A Two-Stage Approach to Pedestrian Detection with a Moving Camera

Miae Kim and Chang-Su Kim

School of Electrical Engineering, Korea University / Seoul, South Korea {miae_kim, changsukim}@korea.ac.kr

***** Corresponding Author: Chang-Su Kim

Received May 13, 2013; Revised May 28, 2013; Accepted June 12, 2013; Published August 31, 2013

** Regular Paper*

***Abstract*:** This paper presents a two-stage approach to detect pedestrians in video sequences taken from a moving vehicle. The first stage is a preprocessing step, in which potential pedestrians are hypothesized. During the preprocessing step, a difference image is constructed using a global motion estimation, vertical and horizontal edge maps are extracted, and the color difference between the road and pedestrians are determined to create candidate regions where pedestrians may be present. The candidate regions are refined further using the vertical edge symmetry features of the pedestrians' legs. In the next stage, each hypothesis is verified using the integral channel features and an AdaBoost classifier. In this stage, a decision is made as to whether or not each candidate region contains a pedestrian. The proposed algorithm was tested on a range of dataset images and showed good performance.

***Keywords*:** Pedestrian detection, Global motion estimation, Color difference

1. Introduction

Pedestrian detection is used in many applications, including driving assistance, robotics, entertainment, and surveillance. Automatic pedestrian detection, however, is a difficult task. The main problem is the significant variations in the visual appearance depending on their clothing, poses, face colors, etc. In addition, it is difficult to discriminate pedestrians from cluttered background or brightness differences over time, weather, occlusion, etc.

Many images for detection pedestrians are taken from fixed monitoring cameras on the streets. In this case, the regions in an image that correspond to pedestrians can be identified easily, which are moving objects, by taking the differences between the captured image and background image or the differences between the consecutive frames. On the other hand, the frame subtraction method is unsuitable for images that are captured from an in-vehicle camera. A moving object is difficult to detect because the background and pedestrians are moving.

The aim of this paper was to develop a system for pedestrian detection using an in-vehicle camera. To this end, a novel frame subtraction method specialized for a moving camera is proposed. Note that it is impossible to remove a moving background using conventional frame subtraction techniques. Therefore, the frame subtraction algorithm, which compensates for background motions using global motion estimation (GME), is suggested.

The outline of this paper is as follows. Section 2 reviews previous work on pedestrian detection. Section 3 presents a two-stage approach to pedestrian detection, as illustrated in Fig. 1. The pedestrian candidate regions are first searched from a difference image using a global motion estimation, edge and color maps, and motion information. In the detection step, the pedestrians are detected using an integral channel feature detector and an AdaBoost classifier. Section 4 assesses the performance of the proposed algorithm on the ETHZ pedestrian dataset [13], consisting of full-length videos with crowded scenes. The study is concluded in Section 5.

2. Related Work

Papageorgiou and Poggio [1] used the Haar wavelet features in combination with a support vector machine (SVM) to detect pedestrians but it is sensitive to intensity changes. Viola and Jones [11] introduced integral images

