Memory-Based Attention Control for Activity Recognition at a Subway Station

Karl F. MacDorman
Indiana University

Hiroshi Nobuta
Wakayama University

Satoshi Koizumi
Osaka University

Hiroshi Ishiguro
Osaka University

We have developed a multicamera system, Digital City Surveillance, which uses a new calibration-free behavior recognition method for monitoring human activity at a subway station. We trained nine support vector machines from operator-classified data to recognize 512 combinations of events. Our method of attention control greatly reduced computation and increased classification accuracy.

Computer vision research is highly applicable to monitoring people in public places.[1,2] Automatic monitoring has become more prevalent in part because the cost of mounting cameras has now been dwarfed by the cost of hiring operators to watch them.

To improve battlefield awareness, for instance, the US Defense Advanced Research Projects Agency (DARPA) sponsored the Video Surveillance and Monitoring (VSAM) project to develop automated technologies for monitoring people and vehicles.[3] These same technologies can be applied to visual surveillance for law enforcement or private security. The project's basic strategy is to detect moving regions in images by background subtraction and to track them with calibrated cameras. Contextual information makes the system more robust.

Along similar lines, several European academic and industrial institutions have developed the ADVISOR (Annotated Digital Video for Intelligent Surveillance and Optimised Retrieval) system to analyze video feeds from cameras at a subway station in real time. The system—tested in Barcelona, Brussels, and London—warns operators of dangerous situations such as crowding and fighting in addition to ongoing acts of vandalism and fare evasion.[4]

Separate systems perform model-based people tracking,[5-7] behavior recognition, and crowd analysis.[8] Because the ADVISOR system is model-based, not memory-based, it relies on models developed specifically for recognizing particular kinds of human activity in a subway station. These models would need to be replaced with different models to recognize other kinds of activity, such as vehicle traffic or wildlife movement. In addition, ADVISOR requires accurate camera calibration to function.

In a significantly different approach to activity recognition, we've developed a calibration-free, memory-based distributed vision system, Digital City Surveillance. Before we explain the specifics of our design, however, we discuss the system context.

Digital City Project

Apart from enhancing security, people-tracking systems can actively support many kinds of human activity, especially in the sensor-rich, computer-networked environments being developed in ubiquitous computing. The Digital City Project of the Japan Science and Technology Agency Core Research for Evolutional Science and Technology (JST CREST) is actively working with people-tracking systems. This project, directed by Toru Ishida, explores the systems' potential for transforming and enhancing people's lives,[9] for example, by realizing evacuation systems for urban disasters.[10] The surveillance system described in this article is part of the Digital City Project and demonstrates one application of people tracking.

We've constructed a distributed vision system for recognizing human activity. Figures 1 and 2 show the multiple camera system installed in the JR Kyoto subway station concourse. The system consists of 28 cameras with special mirrors that

1070-986X/07/$25.00 © 2007 IEEE

Published by the IEEE Computer Society

provide undistorted wide-view images (see Figure 3, next page), 12 in the concourse and 16 on the platform. (The 16 wide-view vision sensors mounted on the platform weren't used by the system we describe.) Figure 3 shows images taken by the system on the concourse. Figure 2 shows the sensors' position and coverage. The system can obtain information about a large area—by means of these vision sensors—including a path, ticket vending machines, ticket gates, and stairs.

Our surveillance system makes several improvements on ADVISOR and other similar systems:

- We developed a new type of wide-view vision sensor for large public spaces that allows the same area to be covered with about one-fourth the number of cameras.
- We replaced the separate subsystems for behavior recognition and crowd estimation with a single general learning mechanism that doesn't require explicitly designed human models or camera calibration.
- We developed a form of attention control based on feature selection to improve classification accuracy and vastly accelerate learning. When classifying behavior, each support vector machine (SVM) pays attention to only those pixels that provide the most information gain with respect to making the correct classification.

Figure 1. Twelve wide-view vision sensors are attached to the ceiling of the concourse at the JR Kyoto subway station. Each sensor consists of a charge-coupled device (CCD) video camera and a nonparametric convex mirror. This system was developed in the Digital City Project supported by the Japan Science and Technology Agency.

Digital City Surveillance focuses on enhancing learning-based behavior recognition by selecting the most informative features in a video. Because the system's "attention" is directed to these features, the mechanism implements a form of attention control.

Figure 4 summarizes our approach. First, the system detects moving pixels by background sub-

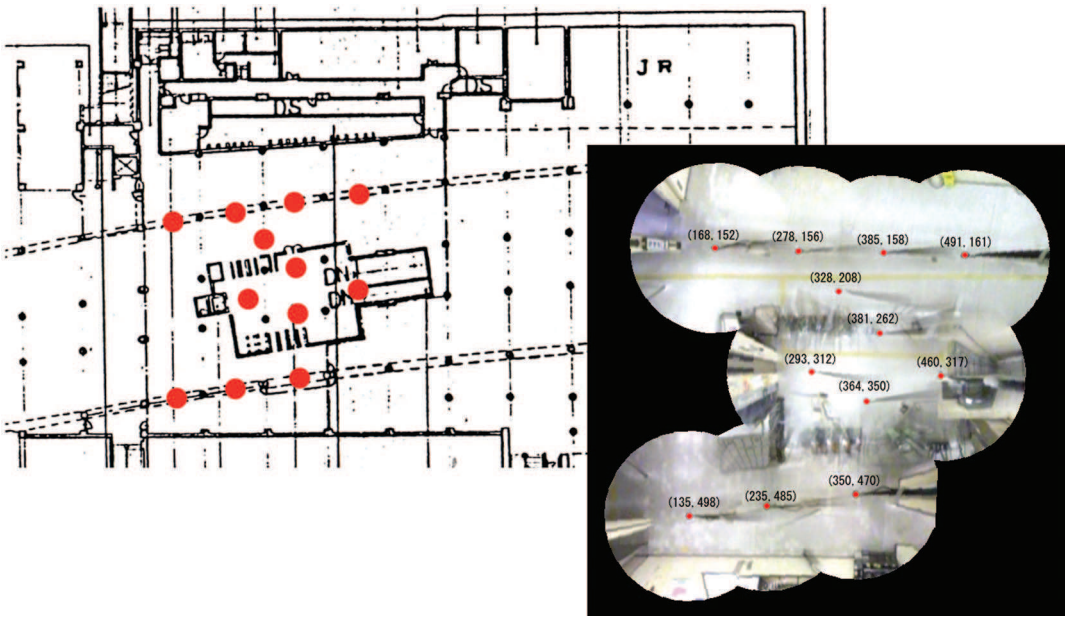


Figure 2. The diagram on the left shows where the wide-view vision sensors are located in the subway concourse. For illustrative purposes, the square on the right shows a mosaic derived from simultaneous recordings from 12 arbitrarily placed vision sensors. The camera positions aren't explicitly represented.

