International Journal of Pattern Recognition
and Artificial Intelligence
Vol. 17, No. 3 (2003) 405–416

World Scientific
www.worldscientific.com

REAL-TIME PEDESTRIAN DETECTION USING SUPPORT VECTOR MACHINES*

SEONGHOON KANG
Department of Computer Science and Engineering, Korea University, Anam-dong, Seongbuk-ku, Seoul 136-701, Korea
shkang@image.korea.ac.kr

HYERAN BYUN
Department of Computer Science, Yonsei University, Shinchon-dong, Seodaemun-gu, Seoul 120-749, Korea
hrbyun@cs.yonsei.ac.kr

SEONG-WHAN LEE†
Department of Computer Science and Engineering, Korea University, Anam-dong, Seongbuk-ku, Seoul 136-701, Korea
swlee@image.korea.ac.kr

In this paper, we present a real-time pedestrian detection method in outdoor environments. It is necessary for pedestrian detection to implement obstacle and face detection which are major parts of a walking guidance system for the visually impaired. It detects foreground objects on the ground, discriminates pedestrians from other noninterest objects, and extracts candidate regions for face detection and recognition. For effective real-time pedestrian detection, we have developed a method using stereo-based segmentation and the SVM (Support Vector Machines), which works well particularly in binary classification problem (e.g. object detection). We used vertical edge features extracted from arms, legs and torso. In our experiments, test results on a large number of outdoor scenes demonstrated the effectiveness of the proposed pedestrian detection method.

Keywords: Pedestrian detection; support vector machines; stereo vision.

1. Introduction

Object detection is essential for a driver assistance system and a walking guidance system for the visually impaired. Especially, the pedestrian is much more important than any other objects. In case of the driver assistance system, the pedestrian is an obstacle, which the driver should avoid. But, in case of the walking guidance

*This research was supported by Creative Research Initiatives of the Ministry of Science and Technology, Korea.
†Author for correspondence.

system for the visually impaired, the pedestrian is a meaningful object to interact. So, pedestrian detection should be performed in real-time before the operator encounters other pedestrians. After a pedestrian is detected, we can extract and recognize a face effectively by reducing the candidate region to search faces.

However, it is very difficult to detect pedestrians in varying outdoor scenes. In general, most previously developed methods for real-time implementation have used motion detection based on image differencing and background subtraction. However, their use is restricted to areas off-limit to people and their performance is often troubled by false alarms due to external environmental effects (e.g. wind blowing, lighting changes, animals wandering around). Therefore, some systems incorporate specific knowledge about human shape and appearance to decrease false alarm. But, previously developed systems have not satisfied both real-time operation and high accuracy due to computational complexity of classification algorithm, which uses knowledge about human shape and appearance (e.g. 2D contours and silhouettes).

In this research, in order to overcome these problems, we applied two techniques for improved detection performance and real-time implementation in an outdoor environment. The proposed method proceeds in two steps: the first is to separate foreground objects from the background. The second is to distinguish pedestrians from other objects. The first step concerns object detection and generate a set of candidates. The second is the step of pedestrian recognition, a binary decision into pedestrian and non-pedestrian. We used stereo-based segmentation for the object detection, and the SVM technique for pedestrian recognition.

This proposed method is the main part of an outdoor walking guidance system for the visually impaired, OpenEyes-II which has been developed in the Center for Artificial Vision Research at Korea University.[6] OpenEyes-II enables the visually impaired to respond naturally to various situations that can happen in unrestricted natural outdoor environments while walking and finally reaching their destination. To achieve this goal, foreground objects (pedestrian, obstacles, etc.) are detected in real-time by using foreground–background segmentation based on stereo vision. Then, each object is classified as pedestrian or obstacles by a SVM classifier. These two main elements make the pedestrian detection system robust and in real-time.