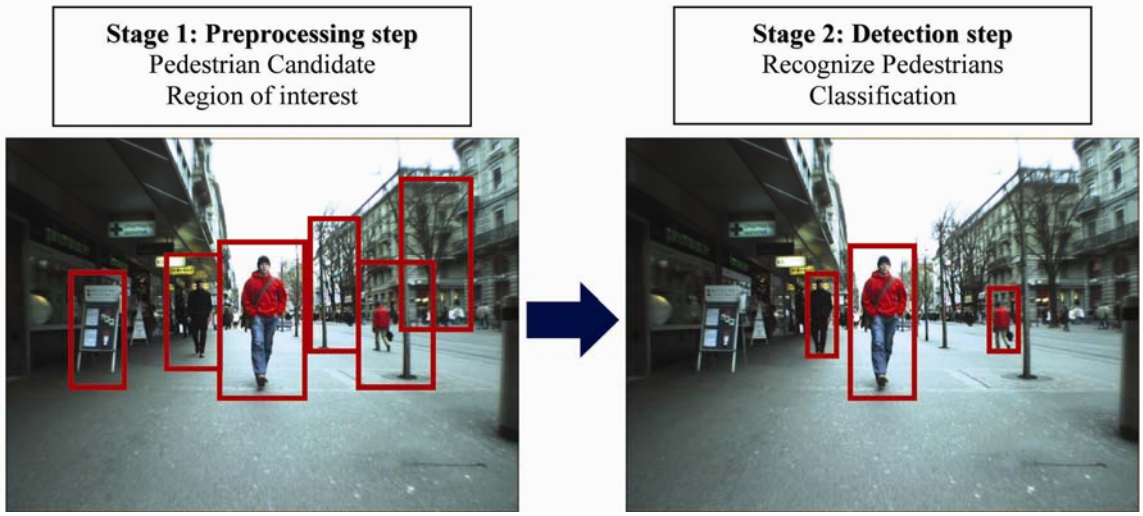


Fig. 1. Two-stage approach for pedestrian detection.

efficient detection, and utilized AdaBoost for fast feature computation and a cascade structure for automatic feature. Dalal and Triggs [2] introduced a histogram of oriented gradient (HOG) features, which represents the distribution of gradient vectors according to their orientations. HOG is effective for pedestrian detection because of its robustness against intensity and color changes. Nevertheless, its complexity is relatively high when evaluating all sub-windows for calculating the HOG descriptors. Felzenszwalb *et al.* [3] reduced the HOG feature vectors, and improved the performance and computation speed using flexible part models. This method enhances the detection performance by combining global person detection with partial component detection. Other motion-based detection methods are available. Viola and Jones [4] applied a motion filter to the integral images for moving pedestrian detection. Their motion features proved to be the most discriminative features. On the other hand, their work was restricted to a static camera setting. Dalal and Triggs [5] suggested a movement feature descriptor, which combined the HOG features with movement information. Gavrila [14] proposed capturing the pedestrians' shape variability using a number of templates. Nevertheless, it incurs considerable computational cost becauseit should consider the entire image including the background regions.

3. Proposed Algorithm

This paper proposes an algorithm for detecting pedestrians in video sequences, which is captured by a moving camera. When detecting pedestrians, it is important to detect separate moving objects from the background accurately. In the case of a fixed camera setting, it is easy to detect moving objects by employing a frame subtraction that identifies a difference between the current frame and the previous frame because the background is invariant. On the other hand, when images are taken from a moving camera on a vehicle, it is difficult to apply the frame subtraction due to the varying background information. Therefore, the variations in the background need to be compensated for by estimating the global motion in a video sequence.

The proposed pedestrian detection algorithm is as follows. The motion vectors are calculated from images taken from an in-vehicle camera using a block matching algorithm. From the motion vectors, the affine model parameters are determined to describe the global motion and compensate for the global motion to make a difference image. The difference image is then converted to an edge map, and pedestrian candidate regions are determined using the color difference information between the road and the pedestrians' feet. The edge map also exhibits strong vertical edges for the leg and/or body parts, and yields symmetrical structures. Based on these features, the pedestrian candidate regions are decided. Finally, the presence of a pedestrian is confirmed or the candidate is discarded using a classifier.

3.1 Difference Image Using GME

A difference image is often used to detect moving objects in video sequences. On the other hand, in video sequences captured from in-vehicle cameras, many edges are extracted from the background regions. These edges are the primary sources of false detection in detecting pedestrians. To reduce the number of false detections that occur in this type of background, the background region should be removed from the difference image using GME.

Global motion in a video sequence, which is caused by camera motion, is modeled as a parametric transform of two-dimensional images. The process of estimating the transform parameters is called GME. The purpose of a motion model is to describe real motion between consecutive frames, F_n and F_{n-1}, in a video sequence at time instances, n and n-1, respectively. The six-parameter affine model was adopted to model global motions. GME was conducted to estimate the six-parameters using the block-based motion vectors. Motion vectors, which are

obtained by a block matching algorithm, describe the displacement of a block in a current frame with respect to the best match in the reference frame.

The affine motion model is estimated using the block-based motion vectors. The affine motion model represents the overall background motion caused by a moving camera. Therefore, by estimating the affine motion model between the subsequent frames, the effects of the background motion can be compensated for and subjects with distinguished local motion can be searched more effectively.

