Chapter 5
Behavior Understanding Based on Intention-Gait Model

Yasushi Yagi, Ikuhisa Mitsugami, Satoshi Shioiri and Hitoshi Habe

Abstract Gait is known as one of biometrics, and there have been many studies on gait authentication. In those studies, it is implicitly assumed that the gait of a certain person is always constant. It is, however, untrue in reality; a person usually walks differently according to their mood and physical/mental conditions, which we call "inertial states." Motivated by this fact, we organized the research project "Behavior Understanding based on Intention-Gait Model", which was supported by JST-CREST from 2010 to 2017. The goal of this project was to map "gait", in the broad sense of the term, to inertial states such as attention, social factors, and cognitive ability. In this chapter, we provide an overview of the three kinds of estimation technologies considered in this project: attention, social factors, and cognitive ability.

Keywords Gait analysis · Eye-head coordination · Visual perception · Gaze estimation · Group segmentation · Dual-task · Cognitive level estimation · Dementia diagnosis · Huge data collection

5.1 Introduction

Walking is a behavior that is fundamental to our daily lives. Because gait (i.e., way of walking) is unique to each person, it is regarded as a biometric property and can be applied to person authentication tasks. There have been many studies

Y. Yagi (✉) · I. Mitsugami
Osaka University, Osaka, Japan
e-mail: yagi@sanken.osaka-u.ac.jp

I. Mitsugami
e-mail: mitsugami@am.sanken.osaka-u.ac.jp

S. Shioiri
Tohoku University, Miyagi, Japan
e-mail: shioiri@riec.tohoku.ac.jp

H. Habe
Kindai University, Osaka, Japan
e-mail: habe@kindai.ac.jp

T. Nishida (ed.), *Human-Harmonized Information Technology, Volume 2*,
DOI 10.1007/978-4-431-56535-2_5

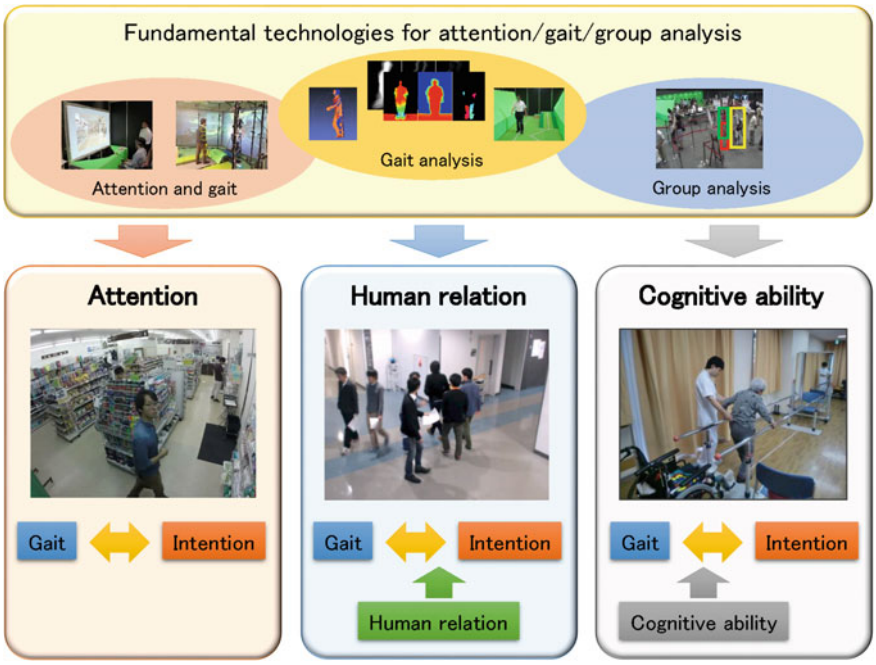


Fig. 5.1 Overview of JST-CREST "Behavior understanding based on Intention-Gait model"

on gait authentication [1–7], and they have revealed that gait is a very effective cue for identifying a person using public gait datasets, such as the CASIA Gait Database [8], the USF Human ID Gait Baseline Database [9], and the OU-ISIR Gait Database [10]. All of these studies, however, implicitly assume that the gait of a certain person is always constant, which is untrue in reality; a person usually walks differently according to their mood and physical/mental conditions, which we call "inertial states." Motivated by this fact, we organized the research project "Behavior Understanding based on Intention-Gait Model", which was supported by JST-CREST from 2010 to 2017. Note that in our previous research, "gait" simply denotes a way of walking, while in this study, we extend this definition to include eye and head motions and walking trajectory. The goal of this project was to map "gait", in the broad sense of the term, to inertial states such as attention, social factors, and cognitive ability. In this chapter, we provide an overview of the three kinds of estimation technologies considered in this project: attention, social factors, and cognitive ability, as shown in Fig. 5.1. Note that in the project we have also developed many fundamental methods for these estimation technologies; calibration of range sensors [11, 12], 3-D reconstruction of human body [13, 14], detection of injured people [15, 16], and gait analysis [12, 17, 18].

Attention is an inertial state that is known to be affected by the interests, feelings, and intention of an individual. For instance, security agencies report that shoplifters often show unique gazing behaviors. Thus, if one were to obtain and analyze gaze

activity of people in shops, it may be possible to detect shoplifters. Similarly, customers usually gaze at products that they are interested in. If this type of gaze information were acquired, it might be possible to use it to effectively give customers salient information about promotions and sales. Considering these examples, gaze information has many potential applications. It is, however, usually impossible to obtain gaze information in real environments because such data collection requires the use of an eye tracker. Thus, it is necessary to consider other ways of collecting gaze information. For this purpose, in this project, we examined the relationship between gaze behavior and whole body behaviors, such as gait. We collected data from many participant to facilitate the development of appropriate models. Section 5.2 describes the details of our investigation.

Section 5.3 describes our approach to social interaction estimation. In the course of daily activities, humans often form social groups, such as families and groups of coworkers. When we are in such groups, our actions are strongly affected by and have a large effect on other group members. Investigating social relationships is thus an important way to understand and/or infer information about our daily activities. This section is focused on explaining our proposed method for identifying social groups by analyzing time-series range data.

Section 5.4 concerns cognitive ability estimation. According to a national investigation report, more than 4.4 million people in Japan present with dementia, which is associated with obstacles in brain function that affect understanding, judgment, and memory. An additional 37 thousand people nation-wide present with early-onset dementia due to cerebral vascular disease. It is important to detect dementia in its early stage and conduct treatments accordingly. To examine methods of detecting early onset dementia, we adopted a "dual-task" procedure, in which participants simultaneously complete a physical task (e.g., walking, stepping, and dancing) and a cognitive task (e.g., counting down numbers, telling words of a certain category). This section describes a measurement systems that we have developed for this purpose, and introduces some recent results.

5.2 Gaze Prediction from Body Movements

Gaze location is one of the most useful ways to glean information about people's interests and can be used to estimate intentions, because gaze location is usually assumed to be at the location of the attention focus, although these are not always consistent with each other. To predict gaze location, a number of models using saliency maps, which topographically represent the visual saliency of a given scene, have been proposed [19, 20]. Saliency maps are based on the bottom-up architecture of visual attention, which involves the hypothesis that the most salient locations in a visual scene tend to attract attention. Visual saliency is calculated by integrating the visual features of a scene, such as color, luminance, and orientation, often with a considerable variety of visual functions, like retinal inhomogeneity [21] and the canceling out of self-motion [22]. However, the accuracy of gaze prediction using

visual saliency alone is limited because it is based on bottom-up factors such as visual features, and does not account for the influence of top-down factors such as the intention of the participant [23–25]. Models that account for top-down factors undoubtedly provide better gaze prediction. Machine learning techniques are one of the best ways to add information about possible gaze locations from empirical data, and are effective when the task and scenes are known, making learning possible beforehand. However, to improve the accuracy of gaze and attention prediction in generalized conditions, other techniques may be required. Previously, a method was proposed to predict gaze locations using head direction [26], following the purpose of the project to link body movements and intention as described hereafter.

Coordinative movements of the eye and head have been shown experimentally, by measuring eye and head movements during simple gaze shifts to targets present in the periphery from the central fixation [27, 28]. These previous studies revealed that eye-head movements are coordinated when gaze shifts are sufficiently large. Although these results suggest the potential use of head direction in gaze estimation, the conditions of these studies are far from natural conditions, where gaze predictions are desired. Therefore, as a first step, we investigated the relationship between eye and head movements when people are continuously shifting their gaze.

5.2.1 Eye-Head Coordination

We conducted three experiments to explore the relationship between the eye and head movements under more natural conditions than single gaze shifts. The first experiment used 360° surrounding display system and a visual search task [29], the second experiment used natural scene pictures on a large display with more than two hundred participants [26], and the third experiment used a movie in a wide field of view display, which was 7,680 pixels wide by 4,320 pixels tall (8 K) [30]. In all of the experiments, eye and head movements were measured while the participants performed the required tasks (Figs. 5.2 and 5.3).

In the first experiment [29], participants searched for a target across six displays, moving their eyes, head and body inside the space surrounded by the displays (Fig. 5.2). The search display consisted of one target, T, and seven distractors, Ls, on one of the six displays, and eight Ls on the other displays. To search for the target, the participant moved their body and head as well as their eyes to look at all the displays. The eye and head movements were recorded during the period of visual search. Figure 5.4 shows an example of the eye and head movements during visual search. There is one head movement in the figure, and coordination between the eyes and head can be seen. The eyes tend to shift in the direction of the head movement, and therefore the gaze location is farther away from the front than the head direction.