2. Related Works

Pedestrian detection is an integrated task of locating and estimating the movement of pedestrians from image sequences. Most human detection and tracking systems employ simple segmentation procedures such as background subtraction or temporal differencing to detect human. However, those existing systems are characterized by various properties such as input camera types and detailed techniques for detection of body parts. Haritaoglu *et al.* introduced the W4 system[2] and the Hydra system[3] for detecting and tracking multiple people or the parts of their bodies. While W4 is an integrated tracking system that uses a monocular, monochrome,

and static camera, Hydra itself is a component, which allows W4 to analyze people moving in a group. The detection and tracking methods used in Hydra make a detailed segmentation of a group of people into individual persons via head detection and distance transformation. The Pfinder system[12] used a multiclass statistical model of a person and the background for person tracking and gesture recognition. This model utilizes stochastic, region-based features, such as blob and 2D contour. Although it performs novel model-based tracking, it is unable to track multiple people simultaneously. Darrell *et al.*[1] used disparity and color information for individual person tracking and segmentation. Their system used stereo cameras, and computes the range from the camera to each person with the disparity. The depth estimate allows the elimination of the background noise, and the disparity is fairly insensitive to illumination changes. Mohan *et al.*[8] presented a more robust human detection system based on the SVM technique. One problem with the system is that it has to search the whole image at multiple-scales to detect many components of human. This would be extremely computationally expensive, and may produce multiple responses from a single detection. To increase reliability, some systems tried to integrate multiple cues such as stereo, skin color, face and shape to detect humans.[1,2] These systems proved that stereo and shape were more reliable and helpful cues than color and face detection.

3. Properties of Support Vector Machines

Basically, a SVM is a linear machine with some very attractive properties. Its main idea is to construct a separating hyperplane between two classes, the positive and negative examples, in such a way that the distance from each of the two classes to the hyperplane is maximized. In other words, a SVM is an approximate implementation of the structural risk minimization method. This induction principle is based on the fact that the actual error rate of a learning machine is upper bounded by the sum of the training error rates. From this property, the SVM can provide a good generalization performance on pattern classification problems without domain knowledge. Figure 1 shows a constructed optimal hyperplane between two classes in a linear space.[4]

If we assume that the input space can be separated linearly for simplicity, a constructed hyperplane is represented by Eq. (1) in a high dimensional space.

$$w \cdot x + b = 0\,, \tag{1}$$

where x is a training pattern vector and w is a normal vector of the hyperplane. Some constraints represented by Eq. (2) must be satisfied to ensure that all training patterns are correctly classified.

$$y_i(w \cdot x_i + b) - 1 \geq 0\,, \quad \forall i\,. \tag{2}$$

Among the hyperplanes constructed under the constraints above, the one from which the distance to its closest point is maximal is called an optimal separating hyperplane (OSH). From the given training data, we would like to find the parameters

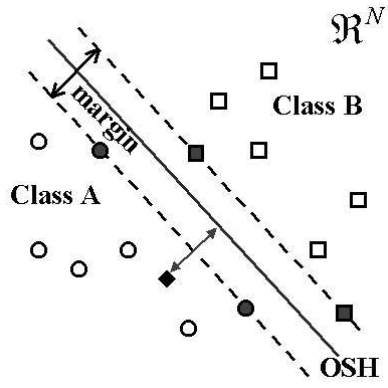

Fig. 1. A constructed hyperplane between two classes in a linear space.

(w_o, b_o) for the OSH, where w_o and b_o are optimal values for w and b, respectively, in Eq. (1). Since the distance to the closest point, the *margin*, is $2/||w||$, w_o can be obtained by minimizing the value $||w||$. This is an optimization problem, and w_o is represented by Eq. (3) with non-negative Lagrange multipliers (α_i)s.

$$w_o = \sum_i^N \alpha_i y_i x_i \,. \tag{3}$$

Using Eqs. (1) and (3), b_o can be obtained.

In classification tasks, the class of an unknown input pattern (x) is determined by the sign of the signed distance of the pattern from the constructed hyperplane. The signed distance is calculated by Eq. (4).

$$\text{sgn}\left(\frac{\sum_i^N \alpha_i y_i x_i \cdot x + b_o}{||w||}\right) . \tag{4}$$

This method can be also applied to a nonlinear input space through a kernel function. Kernel functions map a nonlinear input space to a linear feature space. So, the inner product in Eq. (4) can be replaced by a kernel function. Consequently, a distance measurement can be derived no matter whether the input space is linear or not. However, we are still left with an open question of which kernel function is best to be used for a given problem.