The affine motion model has six parameters,

$$\begin{aligned} x' &= (m_0 + m_2 x + m_3 y) \\ y' &= (m_1 + m_4 x + m_5 y) \end{aligned} \tag{1}$$

where (x, y) denotes a position vector in the current frame, (x', y') is the corresponding position in the reference frame. In addition, $\boldsymbol{m}=[m_0,\ldots,m_5]$ is a vector, the elements of which are the six motion parameter. Let (MVx, MVy) be the motion vector of a block, and $(x'-x, y'-y)$ be the displaced vector. The errors in the horizontal and vertical directions then form an error vector $(MVx-x'+x, MVy-y'+y)$. The sum of squared errors over all blocks can be the calculated using the following equation:

$$\begin{aligned} E &= \sum (ex^2 + ey^2) \\ &= \sum ((MVx - x' + x)^2 + (MVy - y' + y)^2) \end{aligned} \tag{2}$$

The goal of the GME is to minimize the difference in the input motion vectors and estimated ones. In other words, an attempt is made to estimate the affine parameter vector $\boldsymbol{m}$ that minimizes the sum of the squared errors. To obtain the optimal $\boldsymbol{m}$, $\boldsymbol{m}$ is searched to minimize E repeatedly using the least squared method. The conventional difference image is expressed as $I_t(i, j) - I_{t-1}(i, j)$. On the other hand, the difference image after GME is represented by $I'_t(i, j) - I_{t-1}(i, j)$.

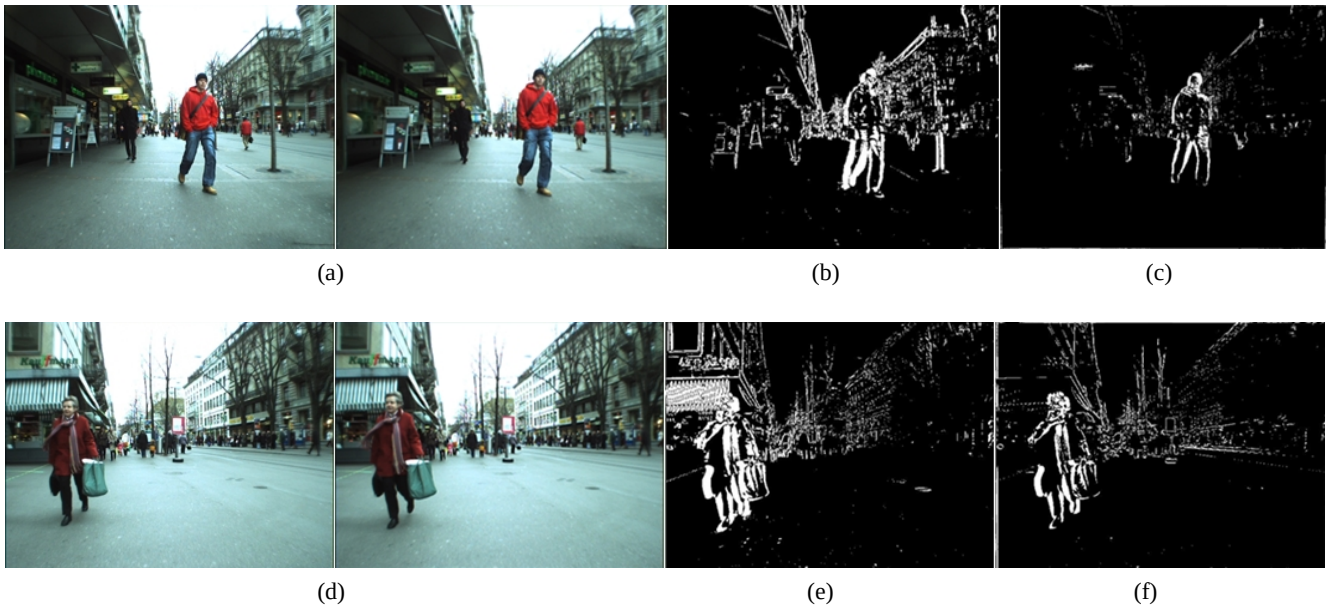
After GME, the effect of the background motions can be reduced. Specifically, in the case of an in-vehicle camera, if the difference image between two sequential images is obtained and the edges are calculated, it is almost impossible to identify the proper edges of the foreground objects due to the heavy differences in the background. On the other hand, after GME, the difference image is composed mostly of foreground object edges. Fig. 2 gives an example that compares the conventional frame subtraction method with the proposed method using GME. Note that that the conventional method detects the background edges and foreground edges, as shown in Figs. 2(b) and (e). On the other hand, the proposed method alleviates the cluttered background information and identifies the foreground edges more effectively, as shown in Figs. 2(c) and (f).