The distribution of the eye orientations was analyzed for different head directions to determine the relationship between the eye and head orientations. Figure 5.5 shows the results of the analysis. Horizontal head orientation is binned and noted at the top of each panel in Fig. 5.5a. The vertical axis shows the percentage of fixations, and the

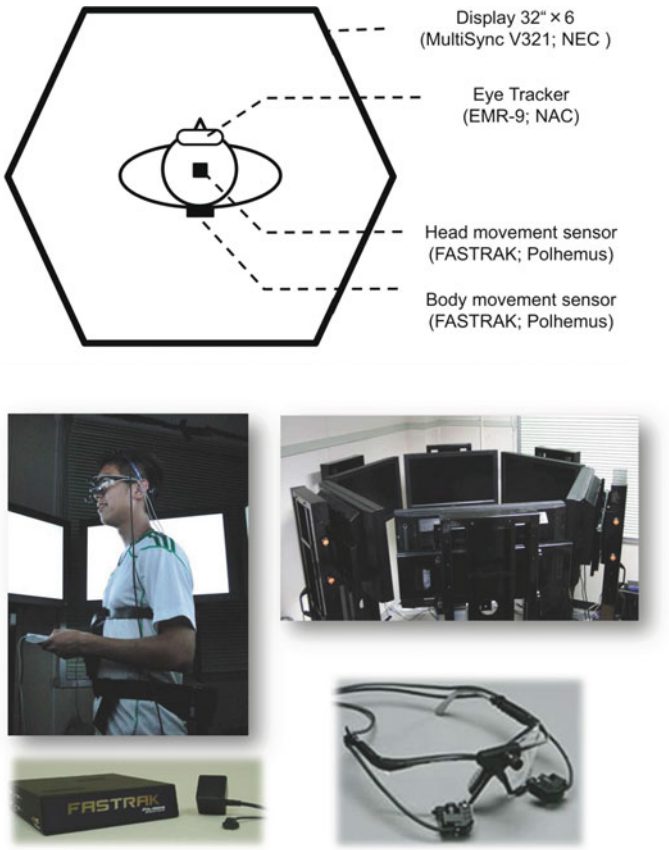


Fig. 5.2 The experiment was performed in a dark room using six 32-inch liquid crystal displays, arranged in the shape of a regular hexagon. Eye-in-head movements were recorded by an eye tracker, and head and body movements were recorded by an electromagnetic motion tracking system

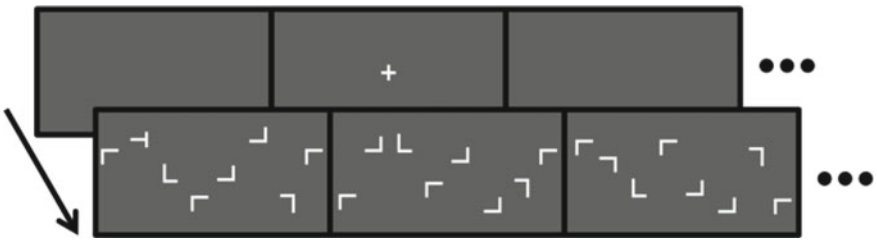


Fig. 5.3 Target letter, T, and distractor letters, Ls, were presented randomly on a *gray* background on one display, and only Ls were presented on the other five displays. Search frames were presented after fixation display

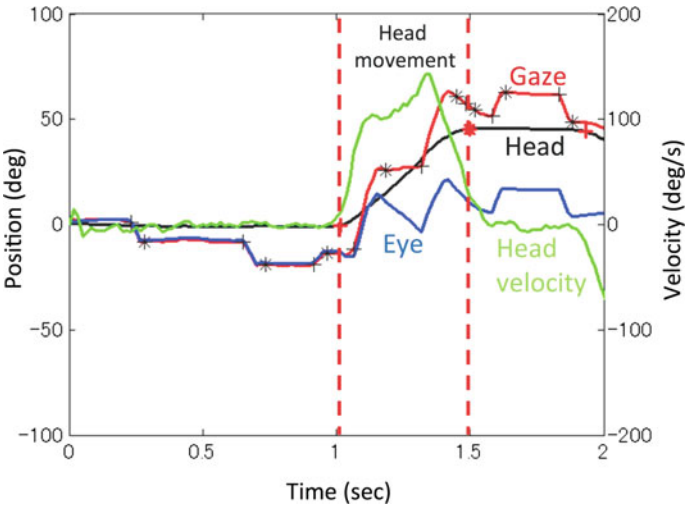


Fig. 5.4 An example of the eye and head movements. The blue line represents eye position, the *black line* represents head position, the *red line* represents gaze location, and the *green line* represents head velocity. The region between the two *red dashed lines* indicates the period of head movement, which was identified based on the head velocity (see the *green line*)

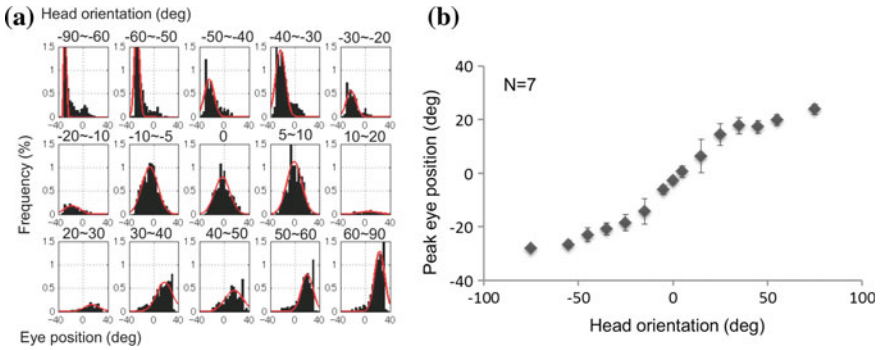


Fig. 5.5 **a** Distribution of eye direction during fixation plotted separately for different head orientations. The *red line* is the Gaussian function fitted to each set of data. **b** Peak eye directions estimated from the Gaussian functions fitted to each participant and averaged over the seven participants. Error bars indicate the standard error of the mean

horizontal axis shows the horizontal eye position relative to the head. To approximate the distribution, we fitted a Gaussian function to each set of data. The red line in each panel shows the function fitted to the average of all participants. Figure 5.5b shows the peak estimated from the Gaussian functions for different head orientations. There is a clear tendency for the eye to be directed in the same orientation as the head relative to the body. That is, when the head was oriented left (or right), the eyes also tended to orient to the left (or right) relative to the head.

Fig. 5.6 Natural scenes used in the second experiment. Outside scenes (*left*) and inside scenes (*right*) are used

In the second experiment [26], we investigated whether there was similar eye and head coordination as that found in the first experiment while viewing natural scenes presented on a 100-inch screen. A total of 30 natural scenes (six indoor and 24 outdoor) containing numerous objects (see Fig. 5.6) were prepared as stimuli and projected onto a large screen. The size of each image was designed to be $57° \times 44°$ from a viewing distance of 125 cm. The eye and head movements were recorded during a 5 s observation period. This experiment was conducted during an outreach activity at the National Museum of Emerging Science and Innovation in Tokyo, Japan, known as "Miraikan." Study participants comprised 228 museum visitors. All the participants had normal or corrected-to-normal vision.

Figure 5.7 shows the peak eye directions estimated from the Gaussian functions fitted to pooled data. There is a similar tendency to that of the results of the first experiment. The eyes tended to be directed in the same orientation as that of the head relative to the body. The eyes also tended to be directed to the left (or right) relative to the head when the head was orientated left (or right) when the participant was looking at natural scenes as well. The experiment also showed that this tendency was maintained for data pooled over more than 200 people.

It was also confirmed that there was similar eye and head coordination when the participants were watching a movie on an 85-inch 8 K ultrahigh-definition television

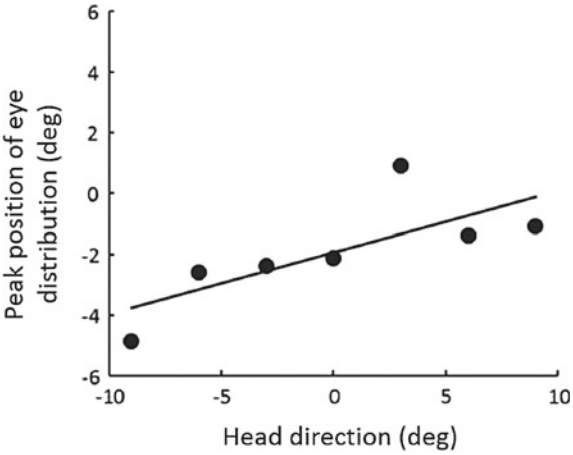


Fig. 5.7 Peak eye directions estimated from the Gaussian functions fitted to each participant and averaged over the seven participants

(UHDTV) in the third experiment [30], where two viewing distances were used to determine the effect of stimulus size on visual angle. The eye and head movements were recorded during the movie, which was 15 min. The movie included scenes of people performing activities such as walking in the streets, surfing, paragliding, and playing football, and nature scenes such as the sun rising, a sea of clouds, flowers, animals. The distribution of eye orientation was biased in the direction of head orientation. When the head was oriented to the left/right, the eyes also tended to orient to the left/right relative to the head. An important finding of the third experiment was that the eye-head coordination in the vertical direction was similar to that in the horizontal direction. In the first and second experiments, the vertical eye direction was distributed around the orientation of the head without showing a bias toward the direction of the head. Figure 5.8 shows the peak eye directions estimated from the Gaussian functions fitted to pooled data as in the previous experiments. In addition to the eye-head coordination in the horizontal direction, there was a similar tendency for the vertical direction. The distribution of eye orientation was biased in the direction of head orientation. When the head was oriented to the top/bottom, the eyes also tended to orient to the top/bottom relative to the head. The reason why we found that the eye orientation was biased toward the head orientation, even for the vertical direction in this experiment, may be because of the image size. The vertical dimensions of the stimuli were was 34, 44 and 53/90° (farther/closer viewing distance) for the first, second and third experiments, respectively. It is not surprising that the head moves more when there is a larger field of view. A question still remains as to why there is no eye-head coordination for field sizes smaller than 44°. Although we have no answer to this question, this is consistent with the finding from a single gaze shift that there is little head movement for a small gaze shift. Thus, this eye-head coordination may be unique to large gaze shifts.

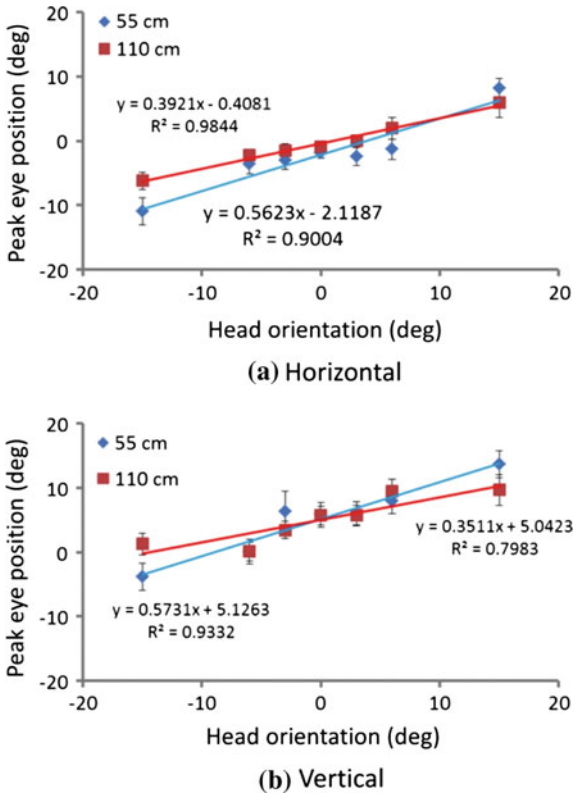


Fig. 5.8 Peak eye directions estimated from the Gaussian functions fitted to each participant and averaged over all participants. The *top* panel shows horizontal movements and the *bottom* panel shows vertical movements. The *red* symbols and line represent the viewing distance of 110 cm and a smaller size of visual angle, and the *blue* symbols and line represent the viewing distance of 55 cm and the larger size of the visual angle