4. Pedestrian Detection System

4.1. *System organization*

The pedestrian detection system developed in this research consists of two parts: a trainer and a detector as shown in Fig. 2. The pedestrian detector model is the core of the entire system. It is the SVM trained with large number of samples. Once created, the model can be used to detect any pedestrians in the natural scenes. Since, however, the natural scenes are highly dynamic and variable, most of

the well-known pattern matching and model based methods cannot be applied to varying natural scenes. So, we used the SVM algorithm[11] in the main engine of the classification. SVM is a well-known tool that can capture the statistical properties of various images by training. It can classify or recognize effectively patterns in a huge data space by using small number of training data. Because pedestrians appear in so many different colors and textures, it is difficult to use color or texture features for detection. Instead we chose a more reliable consistent feature, vertical edges, that can be extracted from arms, legs, and torso.

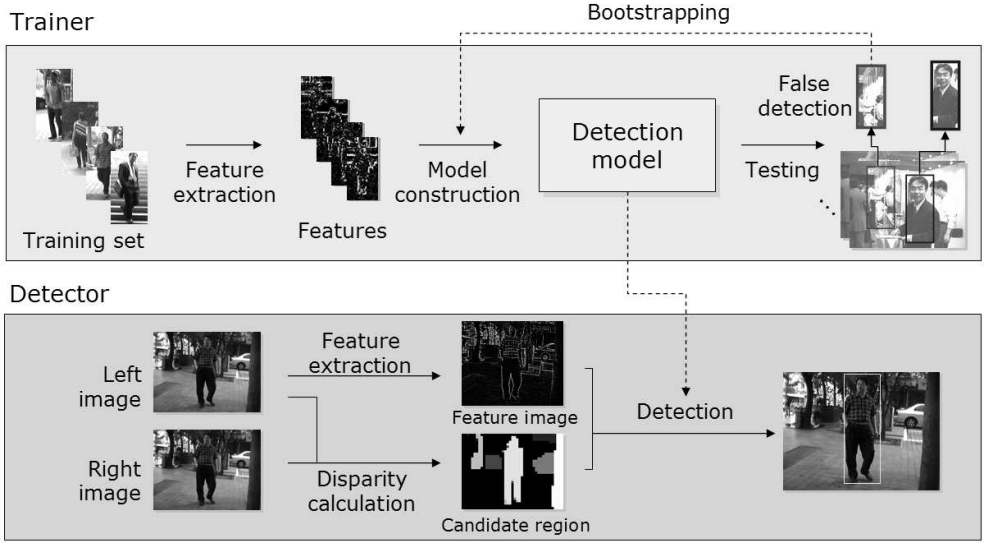

Fig. 2. Block diagram of pedestrian detection system.

4.2. *Training of detection model*

The pedestrian model describes the features of a person moving 4–5 m ahead of the camera. It has been trained using a set of 32×64 images,[a] a mixture of a positive set containing pedestrian and a random negative set containing no pedestrian as shown in Fig. 3.

For further elaboration of the models, we apply "bootstrapping" technique.[10] The combined set of positive and negative samples forms the initial training database for the detector model. At first, the detector model was trained by a 100 data set. After the initial training, we run the system over many images captured at various backgrounds. Any detections clearly identified as false positives are added to the negative training set. The repeated bootstraping training allows

[a]It was calibrated in half scale of input image (160×120 size) for fast and accurate detection.

(a) (b)

Fig. 3. Examples of training data set (a) positive samples and (b) negative samples.

Fig. 4. Bootstrapping training.

the system to construct an incremental refinement of the nonpedestrian class until satisfactory performance is achieved.