3.2 Edge and Color Differences

Pedestrians in standing positions have distinguishing characteristics. First, a pedestrian has strong vertical edges in the body and legs. Second, they have symmetric features in the edges of the legs. These characteristics are exploited to determine the locations of the pedestrian candidates.

First, the horizontal and vertical edges in the difference image are extracted using GME. Subsequently, weak edges are eliminated and relatively strong edges are maintained only. The remaining horizontal and vertical edges are connected by recovering the weak ones within a 10 pixel distance. Labeling of the connected edges is then performed. In addition, the horizontal and vertical edges whose length is too long or too short are eliminated. Too long or too short edges do not belong to pedestrians.

Next, the color image is converted from RGB space to HIS space. The average color values of the upper and the



(a) (b) (c)

(d) (e) (f)

Fig. 2. Comparison of difference images (a), (d) Original images (2 consecutive images); (b), (e) the conventional difference images; and (c), (f) the proposed difference images using GME.

(a) Edge Extraction (b) Removal of Useless Edges (c) Horizontal Edges in the Image

Fig. 3. Edge detection in the difference image. After removing useless edges, we find the leg parts of pedestrians using the color differences across horizontal edges.

lower regions of each horizontal edge in HIS color space are compared. The upper region corresponds to pedestrian's shoes or feet, and the lower region corresponds to sidewalk or roads. Therefore, there tends to be a stark difference between the colors of the upper and lower regions. If the color difference is larger than a threshold, we make a box around the horizontal edge and consider it as a candidate foot region. Within the box, the vertical edges are distributed symmetrically because of the legs. The method of detecting the symmetry of vertical edges can be expressed as

$$if \begin{cases} |x - x_i| = |x - x_j| \\ y_i = y_j \\ G_v(x_i, y_i) > T_v \\ G_v(x_j, y_j) > T_v \end{cases} \quad \text{then } S_k = S_k + 1,\ k = (i+j)/2 \qquad (3)$$

where (x_i, y_i) denotes the i^{th} edge point, x is the x-coordinate of a point within the ROI region, G_v is the vertical component of the gradient, and T_v is a threshold. In addition, S_k is the accumulator array corresponding to the k^{th} column. Note that the vertical symmetry is stronger if S_k is higher. Fig. 4 illustrates the vertical edge symmetry in the legs. The value in the accumulator array represents the strength of symmetry. The strong vertical symmetry indicates the pedestrians' legs.

Using symmetry, a pedestrian candidate region is detected. The size of the pedestrian is estimated based on that of the leg region. The leg region designates the width size of the pedestrian. A pedestrian normally has a 1:3 ratio for the width and height. Therefore, the height size is set to be 3 times larger than the width of the leg region.

3.3 Motion Information

For the color difference method, the motion information is used to determine the pedestrian candidates. Large motion regions are added as pedestrian candidates using the motion vectors and the difference images of an input sequence. The motion vectors are obtained using the block matching algorithm, which searches a similar area to each block window. A large motion vector means that the

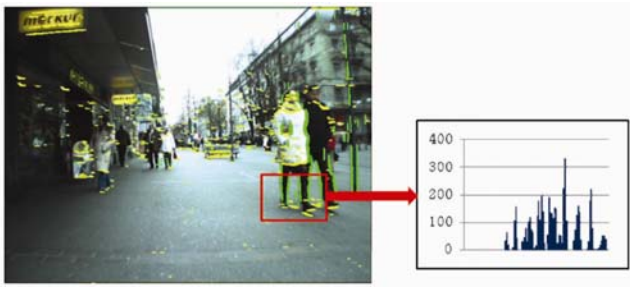


Fig. 4. Example of the symmetry histogram. The vertical symmetry is stronger if the histogram has a higher value.

block has great movement. In addition, a large motion part tends to have a large value in the difference image. Therefore, the common areas with both the large motion vectors and large image differences simultaneously are searched.