Figure 3. On the left are images taken from four wide-view vision sensors, which show people walking in the JR Kyoto subway station. On the right is a compact, dual mirror version of the wide-view vision sensor with identical optics, which we developed as a commercial product in collaboration with Vstone.

traction and obtains binary images (see Figure 5b). The average of a series of binary images, weighted by recency, lets us produce an input vector that includes motion information (Figure 5c). A human operator then classifies these examples as positive or negative instances of nine kinds of events. The system constructs discriminant functions for the given examples using nine support vector machines. One of the merits of support vector machines is that fewer training examples are needed than with other, simpler memory-based approaches (for example, nearest neighbor). The system reduces the size of the parameter space by a feature selection process.

Memory-based recognition: Rationale

The most common approach to detecting human positions and behavior is based on features. In this approach, targets are segmented from the original sensory signal, and extracted features from targets are matched based on models of human behaviors. However, if the environment and behavior to be recognized aren't known, it's difficult to design a robust segmentation algorithm and complex features that will be able to detect these unknown behaviors.

In the model-based approach, such as that of the ADVISOR system, system designers can't anticipate all the features that will be necessary for detecting behaviors, unless they know the behaviors in advance. Therefore, system designers might not be able to implement all the necessary feature detectors beforehand.[11,12] This makes the system susceptible to the symbol grounding problem, because a need arises for symbols but the system is unable to connect them to features of the world. In addition, maintaining a correspondence between a complex, symbolic model and a continuously changing environment might create unnecessary computational demands, which has been identified with the frame problem.[13] Thus, the model-based approach limits behavior recognition to known environments.

A promising alternative is the memory-based approach[14] which uses low-level features and online learning of target behaviors.[15] The system learns discriminant functions for image features obtained after background subtraction. This facilitates segmentation. After a relatively brief period of online instruction, the system can function in environments that were unknown to its designers, but known to its trainer, who doesn't need a technical background.

Digital City Surveillance incorporates a calibration-free behavior recognition method for multiple camera systems. The system maps data from the 12 cameras on the concourse to classes, such as "people are passing through the ticket gates," "there are many people in front of the stairs," and so on (see Table 1 and Figure 6). An announcement system provides this information to station operators in synthesized speech and to subway users through the Internet.

We've adopted a memory-based recognition approach for this application, which only takes a couple of hours for a novice to train. Memory-based approaches derive their recognition results at least in part from classified data points stored in memory. The large amounts of memory and high-speed CPUs now available help us construct the huge parameter space defined by the 12 cameras, thus making memory-based image processing fea-

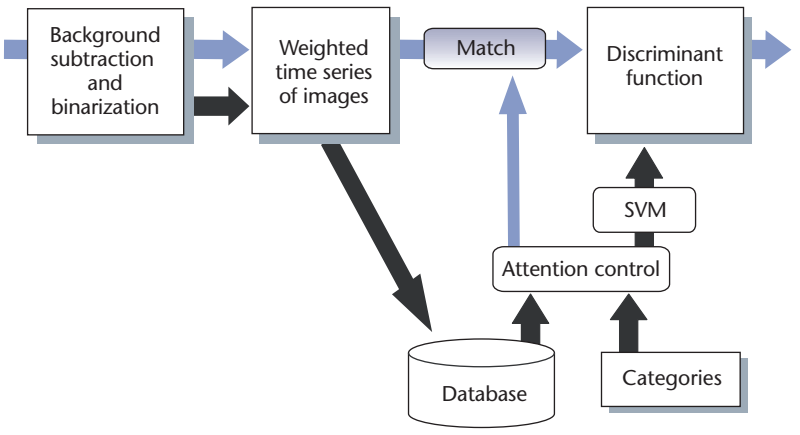


Figure 4. This diagram summarizes the information processing flow of our approach. The black arrows signify the training phase, while the lighter arrows signify the recognition phase. (SVM means support vector machine.)