5.2.2 *Eye-Head Coordination and Visual Perception*

The finding that there is similar eye-head coordination in different conditions suggests that head orientation is directly related to visual perception. Although the eyes do not orient in the same direction as the head, the difference between the eye and head orientation is smaller when the head moves, compared with the condition without head movements; that is, head movement reduces the difference. Studies of single gaze shifts have shown that the head tends to be immobile when the gaze shift is small: an estimation of the range for the immobile head is less than about 30° on average. To investigate the influence of head orientation on visual perception, Nakashima and Shioiri conducted visual search experiments in the present project [31, 32]. They controlled the head orientation: the head was oriented 30° in one condition (lateral viewing condition) and it was oriented straight ahead in the other (front viewing condition). The eye location relative to the stimulus display was virtually the same, and the visual system received identical retinal stimulation between the two conditions. Even with the same retinal stimulation, there was a difference in the reaction time to detect the target (the task was to respond to the orientation of

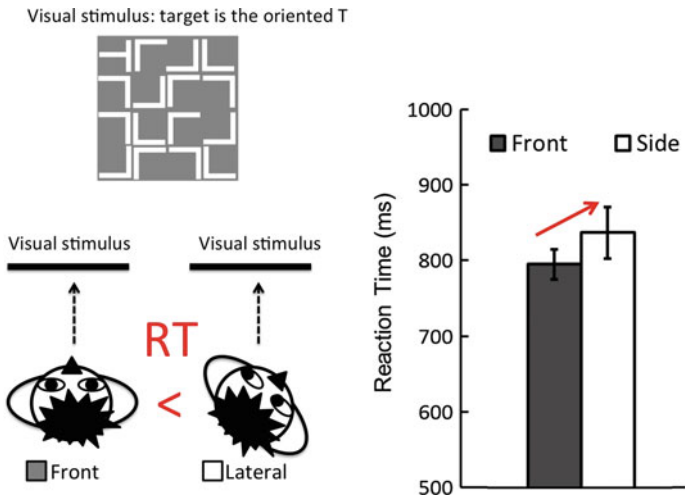


Fig. 5.9 Experimental condition (*left*) and reaction time results. There are two conditions: the head oriented to the visual stimulus in the front viewing condition, and the head oriented 30° from the visual stimulus in the lateral viewing condition. The reaction time is the time to detect and respond to the direction of the target T (*left* or *right*)

the target (left/right) as quickly as possible). A shorter reaction time was found in the front viewing condition, compared with the lateral viewing condition. That is, better performance was found when looking straight ahead (front viewing) than when looking to the side (lateral viewing).

The coordination of the eye and head movements supports the possible facilitation effect of visual processing via head orientation. It was found in the visual search experiment with the 360° surrounding display that there were frequent multiple gaze shifts during a single head movement [29], as shown in Fig. 5.4: the eye jumps twice (two sequential saccadic eye movements) during a single head movement to shift the gaze twice. If there were only one gaze shift in this case, the difference between the eye and head orientations would be larger, which could impair the visual perception (see Fig. 5.9). The analysis of the eye-head coordination revealed that single head movements with multiple saccades constituted as much as 57% of the total head movements with saccades in the visual search experiment in a 360° field of view. These results are different from those of studies using simple tasks, where eye-head coordination was derived from a single head movement with one saccade [27, 28]. No multiple saccades are found in such conditions in general, except for corrective saccades that are much smaller than a primary saccade. Multiple saccades may be critical to the performance of relatively difficult tasks, such as a visual search with sequential gaze shifts. Multiple saccades are one type of eye-head coordination that makes the difference between the eye and head orientations smaller. This smaller difference between the eye and head orientations possibly facilitates visual processing if front viewing is better than lateral viewing [31, 32].

The effect of eye-head coordination on visual perception is important for estimating intention based on gait as in the present project, or on action in general. To focus on an object, the head tends to orient to the object as well as the eye and attention. It is not only the eye, but the head and perhaps the body as well, that is the window to the mind.

5.2.3 Gaze Prediction with Head Orientation

The eye-head coordination results suggest that head orientation can be used to predict gaze location. The distribution of eye position varies systematically dependently on head orientation (Figs. 5.6, 5.7 and 5.8), and the reason that the head orients to the object of interest is to facilitate visual processing (Fig. 5.9). If head orientation can be used to predict gaze location, it has an advantage in terms of the effect on field of application. Usually gaze prediction systems work best for tasks, image types and other conditions specific to the situation of interest. Machine learning techniques are used to build an attention model with a certain bias toward a given task and image type [24], but the model cannot be applied to other tasks and image types easily. The eye-head coordination is general, and can be used without much restriction in terms of tasks and images, although head orientation needs to be measured using a device such as a monitoring camera.

Nakashima et al. proposed a method of gaze prediction using head orientation [26]. The basic idea of the model is to combine the head orientation effect and an attention model of the saliency map, which describes how much an area of the image attracts attention based on low-level visual features (bottom-up attention). The proposed model uses the eye position distribution for a given head orientation to the weighted value of saliency, using the knowledge of eye-head coordination experimentally obtained. The distribution of eye position is estimated based on head orientation, and used as the weighting function. In this study, we apply this model to an experiment using natural scenes presented on a 100-inch screen with 228 museum visitors (the second eye-head-coordination experiment).

Our model uses a saliency map [20], and modulates the map with head direction using weighting functions (Fig. 5.10b). To calculate the saliency map, visual input is decomposed into a set of topographic feature maps, such as those for color and orientation. The feature maps represent the spatial distribution of saliency of the individual features. The information from the individual feature maps is integrated into one map after normalization, and this is the saliency map. The saliency map estimates the activity of the visual cortex, and it is assumed that an area with higher cortical activity attracts attention and, thus, fixation. For each image, a saliency map was calculated following Itti et al. [19].

Assuming that head orientation is given, our model estimates the eye position distribution approximated by a Gaussian function for the head orientation. For this purpose, the relationship between the head orientation and the peak of the eye position distribution was modeled by a linear function as shown by the line in Fig. 5.6. Using

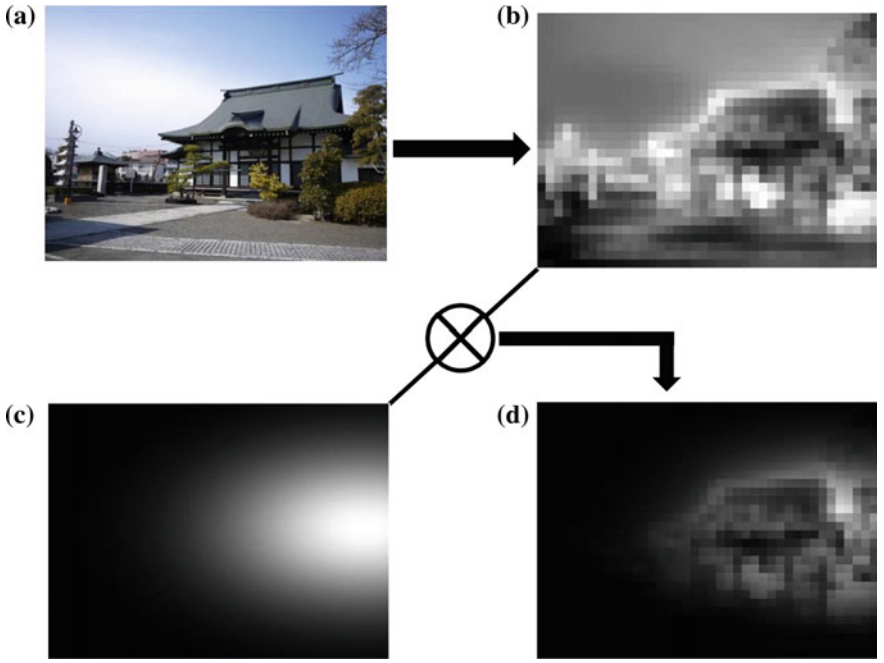


Fig. 5.10 The concept of gaze prediction using head orientation. **b** shows the saliency map of the input image **a**. **c** shows weights of fixation probability estimated from head orientation. **d** shows the gaze prediction map realized by multiplying the saliency map (b) and the weighting function **c**. Reproduced, with permission, from [26]

the linear function, the peak of the eye distribution function was derived for the head orientation. For the spatial spread of the weighting function, the model uses the average of the space constant of Gaussian functions fitted to the experimental results for different head directions. The weighting function was also a Gaussian function, so the peak and the space constant determined the shape. The weighting function was calculated from the head orientation measured experimentally (Fig. 5.10c), at each fixation location during the 5 s of picture viewing. The saliency map was modulated by the weighting function with multiplication for each fixation period to provide the model output of gaze prediction, the prediction map (Fig. 5.10d). The prediction map, therefore, changed with time alongside the head movements. The study used three additional models for gaze prediction for comparison. The first was the original saliency map; the second was our model but without eye-head coordination, which used the head orientation as the peak of the weighting function; and the third was weighted higher at the center of the stimulus scene, which models the tendency to look at the center of the images (the center bias). The last two models used the same spatial spread of the weighting function as the proposed model.

To evaluate model prediction, the study counted the number of fixations within the area defined by the model as the most probable one to attend to, which was

the area with a weighted saliency value (or gaze prediction values) larger than a given threshold. If the prediction of the model is perfect, all the fixations should be within the area, and the percentage of fixations inside the area to all fixations should be 100%, with a lower percentage indicating a less accurate prediction. For the evaluation, the study divided the data into two halves. The eye-head-coordination data were obtained from one of the halves, which was then used for prediction of the other half. The head orientation was used from each participant, and was independent of the model. There was no overlap of participants and stimulus images between the two data sets: one for prediction and the other for testing.

To obtain the general characteristics of the model prediction, we drew receiver operating characteristics (ROC) curves for each model. For this purpose, the percentage of fixations inside the prediction area (hit rate) was calculated for different levels for the prediction values (threshold level): from 10% to 90% with steps of 10%. The top 10% area was determined so that the area size was 10% of the whole field and the gaze prediction values inside the area was larger than the levels outside. The ROC curve is the hit rate as a function of the threshold level (Fig. 5.11a). The study obtained the area under the ROC curve (AUC) to compare accuracy levels

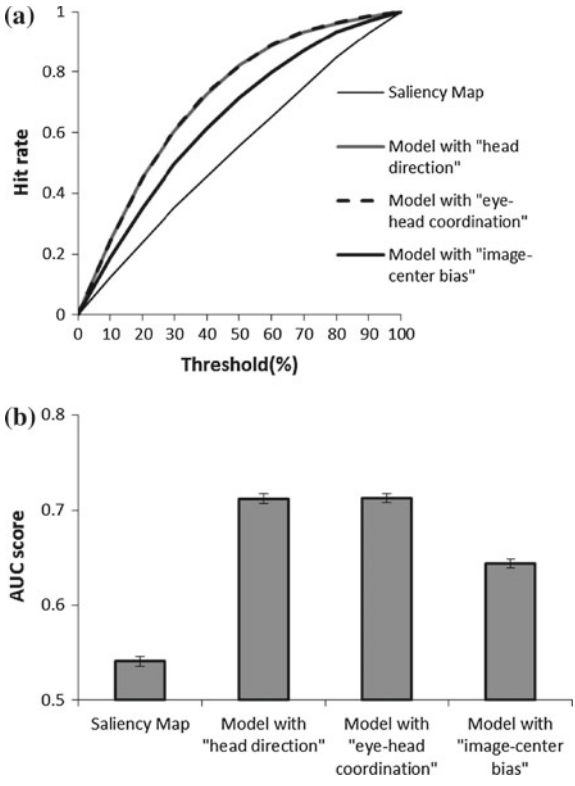


Fig. 5.11 Evaluation results for each gaze prediction model: **a** ROC curves of the models. All of our proposed models are saliency maps with Gaussian distributions whose centers are head direction (head-direction model), the peak of the eye position distribution based on head direction (eye-head-coordination model), and the center of the images (image-center bias model). **b** AUCs of the models. Error bars indicate standard errors across the different scenes. Reproduced, with permission, from [26]

for different models. A larger AUC indicates a better prediction, because the AUC increases with more gaze fixations within higher saliency scores.

