Memory-based Recognition of Human Behavior based on Sensory Data of High Dimensionality

Karl MacDorman,[†] Hiroshi Nobuta,[‡] Takashi Minato, and Hiroshi Ishiguro
Department of Adaptive Machine Systems
Graduate School of Engineering, Osaka University
Suita, Osaka 565-0871 Japan
kfm@ams.eng.osaka-u.ac.jp

***Abstract*— This paper explores memory-based approaches to the recognition of human behavior that rely on a database of previously categorized instances of sensory data. To overcome the curse of dimensionality, we examine two related methods that both rely on a hierarchical division of the sensory space using a decision tree. The first approach iteratively applies linear discriminant analysis to divide the sensory space in half in order to construct a binary tree for recognizing behaviors. We have verified the effectiveness of this approach for real-time behavior recognition using infrared sensors distributed in a desk environment and compared its results to those of Quinlan's C4.5. The second approach applies the well-known ID3 algorithm to the construction of a decision tree based on an information criterion. We use it to recognize browsing behavior at a video rental shop. Inferences are derived directly from the binarized pixel data of four wide-view cameras. Both systems offer behavior recognition rates in excess of 90%.**

I. Introduction

Robust recognition systems that detect complex behaviors are necessary to support human activities. They may be employed, for example, in intelligent rooms [1], [4] or in subways for the prevention of vandalism [3]. They will also enable robots to interact with people in natural settings [8].

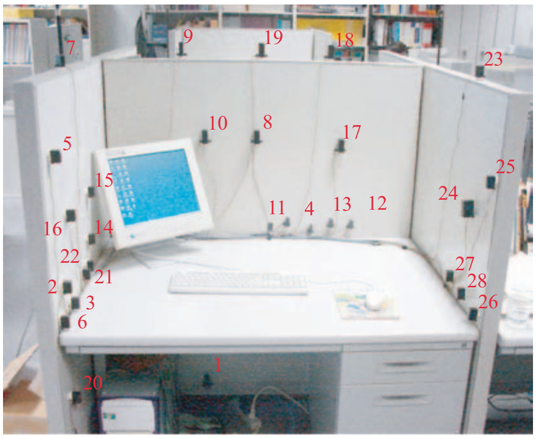

Fig. 1. A system for recognizing a person's work activities based on distributed infrared sensors.

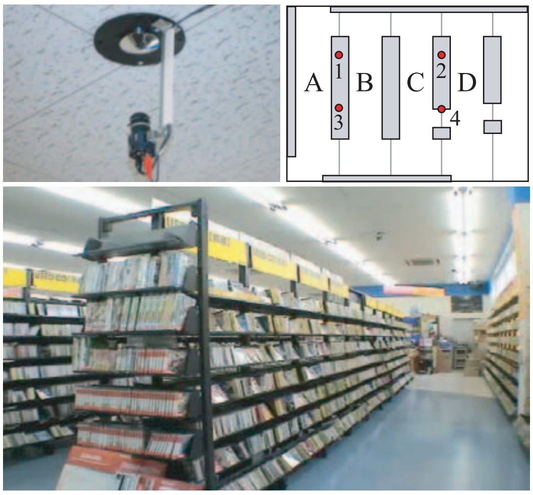

Fig. 2. A system for recognizing browsing activity at a video rental shop. Top left: the wide-view vision sensor composed of a fisheye camera and convex mirror. Top right: the arrangement of the four vision sensors. Bottom: the video rental shop.

Previous work has tended to exploit features of specific environments or situations (e.g., [2], [6], [7]). However, a more general ability to recognize behavior is needed to fully exploit recently developed multimedia technologies. As a step toward this goal, we developed a model-based distributed omnidirectional vision system (ODVS) that covers a wide area [8]. By placing many cameras in the environment, the system obtains redundant visual information. This redundancy makes the image processing of each camera simple and robust.