4.3. *Pedestrian detection*

For the pedestrian detection in outdoor video surveillance, real-time processing is the most important issue for practical application. The SVM module, the main part of this system, is computationally too complex to operate in real-time. In fact, it is very difficult to satisfy both detection accuracy and speed, and we have to compromise the two requirements.

In the proposed solution, we used the stereo vision technique which is common in robot vision area. Stereo vision can provide range information for object segmentation. Using stereo vision to guide pedestrian detection provides some distinct advantages over conventional techniques. First, it allows explicit occlusion analysis and is robust to illumination changes. Second, the real size of an object derived from the disparity map provides a more accurate classification metric than the image size of the object. Third, using stereo cameras can detect both stationary and moving objects. Fourth, computation time is significantly reduced by performing classification on the region only where objects are detected; it is less likely to detect the background area as pedestrian since detection is biased toward areas where objects are detected.[13]

We have employed a video-rate stereo system[7] to provide range information for object detection. This system uses area correlation method for generating disparity

image as shown in Fig. 5. By means of feature for correlation, LOG transform was chosen, because it gives high quality results. Figure 5 shows a typical disparity image. Brighter white regions in disparity result indicate higher disparities (closer objects).

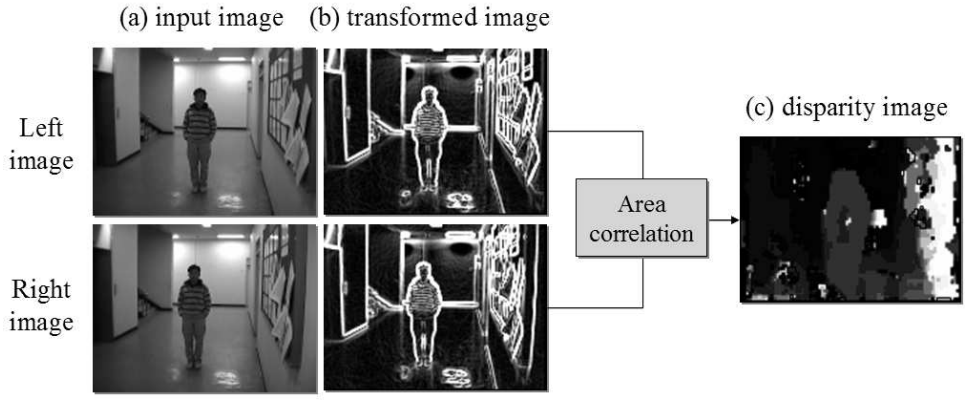

Fig. 5. Examples of disparity image based on area correlation.

Disparity image provides approximated range information. So, we can separate the object region from background with an appropriate filtering. The disparity images are processed in three levels (near distance, middle distance, far distance). We have been interested in only middle distance, because most objects to be detected are in the middle distance. The areas in the far distance are ignored. In this way, the disparity image is binarized given two thresholds corresponding to the middle distance, in order to extract the candidate regions of the objects as shown in Fig. 6.

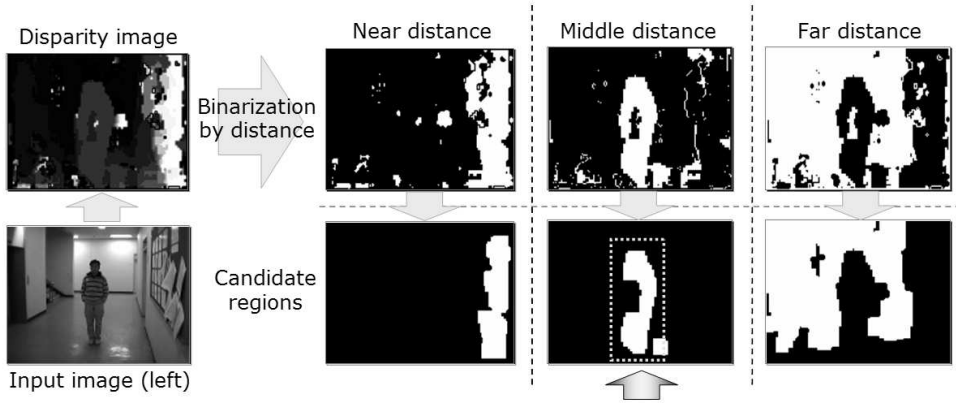

Fig. 6. Candidate regions of detected objects by distance.