Figure 5.11b compares the prediction accuracy among four different models. Accuracy for the models with head direction information (either with or without eye-head coordination) was higher than that of the other two, and the accuracy of the center bias model was higher than that of the saliency map model. There was no significant difference between the two methods that used head direction information. This result indicates that the gaze bias from the head center based on head direction has little effect on gaze prediction accuracy. This is not surprising, because the variance of the Gaussian function is much larger than the peak shift in relation to head direction. Note that this does not imply that the eye-head coordination is completely useless. Recent analysis using the data from watching a movie from an 85-inch 8K UHDTV (the third experiment) [30] shows that the prediction accuracy is higher with eye-head coordination than without it. Although the result is still preliminary, there may be larger effects of eye-head coordination for movies than for static scenes (i.e., a higher slope in Fig. 5.8 than in Fig. 5.7).

The proposed model clearly shows the importance of head orientation for estimating gaze locations. It is interesting to consider the relationship of fixation to center bias. The center bias of gaze is usually consistent with the head orientation bias, because the head directs the center of the stimulus in typical experiments to record gaze shifts on stimuli. The studies of the present project showed the effect of head orientation independently of the center of the stimulus, which suggests that the center bias of the gaze may be a result of the head bias [29], at least partially. A central bias to the display has been reported with success in a model of attention to predict gaze location [33]. The prediction accuracy in our study showed that the model with head orientation (and eye-head coordination) was better than the model with image-center bias. Thus, head orientation likely plays a more important role in gaze estimation compared with the center bias.

5.2.4 Towards Gaze Prediction with Gait Information

We have described so far that gaze can be predicted accurately using head orientation. The goal of the present project is to build a system to predict gaze from gait information. Following studies of eye-head coordination and its application to gaze prediction, we investigated the relationship between gait information and gaze location. To measure gaze information during walking, we conducted the following experiments.

We constructed an immersive environment consisting of a treadmill and surrounding multiple screens and projectors as shown in Fig. 5.12. Figure 5.13 shows its metric. Each participant walked on the treadmill while gazing at the target object projected on the screens. The screens showed a corridor-like virtual space, which flowed based on the treadmill speed to make the participant feel as if they were actually walking in that space. There was also a gaze target (50 cm diameter, green sphere,

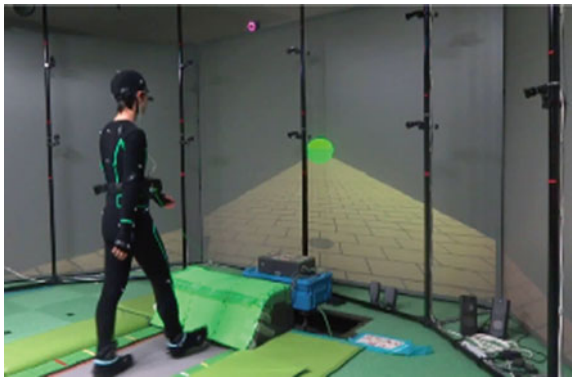


Fig. 5.12 Immersive walking environment

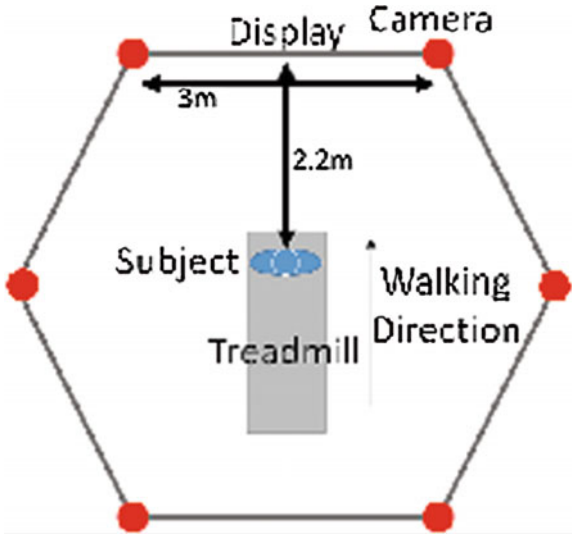


Fig. 5.13 Dimension of the immersive environment

$7.6° \times 7.6°$ of visual angle from the participant). The target randomly appeared at five positions as shown in Fig. 5.14. Six cameras connected to a motion capture system (Bonita 10, Vicon Motion Systems Ltd., UK) were located around the environment. Using the motion capture system, we could obtain all body positions and poses, including those of the head and chest.

Figure 5.15 shows the relationship between the gaze and head directions and between the gaze and chest directions during walking. The horizontal axes denote the gaze directions and the vertical axes denote the head or chest directions. In these graphs, the obtained data and the averages and standard deviations of all participants are shown. The dotted lines denote robust regression results. From these graphs, we find that the angles of the gaze, head and chest have linear relationships similar to those under the non-walking condition discussed in Sect. 2.1, even under the walking condition.

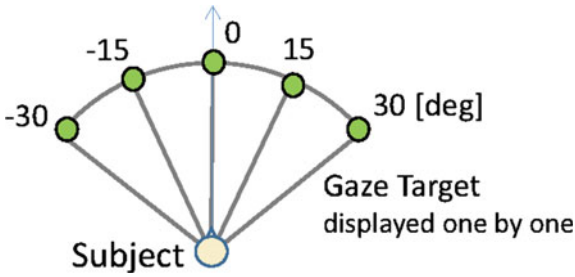


Fig. 5.14 Target positions

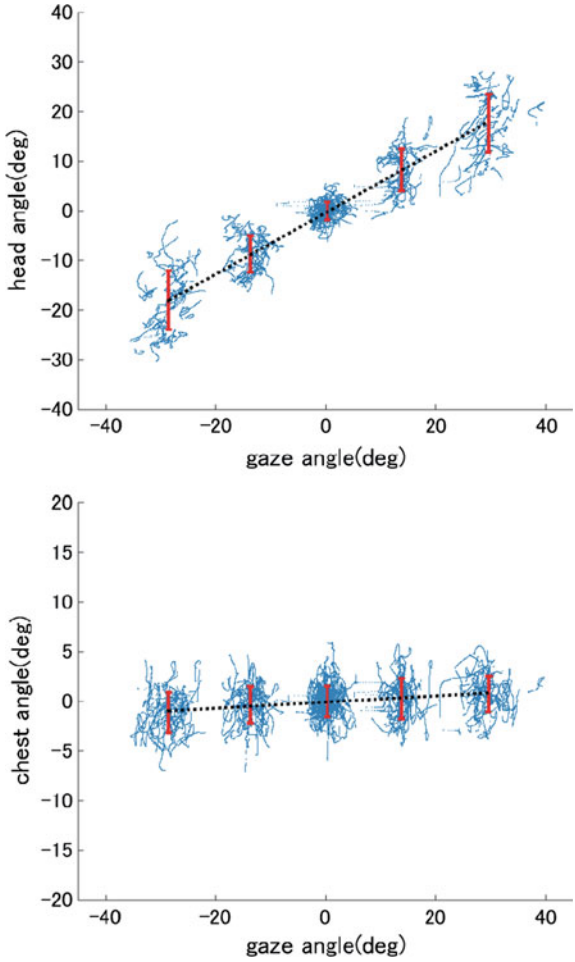


Fig. 5.15 Gaze-head and gaze-chest relationships while walking

Moreover, it was found that not only the head but also the arm and leg movements are related to the gaze location. When the head was oriented to the left relative to the body, the right arm appeared to move more, as did the left leg [34, 35]. These results suggest that the arm and leg movements can be used to predict head orientation, which in turn predicts gaze location.

5.3 Attention-Oriented Pedestrian Group Segmentation

In the course of daily life, we are often with others in social groups such as family and coworkers. When we are in such groups, our actions are strongly affected by and have a strong effect on other members. Therefore, investigating social relationships is an important way to understand and/or infer our daily activities. In this section, we propose a method for determining social groups by analyzing time-series range data.

Many studies have been conducted on segmenting individuals into social groups. Most of them use human trajectories for segmentation. Typically, some features are computed from the trajectories and predetermined criteria or machine learning techniques are applied to the feature values for segmenting groups. Needless to say, additional features should be used if available.

Our idea is to extract a change of attention as a feature for group segmentation. As an example of this idea, consider the situation of walking with friends. We often talk with each other and make gestures for communication. During such interactions, our attention is directed, through speech and/or body language, toward a target, that is, the people with whom we are interacting. In contrast, two people who do not have a social relationship usually do not interact in this way while walking, even if they are in close proximity. This observation implies that we can discern whether two people are interacting with each other by observing their attentional shifts as signaled with speech and/or body language.

Our proposed method uses three types of information to determine whether two people are in the same group. The first is the motion trajectory, which has been commonly used in previous work. We also use chest orientation as a cue for human attention. To obtain these data, we use range sensors placed at chest height. Finally, we use video recording to detect gestures indicating that an interaction is occurring between people. We believe that the combination of this information enables us to detect social groups more accurately.

From the extracted features of motion trajectory, chest orientation, and video, we build a classifier to determine social groups. The classifier is based on multiple instance learning (MIL). When we are walking with a friend or colleague, we do not interact with them all the time. This means that meaningful information for group detection is embedded within only certain parts of the time-series data. To segment social groups accurately, we must detect the meaningful information and ignore the rest. MIL is used for efficiently detecting and focusing on the meaningful features.

To examine the proposed method, we conducted experiments using a practical data set, which was collected on a university campus. The experimental results show that the proposed method outperforms the existing method.

In the following sections, we first introduce related work in Sect. 5.3.1. We outline the proposed method in Sect. 5.3.2. Finally, we present experimental results in Sect. 5.3.3.

5.3.1 Related Work

As previously discussed, the most straightforward and essential information for group detection is a shared motion trajectory [36, 37]. However, additional information has been investigated for more effective group detection. Human attention is one of the most promising clues that reflects interaction among group members [38–40]. Our work is inspired by Chamveha [39], who also used attentional cues; however, we also use range sensors as input devices, videos for detecting interactions, and MIL to add efficiency in group detection.

Laser range sensors are widely used for obtaining reliable data, even outdoors. Scenarios for applying this work include pedestrian tracking [41]. It is relevant to note here that, to the best of our knowledge, existing studies have employed range sensors placed at leg height. This is because these studies have only considered the "footprint" of pedestrians. As mentioned earlier, consideration of human attention will enhance our understanding of human behavior. Our work aims to extend the possibilities of range sensor-based human analysis.