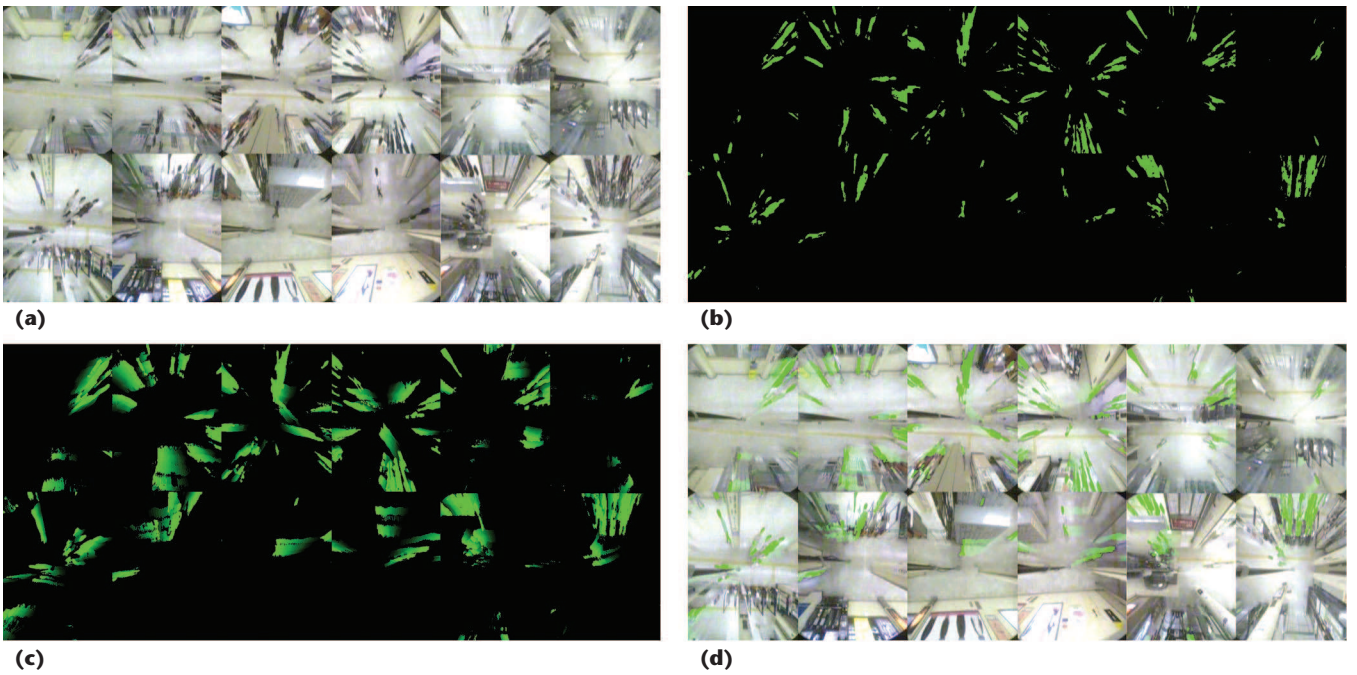


Figure 5. Data preprocessing is achieved as follows: (a) A composite image is formed from each video frame from the 12 wide-view vision sensors in the concourse. There was no attempt to maintain geometrical relations among the cameras. (b) The background is removed and the images are then binarized. (c) An average of composite images, weighted by recency, is computed. It constitutes the vector for categorization. (d) For illustrative purposes, the recency-weighted average is shown overlaying the original image.

Table 1. Recognition tasks for the subway announcement system.

Task	Condition
1	People are (not) going down the stairs.
2	People are (not) passing through the ticket gates.
3	People are (not) walking from the ticket machines to the ticket gates.
4	People are (not) walking by the wall.
5	People are (not) going upstairs.
6	It is (not) crowded.
7	There are (not) a few people.
8	There are (not) many people in front of the stairs.
9	There are (not) many people around the ticket machines.

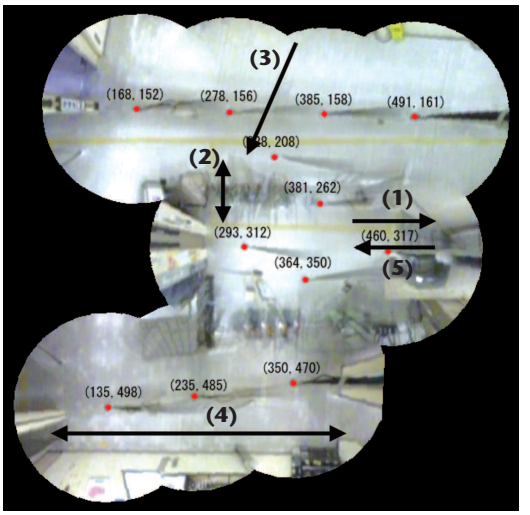


Figure 6. The first five events concern people's movement. The arrows depict their direction of movement. (The camera image mosaic is for illustrative purposes only.)

sible. A memory-based approach lets the system handle camera data without referencing camera positions, viewing angles, or fields of view, giving the nontechnical camera operators unprecedented freedom in placing and moving cameras.

The system informs the operators about what's happening in a subway station by means of a voice synthesizer and sends text-based updates through the Internet. (MPEG video samples are available at http://www.androidscience.org; under Research Projects, see "Distributed Vision.") This kind of system can also act as a perceptual information infrastructure for robot navigation[16] and human–robot interaction. We envision that future subway and train stations will function safely with fewer personnel by employing humanlike robots as passenger guides[17] and special-purpose robots for security, sanitation, and routine maintenance. Prototype systems have already addressed all these applications at the 2005 World Exposition in Aichi, Japan.

The wide viewing field of omnidirectional cameras is suitable for observing targets from various directions.

However, as the sensory space becomes larger as more cameras are used, a memory-based approach requires more time to learn to recognize behaviors. In addition, there are many pillars and other occluding objects in the subway station, rendering parts of the images useless for behavior recognition.

In response to these factors, we introduce a form of attention control[18] that restricts consideration only to those regions of images relevant to the recognition of a particular condition as classified by the teaching signal.[19] (Our approach shouldn't be confused with the biologically inspired method of attention control proposed by Itti et al.[20]) As we'll see, attention control dramatically reduces computation and somewhat increases accuracy by statistically selecting which pixels are worth observing according to the recognition task. Because each of the nine support vector machines detects the presence or absence of a particular condition (see Table 1), each machine is sensitive to movement in different regions of the images.

How the behavior recognition system works

To cover a large subway station with a relatively small number of cameras, we developed a wide-view vision sensor using a nonparametric mirror. The new sensor has an advantage over omnidirectional cameras because it can be ceiling mounted, and it avoids the distortion of fisheye lenses. The system uses attention control to select the pixels from the 12 wide-view cameras in the concourse that provide the highest gain in information for performing each classification task. A novice station operator can train the system in less than three hours.

How the distributed vision sensors work

In most multiple camera applications, increasing viewing areas can cut costs by reducing the number of cameras needed. To this end, we can enhance coverage by using wide-angle, fisheye, or omnidirectional sensors. Previously, we proposed a distributed omnidirectional vision system as a perceptual information infrastructure for monitoring human activity.[21] The wide viewing field of omnidirectional cameras is suitable for observing targets from various directions. We've developed a real-time human tracking system that covers a wide area with relatively few omnidirectional cameras.

We've also developed several algorithms that let the distributed omnidirectional vision system autonomously acquire positional information among cameras based on the fact that the omnidirectional cameras view each other.[22] However, the cameras must be placed approximately at eye level (about 160 cm from the ground), because their field of view is horizontally oriented. Omnidirectional cameras can't be used in a busy public place such as a subway station because that would obstruct the flow of pedestrians. The cameras must be attached to the ceiling.

To address this requirement, we developed a new kind of wide-view vision sensor. Each vision sensor is made of a mirror that has a special nonparametric shape and a color CCD camera. The mirror gives the sensor a broader field of view than a perspective camera. Therefore, comparatively few vision sensors are needed to cover a large area (28 instead of about 100). Strictly speaking, these cameras aren't omnidirectional, because their focus of expansion isn't visible for horizontal movements. In collaboration with Vstone, we developed a compact commercial vision sensor (see Figure 3) with identical optics to the prototype (see Figure 1) by embedding a downward-facing camera in the convex mirror and placing a small, flat mirror immediately below it. Thus, activity below the sensor is reflected by the convex mirror onto the flat mirror and then through the camera lenses.

