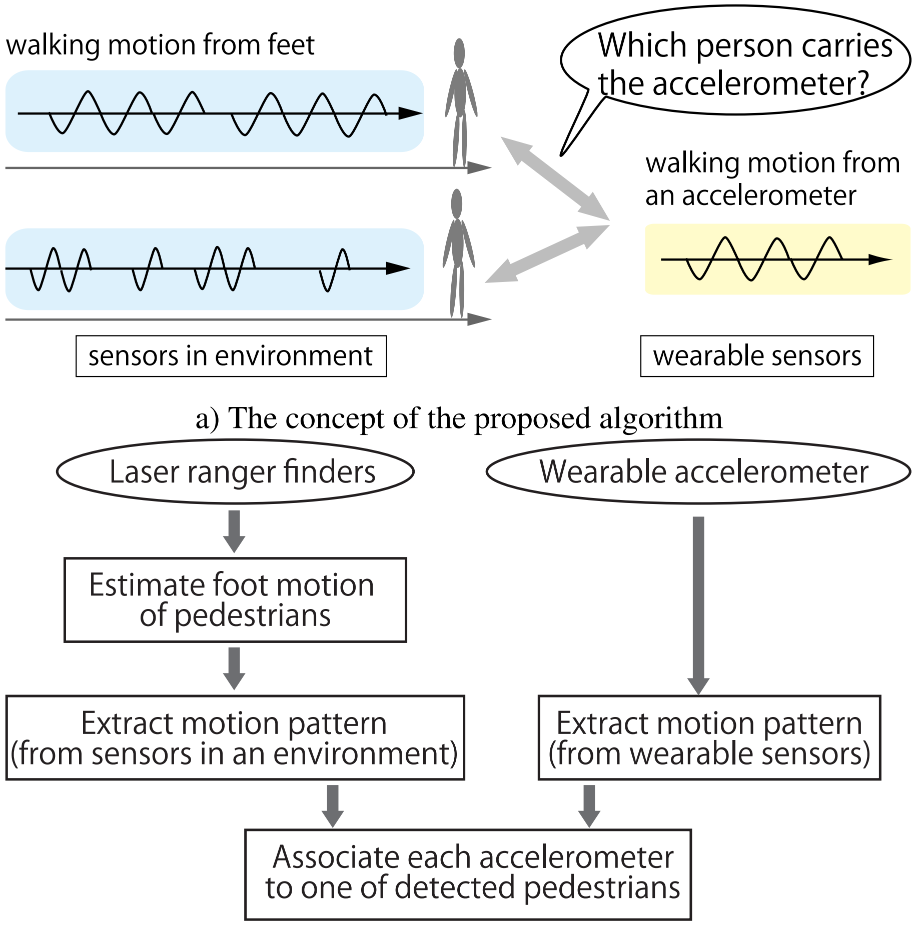
5.2.2 Tracking Biped Foot of Pedestrians by Using LRFs

Zhao and Shibasaki[55)] proposed a pedestrian tracking method by using LRFs at the height of the feet. By observing the feet of pedestrians, not only the positions of pedestrians but also the timing of their footsteps was observed. We modified our tracking system explained in 4.1.

Then we compute velocity and acceleration of each foot from tracked positions:

$$\mathbf{v}(t) = (\tilde{\mathbf{x}}(t) - \tilde{\mathbf{x}}(t-1))/\Delta, \quad \cdots\cdots (5.1)$$

$$a_L(t) = |(\mathbf{v}(t) - \mathbf{v}(t-1))/\Delta|.$$


walking motion from feet
Which person carries the accelerometer?
walking motion from an accelerometer
sensors in environment
wearable sensors
Laser ranger finders
Wearable accelerometer
Estimate foot motion of pedestrians
Extract motion pattern (from sensors in an environment)
Extract motion pattern (from wearable sensors)
Associate each accelerometer to one of detected pedestrians

a) The concept of the proposed algorithm

b) Flow of the proposed method

Figure 5.1 Locating a person carrying a specific wearable device by matching wearable and environmental sensors

where is smoothed position vector, is velocity vector, is acceleration of one foot. Suffix L represents that is an estimation from LRFs. represents sampling period.

To extract walking rhythm from the wearable accelerometer, we focus on the vertical component of the observed acceleration. The vertical acceleration $a_A(t)$ is estimated from three-dimensional acceleration vector $\mathbf{a}(t)$ and unit vector of vertical direction $\mathbf{e}_z(t)$ as explained in section 4.2.

Original and smoothed vertical acceleration signals are shown in Figure 5.3 (a). The accelerometer is attached to the left waist. One footstep of the walk is about 500 milliseconds in the graph, and the timing of the footsteps of both feet is clearly observed. Note that since

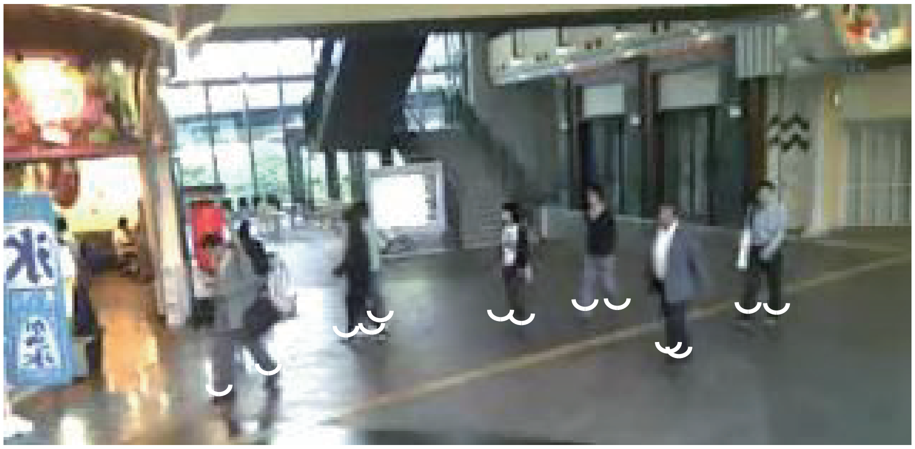


Figure 5.2 Pedestrian walking in a shopping mall. The white marks represent the detected feet of pedestrians. Motion of each pedestrian is observed using wearable accelerometer and LRFs by tracking each foot.

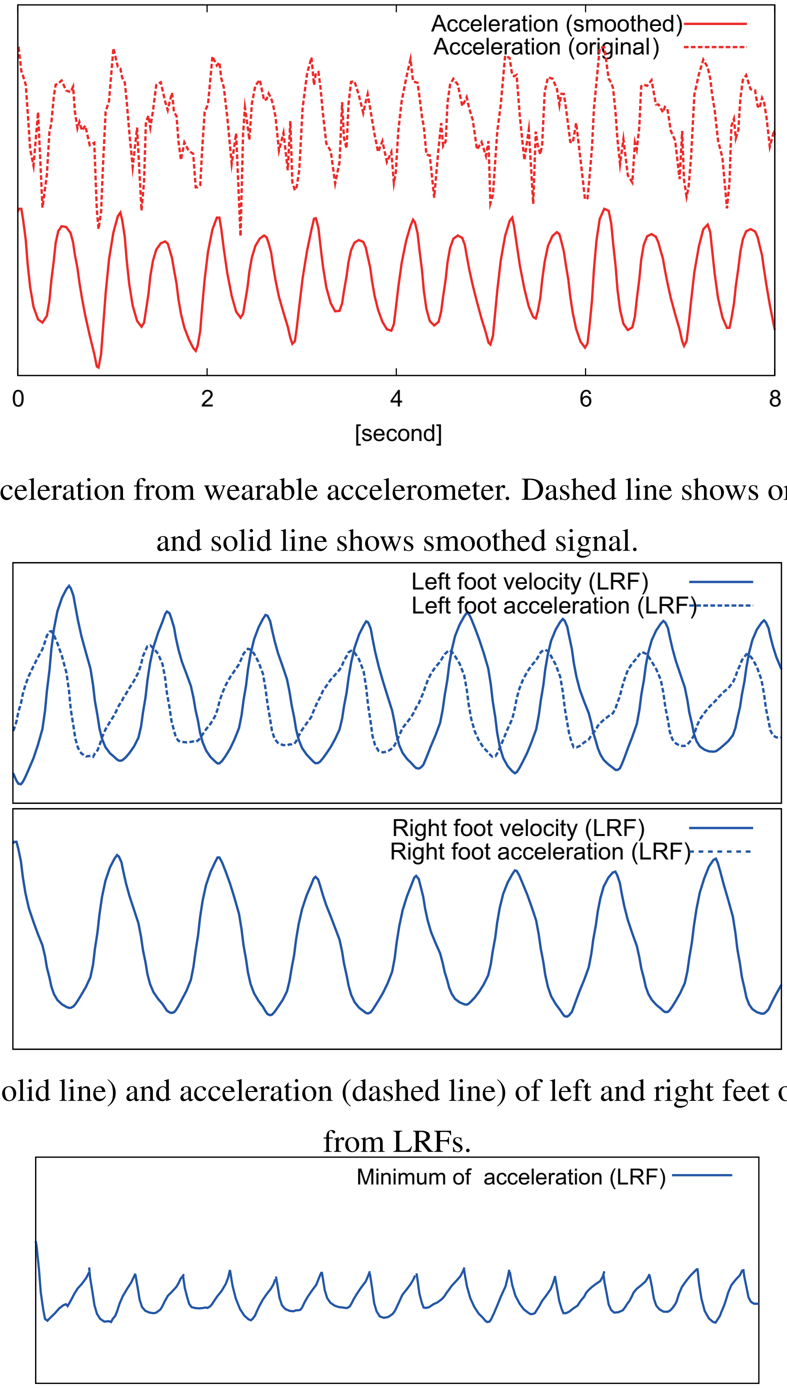
the accelerometer is attached to the left waist, the impact of a footstep of the left foot is clearer.

5.2.3 Associating Motion of Biped Foot and Body Acceleration

Figure 5.3 (b) shows the smoothed velocity and acceleration of each foot estimated by our tracking method. When the speed of a pedestrian's idling foot becomes lower and it finally lands on the ground, a large vertical acceleration is observed. Therefore, we can expect the impact of landing to be observed when the acceleration of the idling foot is negative. Note that since LRFs observe at the height of the leg, the velocity does not become zero when the foot lands.

Figure 5.3 (c) shows minimum of acceleration signals of both feet. This minimum of acceleration (Figure 5.3 (c)) and the vertical acceleration signal (Figure 5.3 (a)) are highly correlated (Figure 5.4).

To evaluate the correlation between the two signals, we propose computing Pearson's correlation function between the minimum foot acceleration from LRFs and the acceleration



a) Vertical acceleration from wearable accelerometer. Dashed line shows original signals and solid line shows smoothed signal.

b) Velocity (solid line) and acceleration (dashed line) of left and right feet of a pedestrian from LRFs.

c) Minimum of acceleration of both feet from LRFs (Equation (4)).

Figure 5.3 Examples of signals from LRFs and an accelerometer taken over eight seconds.

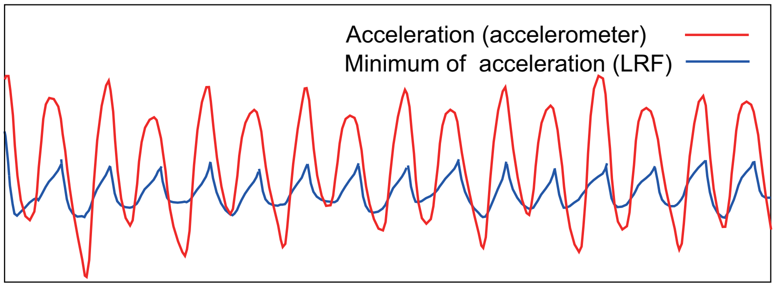


Figure 5.4 Examples of signals from LRFs and an accelerometer taken over eight seconds. Superimposed signal of acceleration from accelerometer Figure 5.3 (a) and LRFs (c)). Signals from same pedestrian shows clear correlation. The vertical axis is adjusted to overlap both signals.

from the accelerometer.

$$f_1(t) = \sum_{\tau=t-T+1}^{t} \hat{a}_A(\tau)\, \hat{a}_{\text{biped}}(\tau)/T, \tag{5.2}$$

where $\hat{a}_A(t)$ is normalized acceleration from wearable accelerometer, $\hat{a}_{\text{biped}}(t)$ is normalized signal of $a_{\text{biped}}(t)$, which is minimum of acceleration of right and left foot $a_{\text{left}}(t)$, $a_{\text{right}}(t)$ that are computed from Equation 5.2:

$$a_{\text{biped}}(t) = \min(a_{\text{left}}(t), a_{\text{right}}(t)). \tag{5.3}$$

For each wearable accelerometer, the trajectory of the person who is carrying the sensor is estimated by selecting the trajectory that maximizes Equation 5.2.

5.3 Evaluating Signal Correlation Depends on the Phase of Walking

In a crowded scene, sometimes only one foot of a pedestrian is observed because of occlusion. However, computing Equation (4) for acceleration from single foot results in low correlation. This is because the acceleration from a wearable device record landing both foot whereas the trajectory records motion of one foot. Figure 5.5 shows relation between acceleration signals from one foot and a wearable accelerometer. In one cycle of acceleration from single feet, signals shows both positive and negative correlation depends on the phase of walking. To cope with this problem, we propose correlation evaluation method that focuses on the phase in cyclic walking behavior.

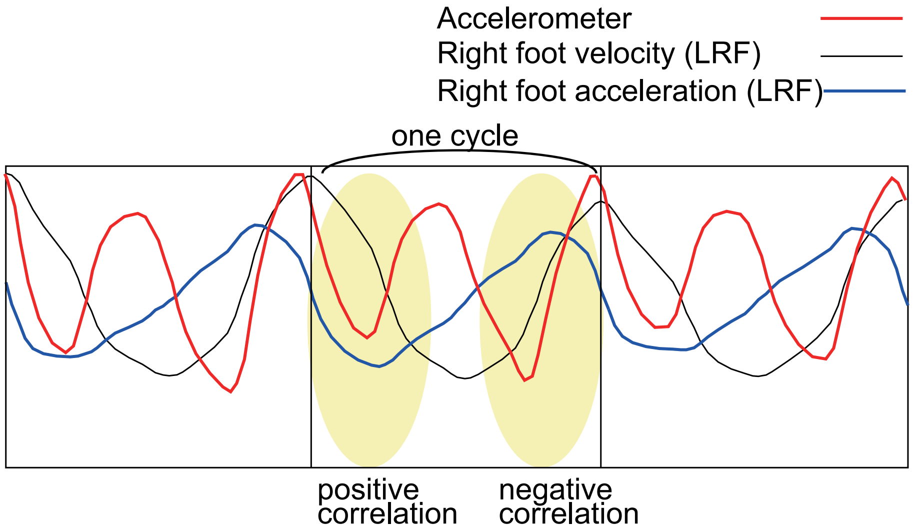


Figure 5.5 Relation between acceleration signals from one foot and a wearable accelerometer.

