Motion Segmentation Using Optical Flow for Pedestrian Detection from Moving Vehicle

Joko Hariyono, Van-Dung Hoang, and Kang-Hyun Jo

Graduate School of Electrical Engineering, University of Ulsan, Ulsan 680–749 Korea
{joko,hvzung}@islab.ulsan.ac.kr, acejo@ulsan.ac.kr

Abstract. This paper proposes a pedestrian detection method using optical flows analysis and Histogram of Oriented Gradients (HOG). Due to the time consuming problem in sliding window based, motion segmentation proposed based on optical flow analysis to localize the region of moving object. A moving object is extracted from the relative motion by segmenting the region representing the same optical flows after compensating the ego-motion of the camera. Two consecutive images are divided into grid cells 14x14 pixels, then tracking each cell in current frame to find corresponding cells in the next frame. At least using three corresponding cells, affine transformation is performed according to each corresponding cells in the consecutive images, so that conformed optical flows are extracted. The regions of moving object are detected as transformed objects are different from the previously registered background. Morphological process is applied to get the candidate human region. The HOG features are extracted on the candidate region and classified using linear Support Vector Machine (SVM). The HOG feature vectors are used as input of linear SVM to classify the given input into pedestrian/non-pedestrian. The proposed method was tested in a moving vehicle and shown significant improvement compare with the original HOG.

Keywords: Pedestrian detection, Optical flow, Motion Segmentation, Histogram of oriented gradients.

1 Introduction

Detecting pedestrian as moving object is one of the essential tasks for understanding environment. In the past few years, moving object and pedestrian detection methods for mobile robots/vehicles have been actively developed. For real-time pedestrian detection system, Gavrila *et al.* [1] were employed hierarchical shape matching to find pedestrian candidates from moving vehicle. Their method uses a multi-cues vision system for the real-time detection and tracking of pedestrians. Nishida *et al.* [2] applied SVM with automated selection process of the components by using AdaBoost. These researches show that the selection of the components and the combination of them are important to get a good pedestrian detector.

Many local descriptors are proposed for object recognition and image retrieval. Mikolajczyk et al. [3], [14] compared the performance of the several local descriptors and showed that the best matching results were obtained by the Scale Invariant

D. Hwang et al. (Eds.): ICCCI 2014, LNAI 8733, pp. 204–213, 2014.

Feature Transform (SIFT) descriptor [4]. Dalal et al. [5] proposed a human detection algorithm using histograms of oriented gradients (HOG) which are similar with the features used in the SIFT descriptor. HOG features are calculated by taking orientation histograms of edge intensity in a local region. They extracted the HOG features from all locations of a dense grid on an image region and the combined features are classified by using linear SVM. They showed that the grids of HOG descriptors significantly out-performed existing feature sets for human detection. Kobayashi et al. [6] proposed selected feature of HOG using PCA to decrease the number of feature. It could reduce the number of features less than half without lowering the performance

Moving object detection and motion estimation methods using the optical flow for a mobile robot also have been actively developed. Talukder et al. [7] proposed a qualitative obstacle detection method was proposed using the directional divergence of the motion field. The optical flow pattern was investigated in perspective camera and this pattern was used for moving object detection. Also real-time moving object detection method was presented during translational robot motion.

Several researchers also developed methods for ego-motion estimation and navigation from a mobile robot using an omnidirectional camera [8], [9]. They used Lucas Kanade optical flow tracker and obtained corresponding features of background in the consecutive two omnidirectional images. Use analyzing the motion of feature points, camera ego-motion was calculated. They obtained camera ego-motion compensated based on an affine transformation of two consecutive frames where corner features were tracked by Kanade-Lucas-Tomasi (KLT) optical flow tracker [10]. However using corner feature for tracking, the detecting moving objects resulted in a problem that only one affine transformation model could not represent the whole background changes. For this problem, our previous work [11] proposed each affine transformation of local pixel groups should be tracked by KLT tracker. The local pixel groups are not a type of image features such as corner or edge. We use grid windows-based KLT tracker by tracking each local sector of panoramic image while other methods use sparse features-based KLT tracker. Therefore we can segment moving objects in panoramic image by overcoming the nonlinear background transformation of panoramic image.