Training a robust classifier

Our method treats a time-weighted image from multiple cameras as an input vector. The series of 160×120-pixel images obtained from 12 cameras are combined into one composite image (see Figure 5a). To reduce its dimensionality, we convert the full color image into a binary image by subtracting the background image and binarizing. Color values range from 0 to 255, and a threshold of 30 was used for the binarization. The initial background image is acquired early each morning when services are halted and the station is guaranteed to be empty. Lighting is

constant because the station is windowless and underground, and objects are secured to the walls and floor at fixed locations. Therefore, it's rarely necessary to update the background image, and a pixel is only updated if, for five minutes, its value continuously deviates from the initial background image by ±10 and remains within ±5 of its five-minute running average ($B_{k+1} = \alpha I_k + (1 - \alpha)B_k$, $\alpha = 0.005$).

Unless specialized hardware is used, most sophisticated methods of background subtraction—for example, a mixture of Gaussians, kernel density and means-shift estimators, and eigen backgrounds—require too much computation to work in a real-time surveillance system with many video cameras operating at a high frame rate, such as 12 to 28 cameras at 30 frames per second (fps). Faster methods, such as computing a running average of recent pixel values, can easily corrupt a background image in a heavily trafficked area like a subway station.

Although the direction of pedestrian flow isn't explicitly represented in the binarized images, to detect it we use the average of a series of past images, weighted by their recency (see Figure 5). We considered this method more robust than computing optical flow, which can be susceptible to noise.[23] The value of pixel i is obtained by the equation

$$L_i(t) = \frac{2}{F(F+1)} \sum_{k=0}^{F-1} I_i(t-k)(F-k) \quad (1)$$

in which t is the frame number of the image, and F is the number of frames in the temporal window. (The value F was set to 20 in the experiments.) $I_i(t) \in \{0, 1\}$ is the result of background subtraction and binarization. The method is similar to Bobick and Davis's[23] in its results, except Digital City Surveillance uses background subtraction instead of image differencing, and recency weighting and a temporal window instead of a decay factor.

Based on a teaching signal, the classifier learns to recognize human activity in the subway station with the composite image. The presence of each condition wasn't defined explicitly (for example, it's crowded if the head count exceeds 100), but was instead based solely on the operator's subjective opinion.

Because various kinds of activity are observed in a complex environment, the system is required to produce multiple recognition results at the same time. Therefore, the support vector machines run in parallel, each monitoring a specific type of activity. (Support vector machines with kernel mapping are introduced in Figure 7 and the "Support Vector Machines" sidebar, next page.) The system was trained in less than three hours. Each support vector machine was trained individually while the operator watched for its corresponding event in randomly selected video segments that played at 30 fps. The operator marked the significant regions by pressing one button at the beginning of the event and another button at its termination.

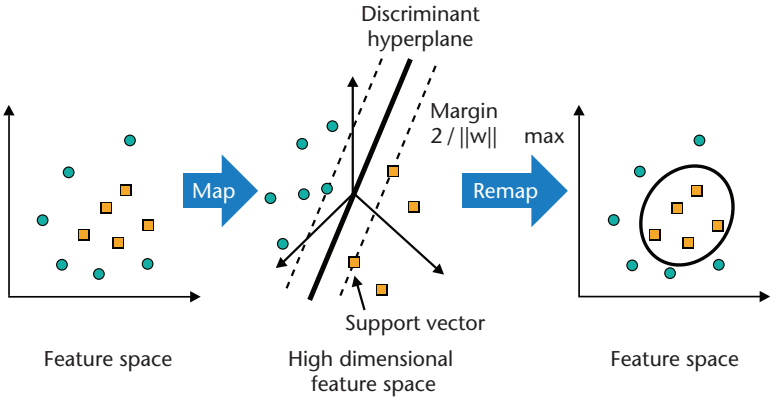


Figure 7. For classification by a support vector machine, kernel functions map data points into a feature space of higher dimensionality in which classes are separated by a hyperplane. This provides nonlinear discrimination in the original feature space. In this application, the features correspond to the N-pixels that provide each support vector machine the highest information gain with respect to the class attribute (for example, "People are walking down the stairs").

Attention control based on information gain

Our method of attention control involves selecting an effective dimension in the input space. This is equivalent to selecting a pixel from the composite image. The selection is based on the information gain from the pixel value and is performed independently for each pixel. Therefore, the input vector to each support vector machine is a one-dimensional array corresponding to the N-pixels with highest information gain (in descending order) for each SVM's particular classification task. In other domains, information gain has proven to be one of the most effective methods of ruling out potential features without sacrificing categorization accuracy.[24] Entropy is computed by the equation

$$H = -\sum_{y=1}^{A} p(y) \log_2 p(y) \quad (2)$$

in which y is the attribute as set by the teaching signal for the class of interest—for example, whether people are present at the ticket gate—and $p(y)$ is its probability of occurrence. A is the total number of attributes. The value $p(y)$ is the number of occurrences of y divided by the total number of training data.

Support Vector Machines

The memory-based system must be able to recognize the specific behaviors that humans are performing. This is represented as a problem of pattern recognition in which an input vector is classified into two categories. We use support vector machines to perform the classification problem.[1,2] Support vector machines encode the relative positions of data points in a feature space of higher dimensionality using their dot products, which can be computed from the data points using a kernel function (see Figure 7 in the main text).[3] Support vector machines combine the nonlinearity of neural networks, the computational efficiency of linear algebra, and the solid theoretical foundation and statistical rigor of regularization methods. The advantages of support vector machines over neural networks include a variable-sized hypothesis space, exact optimization (that is, no local minima), polynomial time convergence (that is, faster training), and no overfitting by adjusting the margin.

The margin of each input vector $x_i (i = 1, 2, \ldots N)$ to a discriminant hyperplane $w^t x + b = 0$ is defined as follows:

$$\gamma = \frac{y_i\left(w^t x_i + b\right)}{\|w\|} \quad \text{(A)}$$

A support vector machine determines w and b based on margin maximization. To determine w and b uniquely, the following constraint is introduced:

$$\min_{i=1,\ldots N} \left|w^t x_i + b\right| = 1 \quad \text{(B)}$$

The formula for margin maximization is derived by applying the Lagrange multiplier α from Equations A and B:

$$\text{maximize} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} y_i y_j \alpha_i \alpha_j x_i^t x_j \quad \text{(C)}$$

$$\text{subject to} \sum_{i=1}^{N} y_i \alpha_i = 0, \alpha_i \geq 0, \left(i = 1,2,\ldots N\right) \quad \text{(D)}$$

The support vectors and the discriminate hyperplane are computed from these equations. In this research, we select the Gaussian kernel, and α is set to 0.5.