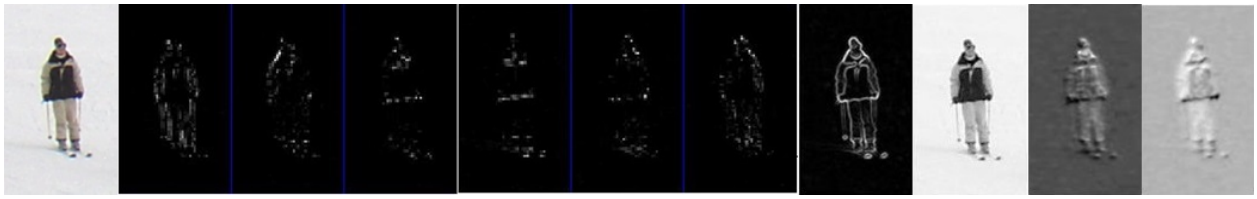
The integral image is used to find the large motion region in the sum image of the motion vector image and the difference image. First, the motion vector image and difference image are combined. The large motion area is then extracted by applying the detector to the integral image of the sum image. The detector uses a rectangle of the width-height ratio of 1:3, because a person typically causes the bounding box of that ratio. The size of the rectangle is varied by multiplying it with the factor, 0.5 ~ 1.3 ratio. The integral detector calculates the average of the sum within the detection window and thresholds the values. Finally, the false detection areas in the detected pedestrian candidate region are removed using the non-maximal suppression (NMS) and the difference in bilateral symmetry.

Although the color difference method is effective in detecting small pedestrian regions, it cannot detect large pedestrian areas well. In such cases, the method normally detects parts of the legs or bodies, instead of the whole persons. In the case of motion picture sequences captured from moving cars, the movements are larger in the subjects that are nearer to the cameras due to the perspective projection. On the other hand, the movement of an object far from the camera is relatively small. Therefore, by

(a)

(b) (c) (d)

Fig. 5. Pedestrian candidate selection using the motion information (a) original images, (b) difference image, (c) motion magnitude image, (d) combined image.



(a) (b) (c) (d)

Fig. 6. Integral channel features (a) original image, (b) gradient histogram (6 channels), (c) gradient magnitude (1 channel), (d) LUV (3 channels).

exploiting the motion information, large pedestrian areas, which are near the cameras, can be detected efficiently. Fig. 5 confirms this property. Fig. 5(b) shows the difference image using the global motion estimation. In Fig. 5(c), the gray level represents the magnitudes of the motion vectors. Fig. 5(d) shows the combined image, representing both motion magnitudes and frame differences, in which the candidate region is detected properly.

3.4 Detection

Dollar *et al.*'s ChnFtrs [7] is used as the baseline detector. The ChnFtrs detector employs the integral channel features. For pedestrian detection, 10 channels in total are used: 6 quantized orientations, 1 gradient magnitude, and 3 LUV color channels. These channels are then used as the input to the AdaBoost classifier. The INRIA pedestrian data and Caltech pedestrian dataset are used as training data. Note that the gradient histogram features represent the distribution of the local edge directions, which are suitable for expressing the object shapes.

The AdaBoost algorithm linearly combines the weak classifiers to construct a strong classifier. Each weak classifier selects one feature to classify the learning data, which is composed of pedestrian images and other images. After learning the weak classifiers, it updates the weighting values repeatedly. The low error weights are assigned to correctly classified data, whereas high error weights are given to incorrectly classified data. The weak learners used for the boosting form depth-2 decision trees, where each node is a simple decision stump, which is defined by a rectangular region, a channel, and a threshold. The final strong classifier is a weighted linear combination of the boosted depth-2 decision trees. Although the ChnFtrs detector searches all the windows, this system uses only the pedestrian candidate regions from the preprocessing step to discriminate efficiently whether or not each candidate is a pedestrian.

