International Journal of Multimedia and Ubiquitous Engineering
Vol.10, No.10 (2015), pp.35-46
http://dx.doi.org/10.14257/ijmue.2015.10.10.04

A Method of Robust Pedestrian Tracking in Video Sequences Based on Interest Point Description

Ming-Shou An and Dae-Seong Kang

Dong-A University, Dept. of Electronics Engineering, 37 Nakdong-daero 550 beon-gil Saha-gu, Busan, Korea
dskang@dau.ac.kr

Abstract

There are several methods proposed for detection and segmentation of object effectively. However, this algorithm struggles to detect an object with a lot of noise and shadows. Therefore, it is difficult to segment the accurate region and information of the object using background modeling only. To solve these problems, this paper introduced a more effective method of object segmentation based on interest point detection and description, which are core SURF theories. As a result, the feature extracted from the region of interest (ROI) was detecTable even with changes in scale, noise, and illumination. We then made the adaptive search window by this feature for ROI. After object detection, we applied the SVM to train the information of the feature from the detected object, and a classifier was built to estimate whether a result was a pedestrian. Therefore, if the result is a pedestrian, we would employ the Camshift algorithm to track the motion of this pedestrian. The experimental results showed the effectiveness of our method through comparison with others.

Keywords: *Background modeling, Interest point detection, SURF, ROI, SVM, Camshift*

1. Introduction

An automatic visual surveillance system analyzes event occurrences from video. It is an important topic in computer vision applications. There are several different research approaches, or research designs, that qualitative researchers have used [1]. In the events that occurred, the target focused on the moving human (pedestrian) in this paper. For the pedestrian surveillance system, a computer vision technique using object detection and tracking is required. The object detection part is the process of detecting the moving objects from video sequences using the image processing algorithm effectively. To extract the region of the object, Stauffer and Grimson proposed a background modeling method based on an adaptive Gaussian mixture [2-3]. Heikkila and Pietikainen proposed a texture-based method based on the texture feature described by local binary patterns for the background model [4]. The object tracking part analyzes the motion patterns of the moving object. Meanshift [5-7] and Camshift algorithms [8] were used for the color probability distribution of the object tracking [9]. Isard and Blake suggested a condensation algorithm for object tracking based on the contour information of objects [10]. It is very important for object tracking to extract and describe the effective features of the object. Jung proposed fast object tracking algorithm by using adaptive background image and dynamic search window [11]. The scale invariant feature transform (SIFT) algorithm is used to detect and describe local features in images that was proposed by Lowe [12]. For any object in an image, interesting points on the object can be extracted to provide a "feature description" of the object [13-14]. However, it takes considerable processing time to extract and describe the features. Speeded up robust features (SURF) is a robust local feature detector that was first presented by Bay *et al*. [15-16]. It was partly inspired by the SIFT descriptor. The standard version of SURF is several times faster than

ISSN: 1975-0080 IJMUE

SIFT, and Bay *et al*. claimed it is more robust against different image transformations than SIFT [17].

The approaches and proposition made here for the surveillance system can be summarized as follows:

The background model based on the Gaussian mixture model (GMM) is proposed to extract the moving object. This is the most common and practical method for pre-processing video analysis.

A feature selection algorithm based on the SURF algorithm is proposed for describing the ROI of the object. Candidate corners are extracted in only the ROI by the core concept of SURF. The outermost corner points on each side of a rectangle are found as selected features. We then take the four corner points and generate an adaptive search window with them.

To determine whether the object is a pedestrian, the SVM algorithm [18] is employed here to filter out the objects that are not pedestrians with physical characteristics [19]. The algorithm is applied with the location and size of the detected object automatically [20]. For accurate tracking, the fast and efficient Camshift algorithm could be put to use.