The proceeding step allows great reduction in search time. For example, the SVM classification must be performed 165 times without reducing the candidate regions of objects using stereo vision, but only two times of the SVM classification are needed with the reduced candidate regions as shown in Fig. 7. When we ignore the computation time for disparity image is negligible, the proposed detection method with stereo vision is about 80 times faster than the detection method without stereo vision.

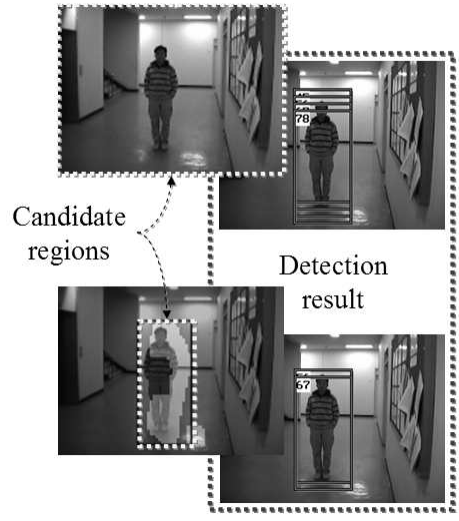

Fig. 7. Detection with reducing candidate regions versus without one.

5. Experimental Result and Analysis

The pedestrian detector model is trained finally as followings. The size of training data set is 528 (140 positive data, 378 negative data are used). Third order polynomial has been used as the kernel function in the SVM. The trained SVM model has 252 support vectors.

An experimental system has been implemented on a Pentium III 800 MHz system under Windows XP with a MEGA-D Megapixel Digital Stereo Head[5] as shown in Fig. 8. It has been designed as a small stand-alone system so that we can carry it easily.

It has been tested extensively on large amounts of outdoor natural scenes including pedestrian. Over 900 instances of pedestrian and other objects have been included in these scenes. The system can detect and classify objects over a 320×240 pixel stereo image at a frame rate ranging from 5 frames/second to 10 frames/second, depending on the number of objects presented in the image.

The performance of any detection system has a tradeoff between the positive detection rate and false detection rate. To capture this tradeoff, we vary the sensitivity

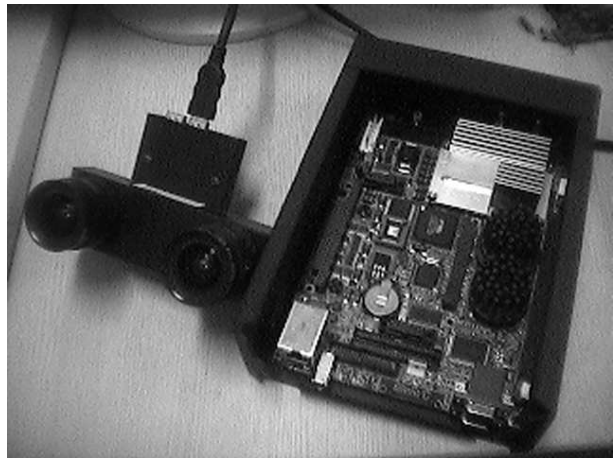

Fig. 8. Stand-alone system with stereo camera for experiments.