4. Experimental Results

This paper proposed an algorithm to search the pedestrian candidate regions using difference images, color differences, and motion information in the preprocessing step. The classifier was then used to validate if each candidate is a pedestrian.

The performance of the proposed algorithm was evaluated using the ETHZ dataset. For comparison, the following four conventional algorithms were also tested: 1) Felzenszwalb *et al.* [3], 2) Schwartz *et al.* [19], 3) Maji *et*

Fig. 7. Experimental Results of the proposed algorithm on the ETHZ dataset.

Table 1. Comparison of the detection rates on the ETHZ dataset.

Detection method	Detection Rate			
	BAHNHOF	JELMOLI	SUNNYDAY	CROSSING
Proposed algorithm	60.5%	52.8%	64.3%	46.0%
P.Felzenszwalb et al.'s method [3]	29.3%	42.3%	42.1%	48.1%
W.Schwartz et al.'s method [19]	60.0%	51.3%	59.4%	45.3%
S.Maji et al.'s method [20]	45.4%	52.1%	68.2%	46.1%
P.Dollar et al.'s method [7]	38.9%	43.3%	56.8%	33.7%

al. [20], and 4) Dollar *et al.* [7]. The ETHZ dataset presents challenging sequences, which were acquired from a stroller moving along a crowded side-walk. Four sequences, BAHNHOF (700 frames), JELMOLI (937 frames), SUNNYDAY (354 frames), and CROSSING (219 frames) were used in the tests. Because the proposed algorithm selects only the likely regions as candidates, its false detection rate is zero. On the other hand, for the other sequences, the false detection rates are not zero in general. Therefore, the parameters of the conventional algorithms were set to have zero false detection.

Fig. 7 shows that the proposed algorithm improves the detection rate compared to the existing integral channel feature method [5]. Fig. 8 compares the proposed algorithm with other state-of-the-art methods. This algorithm is competitive in terms of the detection quality. Table 1 compares the detection rates for the ETHZ dataset, which also confirms that the proposed algorithm provides higher detection rates than the conventional algorithms. Compared to the integral channel feature method, the detection rate of the proposed algorithm is higher in Table 1. As a result, the preprocessing stage enhances the detection rate by looking for the pedestrian candidate regions. In particular, the preprocessing stage helps detect the diverse sizes of the pedestrians compared to the integral channel feature method.

5. Conclusions

This paper presents a two-stage approach to detect pedestrians. The first stage hypothesizes the possible people locations. This step searches the pedestrians' legs using edges, color differences, and motion information. In particular, it reduces the adverse impacts of the background motions by employing GME. The second stage verifies the hypotheses using the AdaBoost classifier with the integral channel features. The experimental results showed that the proposed algorithm provides higher detection rates than the conventional algorithms.

(a) (b) (c) (d) (e)

Fig. 8. Comparison of the proposed algorithm with the conventional algorithms (a) proposed algorithm, (b) Felzenszwalb et al. [3], (c) Schwartz et al. [19], (d) Maji et al. [20], (e) Dollar et al. [7].

Acknowledgement

This work was supported partly by Korea University and partly by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science, ICT & Future Planning (MSIP) (No. 2009-0083495).

References

[1] C. Papageorgiou and T. Poggio, “A trainable system for object detection,” Int’l J. Computer Vision, vol.38, no.1, pp.15-33, 2000. Article (CrossRef Link)

[2] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” Proc. IEEE Conf. Computer Vision and Pattern Recognition, June. 2005. Article (CrossRef Link)

[3] P. Felzenszwalb, D. McAllester, and D. Ramanan, “A discriminatively trained, multiscale, deformable part model,” Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2008. Article (CrossRef Link)

[4] Y. M. Chen and I. V. Bajic, “Motion vector outlier rejection cascade for global motion estimation,” IEEE Signal Process. Lett., vol. 17, no. 2, pp. 197-200, Feb. 2010. Article (CrossRef Link)

[5] Y. M. Chen and I. V. Bajic, “A joint approach to global motion estimation and motion segmentation from a coarsely sampled motion vector field,” IEEE