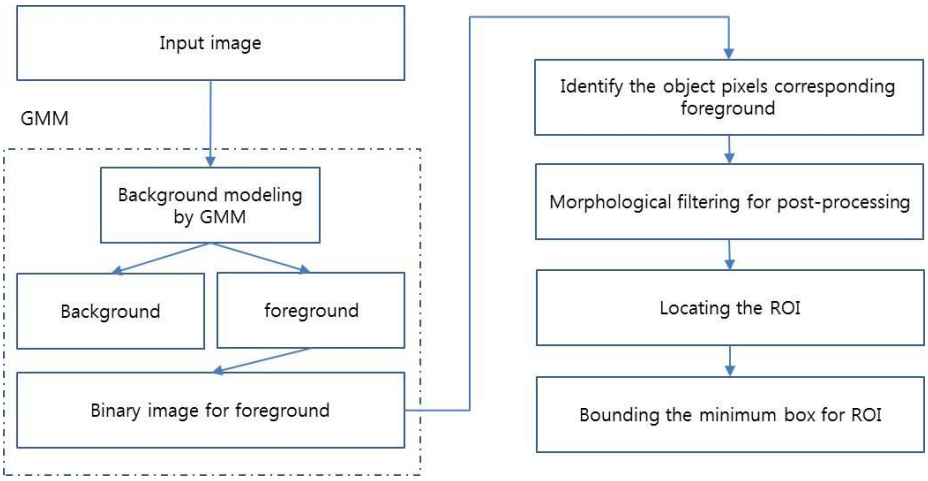


Figure 1. Flowchart for Moving Object Detection Based on Background Modeling

2. Feature Selection for ROI

Most research for object detection had been based on background subtraction. A pixel-based method does not consider the general things in the frame, and therefore shadows and noise must be handled afterwards. In our method, we proposed a method for removing the noise and bounding the optimal box of the object based on feature extraction. A flowchart for the object detection method is shown in Figure 1.

2.1. ROI Detection of Pedestrian

To detect the object, GMM background modeling is used. It is a parametric probability density function that describes a calculated sum of Gaussian densities. This method, which was proposed by Stauffer *et al*. [3], models each pixel as a Gaussian mixture distribution and uses an online approximation for the update. The GMM features detect the tracked object where the progressive light change exists. Observation of the pixel value probability at the current time t is shown in Eq. (1).

$$P(x_t) = \sum_{j=1}^{k} w_{j,t} \cdot \eta(x_i, \mu_{i,t}, \Sigma_{i,t}), \quad (1)$$

Where K is the distribution number, $\omega_{i,t}$, is an estimated weight value, which is a portion of the Gaussian accounting data, $\mu_{i,t}$ is the mean value of the ith at time t, $\sum_{i,t}$ is the covariance matrix at time t, x_t is a random variable vector at time t, and $\eta(X_t, \mu_{i,t}, \sum_{i,t})$ is a parameter of the Gaussian probability density function.

The matching condition of the pixel value X_t is defined as Eq. (2).

$$\begin{cases} maching: & |Xt - \mu i,t| < \lambda \sigma i,t \\ unmaching: & |Xt - \mu i,t| \geq \lambda \sigma i,t \end{cases} \quad i = 1,2,\cdots,K \tag{2}$$

Where λ is the defined pixel value and σ is the standard deviation.

The prior weights of the K distributions at time t are adjusted as shown in equation (3), which shows an iterative process.

$$\omega_{i,t} = (1-\alpha)\omega_{i,t} - 1 + \alpha(M_{i,t}) \tag{3}$$

Where α is the learning rate and $M_{i,t}$ is 1 when the model is matched and 0 when there are remaining models.

The parameters of the mean and the variance for unmatched distributions remain the same. The distribution parameter matches the new observation for updating Eq. (4) and Eq. (5).

$$\mu_{i:} = (1-\rho) + \rho X_t \tag{4}$$

$$\sigma^2{}_{i,t} = (1-\rho)\sigma^2{}_{i,t-1} + \rho(X_t - \mu_{i,t})^T(X_t - \mu_{i,t-1}) \tag{5}$$

where $\rho = \alpha\eta(X_t|\mu_i, \sigma_i)$.

If X_t does not include any Gaussian distributions, the probability distribution will be replaced with a new distribution that has a mean value, high variance, and low prior weight.

When the update is finished, all components in the mixture model are assigned according to the value of ω/σ.

The first B distributions exceeding a certain threshold, T_{bg}, which is a measure of the minimum portion of the data, are then retained for a background distribution, and B can be defined by Eq. (6)

$$B = \operatorname{argmin}_b \left(\sum_{k=1}^{b} \omega_k > T_{bg} \right). \tag{6}$$

2.2. Feature Description for Pedestrian

For bounding the minimum box for ROI, we applied the feature selection with the SURF algorithm, which was proposed by Bay [15], to improve the speed of the feature detection process compared with the SIFT algorithm. It uses a simplified detector, descriptor, and integral image for reducing the computational complexity. The SURF process is divided into two steps: the first is the interest point detection and the next is the interest point description.