5.3.2 Proposed Method

The main features of our proposed method are twofold: (1) using the MIL framework for accurate and efficient group discovery, and (2) using video processing to detect gestural actions that indicate interactions. We describe each feature in the following section.

5.3.2.1 MIL Framework for Group Segmentation

In the latter part of the process, we pick two participants and classify whether or not they are in the same group. This is a common approach used in various studies [39].

Even when we are walking with another individual or a group, we do not interact with the other group member(s) all of the time. For example, each member of a group is sometimes looking at different objects, and at that moment, no interaction is observed among them. Many conventional approaches such as that used by Chamveha [39] employ a histogram of feature values. Histogram-based feature

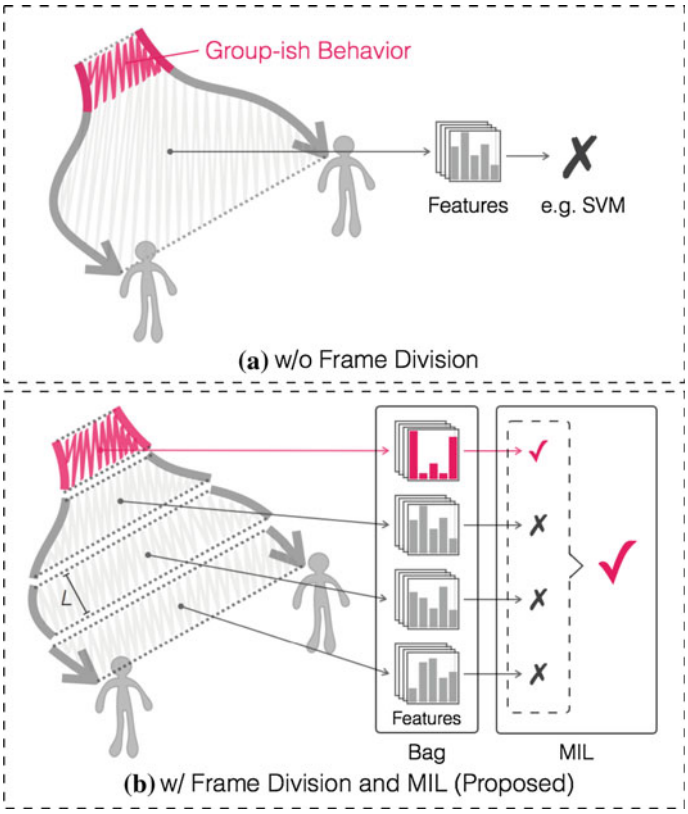


Fig. 5.16 Temporal Frame Division **a** conventional approach (w/o temporal division), **b** proposed approach

representation is commonly used because of its robustness. However, in our application scenarios, the irrelevant behavior contained in histograms conceals the relevant features.

MIL [42] can treat this type of ambiguity efficiently. A set of training data (instances) is treated as a bag, and each bag has a label. If a bag contains at least one positive instance, the bag becomes positive. If all of the contents of a bag are negative, the bag becomes negative.

In our case, as shown in Fig. 5.16, we divide time-series data $D = \{D_t\}$ into several subsets of data with shorter time lengths $D^{(k)}$, $D = \bigcup D^{(k)}$. Each subset is an instance in a bag for MIL. The bag consists of instances extracted from a single pair of walking participants. As mentioned earlier, even when two people are in the same group, not all of the instances will be clues to their social connection. Hence, all instances cannot be treated as positive examples, but at least one instance is positive.

We should note that the benefit of the MIL framework is in the training process. It is quite difficult to manually detect relevant actions for group discovery from

long time-series data that include irrelevant behaviors. MIL automatically finds the relevant instance, i.e., actions, from a set of instances when the set is a positive bag. This significantly reduces the cost of annotation.

We use the same feature values for classification as used by Chamveha [39] because they have yielded reasonable results. Briefly stated, two types of features are used: attention-based features and position-based features. Please see the original paper for details.

5.3.2.2 Gesture Detection for Group Segmentation

Another important feature of the proposed method is the use of video images to detect gestures indicating interactions between a pair of participants. As argued previously, interactions between two people are a clue for determining social groups. Even when the two people are not walking closely together, we can still observe whether the two have a social relationship by observing their interactions.

In this paper, we use a simple processing method for detecting gestures such as pointing and hand-waving. Human action recognition is an active research area in the computer vision community and remains a challenging topic, especially in actual application scenarios. Our goal is not to classify a video into multiple action classes. If we can simply detect the occurrence of gestures, even without understanding what the gestures are, this can be a sufficient cue to understanding how a certain kind of interaction occurs between the two people. With this rationale, we use a simple method for gesture detection.

Figure 5.17 depicts a feature for gesture detection. We compute an optical flow for an observed video. When a person makes a gesture, its motion is larger than that of a normal walking action. Both Fig. 5.17a, b correspond to the optical flow for gestures. We can see large flows for both cases. To measure the likelihood that this

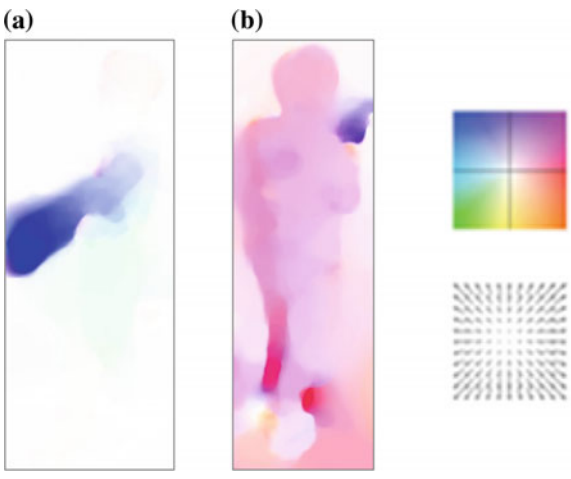


Fig. 5.17 Gesture Detection based on Optical Flow. Hue and saturation correspond to the orientation and strength of flows

flow contains a gesture, we compute the feature $G = V_{max}/V_{min}$, where V_{max} and V_{min} denote the maximum and minimum of the strength of the optical flow. When G is large, the person is more likely to be making a gesture.

After computing the feature G, we simply concatenate it with information from other features introduced in the previous section. Each obtained feature is classified using the MIL framework.

5.3.3 Experiments

To confirm the basic effectiveness of the proposed method, we have carried out experiments using real data captured in a university building and a shopping mall.

5.3.3.1 Data Acquisition in a University Building

As mentioned in the previous section, we used a set of laser range sensors placed at chest height. A 2D range map captured by the range sensors can be integrated into a single map in a unified coordinate system. In the integrated range map, we subtract a background map from the obtained map and apply a conventional segmentation method to the subtracted map. Each segment can be assumed to be a person. Next, we fit an ellipse to the segmented map. The two axes of the ellipse are chosen so that they correspond to the body size. The ellipse has some "thickness" to allow for variations in body size. Finally, we connect the positions of the fitted ellipses to those detected in the previous frame, and obtain the time-series positions of walking people. We can assume that the shorter axis of the fitted ellipse indicates the chest direction. Because we found that the axis direction was not stable in a single frame, we applied temporal filtering to the direction for stabilization.

During data collection, various groups appear in the scene. One or more groups are instructed to move from a start point to a goal point. In some cases, each person in a group has a different start or goal point. Various kinds of instructions enable us to obtain a variety of group actions; however, instructions are only relevant to the start and goal positions. While walking between these two points, the people act naturally, without any instructions.

5.3.3.2 ATC Dataset

To evaluate the proposed method in more realistic scenarios, we used a public data set collected in a shopping mall. The dataset [43–45] hereafter ATC dataset, includes the positions and chest directions of walking pedestrians in a shopping mall. These data are collected using range sensors that are mounted so that walking people in a wide area can be monitored. The ATC dataset also includes manually annotated information about the group membership of the people. The dataset cannot be used

for evaluating group detection using gesture because the data do not include visual information taken by cameras. However, the data can be used for the MIL-based group detection, which only requires trajectories and chest directions. Note that this dataset was collected by the CREST project led by Dr. Takayuki Kanda. Further details of the ATC dataset are described in Drazen [43].

5.3.3.3 Experiment 1—MIL Framework

First, we conducted an evaluation of the effect of the MIL framework described in Sect. 5.3.2.1. Table 5.1 summarizes the quantitative evaluation of the group segmentation results when the whole data set is divided into subsets of time length $L = 210$, where "w/o MIL" means that we did not divide the whole time-series data set. This can be regarded as the result of [39]. The MIL framework yields better results in terms of precision, recall, and F-measure.

To see the effect of the parameter L, we changed the parameter and evaluated performance. The graph in Fig. 5.18 shows the performance changes. The horizontal axis is the time length L. The dotted line shows the results when the whole data set is not divided. According to the graph, the choice of L has a strong effect on performance, so it should be determined for each individual scene.

Figure 5.19 shows typical examples of correctly and incorrectly classified samples. (a), (b), and (c) are successfully detected group pairs. (d) and (e) are false-positive samples, because the two participants are walking or standing close together but are

Table 5.1 Quantitative classification results for the MIL framework

	Proposed ($L = 210$) (%)	w/o MIL [39] (%)
Precision	**87.2**	85.7
Recall	**95.6**	90.3
F-measure	**91.2**	87.9

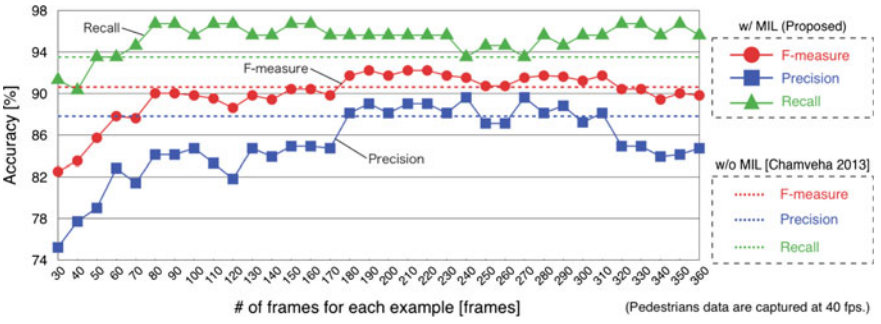


Fig. 5.18 Group Classification Results (Relationship between frame length and classification accuracy)