Unfortunately, our model-based approach has demerits [14], [9]. Generally speaking, it is hard to realize an interactive system based on the model-based approach. It takes time to acquire a new model. The modeling cost is especially serious for a sensory space of high dimensionality. Calibrating many cameras is also complicated. Furthermore, the system cannot adapt to variations in human behavior, which are potentially unlimited in number and

[†] Frontier Research Center, Osaka University.

[‡] Faculty of Systems Engineering, Wakayama University, 930 Sakaedani, Wakayama 640-8510 Japan.

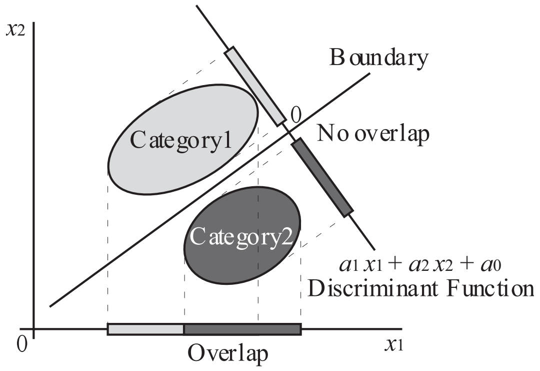

Fig. 3. The application of linear discriminant functions to discriminate analysis.

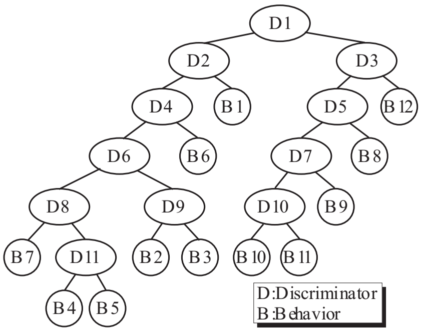

Fig. 4. The recognition tree resulting from using the discriminators.

sensitive to the situation. Therefore, an interactive system that allows us to add or change behavior models online is required.

A simpler solution is for the system to directly process the sensory data without our defining features *a priori* [12]. This memory-based approach uses the sensory data to construct a model for recognizing behavior. Although the memory-based approach is not liable to errors caused by a narrowly defined set of features, it typically requires more memory and computation. However, advances in computer memory and processing speed have made the approach feasible. In addition, it is not necessary for the system designer to anticipate which set of features would be adequate for the task [11].

This paper explores two related memory-based methods for recognizing behaviors in a sensory space of high dimensionality. Owing to the high memory and computational costs of memorizing the complicated boundaries between categories of human behavior, we iteratively divide the sensory space into two subspaces with linear discriminate functions. That is, the sensory space is represented by a hierarchical binary tree. Although Ishiguro et al. [10] have proposed a similar method, it requires a great deal of time to construct the tree; therefore, it is not suitable for an interactive system.

In our first experiment, outputs from infrared sensors distributed in the environment define the sensory space. Each infrared sensor detects that something is present in a certain region of space. We have applied the proposed method to recognizing human posture and behavior in a real environment and verified its effectiveness. The second experiment uses distributed cameras with wideview mirrors to record browsing behavior at a Tsutaya video rental shop. A decision tree is also contructed, but using the ID3 algorithm, which recursively divides the sensory data (viz., binarized images) according to the pixel that offers the highest information gain with respect to the conclusion.

II. Recognition by Hierarchical Linear Discriminant Analysis

A. Basic idea

Linear discriminant analysis determines the hyperplane that divides the parameter space for two given categories and determines to which subspace a new instance belongs (see Fig. 3). If more than two categories exist, linear discriminant analysis is applied to the subspaces divided by the first discriminant function. This process can be applied further, if classification errors remain. As a result, we construct a binary tree whose nodes perform linear discriminant analysis (see Fig. 4).