5.3.1 Relationship Between Acceleration of a Foot and Body

In order to associate cyclic signals that include both positively and negatively correlated part depending on the phase, we propose to learn weight coefficients that models signal correlation in each phase. Figure 5.6 shows computed correlation between observed acceleration signals in each of 16 phase periods, which is division of one cycle defined based on the peak of the foot velocity (See right foot velocity in Figure 5.5). We divided one cycle into 16 phase period. The horizontal axis in Figure 5.6 represents the phase period and the vertical axis is average of $a_A(t)\ a_L(t)$ in each phase period. There are clear positive correlation in earlier phase periods and negative correlation in latter phase periods. Figure 5.6 shows computed results for three subjects. This graph shows the variations among individuals are not significant.

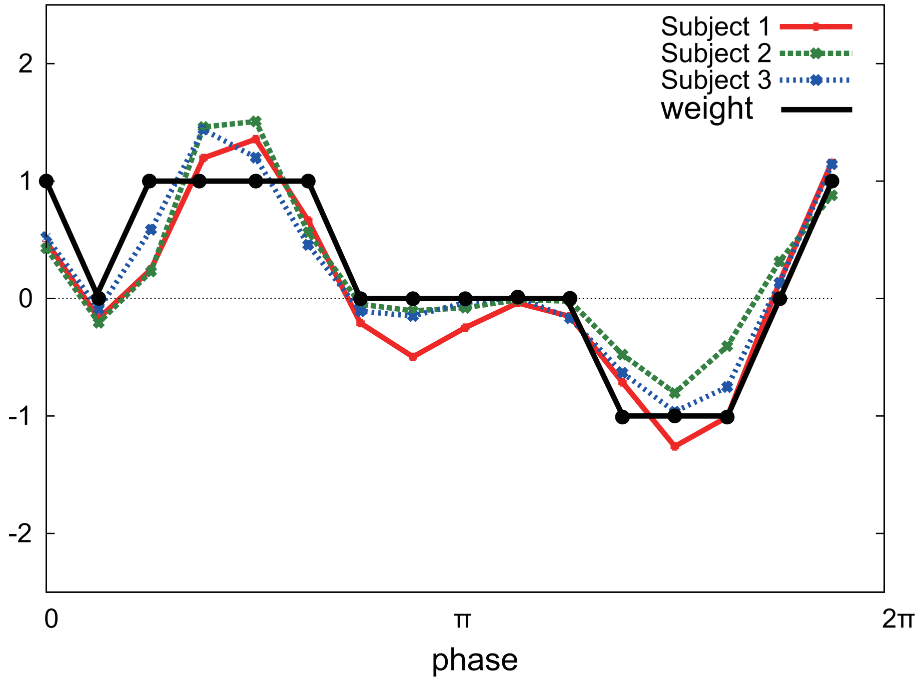


Figure 5.6 Correlation between observed acceleration signals depends on the phase in cyclic motion. The horizontal axis is the phase period which is division of one cycle (See peaks of right foot velocity in Figure 5.5).

5.3.2 Associating Acceleration from a Foot and Body Based on the Weight Depends on the Phase of Walking

Algorithm

1. Smooth observed velocity of a feet and extract local maximal and minimal value. Define one cycle as the period between local maxima. In each time in a cycle, define the phase $\phi(t)$ to zero at the time of local maxima, at local minima, and linearly interpolated phase at other time.

$$\phi(t) = \begin{cases} 0 & \text{velocity is local maxima at } t (= t_1), \\ \pi & \text{velocity is local minima at } t (= t_2), \\ \frac{t-t_1}{t_2-t_1}\pi & t_1 \leq t < t_2, \\ (1 + \frac{t-t_1}{t'_1-t_2}\pi & t_2 \leq t < t'_1, t'_1 \text{ is next local maxima.} \end{cases} \tag{5.4}$$

2. Divide one cycle into M phase periods ϕ_k ($k = 1 \ldots M$). We use $M = 16$ in experiments.

In each period k, compute average of Equation 5.2 in each phase period.

$$\mathrm{avg}_k = \text{average of } \hat{a}_A(t)\ \hat{a}_L(t). \tag{5.5}$$

where t is in ϕ_k and $\hat{a}_A(t)\ \hat{a}_L(t)$ are normalized acceleration from wearable accelerometer and tracking results using LRFs.

3. Define coefficients w according to the average. We use $\theta = 0.25$ in experiments.

$$w(\phi) = \begin{cases} +1 & \phi \text{ is in } \phi_k \text{ and } \mathrm{avg}_k > \theta, \\ -1 & \phi \text{ is in } \phi_k \text{ and } \mathrm{avg}_k \leq \theta, \\ 0 & \text{otherwise.} \end{cases} \tag{5.6}$$

By using the weight function defined in Equation 5.6, evaluate correlation between sensors:

$$f_2(t) = \sum_{\tau=t-T+1}^{t} a_A(\tau)\ a_L(\tau)\ w(\phi(\tau))/T. \tag{5.7}$$

Based on Equation 5.7, the trajectory of the user is estimated by selecting the trajectory that maximizes it. Figure 5.7 shows computed weight function in each phase period.

5.4 Experiments

5.4.1 Experimental Setup

We conducted experiments at a shopping mall in the Asia and Pacific Trade Center, in Osaka, Japan (Figure 5.8). We located people in a 20-m-radius area of the arcade containing many restaurants and shops selling clothing and accessories. People in this area were monitored via a sensor network consisting of six LRFs installed at a height of 20 cm (Figure 5.9). We a system in 4 designed for tracking a biped foot and expanded it to incorporate wearable sensors to locate and identify people.

Each foot of a pedestrian in the environment was detected and tracked with a particle filter. By computing the expectation of the particles, we estimated the position and velocity 25 times per second. This tracking algorithm ran very stably and reliably with a measured position. Three people in the environment each carried one wearable sensor with a three-axis accelerometer (same as 4.4). In the experiments, the observed acceleration signals were

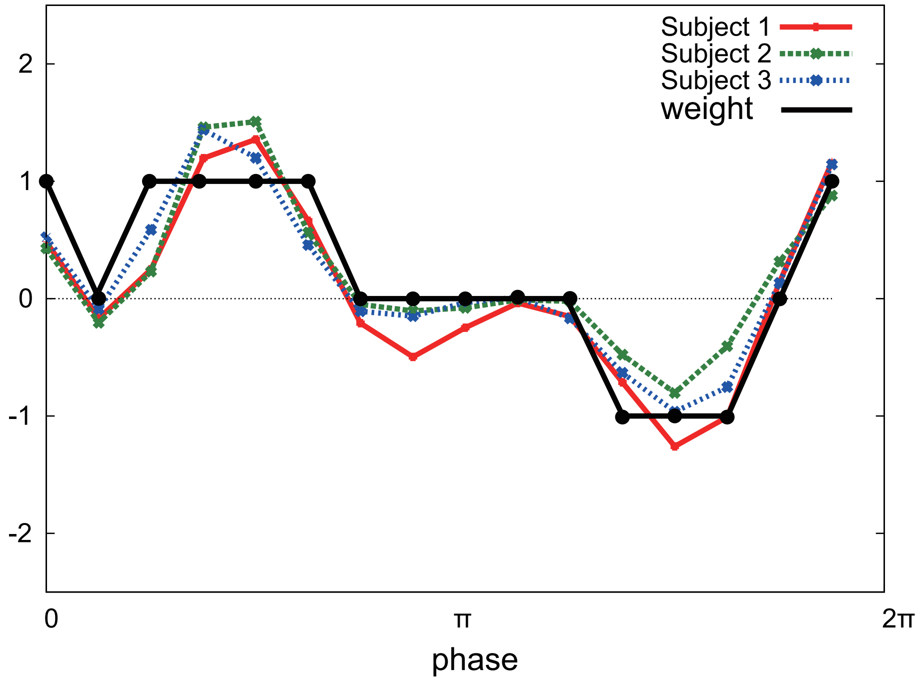


Figure 5.7 Trained weight function in each phase period.

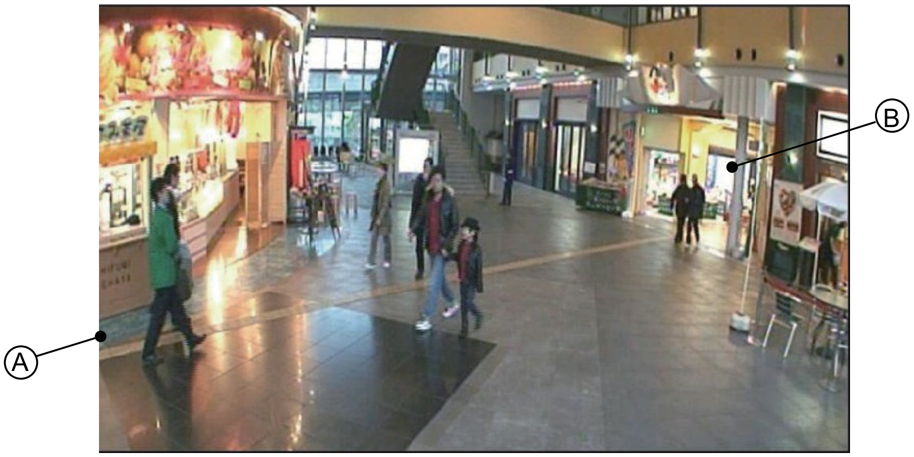


Figure 5.8 Experimental environment in a shopping mall (Figure 5.2 shows the left part in this map).

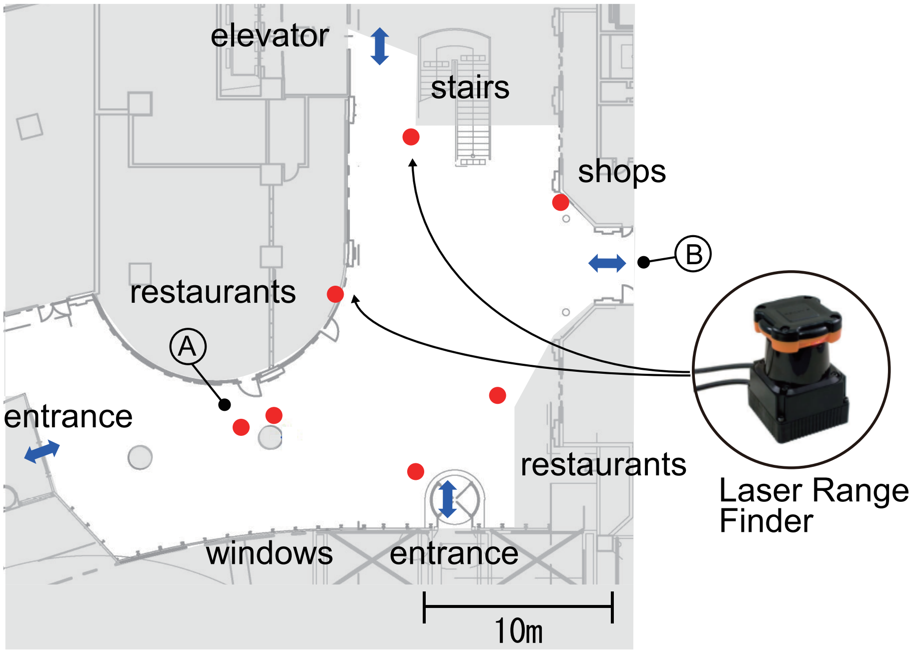


Figure 5.9 Experimental setup. Positions of LRFs in a shopping mall are shown as a red circles. The locations (A)(B) in the figure are also shown in Figure 5.8.

time-stamped and sent to a host PC via Bluetooth. Figure 5.10 shows estimated trajectories of feet in four seconds. Sometimes only one foot of a pedestrian is observed.

5.4.2 Accuracy of Identifying Pedestrian

We tested with three subjects and four trials. Figure 5.11, 5.12, and 5.13 shows the changes of computed correlation for each three wearable sensors at t=3,6,9,12,15. Figure 5.14 shows the first 20 seconds of computed correlation between the wearable sensor on subjects and all tracked foot when the subject was walking. Figure 5.14 (a)(b)(c) show the typical results for subject 1,2 and 3.The colored lines show correlation of the subject. When the colored line is the highest among all trajectories at the same time, the subject is correctly identified. In experiments three subjects carried a wearable accelerometer and walked with other pedestrians in a shopping mall. There are about 10 pedestrians in the environment shown in Figure 5.8.

In 15 experiments, the pedestrian who are carrying the sensor was almost correctly estimated in the sequences. Table 13 shows accuracy of identification at 4, 6, 8, and 10 seconds

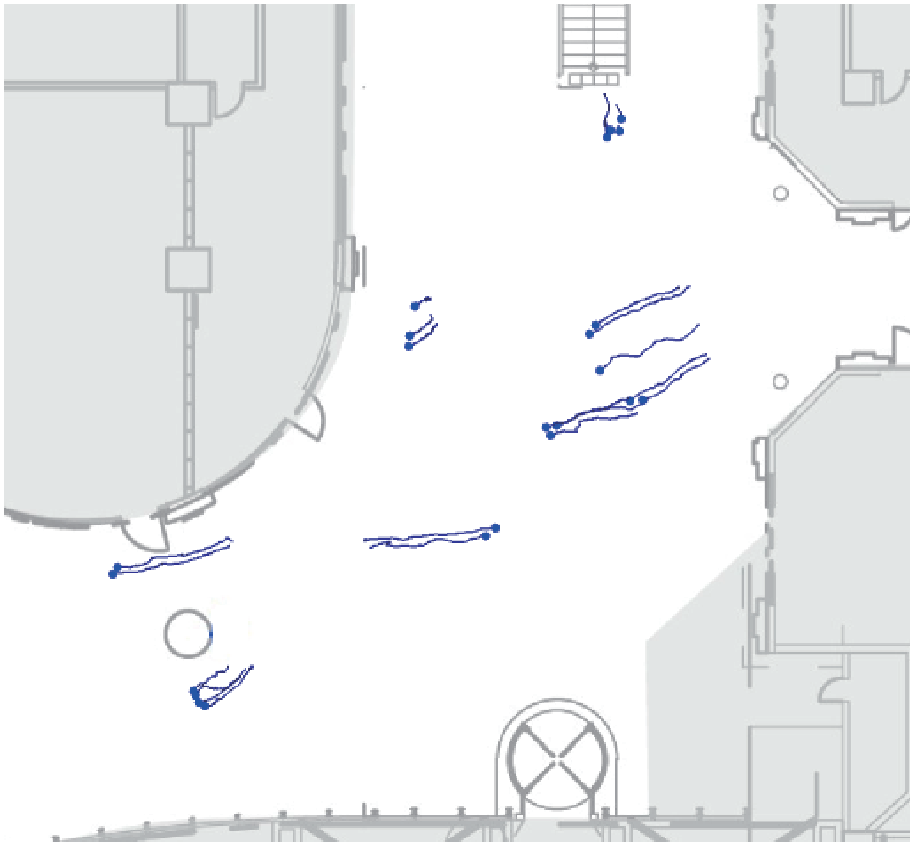


Figure 5.10 Estimated trajectories of feet in four seconds. There were 12 pedestrians in the period. The filled circles represent the latest positions. Sometimes only one foot of a pedestrian is observed because of the effect of occlusion.

after the subject appeared. As the time becomes longer after the subject appeared, more accurate correlation between signals becomes.