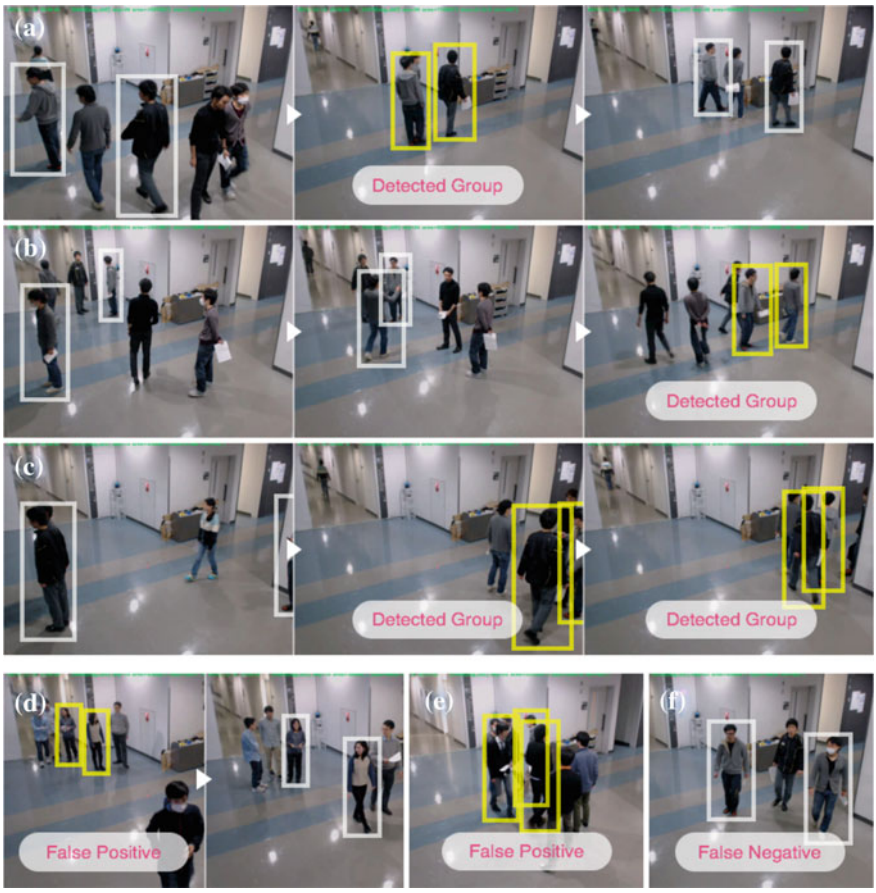


Fig. 5.19 Typical Classification Results Using Frame Division. **a**, **b**, and **c** are successfully detected group pairs even when they are not walking together all the time; **d** is an incorrectly detected pair that is coincidentally walking closely; **e** also shows incorrectly detected pairs who are not in a group but are standing closely; and **f** is an example of misdetection

not in the same group. (f) is a false-negative sample because the two participants are in the same group but have no interaction.

Next, the MIL framework was tested using the ATC dataset. We choose 1090 pedestrians, including 68 group pairs. The data were collected between 10 am and 11 am on January 9th, 2013. Table 5.2 summarizes the quantitative evaluation as in Table 5.1. These results also demonstrate that the MIL framework yields better results even under more realistic scenarios.

Figure 5.20 shows the relationship between time length L and accuracy, as in Fig. 5.18. Although performance varies as the time length changes, it is almost always better than the results without the MIL framework. Figure 5.21 shows the trajectories of a group pair. The pair was not classified as "in the same group" without the MIL

Table 5.2 Quantitative classification results for the MIL framework (ATC dataset)

	Proposed ($L = 240$) (%)	w/o MIL [39] (%)
Precision	**92.6**	87.1
Recall	**92.6**	89.7
F-measure	**92.6**	88.4

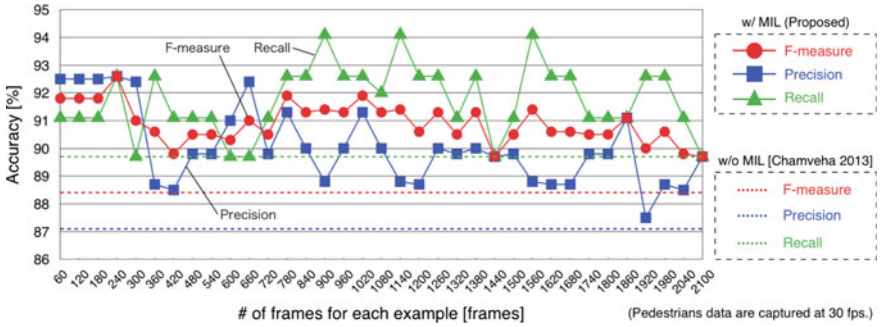


Fig. 5.20 Group classification results of ATC dataset (Relationship between frame length and classification accuracy)

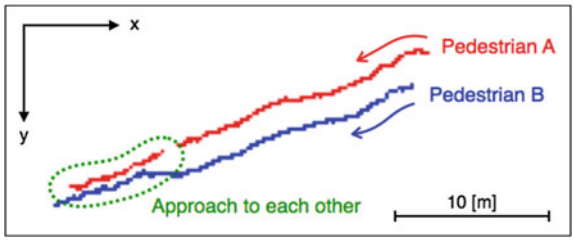


Fig. 5.21 Typical Classification results using frame division (ATC dataset: these two persons are correctly classified as group members)

framework because the pair did not walk closely together. By applying the MIL framework, the pair was classified correctly as a group. This is a typical benefit of the proposed framework.

5.3.3.4 Experiment 2—Group Discovery Using Gesture Detection

When we take into account gesture detection described in Sect. 5.3.2.2, we add a scale factor C to the gesture likelihood G. We evaluated the performance at various values of C. Table 5.3 shows the results. Compared with no gesture detection, the parameter $C = 0.09$ produces the best result. Although we have to choose the parameter carefully, incorporating gesture detection has a positive effect on overall performance.

Table 5.3 Group classification results using gesture detection (Frame Length $L = 100$)

	w/o gesture features	w/ gesture features			
		$C = 0.07$	$C = 0.08$	$C = 0.09$	$C = 0.1$
True-Positive	96.7	94.6	94.6	95.6	92.4
False-Positive	6.5	5.4	4.4	4.4	4.4
F-measure	0.952	0.946	0.951	0.957	0.940

Fig. 5.22 Successful detection of group pair using gesture detection

Fig. 5.23 Successfully suppressed false detection using gesture detection

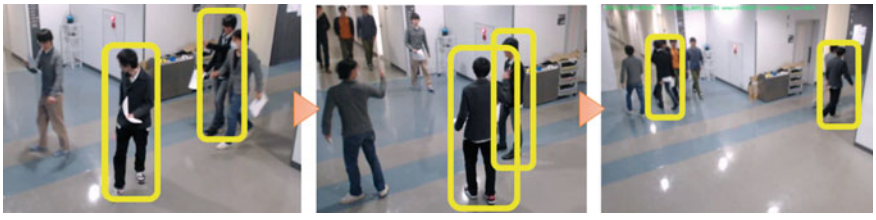


Fig. 5.24 Misdetected group pair using gesture detection

Figures 5.22, 5.23, and 5.24 show typical examples. Figure 5.22 is an example of a case in which gesture detection works well. This pair can only be classified successfully using gesture detection. Figure 5.23 is also an example of a successful result. In this case, the original method incorrectly detects a non-group pair as a group. This is because the two participants walk very closely together. By incorporating gesture detection, we can set an appropriate decision boundary for distance-related

features. Finally, Fig. 5.24 is an example of the negative effects of the proposed method. Originally, the two participants are not detected as a group pair. However, one of the two participants incidentally makes a gesture that does not correspond to an interaction with the other participant. Gesture detection incorrectly detects the gesture, and the two are classified as a group pair.

5.4 Dual-Task Analysis for Cognitive Level Estimation

As mentioned, more than 4.4 million people in Japan present with dementia, while 37 thousand present with early-onset dementia due to cerebral vascular disease. Thus, dementia affects a large percentage of the population, and can threaten any individual as they age. Moreover, as the symptoms of dementia worsen, they can cause secondary damage, such as difficulties with interpersonal communication, co-morbid mental disorders such as depression, and loss of confidence and motivation. It is thus important to detect dementia in its early stage and treat individuals accordingly. However, dementia can be difficult to diagnose because there are often no clear symptoms in the early stages. In other words, by the time that symptoms are visible, dementia has generally progressed to a certain degree. It is therefore important to assess recognition function in our everyday lives.

Although cognitive ability can be assessed using interviews or questionnaires, these are often quite time-consuming and can thus be challenging methods for examining large numbers of people. Thus, assessments that involve simply observing activities in daily life are preferable. Among these, the "dual-task" method (Fig. 5.25) has received attention because it has been useful in identifying the early stages of dementia. Indeed, it has already been introduced in several elderly facilities and hospitals. However, that the assessment is subjective, i.e. it is performed by a doctor or a clinical psychologist, is problematic.

Considering the above situation, our goal was to develop a system that can be used to objectively assess cognitive ability by observing dual-task performance via sensors, such as cameras, range sensors, and microphones. We began by collecting observations from a diverse group of people, ranging from young healthy individuals to older adults with dementia. We extracted features that could be used to assess cognitive abilities. This section describes the measurement systems that we developed for this purpose and introduces some recent results obtained using our systems.

Fig. 5.25 Dual-task method for assessment of cognitive ability

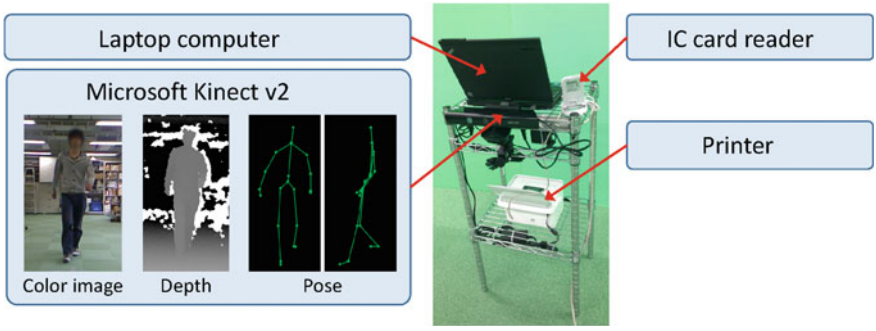


Fig. 5.26 Mobile measurement system

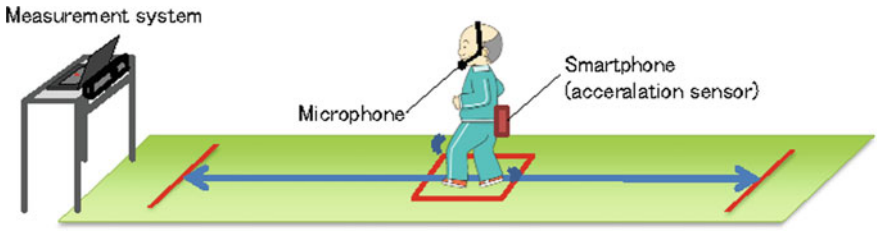


Fig. 5.27 Environmental setting for dual-task observation

5.4.1 Data Collection in an Elderly Care Facility

5.4.1.1 System Overview

We sought to collect data from elderly people with dementia. To this end, we designed a mobile measurement system that could be easily transported to elderly care facilities. Figure 5.26 shows the mobile system we constructed, which comprised a Microsoft Kinect and a laptop computer. An IC card reader was connected to the system for management of participant data.