Hierarchical discriminant analysis is sensitive to the order in which the linear discriminant function are applied. We have to properly decide which linear discriminant function should be assigned to the root node in order to represent a proper abstraction hierarchy in the binary tree. We represent each human behavior as a combination of the discriminators that divide the behavior data into two subspaces by linear discriminant function and assign a weight

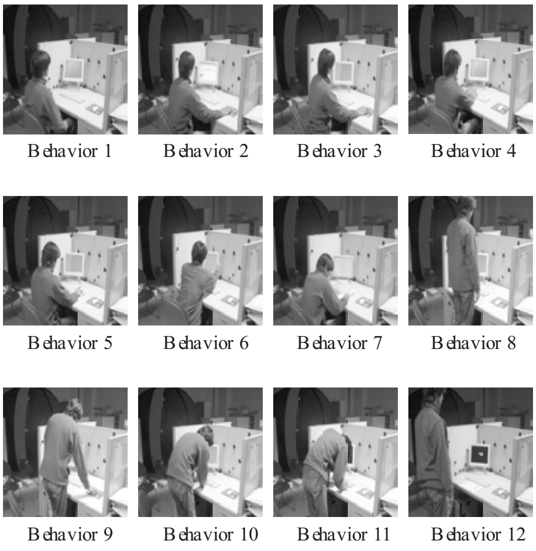
Behavior 1 Behavior 2 Behavior 3 Behavior 4
Behavior 5 Behavior 6 Behavior 7 Behavior 8
Behavior 9 Behavior 10 Behavior 11 Behavior 12

Fig. 5. Behavior categories

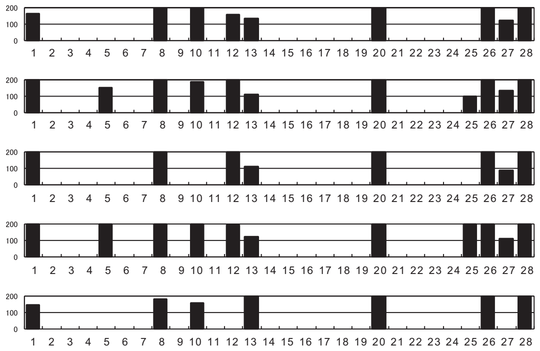

Fig. 6. Sensor pattern (behavior 2)

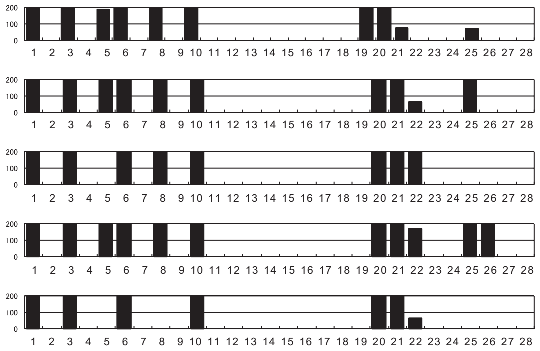

Fig. 7. Sensor pattern (behavior 4)

for each discriminator according to its importance. These two subspaces may differ from the behavior categories. The designer of the system defines their importance. For example, the following three discriminators represent the behavior "using the mouse while taking a seat" in order of importance:

1) Whether the person is taking a seat
2) Whether the person's arm is moving
3) Whether the computer mouse is moving

B. Constructing a recognition tree by linear discriminant analysis

The construction process for the binary decision tree for recognition is as follows (see Fig. 4):

1) Observe behavior and memorize the sensory data and behavior category.
2) The designer defines the discriminators and their importance based on the behavior categories.
3) Divide the data into two categories with the most important discriminator.
4) Repeat step 4 for the next most important discriminators, recursively, in a breadth-first manner, until the classification error becomes less than a constant value or all discriminators are used.

During the recognition phase, the system applies the linear discriminant functions to the observed sensory data, starting at the root node, until it arrives at a leaf node. The system outputs the behavior category that the leaf node belongs to as the result.