5.5 Discussion

5.5.1 Time Synchronization

Since our method locates people by comparing time sequences, it is important to adjust the clocks of the LRFs and wearable sensors. In the following experiments, the wearable sensor clocks were synchronized with the host PC when they initially established a Bluetooth connection.

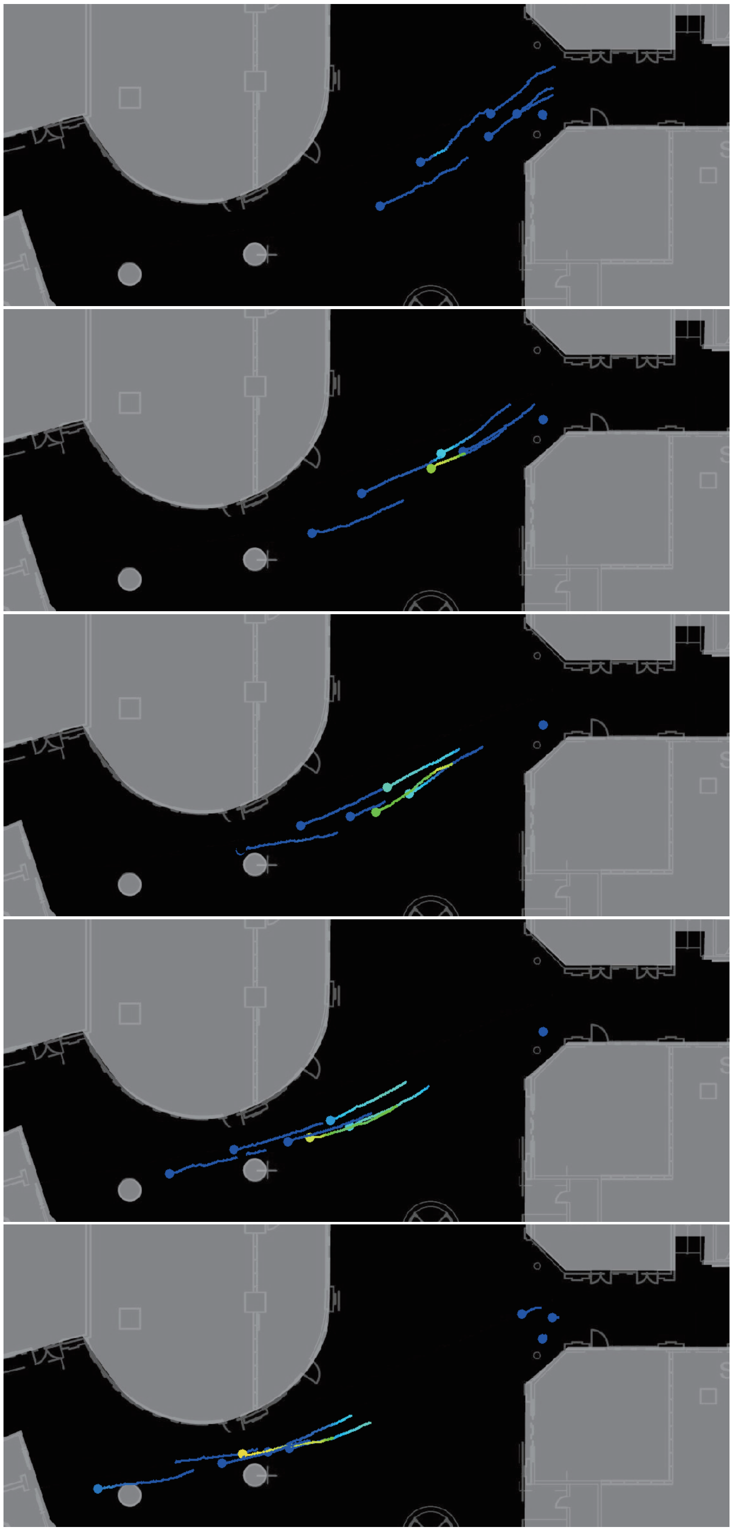


Figure 5.11 Computed correlation function between werable sensor 1 and trajectories of all feet. Filled circles show the current locations of pedestrians and the lines show past trajectories in a few seconds. Brighter color of circles and lines shows higher correlations. Clearly correlation of one of pedestrians becomes higher gradually.

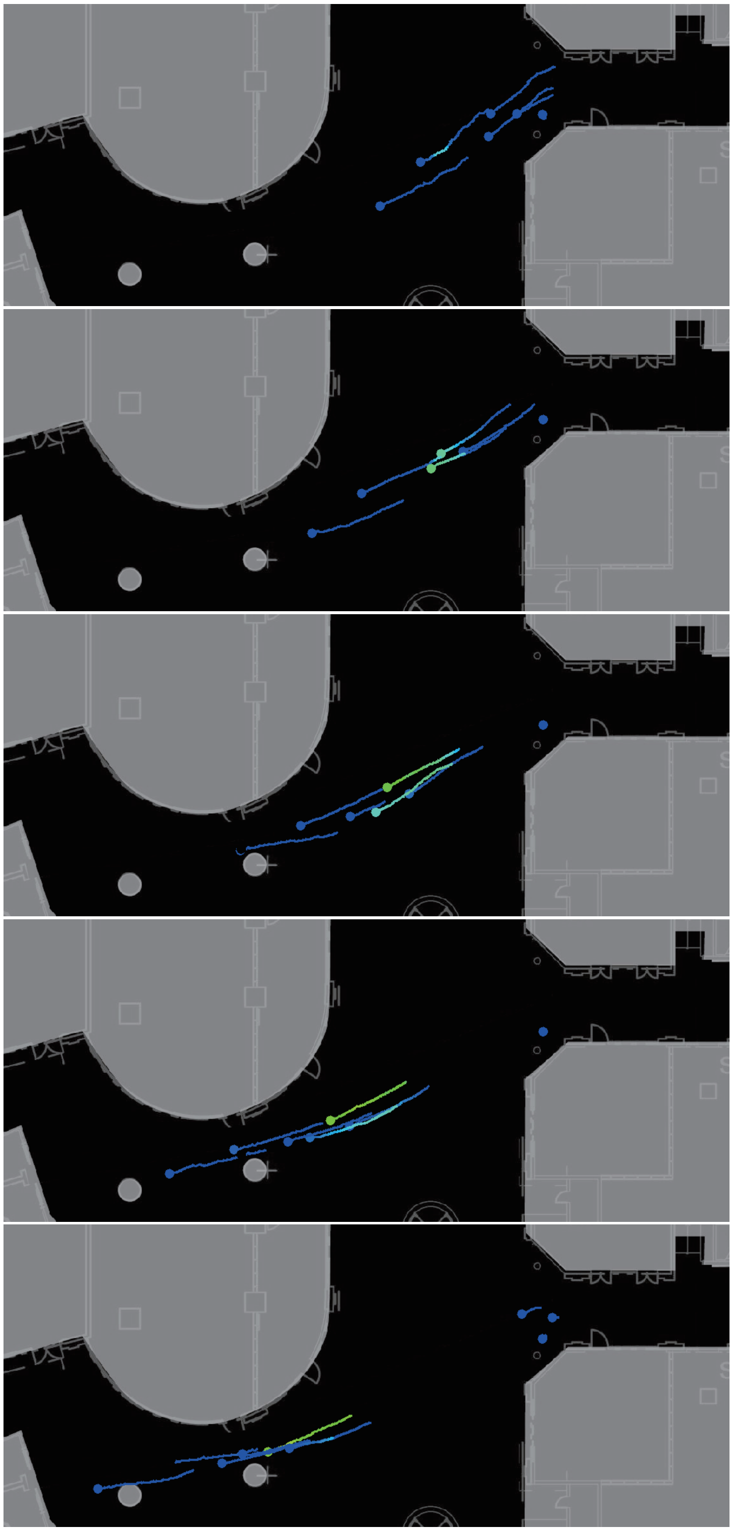


Figure 5.12 Computed correlation function between werable sensor 2 and trajectories of all feet. Correlation of another pedestrian becomes higher gradually.

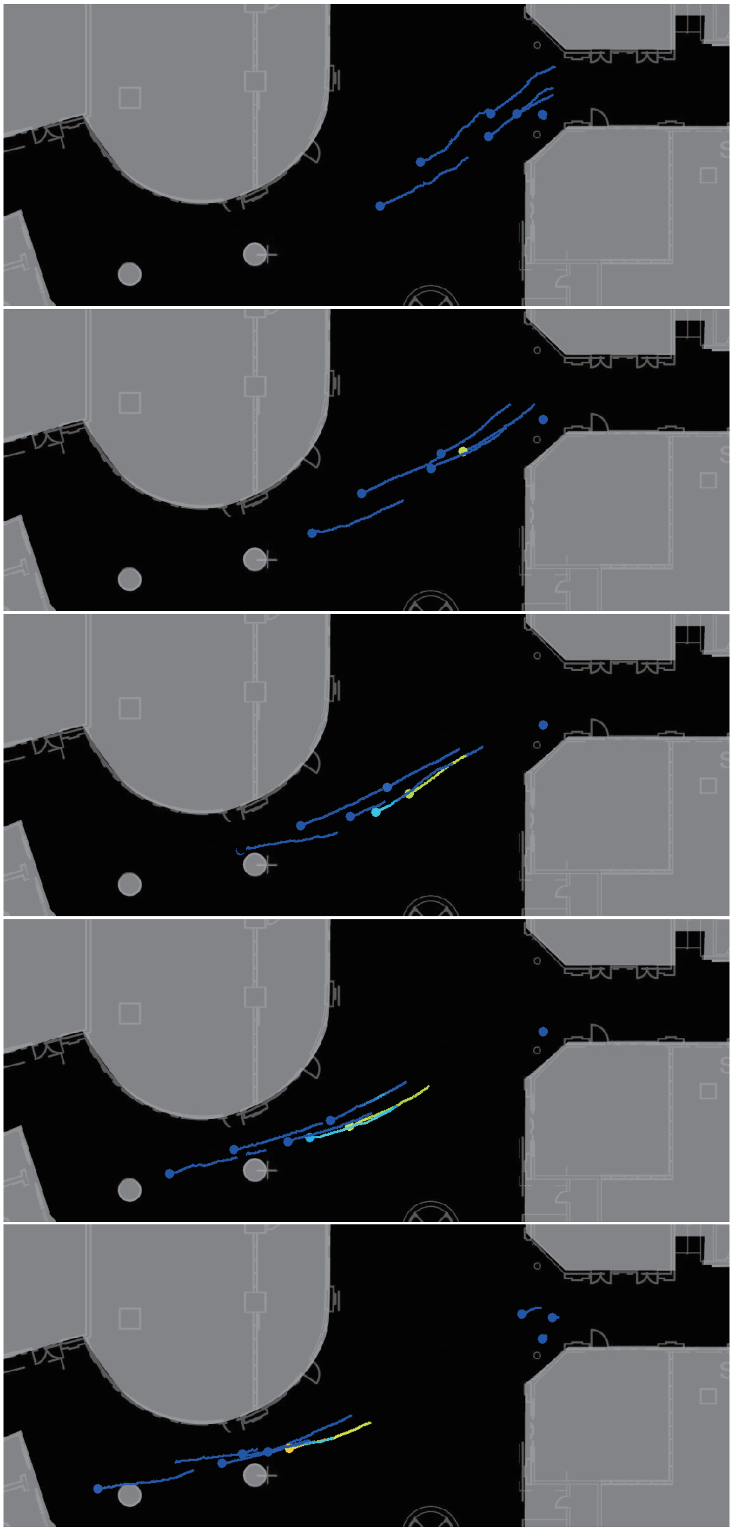
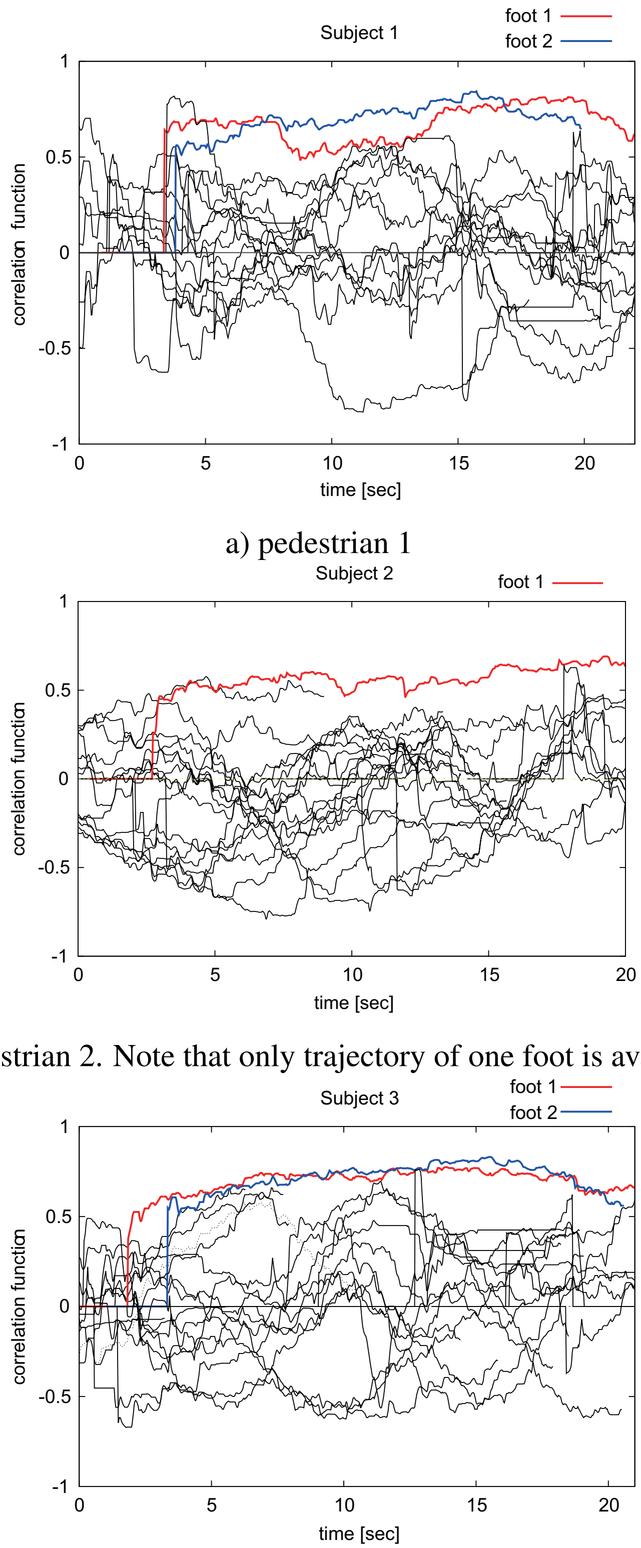


Figure 5.13 Computed correlation function between werable sensor 3 and trajectories of all feet. Correlation of another pedestrian becomes higher gradually.



a) pedestrian 1

b) pedestrian 2. Note that only trajectory of one foot is available.

c) pedestrian 3

Figure 5.14 Correlation function computed between acceleration signal of a wearable accelerometer sensor and each foot tracked in the environment. Colored line shows correlation function of the subject who carries the accelerometer. The correlation

Table 5.1 Accuracy of identification. As the time passed after the subject entered the tracking area, correlation becomes clearer and accuracy becomes higher.