Proposed method is inspired by the works on pedestrian detection from moving vehicle [1], [7], using optical flow [10] and ego-motion estimation [8], which is ego-motion compensated [11]. Pedestrian as a moving object is extracted from the relative motion by segmenting the region representing the same optical flows after compensating the ego-motion of the camera. To obtain the optical flow, feature extracted from an image by divided into grid cells 14x14 pixels. Then, track corresponding cells in the next frame. At least using three corresponding feature cells, affine transformation is performed according to each corresponding cells in the consecutive frame, so that conformed optical flows are extracted. The regions of moving object are detected as transformed objects are different from the previously registered background. Morphological process is applied to get the candidate human region. In order to recognize the object, the HOG features are extracted on the candidate region and classified using linear Support Vector Machine (SVM) [13]. The HOG feature vectors are used as input of linear SVM to classify the given input into pedestrian/non-pedestrian. For the performance evaluation comparative study was presented in this paper.

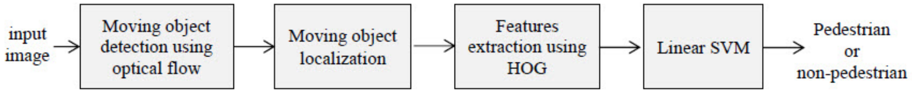

input image
Moving object detection using optical flow
Moving object localization
Features extraction using HOG
Linear SVM
Pedestrian or non-pedestrian

Fig. 1. The overview of the pedestrian detection algorithm

2 Motion Segmentation

This section presents the method to detect object motion from the camera, which is mounted on the moving vehicle. In order to obtain moving object regions from video or sequent images, it is not easy to segment out only moving object region, because it faced two kinds of motion, the first is independence motion caused of object movement and the second is motion caused by camera ego-motion. So, we proposed a method to deal with this situation [11], [15]. We used optical flow analysis to segmenting independent motion of object movement from ego-motion caused by camera, which is ego-motion compensated.

Human eyes will be easy to see and analyze which one is object of interest, such as moving object, and otherwise are static environment. However, robot will understand the environment base on the mathematical model. The optical flow analysis needed to proof that the motion caused of the independent motion of object movement will have difference pattern compare with flow caused by ego-motion from camera. These cues then will give us the region of moving objects and should be localized. Then, this region will be a candidate of detected human/pedestrian after we apply features extraction using HOG and linear SVM as a classifier. The overview of the pedestrian detection algorithm is shown in Fig. 1.

2.1 Ego-motion Compensated

In our previous work [11], we apply KLT optical flow feature tracker [10] in order to deal with several conditions. Brightness constancy which is projection of the same point looks the same in every frame, small motion that points do not move very far and spatial coherence that points move like their neighbors. However, using frame difference will not solve the problem, because it represents all motions caused by the camera ego-motion and moving object in scenes together. It needs to compensate this effect from frame difference to segment out only the independent motion of the object movement, so how much the image background has been transformed in two sequent images. The affine transformation represents the pixel movement between two sequent images as in (1),

$$P' = \mathrm{AP} + \mathrm{t} \tag{1}$$

where P and P' are the pixel location in the first and second image respectively. A is transformation matrix and t is translation vector. Affine parameters can be calculated by the least square method using at least three corresponding features in two images.

In this work, the original input images converted to grayscale images, and obtain one channel intensity pixel value from the input images. Then, using two consecutive images are divided into grid cells size 14x14 pixels, then compare and track each cell in current image to find corresponding cell in the next image. The cell has most similar intensity value in a group will selected as corresponding value. Using method from [10], then find the motion distance of each pixel in a group of cell, the motion d in x and $y-$ axis of each cell $g_{t-1}(i,j)$ by finding most similar cell $g_t(i,j)$ in the next image.

$$g_{t-1}(i,j) = g_t(i + d_x, j + d_y) \tag{2}$$

where d_x and d_y are motion distances in x and $y-$ axis respectively. Using at least three corresponding features in two images, affine parameters can be calculated by the least square method. So, equation (2) can represented as affine transformation of each pixel in the same cell as (3)

$$I_t(x,y) = \mathrm{A}\, I_{t-1}(x,y) + d \tag{3}$$

where $I_t(x,y)$ and $I_{t-1}(x,y)$are vector 2x1 represent pixel location in the current and previous frame respectively, A is 2x2 projection matrix and d is 2x1 translation vector.