III. Behavior recognition using infrared proximity sensors

As an application of the proposed method, we have developed a human behavior recognition system at an office workplace. As shown in Fig. 1, we have distributed 28 infrared proximity sensors around a desk. The digits in the figure represent sensor ID numbers. The sensors were distributed randomly. If a sensor detects something, it returns 1; otherwise, it returns 0. They are only active when someone is present. There are four types of sensors each of which has a different range: 40 cm, 50 cm, 80 cm, and 100 cm. The sensors are oriented perpendicularly to the wall.

A. Behavior category

Fig. 5 lists 12 categories of behavior that were determined *a priori*.

- Behavior 1: Sitting
- Behavior 2: Using the mouse
- Behavior 3: Using the mouse while extending the left elbow
- Behavior 4: Typing on the keyboard
- Behavior 5: Typing on the numeric keypad
- Behavior 6: Touching the monitor
- Behavior 7: Putting one's hands on the desk
- Behavior 8: Standing
- Behavior 9: Using the mouse while standing
- Behavior 10: Typing on the keyboard while standing
- Behavior 11: Putting one's hands on the desk while standing
- Behavior 12: Standing back

B. Acquisition of behavior data

The system acquires data from all infrared sensors every 10 miliseconds. The data represents instantaneous behavior but suffers from noise. To obtain more stable data, the system sums up data sampled during two second intervals. These values are represented by a 28-dimensional feature vector. In this experiment, we stored 300 vectors corresponding to each behavior category.

Fig. 6 and Fig. 7 show instances of behavior while a person uses the mouse and types on the keyboard, respectively. The five graphs represent different iterations with the same person. The horizontal axis represents the ID number of the infrared sensor, and the vertical axis the sum of the sensory data during two seconds. These figures show that the shapes of graphs belonging to the same behavior category are similar, and those belonging

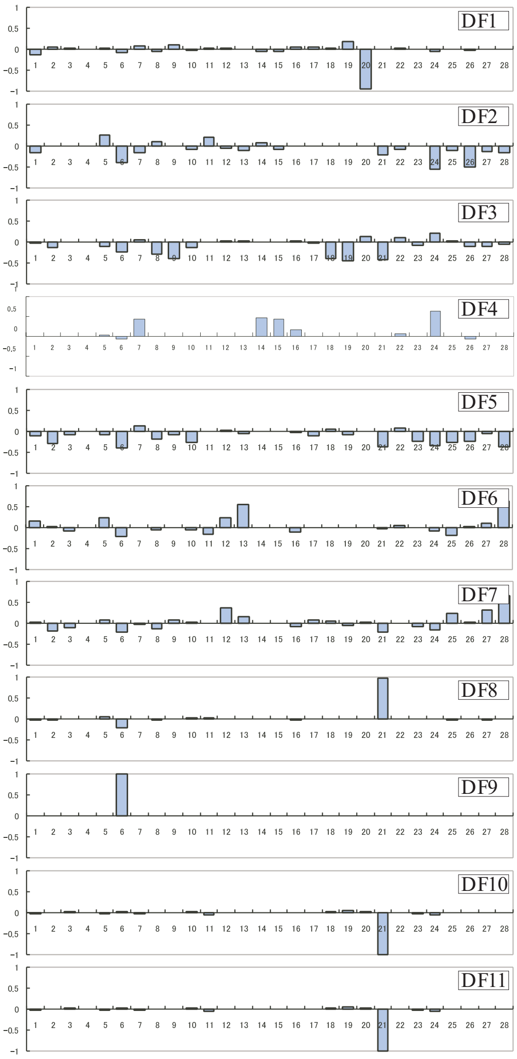

Fig. 8. Weights of each discriminant function

to different behavior categories differ. This means human behaviors can be represented as a combination of positional information, so that it is possible to classify human behaviors based on the stored behavior data.

C. Construction of the recognition tree

We have prepared 11 discriminators as shown in Table I. DF means the value of the linear discriminant function. Some discriminators are the same; however, their importance and discriminant functions vary. Fig. 4 shows a recognition tree obtained by using the discriminators.