Time after the subject appeared [sec]	Accuracy of identification[%]
4	71.0
6	85.0
8	85.0
10	92.0

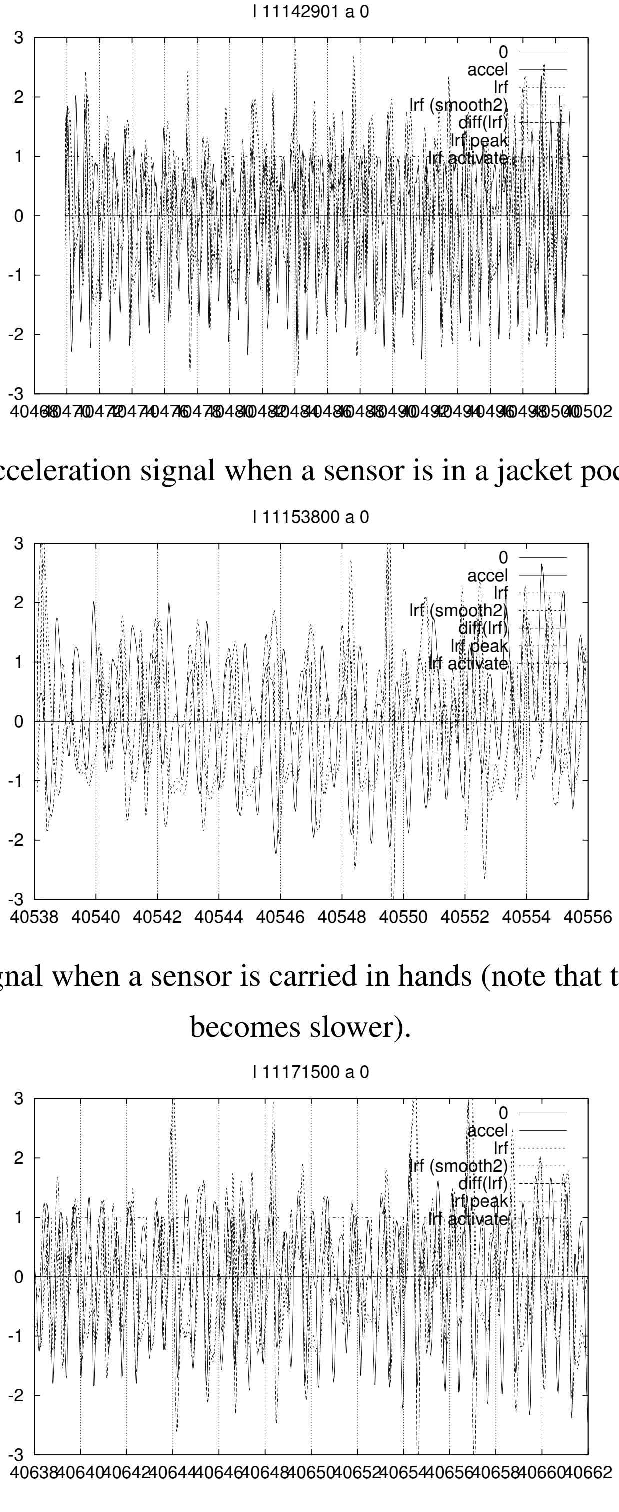
Another problem is the delay in the transmission from the wearable sensors to the host PC. In the experiments, signals were sent with timestamps added by the wearable sensors. If the timestamp were set after the signals had been sent (e.g., by the host PC), the results would be affected by sudden transmission delays.

5.5.2 Privacy Issues

When cameras are installed in public spaces, the problem of invasion of privacy is inevitably raised. Since LRFs do not observe the face or any other information that identifies pedestrians, this issue is irrelevant to our method. The effect of the pose of accelerometer In experiments, we attached wearable acceleration sensors to the waist of pedestrians. By computing vertical component of the acceleration, the pose of the sensors does not affect our method. However, acceleration signals differ depending on the position the sensor is attached. We confirmed the differences that may arise when sensors are carried in different ways: in a pocket, in hands, in a bag (Figure 5.15). The shape of the observed acceleration signals is not completely same, but the detected peaks of acceleration are still clear and there are no significant difference in computing correlation process.

5.6 Conclusion

In this chapter, to estimate both positions and IDs of pedestrians, we propose a method that associates precise position information using sensors in the environment and reliable ID



a) Acceleration signal when a sensor is in a jacket pocket

b) Acceleration signal when a sensor is carried in hands (note that the walking speed becomes slower).

c) Acceleration signal when a sensor is in a bag

Figure 5.15 Examples of acceleration signals in different carrying conditions.

information using wearable sensors. Since the tracking results of biped foot of a pedestrian and the body oscillation of the same pedestrian correlate, we associate these signals from same pedestrian that maximizes correlation between them.

Experimental results for locating people in a shopping mall show the precision of our method. Since LRFs are now becoming common and people are carrying cellular phones that contain accelerometers, we believe that our method is realistic and can provide a fundamental means of location services in public places. In future, we'd like to investigate our method when pedestrians carry cellular phones in different ways. Since we can observe much motions information of pedestrian from accelerometer, we'd like to expand our method to understand pedestrian behavior.

Chapter 6

Media Conversion that Transfers Impression by Keeping Signal Correlation

6.1 Introduction

In recent years, increasing attention has been drawn to media conversion studies that propose algorithms to convert signals in one modality to another. For example, in situations when we can't use specific modality for communication and presentation, it is effective to use different modality in a complementary manner by using media conversion. Another application of media conversion expects a synergistic effect of using multiple modalities by adding another modality to original contents. By focusing on possibilities of media conversion, much work has been proposed especially in the area of KANSEI information processing. There are three kinds of approaches in previous media conversion methods.

A simple approach toward media conversion is pattern recognition based method. This approach defines conversion rules between patterns in both media. When a pattern is detected in one media, the associated pattern in another media is presented. Cronly-Dillion et al.[58)59)] decompose line images into simple geometric shapes such as lines and rectangles then present defined sound patterns based on conversion rules between geometric shapes and sound patterns. Kobayashi and Ohta[60)] also extract landmarks from video and generate three dimensional sounds to navigate pedestrians using only sound information.

Another approach to media conversion is based on words that represent impression. This approach associates patterns in multiple media by evaluating relationship between adjectives and patterns in both media. Uenoyama et al.[61] evaluated relationship between adjectives and impression when subjects listen to drum play. Kumamoto and Ohta[62] proposed to use adjectives to search music clips. Yamawaki and Shiizuka[63] discuss similarity of music and visual color in terms of adjectives.

These approaches could be limited since these approaches associate media via symbolic representation like pre-defined pattern and natural language. Restrictions of pre-defined symbol based association are 1) patterns that are difficult to represent in symbols are discarded in the recognition and abstraction process, 2) the association methods only detect and convert pre-defined pattern and unknown patterns are discarded.

We believe similarity between media is not restricted to symbolic relationship. Signals that represents same target in different media look differently in superficially, but these signals have similarity and correlation. In this paper, we propose new media conversion approach based on signal level relationship. Figure 6.1 summarizes approaches toward media conversion in terms of abstraction level in conversion process.

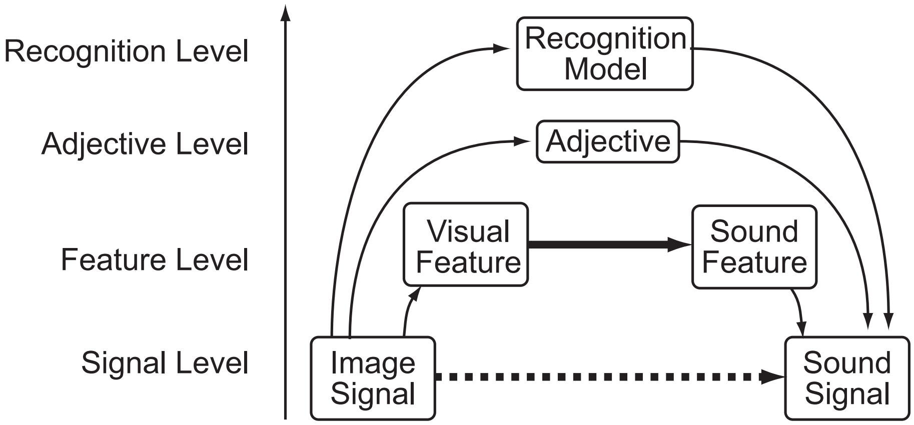


Figure 6.1 The approaches of media conversion

To confirm the possibility of direct media conversion, we have proposed an approach in signal level conversion (dashed arrow in Figure 6.1. The hypotheses are that observed signals contain components that do not depend on each media, and that the listener can understand original visual scene by interpreting the video signal as sound. Based on the idea the method[64] extracts a sequence of intensity in video signal and sends the sequence to sound device directly. This approach is free from any abstraction and recognition process, however, it is painful to listen the generated signal long time since it is similar to random noise and it is difficult to understand the original scene.

Therefore, in this paper we propose a new feature level conversion method (thick solid arrow in Figure 6.1) that generates comfortable sounds to listen to transfer impression of visual scene. We define a set of low-level visual features and musical features, and conversion rule between these features. Since the method does not assume pre-defined entities in the visual scene, the method transfers impression of unseen visual scene with unknown entities. Since the method introduces music constraint in generated sound, it becomes more comfortable to listen.

As related works in signal level media conversion, interactive systems that convert user behavior to different media have been proposed. Yonezawa and Mase[65] proposed a new musical instrument that directly converts interaction to water flow to sounds. Eng et al.[66] proposed an intelligent space that user behaviors are detected and converted into video and sound output. However, they do not focused on the transfer of information from source media to destination media. Nagata et al.[67] asked subjects with/without sound-color synesthesia and investigated their associations between musical features such as key, height, and tone to visual feature such as hue, brightness, and saturation. They suggested subject without synesthesia also have selective combination between musical and visual features. However, they do not evaluate media conversion system in qualitative study. In this paper, we propose feature-level media conversion system that keeps impression in a source media. We propose conversion rule from visual features musical feature that does not assume any predefined symbolic model. Finally we evaluate developed media conversion system in terms of information transfer.

Intermodal relationship in human perception

We usually take it for granted that each sensory modality is separate. However, strong intermodal relationship has been reported as ”synesthesia”. Synesthesia is a neurological condition in which stimulation of one sensory pathway leads to involuntary experiences in a second sensory pathway.[68] One in thousand people report such experience. Baron-Coen[69] proposed a hypothesis that infants until four month old have perception similar to synesthesia. Synesthesia may be interpreted as phenomena that part of unconscious perceptual processes are observed.

Though it is not very clearer compared to synesthesia, many people have same impression to signals in different media. For example, sometimes we use the synesthesic expressions like ”kiiroi koe (shrill voice)” and ”sibui iro (cool color)—” in Japanese, which suggest we have similar impression to visual and auditory, taste and vision perception. The word ”Neiro (tone)” is also visual expression to auditory perception. The presence of such expressions suggests that there is general low level signal interaction in our sensory system.As an example of more general relationship among media, we feel ”intense” impression to all kinds of perception to strong signals. This suggests hypothesis that we feel physical quantity like strength, rhythm, texture, direction as media dependent impression.

People who do not have synesthesia, though not very clearer as synesthesia, show intermodal relationship. McGurk and MacDonald[70)] showed impressive example of visual and auditory interaction at the early stage of human perception that visual signal of speaking ”ba” and auditory signal of ”ga” result in the perception of ”da”. Shimojo and Shams[26)] reported several phenomena of one type of perception affecting to another.

Artificial systems that integrate many sensors at the earlier stage

Inspired by these findings in human perception, integration at the early stage of signal processing has attracted increasing attention. Coen[71)] proposed a concept of signal integration of multi modalities in multi abstraction level. Hershey et al.[45)] proposed locating talking person in video by computing mutual information between pixel intensity in video and volume in sound. Grant and Seitz[72)] showed improvement of automatic speech recognition by using visual information. We also proposed integration of vision and sound to locate moving entities in the scene[73)].

In the area of media conversion, inspiring by these findings we believe there are conversion rules that associate multimodal signals that give similar impression. In this paper, we construct new media conversion method by mapping visual and music features that give similar impression. The visual scenes are observed by using omni-directional camera and visual features from observed images are converted to music features. By using simple visual features without abstraction and pattern recognition, music is generated from any kinds of input visual scene. We also propose to use multiple omni-directional cameras and detect positions of sound source. The position of sound source is also transferred by simple conversion rule of generated music signals. We evaluated the proposed media conversion system in terms of information transfer. We evaluated if the subjects feel same impression in original scene by only listening to the music generated by the system.

In section 6.2, a media conversion method from one omni-directional camera to music is proposed. In section 6.3 the algorithm is expanded to multiple omni-directional cameras that transfer special information of a signal source. In section 6.4 we evaluated the proposed system.