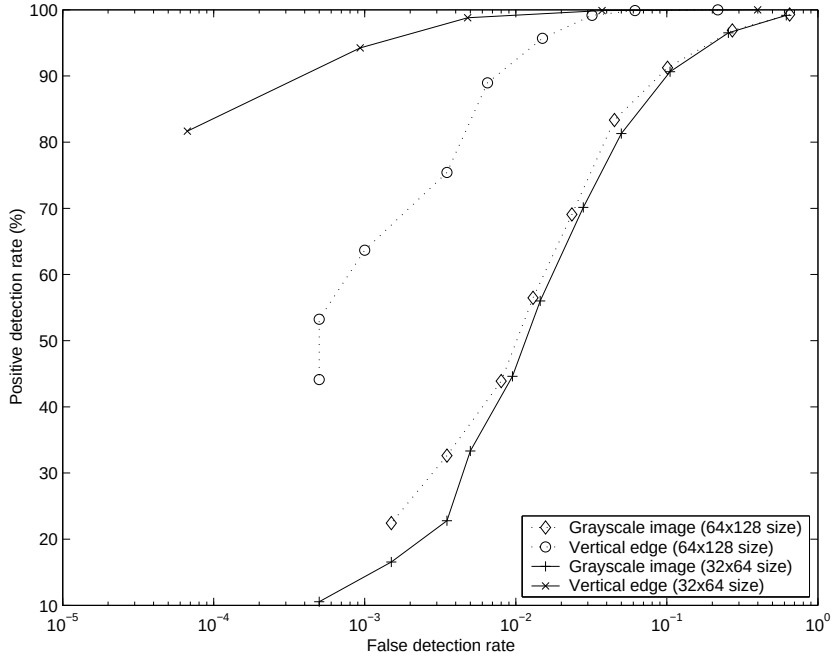

Fig. 9. ROC curves of pedestrian detector using SVM.

of the system by thresholding the output and evaluate the ROC (Receiver Operating Characteristic) curve.[9] As shown in Fig. 9, ROC curves comparing different representations (grayscale images and vertical edge images are used for features) for pedestrian detection. The detection rate is plotted against the false detection rate,

measured on a logarithmic scale. The trained detection system has been run over test images containing 834 images of pedestrian to obtain the positive detection rate. The false detection rate has been obtained by running the system over 2000 images, which do not contain pedestrian. The experiments involve four features: grayscale and vertical edge images of dimension 64×128 and 32×64. As shown in ROC curves, 32×64 size of vertical edge image is superior to the other features. It is faster and more accurate.

Figure 10 shows the results of our pedestrian detection system on some typical urban street scenes. This figure shows that our system can detect pedestrian in different sizes, pose, gait, clothing and occlusion status. However, there are some cases of failure. Most failures have occured when a pedestrian is almost similar in color to the background, or two pedestrian are too close to be separated.

Fig. 10. Detection results.

6. Conclusion and Future Works

This system is part of the outdoor walking guidance system for the visually impaired, OpenEyes-II that aims to enable the visually impaired to respond naturally to various situations that can happen in unrestricted natural outdoor environments while walking and finally reaching the destination. To achieve this goal, the detection method is required that operates in real-time as well as accurately. Previous human detection methods have satisfied these requirements. Some real-time methods have low classification (or recognition) accuracy or no classification functionality. On the other hand, good classifiers sacrificed the real-time processing due to computational complexity of classification (or recognition) algorithm. Moreover, detection methods, which use motion analysis and background subtraction, often fail due to external environmental effects (e.g. wind blowing, lighting changes).

In this paper, in order to overcome these problems, we used foreground–background segmentation based on stereo vision. Then, each detected object is classified as a pedestrian or an obstacle by the SVM classifier. These two main elements make the object detection and classification method robust and real-time.

However, the proposed method has a problem that it becomes slower in proportion to the number of objects to classify in the field of view, due to the complexity of the SVM algorithm. So, as future work, it is necessary to make the SVM classification faster through research in feature vector reduction and hardware implementation. Also, multiobject discrimination and detection properties should be included for other good applications in real life.