We measured recognition rates to verify the decision tree. Three subjects behaved naturally in the environment, and we compared their self-declared behavior with the results of the system, as shown in Table II.

The subjects did not know what behavioral categories the system could recognize; therefore, about 20% of their behavior did not match any of the behavioral categories of the system. This is why the recognition rates are not so high in Table II. However, the rates are about 90% once unexpected behavior categories have been removed. We believe this recognition rate to be sufficiently high for a practical behavior recognition system.

TABLE I
DISCRIMINATORS

	DF< 0	DF> 0
D1	sitting	standing
D2	hand operation	no hand operation
D3	standing forward	standing back
D4	not touching the monitor	touching the monitor
D5	hand operation	no hand operation
D6	not using the mouse	using the mouse
D7	not using the mouse	using the mouse
D8	not typing the keyboard	typing the keyboard
D9	putting the left hand on the desk	putting the left hand under the desk
D10	not typing the keyboard	typing the keyboard
D11	not typing the numeric keypad	typing the numeric keypad

TABLE II
RECOGNITION RATE [%]

	Subject A	Subject B	Subject C	Total
Recog. rate	72.7	72.7	76.1	73.8
Ratio of known behavior data	77.2	81.8	80.9	80.0
Recog. rate under the known behavior data	94.1	88.8	94.1	92.3

D. Analysis

In this experiment, the sensory data is 28 dimensions, and the discriminator function f is represented by

$$f = \sum_{i=1}^{28} a_i x_i + a_0, \quad (1)$$

where x_i is the sensory data.

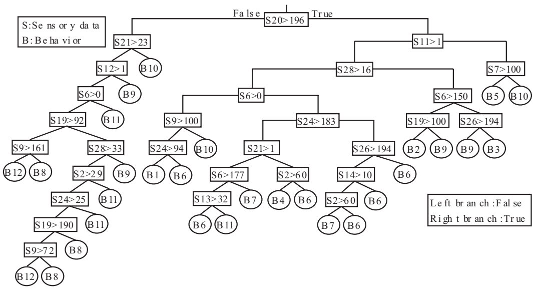

Fig. 9. The decision tree constructed by Quinlan's C4.5 algorithm to recognize behavior at the desk.

Fig. 8 shows the weights a_i of the discriminators that constitute the decision tree. Each weight indicates the contribution of the discriminator to the conclusion. From this figure, we find that only one sensor is dominant in several discriminant functions. This suggests that placing sensors intentionally rather than randomly could offer better performance with fewer sensors for the current set of behavioral categories and environment. However, the random placement of many sensors can be of benefit in situations where the environment can change dynamically or we cannot know the categories in advance.

In the approach proposed here, the designer gives the order of the discriminators. However, if one sensor is sufficient for each discrimination, it is possible to use an information theoretic (or other) criterion to determine the order (cf. ID3 and C4.5 [13]). We have investigated this problem by comparing the performance given by C4.5, which is based on ID3 but allows real valued attributes.

Fig. 9 shows the recognition tree constructed by C4.5. The root discriminator roughly divides all data into two categories: standing and sitting. The decision tree obtained by C4.5 bears some similarity to the results obtained by our method, although the performance of the recognition system of C4.5 is somewhat worse. The judgments of the designer used in our approach roughly match the evaluation based on the information criteria.

IV. Behavior recognition using distributed wideview video cameras

The purpose of this experiment is to evaluate whether a memory-based approach can be applied to a distributed vision system. We mounted four wide angle vision sensors on the ceiling of the display section of a video rental store. The wide field of view of the cameras enables redundant observation, which may enhance the recognition rate of human activity. The mirror is calibrated to avoid the distortion found in alternative systems (e.g., fisheye lenses), though memory-based approaches do not require this. Images from each of the four cameras are combined into a single 640×480 RGB image for simultaneous processing. An image capture board digitizes the analog system, and a PC with a 2.4GHz Pentium IV CPU and 2 gigabytes of memory performs subsequent processing.