6.2 Media Conversion from Omni-directional Video to Music

Overview

Media conversion by using omni-directional cameras

In the proposed media conversion system, omni-directional cameras are used to observe visual scene (Figure 6.2). An omni-directional camera consists of a downward convex mirror and a camera that observes the mirror, which enable it to observe all directions. By this feature of omni-directional cameras, the proposed media conversion system transfers the impression of whole scene. In previous research[64)] we have already proposed a media conversion system that converts omni-directional video images to sound signal. However the sound that the system generates is similar to white noise and the method does not take into account the comfort of the listeners. In this paper, the proposed system generates music features that describe the impression of the observed scene and presents comfortable music based on the computed features 6.3.

Representing impression of the scene by a set of features

The proposed method represents impression of the observed visual scene by using simple video features. Since the method does not depend on pre-defined object model nor object

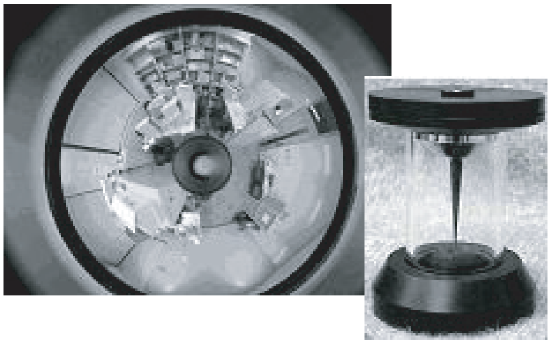


Figure 6.2 An omnidirectional image and an omnidirectional camera

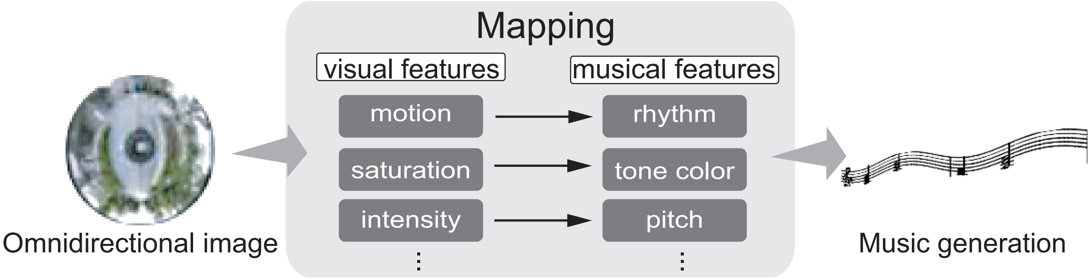


Figure 6.3 Media conversion by mapping visual features to musical features

recognition, it is free from recognition error and can transfer impression of unknown objects in the scene. The method transfers visual information that previous methods do not extract from video images.

Table 6.1 shows the association between visual and music features. To generate music that transfers impression of the visual scene, we associate features that both features will cause same impression. Each association of features in the Table 6.1 is based on experimental results or suggestions that there are relationship between our impressions of these features.

Note that the average background subtraction is the average of the difference between current image and the background image, which is observed when no moving entities in the scene. The foreground image consists of pixels that the difference between the current image and the background image is larger than a threshold. Average frame difference is the average of difference between successive video frames.

Global features and local features

Table 6.1 Mapping from visual features to musical features

	Image feature	→	music feature
(1)	Average intensity		minor / major of tonality
(2)	Average hue		tonality
(3)	Average intensity		height of chord
(4)	Average saturation		timbre
(5)	Average frame difference		tempo
(2)	Average background subtraction		volume
(7)	Difference between hue		chord
(8)	Average intensity in foreground		height of melody

There are two types of features both in visual and music features. One type of feature is global features that represent overall impression of the visual and music scene (Table 6.1 (1)...(7)). Examples are average intensity and average background subtraction in visual features, and tempo and tone in music features. Another type of feature is local features that represent local information in each modality (Table 6.1 (8)). Examples are average foreground intensity and melody (height in melody?). In the conversion rule, we associate between each type of features.

Flow of the conversion process

The conversion process consists of repeated conversion from observation of an omni-directional image to music in a few second. After visual features are extracted from omni-directional images, music features are determined from associated visual features. Tonality, chord, and other global music features are determined from associated global visual features, and melody is selected from a set of fifty short sound series according to the local visual feature (Figure 6.4). Each sound series consists of four music notes and the selected series are transposed according to the selected tonic. To avoid monotony, sometimes random sound series are selected. After the sound series are played, new omni-directional image is observed and music is generated in same manner. Since the music is generated from an observed omni-directional image, there is no long phrase or long term upsurge in music.

The conversion methods from each visual feature to music feature shown in Table 6.1 are

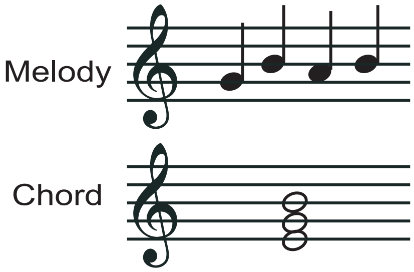


Figure 6.4 An example musical sequence that is generated from an omni-directional image

described here. The evaluation of the proposed media conversion system is shown in section 6.4.

6.2.1 Association Between Visual and Music Features

(1) Major or minor of tonality

Adachi et al.[74] reported that impression of music become brighter as visual intensity becomes higher. The relationship between the impression of visual intensity and brightness of music are also reported in.[75] According to this knowledge, we determine minor or major of the tonality based on the average intensity of video image. By comparing average intensity I of video image and an empirical threshold I_0, we determined the tonality as shown in the Table 6.2.

Table 6.2 Mapping from brightness to major/minor key

Brightness	Major/Minor
$I \geq I_0$	Major key
$I < I_0$	Minor key

(2) Tonality

Nagata et al.[67] suggested the relationship between tonality of music and hue in visual image. According to this knowledge, we determine tonality based on the average hue as shown in the table (Table 6.3).

Table 6.3 Mapping from hue to key

Hue	Key
White	C
Orange	D
Yellow	E
Green	F
Cyan	G
Red	A
Blue	B

(3) The height of the chord

Nagata et al.[67] also reported that subjects have image of brighter color for higher-pitched sound. According to this knowledge, we determine the height of chord based on the average intensity I as shown in the table (Table 6.4).

Table 6.4 Mapping from brightness to octave

Brightness	Octave
$I < 100$	-
$100 \leq I < 130$	+1 octave
$130 \leq I < 150$	+2 octave
$150 < I$	+3 octave

Note: the range of the average velocity I is $0 \leq I < 256$.

(4) Tone

Nagata et al.[67] reported that larger saturation in video images is related to increased high-frequency component in sound. According to this knowledge, we determine music tone based on the average saturation S in video images. For higher saturation in video, we add high frequency component to sine wave (Table 6.5).

(5) Tempo

Sugano[76] reported that subjects feel fast-moving video in harmony for music in fast

Table 6.5 Mapping from saturation to tone

Saturation	Tone
$0.16 \leq S$	the second harmonic
$0.24 \leq S$	the third harmonic
$0.31 \leq S$	the forth harmonic
$0.39 \leq S$	the fifth harmonic
$0.47 \leq S$	the sixth harmonic

Note: The range of S is $0 \leq S < 1$

tempo. Nagashima[77)] reported rhythm in video and music becomes closer and mutually attracted. According to these knowledge, we determine the tempo of music based on total number of pixels in inter frame difference in video (Table 6.6).

Table 6.6 Mapping from frame difference to tempo

Frame difference	Tempo (crotchet)
$D < 2$	40
$2 \leq D < 5$	48
$5 \leq D < 20$	60
$20 \leq D < 60$	80
$60 \leq D$	120

(6) Volume

When there are many moving entities in environments, we find more noise there. We also determine the volume of music based on total amount of observed motion. The volume is determined by a linear equation of the amount of motion.

(7) Chord progression

Kitajima and Doi[78)] reported that a set of color in a visual image gives subjects similar impression as a musical chord. Based on this knowledge, we determine musical chord progressions based on colors appeared in visual images. By considering the difference H of average hue in foreground and background image represent a harmony in color, we determine musical chord as shown in table. We selected stable chord (tonic) for smaller H and unstable chord (subdominant, dominant) for larger H (Table 6.7).

Note that tonic (dominant, subdominant) chord represents chord that is build on the tonic tone (the tone fifth above the tonic, the tone fourth above the tonic) in the key. For example in C major, the tonic chord is C-E-G, the dominant chord is G-H-D, and the subdominant chord is F-A-C.

Table 6.7 Mapping from difference of hue to harmony

Difference of hue	Code
$0 \leq H < 5$	Tonic
$5 \leq H < 20$	Subdominant
$20 \leq H$	Dominant

However, independent selection of musical chord in each video frame sometimes results in non-musical progression of chord. So we introduce following constraints and limit chord progression. To avoid monotonic music, variation is introduced by using substitute codes.

1. Chord must move to a tonic chord after a dominant chord.
2. Though chord may move to any chord from subdominant chord, chord must move from substitute chord II to dominant chord V.

6.3 Media Conversion that Transfers Spatial Information by Using Multiple Omni-directional Video

The media conversion method proposed in the previous section does not transfer the positions of entities in the scene. By using multiple cameras and speakers, we propose to transfer spatial features in the scene. In this section, we expand the media conversion method in the previous section so that it extract positions of entities in the scene from omni-directional camera network and transfer the observed location by difference of volume in multiple speakers.

The characteristic of the proposed method is the conversion is computed in each omni-directional camera independently. Then the computed information is integrated by sum of musical signals generated by each camera. So our method is scalable and we can easily add cameras and speakers to the media conversion system.

6.3.1 Configuration of Cameras and Speakers

Our conversion algorithm supposes two types of configurations:

1. By using devices consisting of an omni-directional camera and a speaker, generate music from recorded visual scene at the location.
2. By installing omni-directional cameras in one place and speakers in another place, transfer the impression of visual scene in the former place to the latter place and generate music. For conducting experiments of media conversion in configuration 2, we developed a device shown in the upper left of Figure 6.5 that an omni-directional camera is put on the top of the device and a speaker is at the bottom.

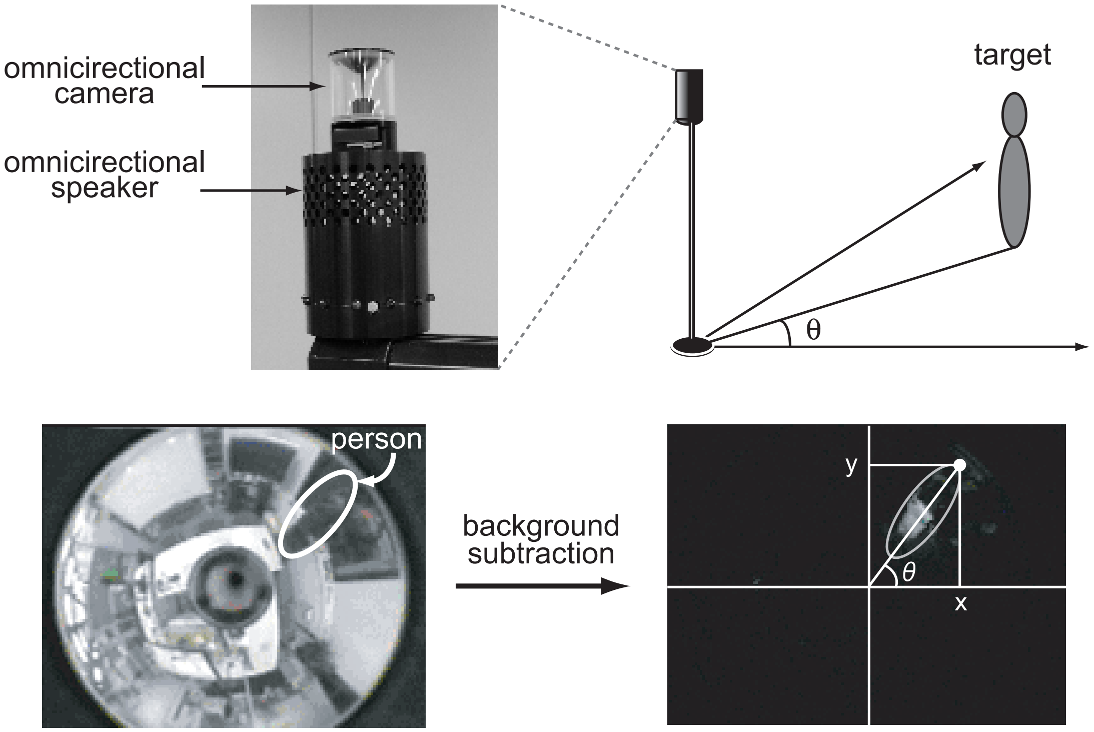


Figure 6.5 Position detection in a omnidirectional image

6.3.2 Presenting Positions of Entities in the Scene by Conversion from Two Omni-directional Cameras to Stereo Speakers

Detecting entities in omni-directional video image

It is straight forward to estimate direction of an entity from the camera by computing background subtraction (see the lower bottom of Figure 6.5). For each area detected in the background subtraction image, we compute the direction of the entity based on the furthest point from the camera.

Transferring direction of the entity by using two cameras and speakers

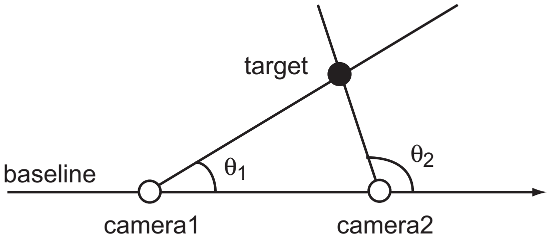


Figure 6.6 Observing a target by two omnidirectional cameras

We integrate a pair of an omni-directional camera and a speaker to transfer the location of an entity in the scene. Here we explain to transfer the approximate location of the entity by using volume difference of the speakers based on the characteristic of omni-directional cameras. Figure 6.6 shows two omni-directional cameras observes an entity, where θ_1 and θ_2 represent angle of the entity relative to each camera from the baseline that connects two cameras. Considering $cos\theta_1$ represents the relative position of the target 1 to the camera 1 (along) the baseline, relative position of the entity is approximated as sum of cosine of these angles:

$$C = \cos\theta_1 + \cos\theta_2 \tag{6.1}$$

C represents the position along the baseline approximately. Figure 6.7 shows the value of C at each location in the scene and brighter color represents smaller value of C. Based on the equation we can determine the volume of each speaker from the omni-directional camera independently without any central integration process. Integration is performed by the sound mixing level. We determine the volume of two speakers as:

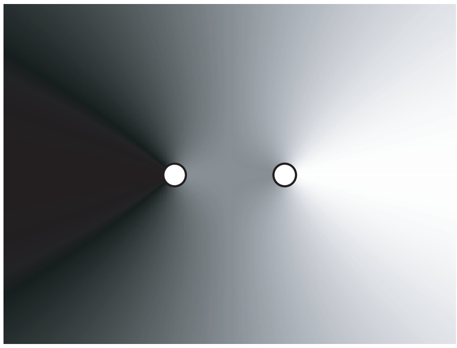


Figure 6.7 Visual display of the equation 6.1

$$Vol_1 = V_0 - V * C \cdots\cdots (6.2)$$
$$Vol_2 = V_0 + V * C$$

Where V_0 and V represents constants that limit the range of volume. The contribution of volume change from a camera is determined based on the detected position in the camera. In the following experiments, we use speakers that accept two input sources and the integration of Equation 6.2 is performed in the mixer of the speakers. For each camera i, we can add generated sound signal independently.

$$\Delta Vol_1 = -V \cos\theta_i \cdots\cdots (6.3)$$
$$\Delta Vol_2 = +V \cos\theta_i$$

Transferring the 2D positions of the entities

To transfer the position perpendicular to the baseline, we integrate sine of the detected angle:

$$S = \sin\theta_1 + \sin\theta_2 \tag{6.4}$$

When a listener of generated music is at a position on the baseline, Equation 6.4 represents the approximate vertical position in Figure 6.6. By expanding equation, we determine the volume of speakers as follows.

$$Vol_1 = V_0 + V * (-C - S) \cdots\cdots (6.5)$$
$$Vol_2 = V_0 + V * (+C - S)$$

In the following experiments, we use further expanded method explained in next section.