Trans. Circuits Syst. Video Technol., vol. 21, no. 9, pp. 1316-1328, Sept. 2011. Article (CrossRef Link)
[6] P. Dollar, S. Belongie and P. Perona, "The fastest pedestrian detector in the west," Proc. British Machine Vision Conf., 2010. Article (CrossRef Link)
[7] P. Dollar, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," Proc. British Machine Vision Conf., 2009. Article (CrossRef Link)
[8] D. Geronimo, A. M. Lopez, A. D. Sappa and T. Graf, "Survey of pedestrian detection for advanced driver assistance systems," IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, No. 7, pp. 1239-1258, Jul. 2010. Article (CrossRef Link)
[9] F. Han, Y. Shan, R. Cekander, H. Sawhney and R. Kumar, "A two-stage approach to people and vehicle detection with hog-based SVM," in Proc. Performance Metrics for Intelligent Systems, pp. 133-140, Aug. 2006. Article (CrossRef Link)
[10] L. Guo, R.-B. Wang, L.-S. Jin, L.-H. Li, and L. Yang, "Algorithm study for pedestrian detection based on monocular vision," Proc. IEEE Vehicular Electronics and Safety, pp. 83-87, Dec. 2006. Article (CrossRef Link)
[11] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," Proc. IEEE Conf. Computer Vision and Pattern Recognition, Dec. 2001. Article (CrossRef Link)
[12] P. Viola, M. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," Proc. Int'l Conf. Computer Vision, pp. 734-741, Oct. 2003. Article (CrossRef Link)
[13] A. Ess, B. Leibe, and L. V. Gool, "Depth and appearance for mobile scene analysis," Proc. Int'l Conf. Computer Vision, Oct. 2007. Article (CrossRef Link)
[14] D. M. Gavrila, "Pedestrian detection from a moving vehicle," Proc. European Conf. Computer Vision, vol. II, pp. 37-49, June. 2000. Article (CrossRef Link)
[15] R. Benenson, M. Mathias, R. Timofte, and L. Van Gool, "Pedestrian detection at 100 frames per second," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp.2903-2910, June. 2012. Article (CrossRef Link)
[16] P. Dollar, C. Wojek, B. Schiele, and P. Lerona., "Pedestrian detection: an evaluation of the state of the art," IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, No. 4, pp. 743-761, Apr. 2012. Article (CrossRef Link)
[17] P. Viola, and M. Jones, "Robust real-time face detection," Int'l J. Computer Vision, vol. 57, no. 2, pp. 137-154, 2004. Article (CrossRef Link)
[18] Y. Su, M.-T. Sun, and V. Hsu, "Global motion estimation from coarsely sampled motion vector field and the applications", IEEE Trans. Circuits Syst. Video Technol., vol. 15, no. 2, pp. 232-242, Feb. 2005. Article (CrossRef Link)
[19] W. Schwartz, A. Kermbhavi, D. Harwood, and L. Davis, "Human detection using partial least squares analysis." Proc. IEEE Int'l Conf. Computer Vision, 2009. Article (CrossRef Link)
[20] S. Maji, A. Berg, and J. Malik, "Classification using intersection kernel support vector machines is efficient," Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2008. Article (CrossRef Link)

Mi-ae Kim received her B.S degree in Computer Engineering from Handong University in 2003. From 2003, she worked at the R&D Team of Digital Imaging Business in Samsung Electronics. Currently, she is a master course student in the School of Electrical Engineering at Korea University. She is interested in image processing.

Chang-Su Kim received his Ph.D degree in Electrical Engineering from Seoul National University (SNU) with a Distinguished Dissertation Award in 2000. From 2003 and 2005, he was an Assistant Professor in the Department of Information Engineering, Chinese University of Hong Kong. In Sept. 2005, he joined the School of Electrical Engineering, Korea University, where he is currently a Professor. His research topics include image, video, and 3D graphics processing and multimedia communications. In 2009, he received the IEEK/IEEE Joint Award for the Young IT Engineer of the Year. He has published more than 160 technical papers in international journals and conferences. He is a Senior Member of IEEE. Dr. Kim is an Editorial Board Member of Journal of Visual Communication and Image Representation and an Associate Editor of IEEE Transactions on Image Processing.