Human activity is isolated in the images by background subtraction and binarization. First the brightness component is isolated in the RGB images: $I = 0.299R + 0.587G + 0.144B$. A parameter is used to threshold the difference between the current and background images: $S_{ij} = 1$, if $|I_{ij} - H_{ij} \geq t|$, otherwise $S_{ij} = 0$.[1] Since environments change over time, the background image is gradually updated by taking a time average across recent images: $H_t = I_t \times \alpha + H_{t-1} \times (1 - \alpha)$, where $0 \geq \alpha \geq 1$. Larger values of α result in quicker updates. The 640×480 image is encoded as a single one dimensional vector.

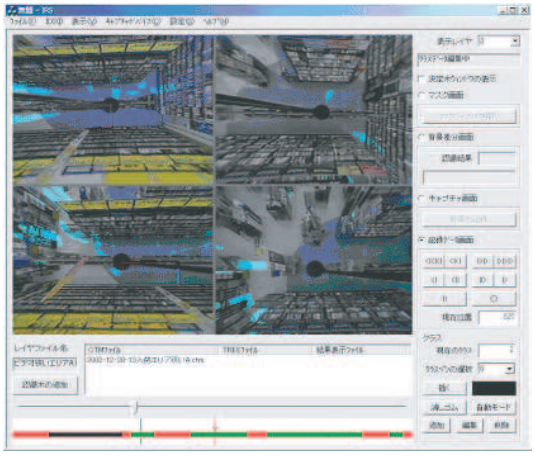

Fig. 10. The software user interface of the behavior recognition system at the video rental shop. Human activity is marked in cyan.

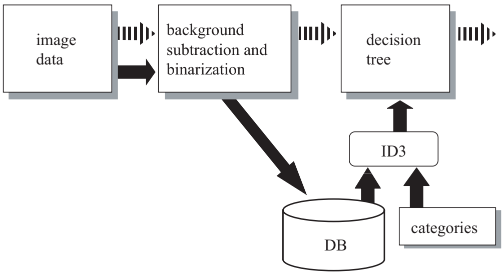

Fig. 11. The black arrows show the processing required to construct the decision tree for behavior recognition, while the dashed arrows show the processing for recognizing behavior once the tree has been constructed.

After this preprocessing stage, the system learns to recognize whether there is no person, one person, or more than one person in each region by using the ID3 algorithm to construct a decision tree. We chose ID3 instead of linear discriminant analysis because it focuses on dimensions independently, which is less problematic in a space of such high dimensionality. Recognition occurs when the current composite image matches a previously processed image.

The display area is divided into the regions A, B, C, and D, and the number of people in each region is quantized into three categories: none, one, or several. The system is to recognize these categories without taking into account camera parameters or the structure of the environment. In typical engineering applications, features that are useful for making discriminations are determined based on the task. However, if the task or environment are not known in advance, an alternative is needed. We draw inspiration from the animal realm, since many species are able to recognize different classes of phenomena by learning to

[1] In the experiments, $t = 50$ and $\alpha = \frac{1}{2048}$.

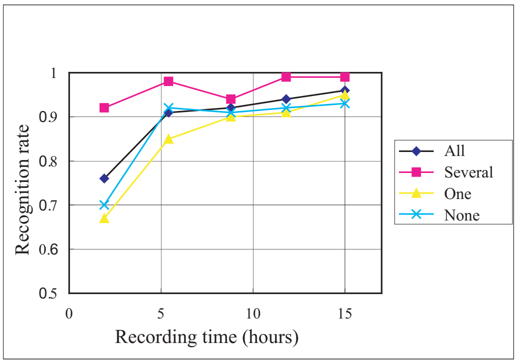

Fig. 12. After processing eight hours of video, the systems achieves a recognition rate in excess of 90% for all categories of behavior.

detect features that are invariant within each class.