To obtain the camera ego-motion compensated, frame difference is applied in two consecutive input images by calculated based on the tracked corresponding pixel cells using (4)

$$I_d(x,\mathrm{y}) = |I_{t-1}(x,y) - I_t(x,\mathrm{y})| \tag{4}$$

where $I_d(x,y)$ is a pixel cell located at (x,y) in the grid cell.

Suppose two consecutive images shown in Fig. 2 (a) and (b) can not segment out moving object using frame difference (c), however when we apply frame difference with ego-motion compensate could obtain moving objects area shown in Fig. 2 (d).

2.2 Motion Segmentation

Each pixel output from frame difference with ego-motion compensated cannot show clearly as silhouette. It just gives information of motion area of object movement. Those moving area are applied morphological process to obtain region of moving object and noise removal. Ideally, we would seek to devise a region segmentation algorithm that accurately locates the bounding boxes of the motion regions in the difference image. Given the sparseness of the data, however, accurate segmentation would involve the enforcement of multiple constraints, making fast implementation difficult. To achieve faster segmentation, we assumed the fact that humans usually appear in upright positions, and conclude that segmenting the scene into vertical strips is sufficient most of the time. In this work we define detected moving objects are represented by the position in width in x axis. Using projection histogram h_x by pixel voting vertically project image intensities into $\mathrm{x-coordinate}$.

Adopting the region segmentation technique by [12], we define the region using boundary saliency. It measures the horizontal difference of data density in the local neighborhood. The local maxima correspond to where maximal change in data density occur, are candidates for region boundaries of pedestrian in moving object detection.

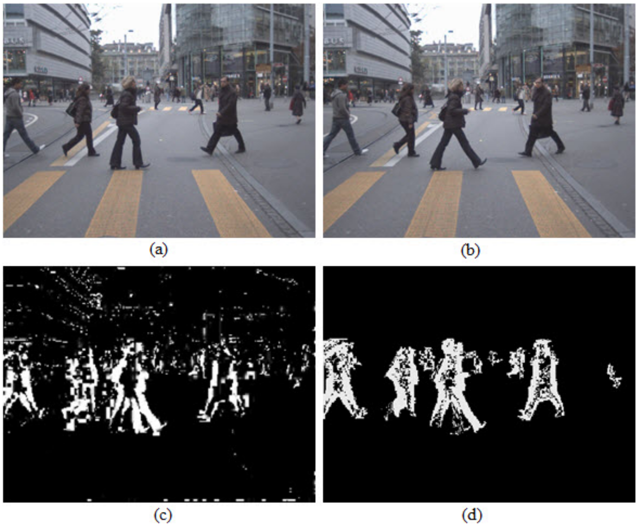


Fig. 2. From two consecutive images (a) and (b), then we applied frame difference (c) and comparing when we applied frame difference with ego-motion compensated (d)

3 Feature Extraction

In this section present how we extract feature from candidate region obtained from previous section. In this work we use Histogram of Oriented Gradients (HOG) to extract features from moving object area localization. Local object appearance and shape usually can be characterized well by the distribution of local intensity gradients or edge direction. HOG features are calculated by taking orientation histograms of edge intensity in local region.

3.1 HOG Features

In this work, we extract HOG features from 16×16 local regions as shown in Fig.3. The first, we use Sobel filter to obtain the edge gradients and orientations were calculated from each pixel in this local region. The gradient magnitude $m(x,y)$ and orientation $\theta(x, y)$ are calculated using directional gradients $d_x(x, y)$ and $d_y(x, y)$ which are computed by Sobel filter as follow (5),

$$m(x, y) = \sqrt{dx(x, y)^2 - dy(x, y)^2} \tag{5}$$

$$\theta(x, y) = \begin{cases} \tan^{-1}\left(\frac{dy(x,y)}{dx(x,y)}\right) - \pi, & if\ dx(x, y) < 0\ and\ dy(x, y) < 0 \\ \tan^{-1}\left(\frac{dy(x,y)}{dx(x,y)}\right) + \pi, & if\ dx(x, y) < 0\ and\ dy(x, y) > 0 \\ \tan^{-1}\left(\frac{dy(x,y)}{dx(x,y)}\right), & otherwise \end{cases} \tag{6}$$

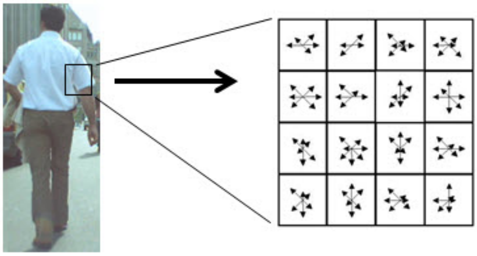


Fig. 3. Extraction Process of HOG features. The HOG features are extracted from local regions with 16 ×16 pixels. Histograms of edge gradients with 8 orientations are calculated from each of 4×4 local cells.