Transferring the 2D positions of the entities by integrating multiple omni-directional cameras and speakers

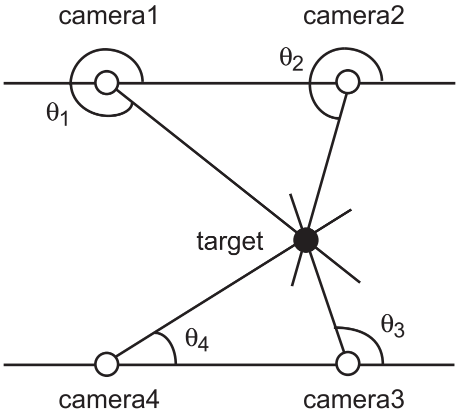


Figure 6.8 Observing a target by four omnidirectional cameras

When there are more omni-directional cameras and speakers, the proposed method transfer the location of the entity by controlling volume of speakers based on equation 3 for each pair of speakers. For example when the number of omni-directional cameras and speakers is four (Figure 6.8),

$$\Delta Vol_1 \propto +\sin\theta_4 + \sin\theta_1 - \cos\theta_1 - \cos\theta_2 \cdots\cdots (6.6)$$
$$\Delta Vol_2 \propto +\cos\theta_1 + \cos\theta_2 + \sin\theta_2 + \sin\theta_3$$
$$\Delta Vol_3 \propto -\sin\theta_2 - \sin\theta_3 + \cos\theta_3 + \cos\theta_4$$
$$\Delta Vol_4 \propto -\cos\theta_3 - \cos\theta_4 - \sin\theta_4 - \sin\theta_1$$

6.4 Experiments

To confirm the effectiveness of the proposed media conversion method, we evaluated if the subjects can imagine original visual scene only by listening to the generated music. We tested two proposed media conversion methods.

6.4.1 Omni-directional Video Stimuli

By using four omni-directional video, we recorded various scene shown in Figure 6.9. Examples of the scene include a bright room without any person, a dark room with a walking person, and a scene close to busy street. We recorded each scene two times by using four omni-directional cameras, and we used one video for naturalization and another for testing. For evaluating the method with an omni-directional camera proposed in section 6.2, we used one of video from four cameras.

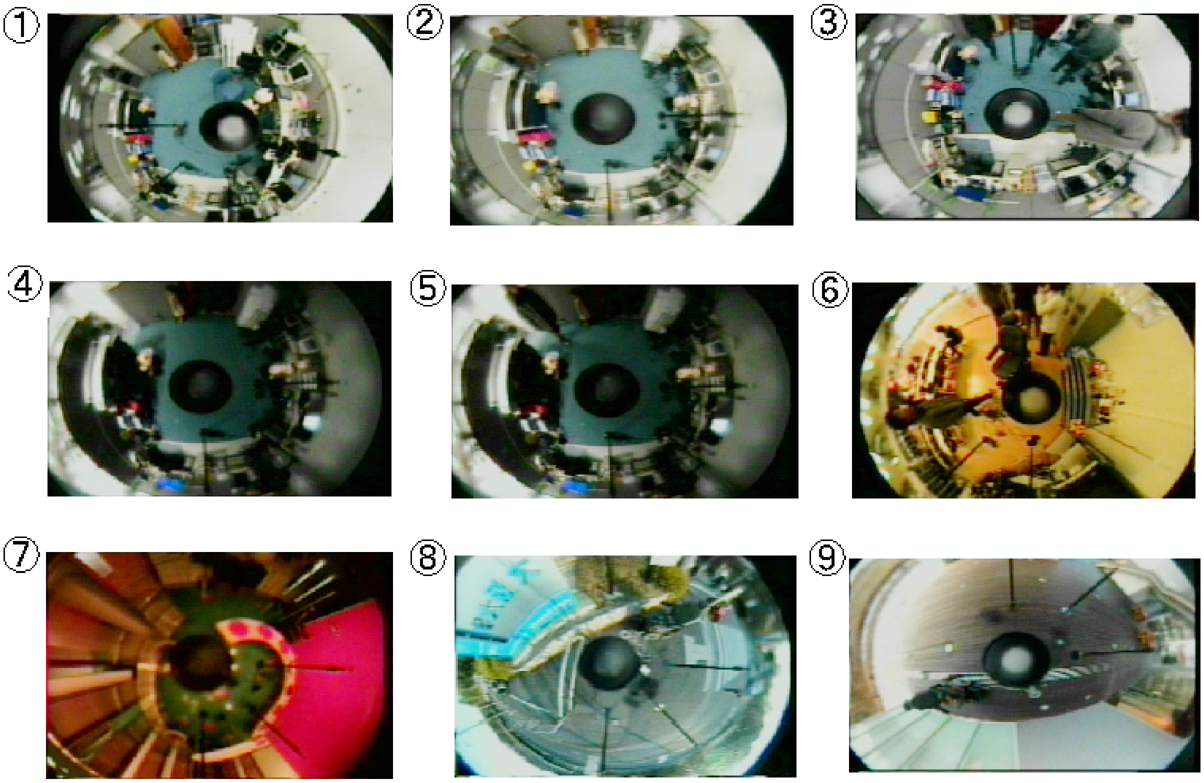
6.4.2 Experimental Setup

We synthesized one video image from four omni-directional cameras. To play with four speakers in synchronation, we have used Hammerfall DSP (RME Co., LTD.). Figure 6.10 shows omni-directional camera used in the experiments. They are placed at the corner of a square of side 1.5m. Speakers are placed at the corner of side 2.0m. We determine volume of four speakers based on the method proposed in section 6.3. Entity in the scene is located by computing the center of gravity of background subtraction. Conversion from omni-directional image to music is performed in real time on a PC by using MAX/MSP and jitter (Cycling'74[79)]).

6.4.3 Experiments

Procedure We explained to subjects that the music is converted from the video. Firstly we presented both an omni-directional video and generated music from the video in each scene for naturalization. Then we presented only music that is generated from omni-directional image for testing. Subjects are asked to select one of omni-directional video of original visual scene.

Experiment 1. (method1, conversion from one omni-directional video) In experiment 1 the subject listen to the music that were generated by using the media conversion method



	scene	moving entities
1	bright room	no people
2	bright room	one person walk
3	bright room	a few people walk
4	dark room	no people
5	dark room	one person walk
6	cafeteria	many people stay and walk
7	elevator hall	no neople
8	busy street	many cars
9	square	no people

Figure 6.9 Example omnidirectional images in the experimental data and brief descripotions of the environments

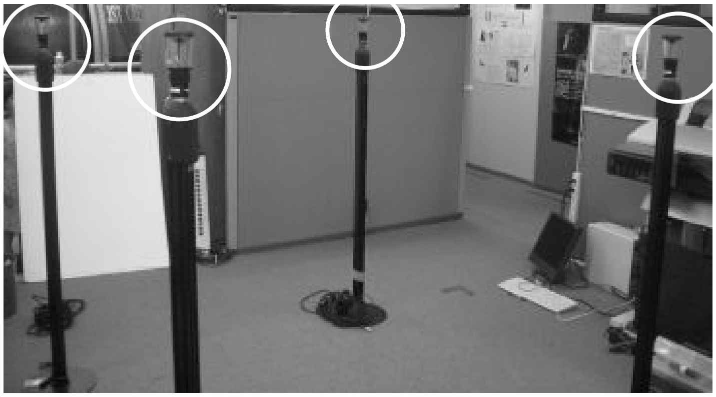


Figure 6.10 Experimental setup

with one omni-directional camera proposed in section 6.2. The participants were 15 adults with age range between 20-30 years (13 males, 2 females). No subject reported the ability to understand tonality and key by listening to music. Table 6.8 shows the accuracy of estimation. For almost all except scene 6, the accuracy is more than 50%. Average accuracy is 58.5%.

Experiment 2. (methos2, conversion from four omni-directional video).

In experiment 2 the subject listen to the music that were generated by using the media conversion method with four omni-directional camera proposed in section 6.3. The participants were 13 adults with age range between 20-30 years (11 males, 2 females). No subject reported the ability to understand tonality and key by listening to music. Table 6.8 shows the accuracy of estimation. For almost all except scene 7 and 8, the accuracy is more than 70%. Average accuracy is 66.7%.

6.5 Discussion and Conclusion

6.5.1 Discussion

Comparison with the previous method

Since the generated music by using the previous media conversion method[64] does not

Table 6.8 Correct answer rate in the experiment

	scene	accuracy [%]	
		method 1	method 2
1	bright room (no peole)	60	77
2	bright room (one person)	47	70
3	bright room (a few people)	60	77
4	dark room (no people)	94	70
5	dark room (one person)	94	92
6	cafeteria	20	54
7	elevator hall	47	15
8	busy street	47	70
9	square	60	77
average		58.5	66.7

depend on the brightness and the color, it is limited in its ability to transfer impression of the scene. The proposed method can transfer the impression of the scene by combining many visual features.

Transfer of spatial information by music

In two experiments, significant decrease of accuracy appeared in scene 4 and 7, whereas significant increase appeared in other scene except 5. Increase of accuracy appeared mainly in the scene that there are moving entities. It suggests that the media conversion method proposed in section 6.3 transfers motion information clearly and results in better accuracy. On the other hand, decrease of accuracy appeared in the scene that there are no moving entities. When there are not any motion in the scene, composed four omni-directional images may not present clear impression of the scene.

Easiness of listening

By introducing musical constraints, subjects reported that the proposed media conversion method generates more comfortable sound compared to the previous method.[64)] Subjects also reported that the generated music is different from the expected music like sound. Since the proposed method generates sound by connecting fragments of melody, the lack of long and consistent melody may results in this impression. Also since the proposed method al-ters the tone by adding high frequency component to sine wave, the tone is different from standard instruments. This point may cause a sense of discomfort. To generate more natural music sounds, introducing rules that keeps consistent melody and sounds of standard instruments may decrease the discomfort.

Application of media conversion system

By introducing musical constraints, now it is possible to the generated music as background music that transfers the scene in the next room. That means the proposed media conversion enable us to work in a room by watching displays while watch the state of the next room by music.

However, the proposed associations between vision and music are, though they are based on previous experimental knowledge, may not intuitive for listener. After we listen to the original visual scene and generated music for a while, it is expected that we learn relationship between these associations and we are able to understand the scene only by listening to the music.

6.5.2 Summary and Future Work

We proposed new media conversion method from omni-directional video to music by converting a set of visual features to musical features. Since the proposed method convert at the early level of signal processing, which is based on simple signal features and not based on symbolic representation, the method convert any kinds visual scene to music by keeping impression of the scene. To transfer two-dimensional location of an entity in the scene, we expanded to use multiple omni-directional cameras. We construct media conversion systems and evaluate if the methods transfer impression of the original visual scene. Subjects listened generated music and selected video of the original visual scene based on similarity of impression. They correctly select original visual scene for 58.5% on average for the conversion method with an omni-directional camera. The accuracy improved to 66.7% for the extended method with multiple omni-directional cameras. These results suggests the impression of original visual scene is transferred by the proposed media conversion method and the effect of using multiple omni-directional system for transferring the location of an entity in the scene.

In future, we'd like to expand the method to handle multiple entities in the scene and transfer their location independently. In the proposed method, we associated visual and music feature that relationships are suggested in previous studies. However, the associations sometimes differ in different studies. We'd like to confirm the effect when we select different association between features. In this paper our focus of experiments is to confirm possibility of media conversion in an application of transferring impression of visual scene to music and estimate original scene from generated music. We'd like to further investigate contribution of each feature and combination of features to transfer impression of visual scene.

Chapter 7

Conclusion

Recently high-performance sensors and processors have become more common, and many kinds of sensors are installed in public locations and on wearable devices. In such social environments, observations using such sensors in our daily life have become commonplace and many multimedia databases are now available. Since we cannot understand such huge data as they are, we must extract and collate meaningful components from them.

This research extracts such components from the data by focusing on the signal level correlations among the large amount of sensory data inherently contained in multisensory observations. In previous approaches to multisensory integration, the observed signals are processed in each sensor and their features and patterns are extracted. Then the extracted features are integrated in pre-defined common coordinates or symbol systems that are defined by the designers of perceptual systems. Compared to such an *integration after recognition* approach, our *signal level integration* approach has the following merits:

- The signal level association method associates observations based on signal correlations without knowledge of common coordinates or symbol systems. Previous association methods require that signals be on common coordinate systems.
- It associates the observations of many kinds of sensors including binary touch and inertial sensors that do not observe position information. Previous association methods require that signals have identical physical quantities and be mainly limited to positions.
- It extracts signal correlations that are abstracted away during the independent fea-

ture extraction and recognition processes. Previous association methods discard such critical correlation information. In this thesis, we expand the signal level correlation method and apply it to many kinds of sensors and in many situations to tackle the above research issues.

In this thesis, we expanded signal level correlation method and applied to many kinds of sensors in many situations.

Signal level association in various situations

We expanded methods to associate moving targets in array sensors.