References

1. C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning,* vol. 20, 1995, pp. 273-297.
2. V. Vapnik, *Statistical Learning Theory,* Wiley, 1998.
3. J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge Univ. Press, 2004.

The system calculates the entropy H_0 when no pixel values are known. For each pixel i it computes the entropy H_i when the value of pixel i is known. So, the information gain of pixel i is $H_i - H_0$. The system selects the pixels with highest information gain and classifies behavior based on those selected pixels. This selection process continues until the system classifies correctly.

Figure 8a shows the relationship between classification accuracy and the number of pixels used to detect people walking by the wall. Figure 8b shows the relationship between classification accuracy and the number of pixels used to detect people descending the stairs. Figure 8c shows the pixels selected by attention control when the system recognizes people walking along the wall, and Figure 8d when the system recognizes that people are walking down the stairs. This confirms that the distributed vision system was attending to the area by the wall or around the stairs.

System reliability

We assume that the system is applied to surveillance. In a complex environment, it's sometimes difficult for even a person to recognize what's going on. In such an ambiguous situation, or for borderline states, the system isn't necessarily required to produce output. For example, if a region is only moderately crowded, the system shouldn't be forced to make either the announcement that it's crowded or it's not. In this study, the system notifies us only when the reliability value is above 0.7. (We explain how the choice of this value was determined empirically later.) We use the distance between unknown data and a discriminant hyperplane to calculate a value analogous to the reliability of the judgment (see Figure 9, page 46). If this reliability value is lower than the threshold, the system outputs nothing.

Evaluating system accuracy

We conducted several experiments to verify the performance of the recognition system in a large subway station. We evaluated the system's recognition accuracy in terms of the number of training samples. The influence of attention control on accuracy and recognition speed is experimentally determined as is the relationship between accuracy and the number of pixels selected by attention control for use by the support vector machines.

Preparing the system for attention control

In preparation for further experiments, we examined the relationship between classification accuracy and the number of pixels selected by attention control. Each support vector machine was trained, first using the pixel with highest information gain, and then the two pixels with highest information gain, and so on, until we

achieved 99 percent accuracy. We used the same data for both training and testing because our goal was to evaluate attention control and not classification performance. In all other experiments, we based performance evaluation on a data set for testing, which didn't include any training data.

Figure 8 shows the results for two different tasks. The number of pixels required to achieve 99 percent performance is task dependent. Recognizing people who were descending the stairs required only 32 pixels, but recognizing people walking by the wall required 130 pixels. Other tasks varied between these extremes, except for the two support vector machines that detected the general level of crowding in the concourse area. If too many pixels are used, the accuracy declines because of noise from irrelevant data. This demonstrates attention control's power to exclude the irrelevant. We recorded data from seven to eight o'clock on a weekday morning to evaluate the accuracy of the recognition system.

Accuracy of the support vector machines

The composite images were presented to the system in random order without replacements, so that the system was evaluated only on novel images. The number of pixels used by attention control was determined by the previous experiment. For most tasks, learning converged after 100 to 5,000 training samples, with accuracy rates ranging from 75 percent to 90 percent on unseen testing data with 81 percent being the average. However, for task 3 in Table 1 ("People are walking from the ticket machines to the gates"), learning hadn't converged even after 10,000 training samples, at which point the accuracy was only 68 percent, thus bringing average performance down to just below 80 percent.

The low performance for task 3 might have something to do with the fact that there are many ways of walking from the ticket machines to the gates. In addition, many situations appear ambiguous even to the operator who must determine the training signal, illustrating the difficulty of this particular task.

Figure 10 shows the accuracy of recognition with respect to the number of training samples for the first three tasks in Table 1:

- Task 1. People are walking down the stairs.
- Task 2. People are going through the ticket gates.
- Task 3. People are walking from the ticket-vending machines to the gates.

The performance of the system generally improves as the number of training samples increases.

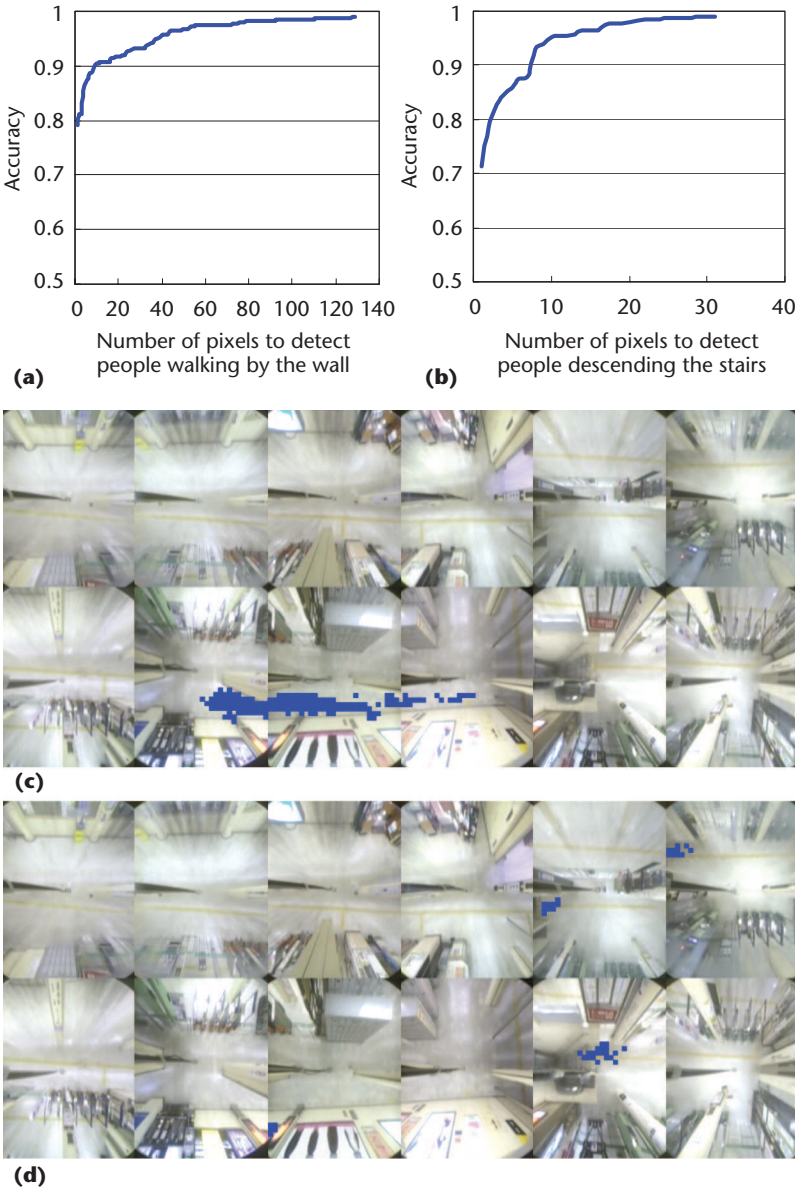


Figure 8. (a, b) Classification accuracy improves with the number of pixels selected by attention control. Pixels are selected based on information gain. (c) For training data, 130 pixels are required to achieve 99 percent accuracy in recognizing people walking by the wall but (d) only 32 pixels in recognizing people going down the stairs.