The local region is divided into small spatial cell, each cell size is 4×4 pixels. Histograms of edge gradients with 8 orientations are calculated from each of the local cells. The total number of HOG features is 128 = 8 × (4 × 4) and they constitute a HOG feature vector. To avoid sudden changes in the descriptor with small changes in the position of the window, and to give less emphasis to gradients that are far from the center of the descriptor, a Gaussian weighting function with σ equal to one half the width of the descriptor window is used to assign a weight to the magnitude of each pixel.

A vector of the HOG features represent local shape of an object, it have edge information at plural cells. In flatter regions like a ground or a wall of a building, the histogram of the oriented gradients has flatter distribution. On the other hand, in the border between an object and background, one of the elements in the histogram has a large value and it indicates the direction of the edge. Even though the images are normalized to position and scale, the positions of important features will not be registered with same grid positions. It is known that HOG features are robust to the local geometric and photometric transformations. If the translations or rotations of the object are much smaller than the local spatial bin size, their effect is small.

Dalal *et al.* [5] extracted a set of the HOG feature vectors from all locations in an image grid and are used for classification. In this work, we extract the HOG features from all locations on the candidate region from an input image as shown in Fig. 4.

3.2 Linear SVM Classifier

In the human detection algorithm proposed by [5], the HOG features are extracted from all locations of a dense grid and the combined features are classified using the linear SVM. The HOG shows significantly outperformed existing feature sets for human detection. This work also used the linear SVM to perform work in various data classification tasks. Let $\{\boldsymbol{f}i, ti\}_{i=1}^{N}$ $(\boldsymbol{f_i} \in R^D,\ t_i \in \{-1, 1\})$ be the given training sample in D-dimensional feature space. The classification function is given as

$$z = sign(\omega^T f_i - h) \tag{7}$$

where $\boldsymbol{w}$ and h are the parameters of the model. For the case of soft-margin SVM, the optimal parameters are obtained by minimizing

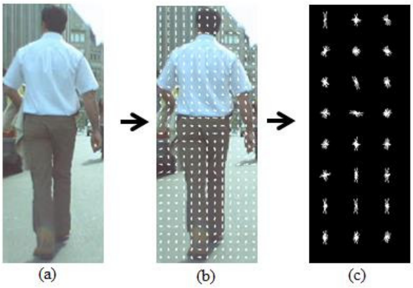


Fig. 4. From a candidate input image size 150x382 (a), HOG features are extracted from all locations on the candidate region of an input image with 16x16 pixels region (b), and the result shown in (c)

$$L(\omega, \xi) = \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{N} \xi_i \tag{8}$$

under the constraints

$$\xi_i \geq 0, t_i(\omega^T f_i - h) \geq 1 - \xi_i (i = 1, \ldots, N) \tag{9}$$

where ξi (≥ 0) is the error of the i-th sample measured from the separating hyperplane and C is the hyper-parameter which controls the weight between the errors and the margin. The dual problem of (8) is obtained by introducing Lagrange multipliers $\boldsymbol{\alpha} = (\alpha 1, \ldots., \alpha N)$, $\alpha k \geq 0$ as

$$L_D(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{N} \alpha_i\, \alpha_j t_i t_j f_i^T f_i \tag{10}$$

under the constraints

$$\sum_{i=1}^{N} \alpha_i t_i = 0, \quad 0 \leq \alpha_i \quad (i = 1, \ldots N) \tag{11}$$

By solving (10), the optimum function is obtained as

$$z = sign(\sum_{i \epsilon S} \alpha_i^* t_i f_i^T f_i - h^*) \tag{12}$$

where *S* is the set of support vectors. To get a good classifier, we have to search the best hyper-parameter *C*. The cross-validation is used to measure the goodness of the linear SVM classifier.