References

1. T. Darrell, G. Gordon, M. Harville and J. Woodfill, "Integrated person tracking using stereo, color, and pattern detection," *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, Santa Barbara, CA, 1998, pp. 601–608.
2. I. Haritaoglu, D. Harwood and L. S. Davis, "W4: Who? When? Where? What? A real time system for detecting and tracking people," *Proc. Int. Conf. Face and Gesture Recognition*, Nara, Japan, April 1998, pp. 222–227.
3. I. Haritaoglu, D. Harwood and L. S. Davis, "Hydra: multiple people detection and tracking using silhouettes," *Proc. 2nd IEEE Workshop on Visual Surveillance*, Fort Collins, Colorado, June 1999, pp. 6–13.
4. S. Haykin, *Neural Networks*, Prentice Hall, NJ, 1998.
5. http://www.videredesign.com
6. S. Kang and S.-W. Lee, "Hand-held computer vision system for the visually impaired," *Proc. 3rd Int. Workshop on Human-Friendly Welfare Robotic*, Daejeon, Korea, January 2002, pp. 43–48.
7. K. Konolige, "Small vision systems: hardware and implementation," *Proc. 8th Int. Symp. Robotics Research*, Hayama, October 1997.
8. A. Mohan, C. Papageorgiou and T. Poggio, "Example-based object detection in images by components," *IEEE Trans. Patt. Anal. Mach. Intell.* **23**, 4 (2001) 349–361.
9. M. Oren, C. Papageorgiou, P. Sinha, E. Osuna and T. Poggio, "Pedestrian detection using wavelet templates," *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, San Juan, Puerto Rico, 1997, pp. 193–199.
10. K.-K. Sung and T. Poggio, "Example-based learning for view-based human face detection," A.I. Memo 1521, AI Laboratory, MIT, 1994.
11. V. Vapnik, *Statistical Learning Theory* (Wiley, NY, 1998).
12. C. Wren, A. Azarbayejani, T. Darrell and A. Pentland, "Pfinder: real-time tracking of the human body," *IEEE Trans. Patt. Anal. Mach. Intell.* **19**, 7 (1997) 780–785.
13. L. Zhao and C. Thorpe, "Stereo- and neural-based human detection," *Proc. Int. IEEE Conf. Intelligent Transprtation Systems*, Tokyo, Japan, October 1999, pp. 5–10.

Seonghoon Kang received his B.S. and M.S. degrees in control and instrumentation engineering from Korea University, Seoul, in 1995 and 1997, respectively; and Ph.D. in computer science and engineering from Korea University in 2003.

His research interests include motion and object detection, active vision and mobile visual aids.

Hyeran Byun received the B.S. and M.S. degrees in mathematics from Yonsei University, Korea. She received her Ph.D. in computer science from Purdue University, West Lafayette, Indiana. She was an Assistant Professor in Hallym University, Chooncheon, Korea from 1994–1995. Since 1995, she has been an Associate Professor of computer science at Yonsei University, Korea.

Her research interests are multimedia, computer vision, image and video processing, artificial intelligence and pattern recognition.

Seong-Whan Lee received his B.S. degree in computer science and statistics from Seoul National University, Korea, in 1984; and M.S. and Ph.D. degrees in computer science from the Korea Advanced Institute of Science and Technology in 1986 and 1989, respectively. From February 1989 to February 1995, he was an Assistant Professor in the Department of Computer Science at Chungbuk National University, Cheongju, Korea. In March 1995, he joined the faculty of the Department of Computer Science and Engineering at Korea University, Seoul, and now he is a full professor. Prof. Lee is also the Director of National Creative Research Initiative Center for Artificial Vision Research (CAVR) supported by the Korean Ministry of Science and Technology and the visiting professor of the Artificial Intelligence Laboratory at MIT. Prof. Lee was the winner of the Annual Best Paper Award of the Korea Information Science Society in 1986. He obtained the Outstanding Young Researcher Paper Award at the 2nd International Conference on Document Analysis and Recognition in 1993, and the First Distinguished Research Professor Award from Chungbuk National University in 1994. He obtained the Outstanding Research Award from the Korea Information Science Society in 1996. He also received an Honorable Mention of the Annual Pattern Recognition Society Award for an outstanding contribution to the *Pattern Recognition Journal* in 1998. He is a fellow of International Association for Pattern Recognition, a senior member of the IEEE Computer Society and a life member of the Korea Information Science Society.

He has more than 200 publications on computer vision and pattern recognition in international journals and conference proceedings, and has authored 10 books.

His research interests include computer vision, pattern recognition, and neural networks.