As shown in Fig. 5.27, for the data collection, we prepared a pathway and located the measurement systems at its end points. Each participant was asked to perform stepping and walking back and forth along the pathway, and to perform the stepping/walking while engaged in cognitive tasks (dual-task). There were two kinds of cognitive tasks: a "count-down" task (reciting numbers from one hundred in descending order in a certain amount of time) and a "word-frequency" task (saying words which begin with a certain letter), which are standard tasks in clinical settings. For data analysis, we needed ground-truth score information about participant cognitive abilities in addition to our observations. We adopted the Mini-Mental State Examination (MMSE), which is a standard metric often used in elderly facilities. It ranges from 0 to 30; 30 means that a person has normal cognitive ability, and less than 23 signifies possible dementia.

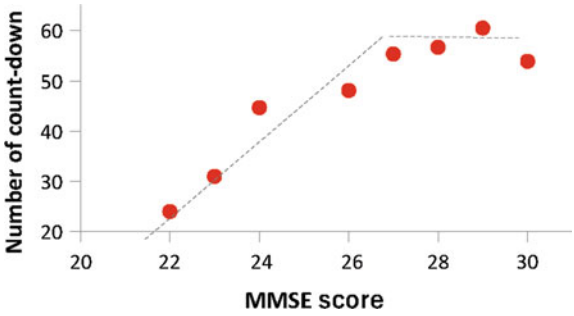


Fig. 5.28 Relationship between MMSE score and the number of "count-down"

5.4.1.2 Results

We thoroughly investigated the relationships between the MMSE score and any features extracted from the observation to search for features that would predict the MMSE score. Among such features, we used the average number of "count-down" numbers given in Fig. 5.28 The figure shows a curve, indicated by a dotted line. For MMSE scores ranging from 20 to 27, the number and the score are well correlated, and this converges with a "count-down" score of about 60, where the MMSE score is more than 27. Thus, the better the cognitive ability of a participant, the more numbers he/she can tell in a certain period. However the number cannot be larger than about 60.

5.4.2 Huge Data Collection in Miraikan

Throughout the process of data collection in elderly care facilities, as described in Sect. 5.4.1, we found several features that predict MMSE score. Thus, it should be possible to assess cognitive ability from the dual-task observation. On the other hand, we also found that we needed an increasingly large number of participants to ensure the reliability of the results. As we performed the data collection using identical procedures, doubling or tripling the number of participants doubled or tripled the length of the data collection period. Thus, the experiment was very time-consuming and a large effort was required to obtain a large dataset.

To address this problem, we adopted a new experimental method: We designed a fully automatic demonstration system (Figs. 5.29 and 5.30) that could administer the dual-task procedure and collect data and positioned the system in the National Museum of Emerging Science and Innovation, known as "Miraikan". Miraikan is located in the Tokyo bay area and receives several thousands of visitors per day. The demonstration period was approximately one year; from 15 July 2015 to 27 June 2016. With the help of Miraikan, the demonstration was optimized so that participants could enjoy it as if it were a game. Figure 5.31 shows the statistics of

Fig. 5.29 Demonstration system in Miraikan

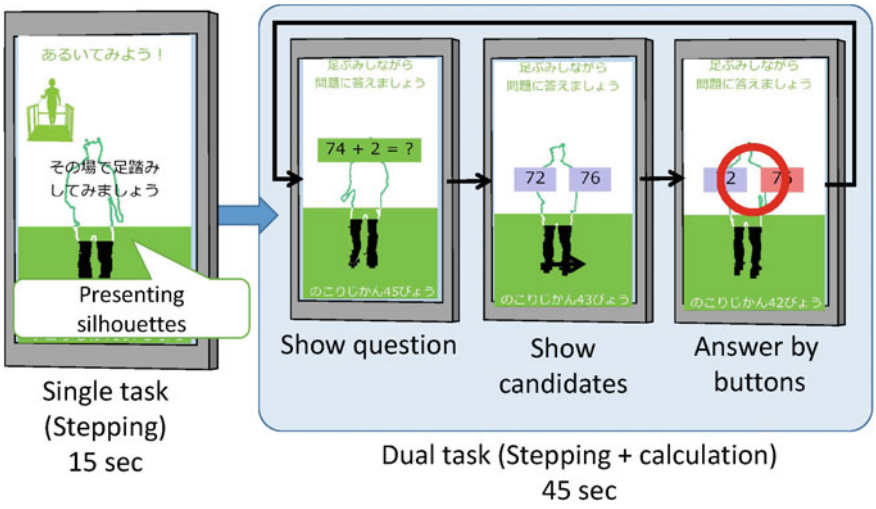
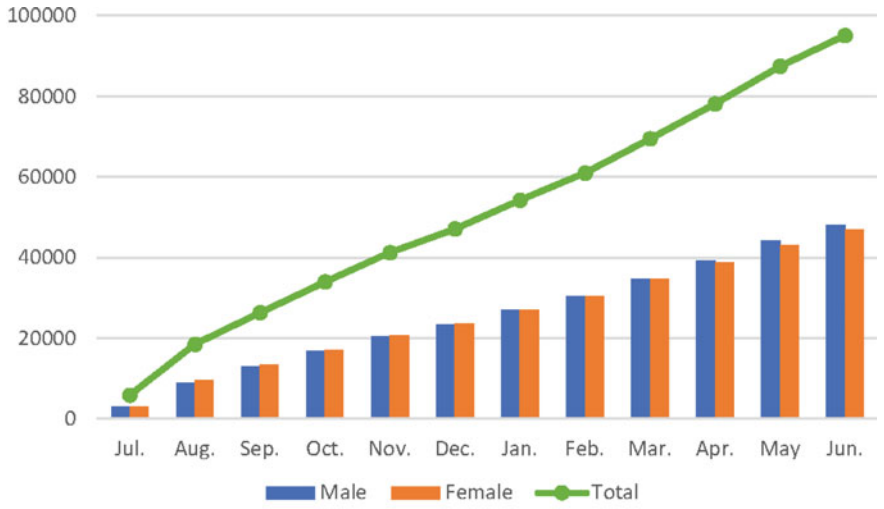


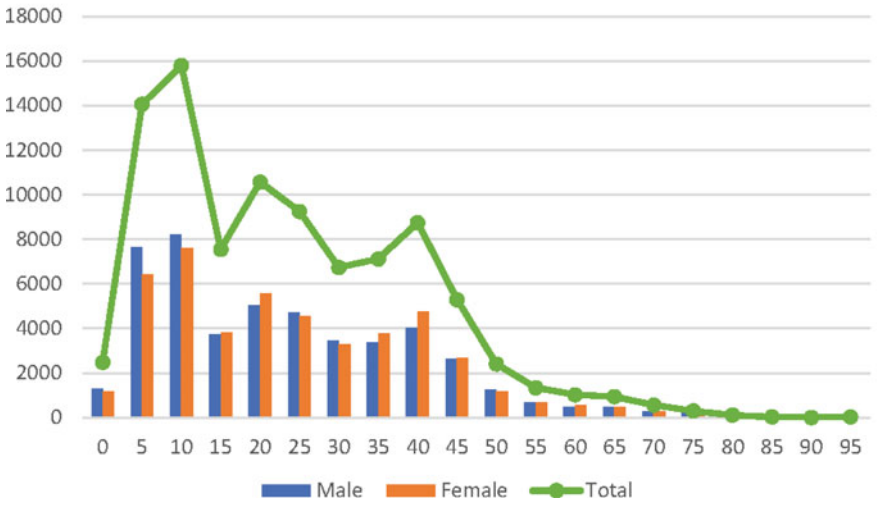
Fig. 5.30 Flow of demonstration

the participants. As a result, we collected data from more than 95,000 participants, making it one of the largest datasets in the world.[1]

[1] We also placed another system in Miraikan. It captures videos of a person walking from multiple viewpoints. The video dataset obtained by the system is not for assessment of cognitive ability but for gait authentication purpose, which is an most important research topic for us. This dataset is also among the largest in the world.



(a) Accumulation



(b) Age distribution

Fig. 5.31 Statistics of demonstration in Miraikan

5.5 Discussion and Future Work

The achievements in Sects. 5.2 and 5.3 have great potential to be applied in our daily lives, especially in commercial facilities. By applying the models and knowledge in Sect. 5.2 and extending them to more natural cases, it should be possible to estimate

attention of each customer from his/her behaviors captured by security cameras, which are located anywhere in such facilities nowadays. When it is realized, it might be possible to detect shoplifters from his/her suspicious gaze behavior. It should be also possible to find customers who are interested in a certain product and effectively give them salient information about promotions and sales. Indeed, there have already been several works that analyzed the relation between person's intention or interest and his/her gaze behavior motivated by the similar consideration to ours. They are, however, still quite far from real applications, because in all those studies they need eye-trackers, which is in fact an unsatisfied condition; customers and shoplifters will never wear wearable eye-trackers, and a stationary eye-tracker usually has very limited view area so that it is impossible to measure gaze of a person walking freely around the space. We can conclude, therefore, our techniques for gaze estimation without eye-tracker have a really great impact for reaching the expected future. In addition, the achievements in Sect. 5.3 also take on an important role in the commercial facility scenario. In the course of daily activities, humans often form social groups, such as families and groups of coworkers. When we are in such groups, our actions are strongly affected by and have a large effect on other group members. If we can know how many groups there are in the space, which group each person belongs to, what kind of group each of them is, which person in a group is taking initiative, for example, it is helpful for a facility manager to analyze purchasing behaviors of customers and offer commercial advertisement effectively.

The systems for cognitive level assessment and experiences in the elderly facilities are also important achievements in this project. The design/process/stability of the systems have been evaluated from several viewpoints; system developers (the members in the project including the authors) who chose sensors and implemented software, staffs in the elderly facilities who observe the elderly/dementia people every day, and the elderly people who are in fact the participants of the systems. Especially the long-term exhibition in Miraikan, whose details are mentioned in Sect. 5.4.2, was a meaningful opportunity for establishing a good system. Through the one-year exhibition, the system was well brushed up and became stable, safe, and interesting for the participants. Fortunately, the improved system was decided to be installed in the elderly facilities for long-term (more than three years) data collection. By the data collection, we will obtain another kind of data; long-term history of the dual-task performance of a certain person combined with cognitive level assessed by clinical psychotherapists. It is really interesting data because it would be possible to model degradation of cognitive level, which is meaningful to realize "prediction" of dementia.

5.6 Conclusion

Gait has been regarded as one of biometrics and thus mainly studies for authentication task. In that scenario it is implicitly assumed that the gait of a certain person is always constant. It is, however, untrue in reality; a person usually walks differently

according to their mood and physical/mental conditions. In our project "Behavior Understanding based on Intention-Gait Model" supported by JST-CREST, therefore, we focused on the variation of gait within a person, and have studied the relation between the gait variation to the inertial states. This chapter introduced some of our achievements related to three kinds of the inertial states; attention (gaze direction), human relation (group segmentation), and cognitive level (assessment of dementia). Though those about attention and human relation estimation are on still developing stage, they indicate possibility of realizing skill of "reading minds." Those about cognitive level assessment are also important contribution for the coming "super-aging society." We would be happy if our achievements would contribute to make our daily lives safer, more convenient, and more agreeable.