- Association of a moving target in video images: When the signals source moves, previous method[16)] cannot associate observations. We simultaneously detected and tracked a sound source based on the criteria of mutual information maximization. The problem of detecting and tracking a sound source is solved as an optimization problem to find the path that maximizes the mutual information between video and audio signals. We described a sensor fusion algorithm based on mutual information maximization and apply it to the problem of sound source localization by combining audio and visual signals.
- Association of moving targets in binary touch sensors: Observations of sensors like touch sensor and event detection sensor are binary. Previous method assumed continuous observation signals. We proposed a method that associates binary signals based on signal correlation and explained an integration method of wearable and floor sensors that detects the positions of people. Floor sensors consist of small unit sensors, each of which returns ‘ 1 ’ when someone is standing on one of them and ‘ 0 ’ otherwise. To integrate these binary and acceleration signals from wearable sensors, we proposed an integration method that evaluates signal correlation based on a statistical test.

Signal level association based on unreliable observations

We proposed methods that reliably extract signal level correlations.

- Association of unreliable observations: When observation confidence changes according to situation and it affects to signal correlation, it is difficult to associate observations in stable manner. We proposed an association method from among different kinds of sensors that considered confidence in observation. Different kinds of sensors have different reliability characteristics depending on the situation. We focused on the association of laser range finders (LRFs) and wearable gyroscopes to track and identify each person and proposed an association method that considers the reliability of LRF observations.
- Construct time-dependent correlation model: When observations are limited, computed correlation among sensors is sometimes not reliable. We proposed a method that associated the leg motion of pedestrians and wearable accelerometers based on time-dependent correlation model. LRFs observed pedestrians at the height of their feet and extract features from a bipedal walking pattern. Wearable accelerometers also observed walking patterns. Since walking rhythms differ from person to person, our proposed method distinguished pedestrians walking in a line. Another characteristic was that it only used an accelerometer in the wearable devices.

Applications of signal level relationship

We used the signal level relationship in the area of tracking and identification and media conversion.

- Application to people tracking with identification: Recently people behavior in public locations is observed and statistically analyzed and the results are attracting attention in the area of environment design and marketing. We proposed a method that not only estimates positions but also identifies each person who carries wearable sensors.
- Application to convert signals between different media: We proposed a new feature level media conversion method that generates comfortable sounds to listen to the transfer impressions of visual scenes. We defined a set of low-level visual and musical features and conversion rules between them. Since the method did not assume predefined entities in visual scenes, it transfered the impressions of unseen visual scenes with unknown entities. By introducing music constraints in the generated sound, lis-

tening became more comfortable.

In future researches in recognition, one of fundamental problems is closely associating multimodal observations with each other in the various abstraction levels. Sensor information processing has developed in the field of each type of sensor. However, humans have developed our perceptual system by observing multimodal information. We believe that we can find rich information in relationship among sensory observations. In order to understand something, there is a need to associate something. Information is present in the association. A promising way to realize an intelligent system that recognizes a scene is to closely associate multimodal observations in flexible manner at many levels of abstraction, from the signal level to the symbol level.

Acknowledgements

I would like to express my sincere gratitude to Professor Minoru Asada for his thorough discussion and support. I gratefully acknowledge valuable comments of other members of my thesis committee, Professor Koh Hosoda and Professor Hideyuki Nakanishi at Osaka University. I would also like to thank Professor Hiroshi Ishiguro for his patient guidance and continuous support to my reserch. I would like to thank Dr. Takuichi Nishimura whose comments made enormous contribution to my research.

I also greatly appreciate Dr. Norihiro Hagita, Dr. Takahiro Miyashita, Dr. Kazuhiko Shinozawa, Dr. Tadahisa Kondo, Dr. Masahiro Shiomi, my supervisors, and Dr. Dylan F. Glas. Without their patience, discussion and warmful support, I could not have completed the work.

I owe a great deal to the members of Professor Asada's Laboratory. I had many suggestions from Dr. Eiji Uchibe, Dr. Yasutake Takahashi, Dr. Noriaki Mitsunaga, Dr. Takashi Minato, Dr. Yasunori Tada, Dr. Masaki Ogino, and Dr. Yukie Nagai in thorough discussion. I also owe to the members of Professor Ishiguro's Laboratory. I had many suggestions from Dr. Yutaka Nakamura. I am also indebted to Mr. Takeshi Ishida, Mr. Kengo Murota for their cooperation in the work.

I would like to thank Professor Shuji Doshita, Professor Tatsuya Kawahara, and Professor Masahiro Araki who introduced me to the research activities.

Finally, I wish to thank my family and friends for their warm support.

Bibliography

1) 浅田 稔. 情報統合. 甘利 俊一, 外山 敬介（編）, **脳科学大事典**, pp. 479–482. 朝倉書店, 2000.

2) 山崎 弘郎 and 石川 正俊 (編). **センサフュージョンー実世界の能動的理解と知的再構成ー**. コロナ社, 1992.

3) 石川 正俊 (編). 特集「センサフュージョン」. **日本ロボット学会誌**, 12(5), 1994.

4) 片山 徹. **応用カルマンフィルタ**. 朝倉書店, 2000.

5) Dieter Fox, Jeffrey Hightower, Lin Liao, Dirk Schulz, and Gaetano Borriello. Bayesian filtering for location estimation. *IEEE Pervasive Computing*, 2(3):24–33, 2003.

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

7) R. A. Brooks. Intelligent room project. In *Proc. 2nd Int. Cognitive Technology Conf.*, 1997.

8) Cory D. Kidd, Robert Orr, Gregory D. Abowd, Christopher G. Atkeson, Irfan A. Essa, Blair MacIntyre, Elizabeth D. Mynatt, Thad Starner, and Wendy Newstetter. The aware home: A living laboratory for ubiquitous computing research. In *Proc. the Second International Workshop on Cooperative Buildings, Integrating Information, Organization, and Architecture (CoBuild '99)*, pages 191–198, 1999.

9) K. Mase, Y. Sumi, T. Toriyama, M. Tsuchikawa, S. Ito, S. Iwasawa, K. Kogure, and N Hagita. Ubiquitous experience media. *IEEE Multimedia*, 13(4):20–29, 2006.

10) J. Hightower and G. Borriello. Location systems for ubiquitous computing. *Computer*, 34(8):57–66, 2001.

11) M. Kourogi and T. Kurata. Personal positioning based on walking locomotion analysis with self-contained sensors and a wearable camera. In *Proc. 2nd Int. Symposium on Mixed and Augmented Reality (ISMAR03)*, pages 103–112, 2003.

12) Eric Foxlin. Pedestrian tracking with shoe-mounted inertial sensors. *IEEE Computer Graphics and Applications*, 25(6):38–46, 2005.

13) D. Schulz, D. Fox, and J. Hightower. People tracking with anonymous and id-sensors using rao-blackwellised particle filters. In *Proc. 18th Int. Joint Conf. Artificial Intelligence (IJCAI'03)*, pages 921–928, August 2003.

14) K. Nakadai, K. Hidai, H. Mizoguchi, H. G. Okuno, and H. Kitano. Real-time auditory and visual multiple-object tracking for robots. In *Proc. Int. Joint Conf. Artificial Intelligence (IJCAI'01)*, 2001.

15) D. Roy. Grounded spoken language acquisition: experiments in word learning. *IEEE Trans. on Multimedia*, 5(2):197–209, 2003.

16) J. Hershey, J. R. Movellan, and H. Ishiguro. Audio vision: Using audio-visual synchrony to locate sounds. In *Proc. Neural Information Processing Systems (NIPS'99)*, pages 813–819, 1999.

17) J. W. Fisher III, T. Darrell, W. T. Freeman, and P. Viola. Learning joint statistical models for audio-visual fusion and segregation. In *Proc. Neural Information Processing Systems (NIPS'00)*, pages 772–778, 2000.

18) M. Slaney and M. Covell. Facesync: A linear operator for measuring synchronization of video facial images and audio tracks. In *Proc. of the Neural Information Processing Systems (NIPS'00)*, pages 814–820, 2000.

19) H. J. Nock, G. Iyengar, and C. Neti. Assessing face and speech consistency for monologue detection in video. In *Proc. ACM Int. Conf. Multimedia 2002*, pages 303–306, 2002.

20) D. Li, N. Dimitrova, M. Li, and I. K. Sethi. Multimedia content processing through cross-modal association˙ In *Proc. ACM Int. Conf. Multimedia 2003*, pages 604–611, 2003.

21) E. Kidron, Y. Schechner, and M. Elad. Pixels that sound. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR 2005)*, pages 88–95, June 2005.

22) J. W. Fisher III and T. Darrell. Speaker association with signal-level audiovisual fusion. *IEEE Trans. on Multimedia*, 6(3):406–413, 2004.

23) H. J. Nock, G. Iyengar, and C. Neti. Speaker localisation using audio-visual synchrony : An empirical study. In *Proc. the 4th Int. Conf. Image and Video Retrieval (CIVR2003)*,

pages 488–499, 2003.

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

25) T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 1991.

26) S. Shimojo and L. Shams. Sensory modalities are not separate modalities: plasticity and interactions. *Current Opinion in Neurobiology*, 11(4):505–509, Aug 2001.

27) A. Harter, A. Hopper, P. Steggles, A. Ward, and P. Webster. The anatomy of a context-aware application. In *Proc. 5th Annual ACM/IEEE Int. Conf. Mobile Computing and Networking (Mobicom '99)*, August 1999.

28) Weiming Hu, Tieniu Tan, Liang Wang, and S. Maybank. A survey on visual surveillance of object motion and behaviors. *IEEE Trans. on Systems, Man and Cybernetics, Part C*, 34(3):334–352, 2004.

29) M.D. Addlesee, A.H. Jones, F. Livesey, and F.S. Samaria. The orl active floor. *IEEE Personal Communications*, 4(5):35–41, 1997.

30) R. J. Orr and G. D. Abowd. The smart floor: A mechanism for natural user identification and tracking. In *Proc. Conf. Human Factors in Computing Systems (CHI 2000)*, pages 303–306, 2000.

31) 佐藤 知正, 森 武俊, 原田 達也. ロボティックルームの知能ーユービキタス知能ー. **日本ロボット学会誌**, 20(5):482–486, 2002.

32) T. Murakita, T. Ikeda, and H Ishiguro. Human tracking using floor sensors based on the markov chain monte carlo method. In *Proc. 17th Int. Conf. Pattern Recognition (ICPR 2004)*, volume 4, pages 917–920, 2004.

33) 神田 崇行（編）. 特集「ネットワークロボット」. **日本ロボット学会誌**, 23(6):1–56, 2005.

34) G. C. De Silva, T. Yamasaki, and K. Aizawa. Ubiquitous home: Retrieval of experiences in a home environment. *IEICE Trans. on Information and Systems*, E91-D(2):330–340, 2008.

35) Wen-Hau Liau, Chao-Lin Wu, and Li-Chen Fu. Inhabitants tracking system in a cluttered home environment via floor load sensors. *IEEE Trans. on Automation Science and*

Engineering, 5(1):10–20, 2008.

36) 山西 利明, 森 武俊, 森下 広, 原田 達也, 佐藤 知正. 時系列床面圧力分布情報を用いた個人識別システムの開発. **電子情報通信学会技術報告**, volume PRMU2005-167, pages 105–110, 2006.

37) Roy Want, Andy Hopper, Veronica Falcão, and Jonathan Gibbons. The active badge location system. *ACM Trans. Inf. Syst.*, 10(1):91–102, 1992.

38) Tomohiro Amemiya, Jun Yamashita, Koichi Hirota, and Michitaka Hirose. Virtual leading blocks for the deaf-blind: A real-time way-finder by verbal-nonverbal hybrid interface and high-density RFID tag space. In *Proc. IEEE Virtual Reality Conf. 2004*, pages 165–172, March 2004.

39) Lionel M. Ni, Yunhao Liu, Yiu Cho Lau, and Abhishek P. Patil. Landmarc: Indoor location sensing using active rfid. *Wireless Networks*, 10(6):701–710, 2004.

40) P. Bahl and V. N. Padmanabhan. RADAR: An in-building RF-based user location and tracking system. In *Proc. IEEE INFOCOM 2000*, volume 2, pages 775–784, March 2000.

41) K. Mizugaki, R. Fujiwara, T. Nakagawa, G. Ono, T. Norimatsu, T. Terada, M. Miyazaki, Y. Ogata, A. Maeki, S. Kobayashi, N. Koshizuka, and K. Sakamura. Accurate wireless location/communication system with 22-cm error using uwb-ir. In *IEEE Radio and Wireless Symposium*, pages 455–458, 2007.

42) L. Bao and S. S. Intille. Activity recognition from user-annotated acceleration data. In *Proc. of PERVASIVE 2004, vol. LNCS 3001, A. Ferscha and F. Mattern, Eds. Berlin Heidelberg: Springer-Verlag*, pages 1–17, 2004.

43) O. Woodman and R. Harle. Pedestrian localisation for indoor environments. In *Proc. 10th Int. Conf. Ubiquitous Computing (UbiComp '08)*, pages 114–123, 2008.

44) T. Mori, Y. Suemasu, H. Noguchi, and T. Sato. Multiple people tracking by integrating distributed floor pressure sensors and RFID system. In *Proc. of IEEE Int. Conf. Systems, Man and Cybernetics*, volume 6, pages 5271–5278, Oct 2004.

45) J. Hershey, J. R. Movellan, and H. Ishiguro. Audio vision: Using audio-visual synchrony to locate sounds. In *Proc. Neural Information Processing Systems (NIPS'99)*, pages 813–819, 1999.

46) 池田 徹志, 石黒 浩, 浅田 稔. 相互情報量最大化に基づく信号情報源の移動軌跡の推定.

電子情報通信学会誌, J90-D(2):535–543, 2007.

47) 庄司 武, 中村 聡史, 塚本 昌彦, 西尾 章治郎. ウェアラブル計算環境での環境変化を考慮した入力インタフェース. **インタラクション** *2004*, 2004.

48) D. B. Reid. An algorithm for tracking multiple targets. *IEEE Trans. Automatic Control*, 24(6):843–854, 1979.

49) Y. Bar-Shalom and T. E. Fortmann. *Tracking and Data Association*. Academic Press, 1988.

50) I.J. Cox. A review of statistical data association techniques for motion correspodence. *Int. J. Computer Vision*, 10(1):53–66, 1993.

51) 東京大学教養学部統計学教室 (編). **自然科学の統計学**. 東京大学出版会, 1992.

52) ヴィストン (株). http://www.vstone.co.jp/.

53) H. Fujiyoshi, A. Lipton, and T. Kanade. Real-time human motion analysis by image skeletonization. *IEICE Trans. Inf. & Syst.*, E87-D(1):113 – 120, January 2004.

54) Jinshi Cui, Hongbin Zha, Huijing Zhao, and Ryosuke Shibasaki. Laser-based detection and tracking of multiple people in crowds. *Computer Vision and Image Understanding*, 106(2-3):300–312, 2007.

55) H. Zhao and R. Shibasaki. A novel system for tracking pedestrians using multiple single-row laser range scanners. *IEEE Trans. Syst., Man, Cybern. Part A*, 35(2):283–291, 2005.