Attention control increases accuracy and speed

In the next experiment, we inspected the efficacy of attention control. Attention control increased accuracy for all tasks. The minimum

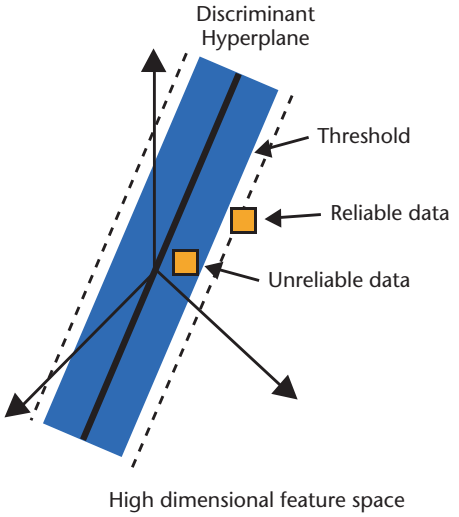


Figure 9. The support vector machines suppress their output when input data is borderline to avoid the risk of making a misclassification. The reliability metric is set to 70 percent of the distance from the discriminant hyperplane and the outer hyperplane.

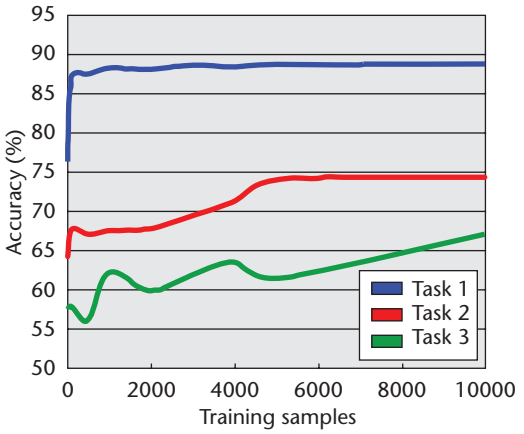


Figure 10. For learning the support vector machine classifiers, this experiment used samples that were extracted randomly from the data recorded during one hour. This experiment verifies improvement in classification accuracy with respect to the number of training samples for randomly selected testing data that wasn't included in the training set.

increase in accuracy was 2 percent for task 3, and the maximum was 23 percent for task 1. Figure 11 shows the accuracy for the first three tasks with or without attention control. In this experiment, the number of training samples is only 1,000. The accuracy of recognition with attention control is higher than without attention control.

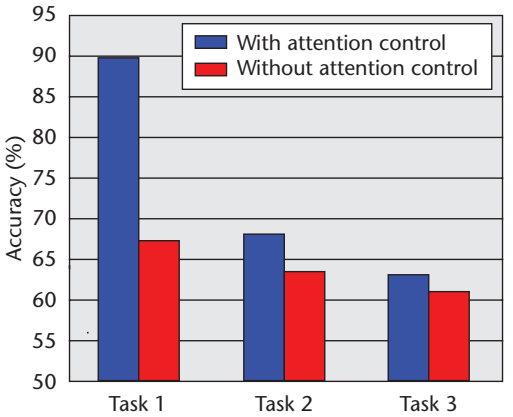


Figure 11. The accuracy of recognition with and without attention control for unseen testing data.

We also compared the average computation time for training the support vector machines with 1,000 training samples: the computation time is 5 seconds with attention control and 961 seconds without it. In the absence of attention control, 200 times more computation is required, which isn't surprising since introducing attention control significantly reduces the dimensionality of the sensor space. Thus, attention control based on information gain is effective for both improving accuracy and decreasing the computation time.

Reliability metric improves overall performance

In another experiment, we evaluated the effect of introducing a reliability metric, which prohibits the system from producing output when the reliability value is lower than a threshold. We investigated the recognition accuracy and the system's response rate while varying the threshold from 0 to 1, where 0 corresponds to the location of the discriminant hyperplane and 1 corresponds to the outer hyperplane.

Figure 12 shows the accuracy of recognition with respect to the threshold. For most tasks, accuracy peaks at a reliability value of 0.8. Figure 13 shows the response rate of the system with respect to the threshold. The response rate for most conditions declines markedly for a threshold above 0.75. From these results, we can see that the accuracy rate and the response rate are well balanced when the threshold is set to about 0.7.

The fact that announcements are suppressed about 15 percent of the time at this value didn't impair the performance of the announcement system; since announcements are made sequentially, there's time to obtain more reliable classifications while other announcements are being made. So, by implementing a reliability metric, the system's overall performance could be increased from between 5 to 83 percent.

Evaluating the announcement system

We would like to briefly mention another experiment using the JR Kyoto subway station data that was intended only to evaluate the effectiveness of different methods of applying the announcement system to practical use. Because Digital City Surveillance can detect up to nine events or their absence, it can at any given moment select among as many as 9 of the 18 possible announcements Table 1 lists for a total of 512 combinations. How should these announcements

be ordered when reporting them to the human operators? To find out, we constructed two announcement systems. The first simply selects an event on which to report at random and then makes the announcement; the second orders the reporting of events from those that describe global conditions to more specific statements.

To evaluate the announcement system, we presented four participants with four video images. The system generated an announcement for one of them, either by randomly selecting which event to report or by following the general-to-specific ordering. We asked each participant to select which video corresponds to the sequence of announcements. If the participant could make a rapid and accurate selection, we deemed the announcement to be effective. Figure 14 shows the precision of the participants' selections, and Figure 15 shows the reaction time. The precision is almost the same for both orderings, but the reaction time is an average of 9 seconds faster for the general-to-specific ordering. This demonstrates that the ordering rules increase the announcements' usefulness.

Discussion and conclusion

In this article, we proposed a memory-based classification method for recognizing human activities in a complex, real-world environment. This method is flexible because it doesn't require sensor calibration, and although it's trained by a human operator, the system designer isn't dependent on prior knowledge about the environment to be monitored.