Acknowledgements We appreciate Honorary Professor Masatsugu Kidode in Nara Institute of Science and Technology for his great contribution as an editorial supervisor of this chapter.
Section 5.2 describes achievements by Mr. Yu Fang, Dr. Ryoichi Nakashima, Dr. Yasuhiro Hatori, and Associate Professor Kazumichi Matsumiya in Tohoku University. Section 5.3 is about the collaborative research with Professor Kazuhiko Sumi in Aoyama Gakuin University and Mr. Ryota Sato in Kindai University. The dual-task system in Sect. 5.4 was constructed in collaboration with Dr. Mitsuru Nakazawa, Dr. Masataka Niwa, Dr. Kota Aoki and Assistant Professor Fumio Okura in Osaka University, and Lecturer Hirotake Yamazoe in Ritsumeikan University. We appreciate these members for their efforts in the project.
This project was also supported by some institutes. Dr. Takayuki Kanda in ATR, who is a leader of another CREST project, and his colleagues Dr. Drazen Brscic and Dr. Satoshi Satake had discussion with our projects members to share ideas several times during our project period. Moreover, They kindly shared their datasets captured in a commercial facility. The ideas and datasets helped us to get better achievements. Mr. Masuhiro Okuda, who is the president of Social Welfare Corporation "Misasagikai," kindly believed the future we designed about elderly care systems, and gave us many opportunity to capture data of elderly people in his facilities, and accepted us to locate our dual-task system in each building. Staffs in his facilities were also kind and contributed professionally for our project. This project would never achieve the success without his understanding and help. We sincerely appreciate him for the contribution.
We also acknowledge the discussion to initiate the study in the meeting of the Cooperative Research Project of the Research Institute of Electrical Communication in Tohoku University and the Institute of Scientific and Industrial Research in Osaka University.

References

1. N. Lynnerup, J. Vedel, Person identification by gait analysis and photogrammetry. J. Forensic Sci. **50**(1), 112–118 (2005)
2. J. Han, B. Bhanu, Individual recognition using gait energy image. IEEE Trans. Pattern Anal. Mach. Intell. **28**, 316–322 (2006)
3. Y. Makihara, R. Sagawa, Y. Mukaigawa, T. Echigo, Y. Yagi, Gait recognition using a view transformation model in the frequency domain, in *Proceedings of the 9th European Conference on Computer Vision*, pp. 151–163, 2006
4. Y. Makihara, R. Sagawa, Y. Mukaigawa, T. Echigo, Y. Yagi, Adaptation to walking direction changes for gait identification. IEEE Int. Conf. Pattern Recognit. **2**, 96–99 (2006)
5. P.K. Larsen, E.B. Simonsen, N. Lynnerup, Gait analysis in forensic medicine. J. Forensic Sci. **53**(5), 1149–1153 (2008)

6. T.H.W. Lam, K.H. Cheung, J.N.K. Liu, Gait flow image: A silhouette-based gait representation for human identification. Pattern Recognit. **44**, 973–987 (2011)
7. I. Bouchrika, M. Goffredo, J. Carter, M. Nixon, On using gait in forensic biometrics. J. Forensic Sci. **56**(4), 882–889 (2011)
8. S. Zheng, J. Zhang, K. Huang, R. He, T. Tan, Robust View transformation model for gait recognition, in *Proceedings of the IEEE International Conference on Image Processing*, 2011
9. S. Sarkar, P. Jonathon Phillips, Z. Liu, I. Robledo, P. Grother, K.W. Bowyer, The human ID gait challenge problem: data sets, performance, and analysis. IEEE Trans. Pattern Anal. Mach. Intell. **27**(2), 162–177 (2005)
10. H. Iwama, M. Okumura, Y. Makihara, Y. Yagi, The OU-ISIR gait database comprising the large population dataset and performance evaluation of gait recognition. IEEE Trans. Inf. Forensics Secur. **7**(5), 1511–1521 (2012)
11. H. Yamazoe, H. Habe, I. Mitsugami, Y. Yagi, Easy depth sensor calibration. Int. Conf. Pattern Recognit. (2012)
12. M. Nakazawa, I. Mitsugami, H. Habe, H. Yamazoe, Y. Yagi, Calibration of multiple kinects with little overlap regions. IEEJ Trans. Electr. Electron. Eng. **10**(S1) (2015)
13. H. Nakajima, Y. Makihara, H. Hsu, I. Mitsugami, M. Nakazawa, H. Yamazoe, H. Habe, Y. Yagi, Point cloud transport. Inte. Conf. Pattern Recognit. (2012)
14. M. Nakazawa, I. Mitsugami, Y. Makihara, H. Nakajima, H. Yamazoe, H. Habe, Y. Yagi, Dynamic scene reconstruction using asynchronous multiple kinects. Int. Conf. Pattern Recognit. (2012)
15. C. Zhou, I. Mitsugami, Y. Yagi, Detection of elderly gait impairment by Patch-GEI. IEEJ Trans. Electr. Electron. Eng. **10**(S1) (2015)
16. H. Yamazoe, T. Ogawa, I. Mitsugami, Y. Yagi, Gait analysis of simulated left knee disorder, in *9th International Conference on Bio-inspired Information and Communications Technologies* (2015)
17. H. Nakajima, I. Mitsugami, Y. Yagi, Depth-based gait feature representation. IPSJ Trans. Comput. Vis. Appl. **5**, 94–98 (2013)
18. T. Ikeda, I. Mitsugami, Y. Yagi, Depth-based gait authentication for practical sensor settings. IPSJ Trans. Comput. Vis. Appl. **7**, 94–98 (2015)
19. L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. **20**, 1254–1259 (1998)
20. L. Itti, C. Koch, Computational modelling of visual attention. Nat. Rev. Neurosci. **2**, 194–203 (2001)
21. H. Kubota, Y. Sugano, T. Okabe, Y. Sato, A. Sugimoto, K. Hiraki, Incorporating visual field characteristics into a saliency map, in *Symposium on Eye Tracking Research and Applications*, pp. 333–336, 2012
22. A. Hiratani, R. Nakashima, K. Matsumiya, K. Kuriki, S. Shioiri, Considerations of self-motion in motion saliency. International Joint Workshop on Advanced Sensing/Visual Attention and Interaction. presented at the International Joint Workshop on Advanced Sensing/Visual Attention and Interaction-Toward Creation of Human-Harmonized Information Technology-, Okinawa, Japan
23. J. Henderson, J.R. Brockmole, M.S. Castelhano, M. Mack, Visual saliency does not account for eye movements during visual search in real-world scenes, in *Eye movements: a window on mind and brain*, ed. by R. van Gompel, M. Fischer, W. Murray, R. Hill (Elsevier, 2007), pp. 537–562
24. A. Torralba, A. Oliva, M.S. Castelhano, J.M. Henderson, Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. Psychol. Rev. **113**, 766–86 (2006)
25. A. Kimura, R. Yonetani, T. Hirayama, Computational models of human visual attention and their implementations: a survey. IEICE Trans. Inf. Syst. **96-D**, 562–578 (2013)
26. R. Nakashima, Y. Fang, Y. Hatori, A. Hiratani, K. Matsumiya, I. Kuriki et al., Saliency-based gaze prediction based on head direction. Vis. Res. **117**, 59–66 (2015)

27. J.S. Stahl, Amplitude of human head movements associated with horizontal saccades. Exp. Brain Res. **126**, 41–54 (1999)
28. A.L. Cecala, E.G. Freedman, Amplitude changes in response to target displacements during human eye-head movements. Vis. Res. **48**, 149–66 (2008)
29. Y. Fang, R. Nakashima, K. Matsumiya, I. Kuriki, S. Shioiri, Eye-head coordination for visual cognitive processing. PLoS One **10**, e0121035 (2015)
30. Y. Fang, M. Emoto, R. Nakashima, K. Matsumiya, I. Kuriki, S. Shioiri, Eye-position distribution depending on head orientation when observing movies on ultrahigh-definition television. ITE Trans. Media Technol. Appl. **3**, 149–154 (2015)
31. R. Nakashima, S. Shioiri, Facilitation of visual perception in head direction: visual attention modulation based on head direction. PLoS One **10**, e0124367 (2015)
32. R. Nakashima, S. Shioiri, Why do we move our head to look at an object in our peripheral region? Lateral viewing interferes with attentive search. PLoS One **9**, e92284 (2014)
33. C.H. Tseng, Z. Vidnyanszky, T. Papathomas, G. Sperling, Attention-based long-lasting sensitization and suppression of colors. Vis. Res. **50**, 23–416 (2010)
34. T. Okada, H. Yamazoe, I. Mitsugami, Y. Yagi, Preliminary analysis of gait changes that correspond to gaze directions, in *International Joint Workshop on Advanced Sensing/Visual Attention and Interaction*, pp. 788–792, 2013
35. I. Mitsugami, Y. Nagase, Y. Yagi, Primary analysis of human's gait and gaze direction using motion sensors, in *Asian Conference on Pattern Recognition*, 2011
36. M. Manfredi, R. Vezzani, S. Calderara, R. Cucchiara, Detection of static groups and crowds gathered in open spaces by texture classification. Pattern Recognit. Lett. **44**, 39–48 (2014)
37. M. Zanotto, L. Bazzani, M. Cristani, V. Murino, Online bayesian non-parametrics for social group detection, in *Proceedings of the British Machine Vision Conference* (BMVA Press, 2012), pp. 111.1–111.12
38. S. Calderara, R. Cucchiara, A. Prati, Group detection at camera handoff for collecting people appearance in multi-camera systems, in *Proceedings—IEEE International Conference on Video and Signal Based Surveillance 2006, AVSS 2006*, 2006
39. I. Chamveha, Y. Sugano, Y. Sato, A. Sugimoto, Social group discovery from surveillance videos: a data-driven approach with attention-based cues, in *BMVC 2013*, 2013
40. F. Setti, H. Hung, M. Cristani, Group detection in still images by F-formation modeling: a comparative study, in *2013 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)* (IEEE, 2013), pp. 1–4
41. H. Zhao, R. Shibasaki, A novel system for tracking pedestrians using multiple single-row laser-range scanners. IEEE Trans. Syst. Man Cybern. Part A: Syst. Hum. **35**(2), 283–291 (2005)
42. G. Doran, S. Ray, A theoretical and empirical analysis of support vector machine methods for multiple-instance classification. Mach. Learn. **97**(1–2), 1–24 (2013)
43. D. Brscic, T. Kanda, T. Ikeda, T. Miyashita, Person tracking in large public spaces using 3D range sensors. IEEE Trans. Hum.-Mach. Syst. (2013)
44. F. Zanlungo, D. Brscic, T. Kanda, Spatial-size scaling of pedestrian groups under growing density conditions. Phys. Rev. E **91**(6), 062810 (2015)
45. Pedestrian Group Dataset: http://www.irc.atr.jp/sets/groups/