4 Experimental Results

In this work, our vehicle system is run in outdoor environment with varies speed and detected object moving surround its path. Proposed algorithm was programmed in MATLAB and executed on an Intel Pentium 3.40 GHz, 32-bit operating system with 8 GB Random Access Memory. The proposed algorithm was evaluated by using five images sequences from ETHZ pedestrian datasets which contains around 5,000 images of pedestrians in city scenes [12]. It contains only front or back views with relatively limited range of poses and the position and the height of human in the image are almost adjusted. The size of the image is 640 × 480 pixels. For the training process, we used person INRIA datasets in [5]. These images were used for positive samples in the following experiments. The negative samples were originally collected from images of sky, mountain, airplane, building, etc. The number of negative images is 3,000. From these images, 1,000 person images and 2,000 negative samples were used as training samples to determine the parameters of the linear SVM. The remaining 100 pedestrian images and 200 negative samples were used as test samples to evaluate the recognition performance of the constructed classifier. When we implemented the original HOG, which proposed by Dalal *et. al* using those dataset, the recognition rate for test dataset is 98.3%. We used ego-motion compensated and HOG feature to evaluate performance improvement.

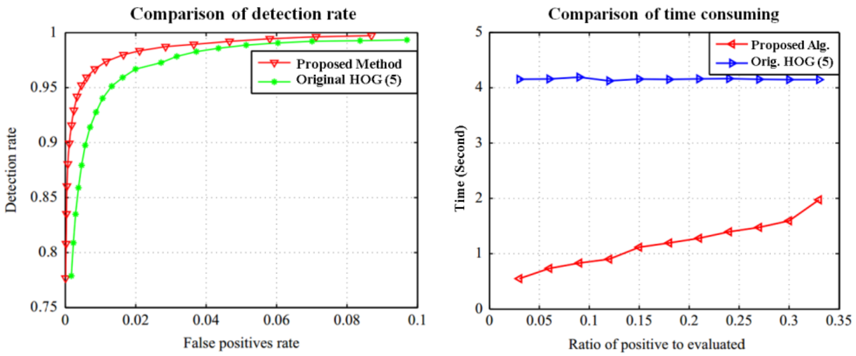


Fig. 5. Comparison result when we tested our proposed method and Original HOG by Dalal *et. al.* (a) comparison of detection rate and (b) comparison of time consuming

HOG feature vectors were extracted from all locations of the grid for each training sample. Then, the selected feature vectors were used as input of the linear SVM. The selected subsets were evaluated by cross validation. Also we evaluated the recognition rates of the constructed classifier using test samples. The relation between the detection rates and the number of false positive rate are shown in Fig. 5. The best recognition rate 99.3 % was obtained at 0.09 false positive rates. It means that we obtain higher detection rate with smaller false positives rate. The computational cost also reduces eight times better when we use small ratio of positive to evaluated data. However, if we increase the number of ratio it also reduces time consuming significantly. The results are shown in Fig 6 and false detection shown in Fig. 7

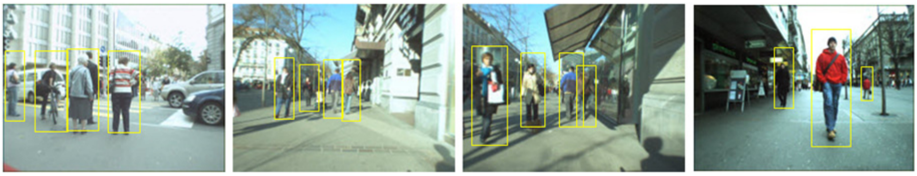
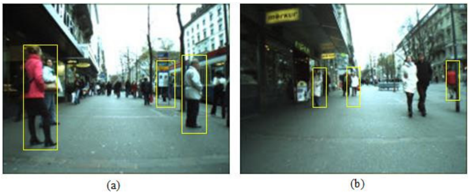


Fig. 6. Successful moving objects detection results



(a)

(b)

Fig. 7. (a) False positives detection and (b) False negative detection

HOG feature vectors were extracted from all locations of the grid for each training sample. Then, the selected feature vectors were used as input of the linear SVM. The selected subsets were evaluated by cross validation. Also we evaluated the recognition rates of the constructed classifier using test samples. The relation between the detection rates and the number of false positive rate are shown in Fig. 7. The best recognition rate 99.3 % was obtained at 0.09 false positive rates. It means that we obtain higher detection rate with smaller false positives rate. The computational cost also reduces eight times better when we use small ratio of positive to evaluated data. However, if we increase the number of ratio it also reduces time consuming significantly. The results are shown in Fig 8 and false detection shown in Fig. 9.