Our method can be applied to large-scale sensor networks. As a first step, we attached multiple vision sensors in a large station where people perform daily activities and confirmed the usefulness of the proposed method. As a result, we've shown that the system can recognize human activities robustly with average success rates exceeding 80 percent for testing data. The average computation time for training the support vector machines with attention control is only about 0.5 percent of the average computation time for learning without attention control—a tremendous efficiency gain.

The ADVISOR system[4,25] is most like Digital City Surveillance in terms of target domain: the visual surveillance of subway stations. However, it's difficult to compare the two systems, because they're intrinsically different. ADVISOR'S approach is model-based and purpose-built for specific recognition tasks. In contrast, we could have used our system to monitor such nonhuman targets as animals at a zoo or cars in a parking lot without redesigning the algorithms.

In addition, the ADVISOR system is designed to recognize events that occur infrequently but cre-

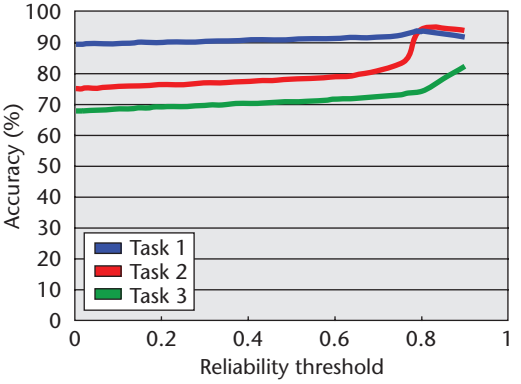


Figure 12. The accuracy of the recognition system increases with the reliability threshold until it peaks at about 0.8 for most tasks.

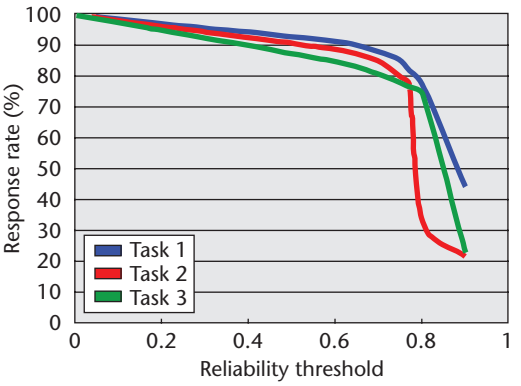


Figure 13. The percentage of time the recognition system is allowed to make a response drastically decreases as the reliability threshold increases beyond 0.75.

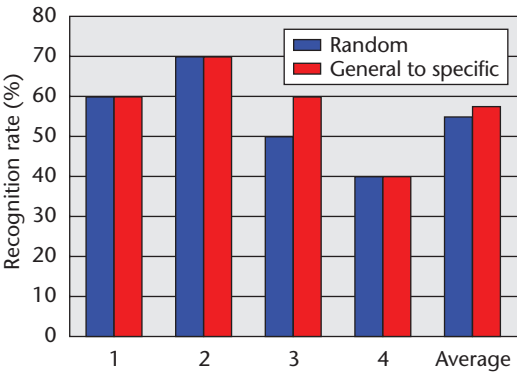


Figure 14. In this experiment, the four participants' accuracy in recognizing human activity in the subway concourse only improves slightly, when rules are used to order announcements from the more general to the more specific.

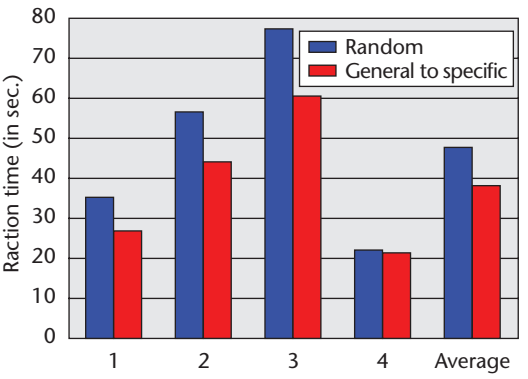


Figure 15. The general-to-specific ordering of announcements results in significantly faster reaction times for the four participants who were attempting to recognize the activity at the subway concourse.

The ADVISOR system is designed to recognize events that occur infrequently but create potential liabilities for the subway station.

ate potential liabilities for the subway station. That system's recognition rates were as follows: fighting, 95 percent; blocking, 78 percent; jumping over the barrier, 88 percent; vandalism, 100 percent; and overcrowding, 80 percent.[26] The average recognition rate was 88 percent. In addition, false positives occurred less than 1 percent of the time.

By contrast, the recognition rate of our system was about 5 percent lower for testing data. False positives and false negatives were equally prevalent at about 13 percent for testing data. Obtaining necessary training data to recognize the same events as the ADVISOR system would be difficult owing to the rarity of these events at the JR Kyoto subway station. However, our system can detect and inform the operator that an unusual event is occurring, even though the event isn't included in the training data, based on the distance of an input vector to an SVM from any prior data.

Note that human factors are responsible for some of the errors made by our system. Training conditions such as "it is crowded" weren't explicitly defined—for example, by using an objective measure such as a head count. They were instead based solely on the operator's subjective opinion. Different operators might not have agreed on how to classify a given composite image, and the same operator might have given different classifications at different times. Frequently, ambiguous conditions occur for which there are no clear, correct answers, and the operator must depend on intuition. However, if the system can reproduce subjective value judgments, that in itself is of value. **MM**

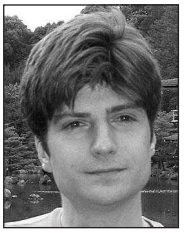
Acknowledgments

This research was supported by the JST CREST Digital City Project. Much gratitude is due to Toru Ishida, the project leader. We would also like to thank Junpeng Gong, Tetsushi Ikeda, and Takashi Minato for their assistance in preparing this article.