Owing to the prohibitive computational demands of directly comparing the current image to those previously categorized, Quinlan's ID3 is used to reconstitute the database of past images in the form of a decision tree. ID3 recursively divides the images according to the attribute (i.e., pixel) that has the least entropy (i.e., degree of doubt) with respect to the conclusion, forming two subtrees. The data in the subtrees are further divided until all attributes have been exhausted or the conclusion for all the data in a subtree is the same.

In our experiment, the attributes are the 0 or 1 values of the image vector, and the conclusion is whether there are zero, one, or several persons in each of the four regions of the image (see Fig. 2). Training data were taken from approximately 15 hours of recordings over the course of three days. Recognition rates are based on one hour of observation, and this data was not included in the training data. The frame rate is 10 fps. Fig. 12 plots the recognition rate versus the amount of training data as measured in hours of recording time. Eventually the system exceeds a 90% rate of accuracy for all attributes.

V. Conclusion

This paper proposed a robust, real-time method of recognizing everyday human activity by means of a memory-based approach that uses many infrared sensors. The developed system performs sufficiently well to be practical. Furthermore, we could verify that the order of the discriminators given by the designer matches those selected by the information criterion to some extent.

In addition, we applied distributed wideview cameras. The wideview cameras provide much richer information about human behavior and have the potential to recognize more sophisticated and detailed behavior. Owing to the difficulty in applying linear discriminant analysis to the high dimensional space of the camera image, we selected ID3, which focuses on local dimensions independently.

The advantages of our approach over traditional model-based approaches are that we do not need to define a model *a priori* and the computational cost is low. One problem with model-based approaches is that if the system fails in constructing a model, behavior recognition also fails. Our approach is robust, as exhibited by a successful recognition rate of behavior in excess of 90%.

VI. References

[1] A. Bobick, S. Intille, J. Davis, F. Baird, C. Pinhanez, L. Campbell, Y. Ivanov, A. Schutte, and A. Wilson. The KidsRoom: A Perceptually-Based Interactive and Immersive Story Environment. Technical Report 398, E15, 20 Ames Street, Cambridge, MA 02139, December 1996.

[2] A. Bobick and A. Wilson. A state-based technique for the summarization and recognition of gesture. In *Proc. ICCV*, pages 382–388, 1995.

[3] N. Chleq, F. Bremond, and M. Thonnat. Image understanding for prevention of vandalism in metro stations.

[4] M. H. Coen. Design principles for intelligent environments. In *AAAI/IAAI*, pages 547–554, 1998.

[5] J. Cooperstock, K. Tanikoshi, N. T. G. Beirne, G., and Buxton. Evolution of a reactive environment. pages 170–177, May 1995.

[6] J. Davis and M. Shah. Recognizing hand gestures. In *Proc. ECCV*, pages 331–340, 1994.

[7] Y. Guo, G. Xu, and S. Tsuji. Understanding human motion patterns. In *Proc. ICPR*, pages 325–329, 1994.

[8] H. Ishiguro. Distributed vision system: A perceptual information infrastructure for robot navigation. In *Proc. IJCAI*, pages 36–41, 1997.

[9] H. Ishiguro and T. Nishimura. Vambam: View and motion based aspect models for distributed omnidirectional vision systems. In *Proc. IJCAI*, pages 1375–1380, 2001.

[10] H. Ishiguro, R. Sato, and T. Ishida. Robot oriented state space construction. In *Proc. IROS*, pages 1496–1501, 1996.

[11] K. F. MacDorman. Feature learning, multiresolution analysis, and symbol grounding: A peer commentary on 'the development of features in object concepts'. *Behavioral and Brain Sciences*, 21:32–33, 1998.

[12] K. F. MacDorman, K. Tatani, Y. Miyazaki, and M. Koeda. Proto-symbol emergence. In *Proc. IROS*, 2000.

[13] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[14] T. Sogo, H. Ishiguro, and M. Trivedi. Real-time target localization and tracking by n-ocular stereo. In *Proc. IEEE Workshop on Omnidirectional Vision*, pages 153–160, 2000.