5 Conclusion

This paper presents pedestrian detection method using optical flow based on moving vehicle properties. The moving object is segment out through the relative evaluation of the optical flow to compensate ego-motion of camera. In order to recognize the object, the HOG features were extracted on a candidate region and classified using the SVM. The HOG feature vectors are used as an input of linear SVM to classify the given input into pedestrian/non-pedestrian. The proposed algorithm achieved comparable results comparing with the original HOG, and also reduces computational cost significantly using moving object localization.

Acknowledgement. This research was supported by the MOTIE (The Ministry of Trade, Industry and Energy), Korea, under the Human Resources Development Program for Convergence Robot Specialists support program supervised by the NIPA (National IT Industry Promotion Agency) (H1502-13-1001)

References

1. Gavrila, D.M., Munder, S.: Multi-cue Pedestrian Detection and Tracking from a Moving Vehicle. International Journal of Computer Vision 73(1), 41–59 (2007)
2. Nishida, K., Kurita, T.: Boosting soft-margin SVM with feature selection for pedestrian detection. In: Oza, N.C., Polikar, R., Kittler, J., Roli, F. (eds.) MCS 2005. LNCS, vol. 3541, pp. 22–31. Springer, Heidelberg (2005)
3. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 27, 1615–1630 (2005)
4. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
5. Dalal, N., Triggs, B.: Histograms of Oriented Gradients for Human Detection. In: IEEE Conference on Computer Vision and Pattern Recognition, San Diego, pp. 886–893 (2005)
6. Kobayashi, T., Hidaka, A., Kurita, T.: Selection of Histograms of Oriented Gradients Features for Pedestrian Detection. In: Ishikawa, M., Doya, K., Miyamoto, H., Yamakawa, T. (eds.) ICONIP 2007, Part II. LNCS, vol. 4985, pp. 598–607. Springer, Heidelberg (2008)
7. Talukder, S.G., Matthies, L., Ansar, A.: Real-time detection of moving objects in a dynamic scene from moving robotic vehicles. In: Proc. of Int. Conf. Intelligent Robotics and Systems, pp. 1308–1313 (2003)
8. Vassallo, R.F.: Santos-Victor and H. Schneebeli, A General Approach for Egomotion Estimation with Omnidirectional Images. In: Proceedings of the Third Workshop on Omnidirectional Vision, Copenhagen, pp. 97–103 (2002)
9. Liu, H., Dong, N., Zha, H.: Omni-directional Vision based Human Motion Detection for Autonomous Mobile Robots. Systems Man and Cybernetics 3, 2236–2241 (2005)
10. Tomasi, C., Kanade, T.: Detection and Tracking of Point Features. International Journal of Computer Vision 9, 137–154 (1991)
11. Hariyono, J., Hoang, V.-D., Jo, K.-H.: Human detection from mobile omnidirectional camera using ego-motion compensated. In: Nguyen, N.T., Attachoo, B., Trawiński, B., Somboonviwat, K. (eds.) ACIIDS 2014, Part I. LNCS, vol. 8397, pp. 553–560. Springer, Heidelberg (2014)
12. Ess, A., Leibe1, B., Gool, L.V.: Depth and Appearance for Mobile Scene Analysis. In: IEEE International Conference on Computer Vision, ICCV 2007 (2007)
13. Hoang, V.-D., Le, M.-H., Jo, K.-H.: Hybrid Cascade Boosting Machine using Variant Scale Blocks based HOG Features for Pedestrian Detection. Neurocomputing 135, 357–366 (2014)
14. Lu, Y.-Y., Huang, H.-C.: Adaptive reversible data hiding with pyramidal structure. Vietnam Journal of Computer Science, 1–13 (2014)
15. Hariyono, J., Kurnianggoro, L., Wahyono, Hernandez, D.C., Jo, K.-H.: Ego-motion compensated for moving object detection in a mobile robot. In: Ali, M., Pan, J.-S., Chen, S.-M., Horng, M.-F. (eds.) IEA/AIE 2014, Part II. LNCS, vol. 8482, pp. 289–297. Springer, Heidelberg (2014)