56) D. F. Glas, T. Miyashita, H. Ishiguro, and N. Hagita. Laser-based tracking of human position and orientation using parametric shape modeling. *Advanced Robotics*, 23(4):405–428, 2009.

57) O. J. Woodman. An introduction to inertial navigation. Technical Report UCAM-CL-TR-696, University of Cambridge, Computer Laboratory, Aug 2007.

58) J. Cronly-Dillon, K. Persaud, and R. P. F. Gregory. The perception of visual images encoded in musical form: a study in cross-modality information transfer. *Proc. R. Soc. Lond. B*, 266:2427–2433, 1999.

59) J. Cronly-Dillon, K. C. Persaud, and R. Blore. Blind subjects construct conscious mental images of visual scenes encoded in musical form. *Proc. R. Soc. Lond. B*, 267:2231–2238, 2000.

60) 小林 真, 太田 道男. 全方位センサと 3 次元音響を利用した視覚障害者用歩行誘導システム. **バイオメカニズム学会誌**, 24(2):123–125, 2000.

61) 上野山 努, 樫村 雅章, 小沢 慎治. ドラム音の音色における感性情報と工学的パラメータの対応づけ. **日本音響学会誌**, 49(10):671–681, 1993.

62) 熊本 忠彦, 太田 公子. 印象に基づく検索のための印象語選定法の提案. **情報処理学会論文誌**, 44(7):1808–1811, 2003.

63) 山脇 一宏, 椎塚 久雄. 音楽とカラーイメージの類似性について. **情報学音楽情報科学研報**, 2002-MUS-047, pages 105–109, 2002.

64) 港 隆史, 関戸 智史, 石黒 浩, 河原 英紀. 全方位視覚の特性を利用した画像から音信号への変換. **日本ロボット学会学術講演会予稿集**, 2002.

65) 米澤 朋子, 間瀬 健二. 流体による楽器インタラクションの考察〜tangible sound #2 における展開. **日本バーチャルリアリティ学会論文誌**, 5(1):755–762, 2000.

66) Kynan Eng *et al.* Ada - intelligent space: An artificial creature for the swiss expo.02. In *Proc. IEEE Int. Conf. Robotics and Automation (ICRA2003)*, pages 4154–4159, 2003.

67) 長田 典子, 岩井 大輔, 津田 学, 和氣 早苗, 井口 征士. 音と色のノンバーバルマッピング−色聴保持者のマッピングルール抽出とその応用−. **電子情報通信学会誌**, J86-A(11):1219–1230, 2003.

68) Richard E. Cytowic. Synesthesia: Phenomenology and neuropsychology. *PSYCHE*, 2(10), Jul 1995.

69) Simon Baron-Cohen. Is there a normal phase of synaesthesia in development? *PSYCHE*, 2(27), Jun 1996.

70) H. McGurk and J. W. MacDonald. Hearing lips and seeing voices. *Nature*, 264:746–748, 1976.

71) M. H. Coen. Multimodal integration - a biological view. In *Int. Joint Conf. Artificial Intelligence*, volume 2, pages 1417–1424, 2001.

72) K. W. Grant and P.F. Seitz. The use of visible speech cues for improving auditory detection of spoken sentences. *J. Acoust. Soc. Am.*, 108(3):1197–1208, Sep 2000.

73) T. Ikeda, H. Ishiguro, and A. Asada. Sensor fusion as optimization: maximizing mutual information between sensory signals. In *Proc. 17th Int. Conf. Pattern Recognition (ICPR 2004)*, volume 2, pages 501–504, 2004.

74) 安達 太郎, 岩宮 眞一郎. 色彩と音楽とが互いに及ぼす影響–ショパンのエチュードを手がかりに. **日本音響学会九州支部学生のための研究会**, 2003.

75) 梅本 尭夫. **音楽心理学**. 誠信書房, 1966.

76) 菅野 禎盛, 岩宮 眞一郎. 音楽のリズムと映像の動きの同期が音楽と映像の調和に及ぼす効果. **音楽知覚認知研究**, 5(1):1–10, 1999.

77) 長嶋 洋一. 音楽的ビートが映像的ビートの知覚に及ぼす引き込み効果. **芸術科学会論文誌**, 3(1):108–148, 2004.

78) 北島 吾郎, 土居 元紀. 画像の構成を手掛かりとした音楽の検索. **情報処理学会関西支部大会ビジュアルインフォメーション研究会**, A-15, 2003.

79) Cycling’74. http://www.cycling74.com/.

Publications by author

— Journal papers —

J1) 池田 徹志, 石黒 浩, 浅田 稔. 相互情報量最大化に基づく信号情報源の移動軌跡の推定. **電子情報通信学会誌**, J90-D(2):535–543, 2007.

J2) 池田 徹志, 室田 健吾, 石黒 浩. 全方位映像から音楽情報へのメディア変換に基づく視覚情報の伝達. **情報処理学会論文誌**, 48(1):274–283, 2007.

J3) 池田 徹志, 石黒 浩, 西村 拓一. 床センサと加速度センサの統計的統合による複数人間同定追跡. **計測自動制御学会論文集**, 45(1):60–68, 2009.

J4) F. Zanlungo, T. Ikeda, and T. Kanda. Social force model with explicit collision prediction. *EPL (Europhysics Letters)*, 93(68005), 2011.

J5) K. Kamei, T. Ikeda, M. Shiomi, H. Kidokoro, A. Utsumi, K. Shinozawa, T. Miyashita, and N. Hagita. Cooperative customer navigation between robots outside and inside a retail shop - an implementation on the ubiquitous market platform -. *Annales des Telecommunications*, 67(7-8):329–340, 2012.

J6) F. Zanlungo, T. Ikeda, and T. Kanda. A microscopic social norm model to obtain realistic macroscopic velocity and density pedestrian distributions. *PLoS ONE*, 7(12)(e50720), 2012.

J7) Z. Yücel, F. Zanlungo, T. Ikeda, T. Miyashita, and N. Hagita. Deciphering the crowd: Modeling and identification of pedestrian group motion. *Sensors 2013*, 13(1):875–897, 2013.

J8) T. Ikeda, H. Ishiguro, T. Miyashita, and N. Hagita. Pedestrian identification by associating wearable and environmental sensors based on phase dependent correlation of human walking. *Journal of Ambient Intelligence and Humanized Computing*, 2013.

J9) D. Brščić, T. Kanda, T. Ikeda, and T. Miyashita. Person tracking in large public spaces using 3d range sensors. *IEEE Trans. on Human-Machine Systems*, 43(6):522–534, 2013.

J10) 池田 徹志, 石黒 浩, Dylan F. Glas, 塩見 昌裕, 宮下 敬宏, 萩田 紀博. 信号の動きの相関性に注目したセンサ非依存の知覚の結び付け手法. **電子情報通信学会論文誌** *D* **（***in press)*, J97-D(3), 2014.

J11) F. Zanlungo, T. Ikeda, and T. Kanda. A potential for the dynamics of pedestrians in a socially interacting group. *Physical Review E* 89, 012811, 2014.

— Major publications —

C12) T. Ikeda, H. Ishiguro, and M. Asada. Attention to clapping - a direct method for detecting sound source from video and audio -. In *Proc. of IEEE Conf. Multisensor Fusion and Integration for Intelligent Systems (MFI2003)*, pages 264–268, 2003.

C13) T. Ikeda, H. Ishiguro, and M. Asada. Adaptive fusion of sensor signals based on mutual information maximization. In *Proc. of IEEE Int. Conf. Robotics and Automation (ICRA2003)*, pages 4398–4402, 2003.

C14) T. Ikeda, H. Ishiguro, and A. Asada. Sensor fusion as optimization: maximizing mutual information between sensory signals. In *Proc. 17th Int. Conf. Pattern Recognition (ICPR 2004)*, volume 2, pages 501–504, 2004.

C15) T. Murakita, T. Ikeda, and H Ishiguro. Human tracking using floor sensors based on the markov chain monte carlo method. In *Proc. 17th Int. Conf. Pattern Recognition (ICPR 2004)*, volume 4, pages 917–920, 2004.

C16) T. Ikeda, T. Ishida, and H. Ishiguro. Framework of distributed audition. In *Proc. of 13th IEEE Int. Workshop on Robot and Human Interactive Communication (RO-MAN 2004)*, page 1B4, 2004.

C17) K. F. MacDorman, H. Nobuta, T. Ikeda, S. Koizumi, and H. Ishiguro. A memory-based distributed vision system that employs a form of attention to recognize group activity at a subway station. In *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS 2004)*, pages 571–576, 2004.

C18) T. Ikeda, H. Ishiguro, and T. Nishimura. People tracking by fusing different kinds of sensors, floor sensors and acceleration sensors. In *Proc. of IEEE*

Multisensor Fusion and Integration for Intelligent Systems (MFI2006), pages 530–535, 2006.

C19) T. Ikeda, H. Ishiguro, and T. Nishimura. People tracking by cross modal association of vision sensors and acceleration sensors. In *Proc. of IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS2007)*, pages 4147–4151, 2007.

C20) T. Ikeda, H. Ishiguro, D. F. Glas, M. Shiomi, T. Miyashita, and N. Hagita. Person identification by integrating wearable sensors and tracking results from environmental sensors. In *Proc. of IEEE Int. Conf. Robotics and Automation (ICRA2010)*, pages 2637–2642, 2010.

C21) K. Kamei, K. Shinozawa, T. Ikeda, A. Utsumi, T. Miyashita, and N. Hagita. Recommendation from robots in a real-world retail shop. In *12th Int. Conf. on Multimodal Interfaces and 7th Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI 2010)*, pages 19:1–19:8, 2010.

C22) K. Okamoto, A. Utsumi, T. Ikeda, H. Yamazoe, T. Miyashita, S. Abe, and N. Hagita. Classifcation of pedestrian behavior in a shopping mall based on lrf and camera observations. In *12th IAPR Conf. on Machine Vision Applications (MVA2011)*, pages 1–5, 2011.

C23) T. Ikeda, Y. Chigodo, T. Miyashita, F. Kishino, and N. Hagita. A method to recognize 3d shapes of moving targets based on integration of inclined 2d range scans. In *Proc. IEEE Int. Conf. on Robotics and Automation (ICRA 2011)*, pages 3575–3580, 2011.

C24) Z. Yücel, T. Ikeda, T. Miyashita, and N. Hagita. Identification of mobile entities based on trajectory and shape information. In *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS2011)*, pages 3589–3594, 2011.

C25) K. Kamei, T. Ikeda, H. Kidokoro, M. Shiomi, A. Utsumi, K. Shinozawa, T. Miyashita, and N. Hagita. Effectiveness of cooperative customer navigation from robots around a retail shop. In *Third IEEE International Conference on Social Computing (SocialCom 2011)*, pages 235–241, 2011.

C26) K. Lee, T. Ikeda, T. Miyashita, H. Ishiguro, and N. Hagita. Separation of tactile information from multiple sources based on spatial ica and time series clustering. In *IEEE/SICE International Symposium on System Integration (SII 2011)*, pages 791–796, 2011.

C27) T. Ikeda, H. Ishiguro, T. Miyashita, and N. Hagita. Pedestrian identification by associating walking rhythms from wearable acceleration sensors and biped tracking results. In *Proc. of Int. Conf. Pervasive and Embedded Computing and Communication Systems (PECCS 2012)*, pages 21–28, 2012.

C28) T. Ikeda, Y. Chigodo, D. Rea, F. Zanlungo, M. Shiomi, and T. Kanda. Modeling and prediction of pedestrian behavior based on the sub-goal concept. In *Proc. Robotics Science and Systems (RSS)*, 2012.

C29) Z. Yücel, F. Zanlungo, Tetsushi Ikeda, Takahiro Miyashita, and Norihiro Hagita. Modeling indicators of coherent motion. In *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS2012)*, pages 2134–2140, 2012.

C30) F. Zanlungo, Y. Chigodo, T. Ikeda, and T. Kanda. Experimental study and modelling of pedestrian space occupation and motion pattern in a real world environment. In *6th Int. Conf. on Pedestrian and Evacuation Dynamics*, 2012.

C31) A. Kanemura, Y. Morales, M. Kawanabe, H. Morioka, N. Kallakuri, T. Ikeda, T. Miyashita, N. Hagita, and S. Ishii. A waypoint-based framework in brain-controlled smart home environments: Brain interfaces, domotics, and robotics integration. In *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS2013)*, page 865–870, 2013.

— Other publications —

M32) 池田 徹志, 石黒 浩, 浅田 稔. ユビキタスセンサ環境におけるカメラと床面上の接触センサを統合した人間追跡. **情報処理学会関西支部 環境知能研究会報告**, number B07, pages 117–120, 2003.

M33) 池田 徹志, 石黒 浩, 浅田 稔. 足音を見る − 画像と音の直接統合による移動音源の位置同定,. **日本ロボット学会第** *21* **回学術講演会講演予稿集**, 2003.

M34) 池田 徹志, 石黒 浩, 浅田 稔. ユビキタスセンサ環境における音と画像の直接統合. **人工知能学会 第** *18* **回** *AI Challenge* **研究会予稿集**, SIG-Challenge-0318-8, pages 45–50, 2003.

M35) 池田 徹志, 室田 健吾, 石黒 浩. 画像から音へのメディア変換に基づく音による視覚情報伝達. **第5回計測自動制御学会システムインテグレーション部門講演会** *(SI2004)*, 2004.

M36) 石田 武志, 池田 徹志, 石黒 浩. 分散聴覚システムの基礎研究. **日本ロボット学会第*22*回学術講演会講演予稿集**, 3D17, 2004.

M37) 室田 健吾, 池田 徹志, 石黒 浩. 画像情報から音楽情報へのメディア変換による環境認識. **情報処理学会関西支部 環境知能研究会報告**, B-07, 2004.

M38) 池田 徹志, 石黒 浩, 西村 拓一. 床センサと加速度センサの統合による複数人間追跡. **電子情報通信学会技術研究報告**, volume 106, pages 7–12, 2006.

M39) 池田 徹志, 石黒 浩, Dylan F. Glas, 塩見 昌裕, 宮下 敬宏, 萩田 紀博. 環境設置型センサとウェアラブルセンサを統合した個人同定. **電子情報通信学会技術研究報告**, volume 109 of *PRMU2009-191*, pages 243–248, 2010.

M40) 池田 徹志, 児堂 義弘, D. Rea, F. Zanlungo, 塩見 昌裕, 神田崇行. 街角における歩行者のサブゴール遷移モデル. **日本ロボット学会第*31*回学術講演会講演予稿集**, RSJ2013, 3I2-03, 2013.