References

1. I. Haritaoglu, D. Harwood, and L. S. Davis, "W4: Real-Time Surveillance of People and Their Actions," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, 2000, pp. 809-830.
2. J. Ohya, A. Utsumi, and J. Yamato, *Analyzing Video Sequences of Multiple Humans,* Kluwer Academic Publishers, 2002.
3. R.T. Collins et al., "A System for Video Surveillance and Monitoring: VSAM Final Report," tech. report CMU-RI-TR-00-12, Robotics Inst., Carnegie Mellon Univ., May 2000; http://www.cs.cmu.edu/~vsam.
4. G. Medioni et al., "Event Detection and Analysis from Video Streams," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 23, no. 8, 2001, pp. 873-888.
5. J. O'Rourke and N. Badler, "Model-Based Image Analysis of Human Motion Using Constraint Propagation," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 2, 1980, pp. 522-536.
6. D. Hogg, "Model-Based Vision: A Program to See a Walking Person," *Image and Vision Computing,* vol. 1, 1983, pp. 5-20.
7. C.R. Wren, A. Azarbayejani, T. Darrell, and A.P. Pentland, "Pfinder: Real-Time Tracking of the Human Body," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, 1997, pp. 780-785.
8. J.H. Yin, S.A. Velastin, and A.C. Davies, "Measurement of Crowd Density Using Image Processing," *Proc. 7th European Signal Processing Conf.,* vol. 3, Elsevier, 1994, pp. 1397-1400.
9. T. Ishida, "Digital City Kyoto: Social Information Infrastructure for Everyday Life," *Comm. ACM,* vol. 45, no. 7, 2002, pp. 76-81.
10. H. Nakanishi et al., "Transcendent Communication: Location-Based Guidance for Large-Scale Public Spaces," *Proc. SIGCHI Conf. Human Factors in Computing Systems,* ACM Press, 2004, pp. 655-662.
11. K.F. MacDorman, "Feature Learning, Multiresolution Analysis, and Symbol Grounding: A Peer Commentary on Schyns, Goldstone, and Thibaut's 'The Development of Features in Object Concepts,'" *Behavioral and Brain Sciences,* vol. 21, no. 1, 1998, pp. 32-33.
12. K.F. MacDorman, "Grounding Symbols through Sensorimotor Integration," *J. Robotics Soc. Japan,* vol. 17, no. 1, 1999, pp. 20-24.
13. L.-E. Janlert, "The Frame Problem: Freedom or Stability? With Pictures We Can Have Both," *The Robot's Dilemma Revisited,* K.M. Ford and Z.W. Pylyshyn, eds., Ablex, 1996.

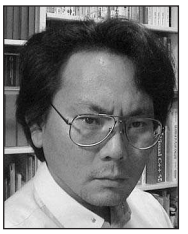
14. C. Stanfill and D. Waltz, "Toward Memory-Based Reasoning," *Comm. ACM,* vol. 29, no. 12, 1986, pp. 1213-1228.
15. H. Ishiguro, R. Sato, and T. Ishida, "Robot Oriented State Space Construction," *Proc. IEEE/Robotics Society of Japan Int'l Conf. Intelligent Robots and Systems,* IEEE Press, 1996.
16. K. Kato, H. Ishiguro, and M. Barth, "Identifying and Localizing Robots in a Multi-Robot System," *Proc. Int'l Conf. Intelligent Robots and Systems,* IEEE Press, 1999, pp. 966-972.
17. K.F. MacDorman and H. Ishiguro, "The Uncanny Advantage of Using Androids in Social and Cognitive Science Research," *Interaction Studies*, vol. 7, 2006, pp. 297-337.
18. H. Ishiguro, M. Kamiharako, and T. Ishida, "State Space Construction by Attention Control," *Proc. 16th Int'l Joint Conf. Artificial Intelligence* (IJCAI), vol. 2, Morgan Kaufmann, 1999, pp. 1131-1139.
19. K. MacDorman et al., "A Memory-Based Distributed Vision System that Employs a Form of Attention to Recognize Group Activity at a Subway Station," *Proc. IEEE/Robotics Society of Japan Int'l Conf. Intelligent Robots and Systems,* IEEE Press, 2004, pp. FA1-J3.
20. L. Itti, C. Koch, and E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 20, no. 11, 1998, pp. 1254-1259.
21. H. Ishiguro, "Distributed Vision System: A Perceptual Information Infrastructure for Robot Navigation," *Proc. Int'l Joint Conf. Artificial Intelligence* (IJCAI), IEEE Press, 1997, pp. 36-41.
22. H. Ishiguro and T. Nishimura, "VAMBAM: View and Motion-Based Aspect Models for Distributed Omnidirectional Vision Systems," *Proc. Int'l Joint Conf. Artificial Intelligence* (IJCAI), IEEE Press, 2001, pp. 1375-1380.
23. A. Bobick and J. Davis, "The Representation and Recognition of Action Using Temporal Templates," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 23, no. 3, 2001, pp. 257-267.
24. Y. Yang and J.O. Pedersen, "A Comparative Study on Feature Selection in Text Categorization," *Proc. 14th Int'l Conf. Machine Learning* (ICML), Morgan Kaufmann, 1997, pp. 412-420.
25. N.T. Siebel and S. Maybank, "Fusion of Multiple Tracking Algorithms for Robust People Tracking," *Proc. 7th European Conf. Computer Vision* (ECCV), vol. 4, LNCS 2353, Springer Verlag, 2002, pp. 373-387.
26. N.T. Siebel and S. Maybank, "The Advisor Visual Suirveillance System," *Proc. European Conf. Computer Vision (ECCV) Workshop Applications of Computer Vision,* Czech Technical Univ., 2004, pp. 103-111.



Karl F. MacDorman is an associate professor of informatics at Indiana University. His research interests include robotics, machine learning, and cognitive science. MacDorman received a BA in computer science from the University of California, Berkeley, and a PhD in machine learning and robotics from Cambridge University. He is a senior member of the IEEE.

Hiroshi Nobuta participated in the development of Digital City Surveillance as part of his master's research at Wakayama University. His research interests include sensor networks for behavior recognition. Nobuta received a BEng and an MEng in systems engineering at Wakayama University, Japan.

Satoshi Koizumi is an associate professor at Osaka University. His research interests include computer vision and virtual reality. Koizumi received a BEng and MEng in mechanical engineering from Aoyama Gakuin University, Japan, and a DEng in computational intelligence and systems science from the Tokyo Institute of Technology.



Hiroshi Ishiguro is a professor at Osaka University. He's also a visiting project leader at ATR, and a researcher of the Japan Science and Technology Agency Core Research for Evolutional Science and Technology (JST CREST). His research interests include distributed vision systems, robotics, and android science. Ishiguro received a BEng and MEng in computer science from Yamanashi University, Japan, and a DEng in systems engineering from the Osaka University. He is a member of the IEEE.

Readers may contact Karl MacDorman at the Indiana University School of Informatics, IT 487, 535 West Michigan St., Indianapolis, IN 46202; http://www.macdorman.com.