Interest point detection:

In the point detection step, the integral image is generated to obtain the object region in the original image. To simplify the calculations, the approximate second-order Gaussian filter is used. Figure 1 shows the approximated Hessian matrix filters. By using the non-maximal suppression and up-scaling the hessian matrix, the interest point is detected by comparing the neighbor points.

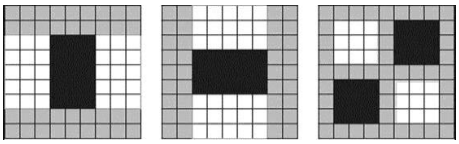


Figure 2. Filters of Approximated hessian Matrix ($L_{xx}(X,\sigma)$, $L_{xy}(X,\sigma)$, $L_{yy}(X,\sigma)$) [21]

Interest point description:

In the point description step, feature point interpolation is conducted to obtain a spin invariant feature. Lastly, 32, 64, and 128-dimensional descriptors are composed by using the Haar Wavelet Feature, which uses a 4*4 detailed region as the center of the feature point. Figure 2 shows a point detection result and the point description.

Figure 3 shows the processing of the generation of the bounding box for ROI from the input frame. First, the input frame was analyzed by performing background subtraction and the ROI of the moving object was detected. Second, the local feature points were extracted using the SURF algorithm and the minimum bounding box was extracted. The first image in Figure 3 is the original image, the second is the binary image of the foreground, the third shows the processing for extracting the local features and bounding box, and the last shows the optimized features of the bounding box.

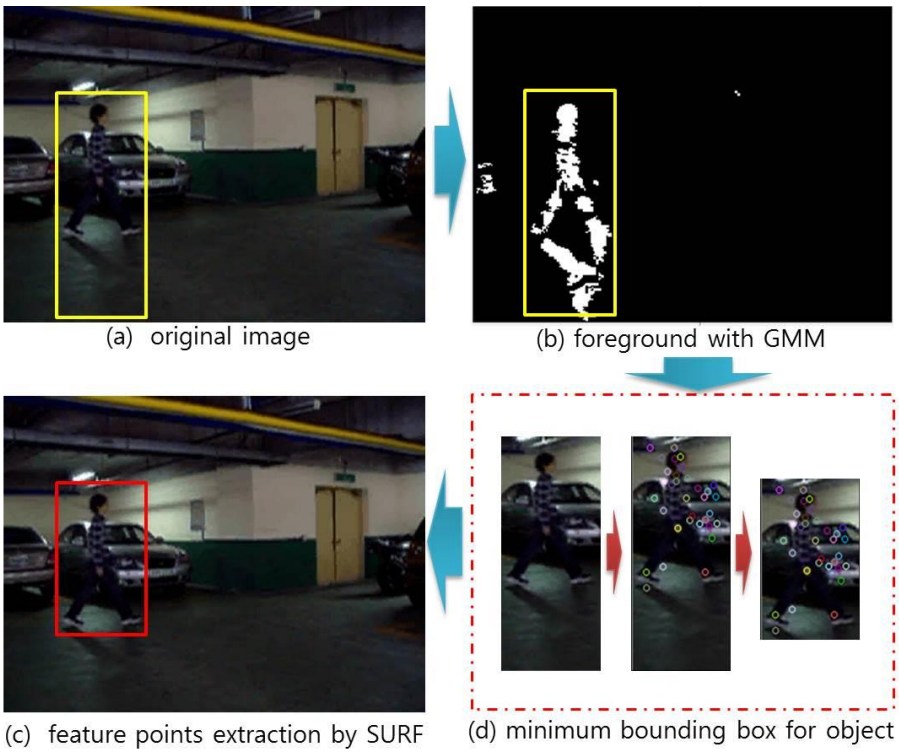


Figure 3. Processing of Generation Feature of Bounding Box from Input Frame (Samples from Video #)

3. Pedestrian Detection and Tracking

3.1. Pedestrian Detection

SVM is a novel kind of learning method which originally developed from the linearly separable problem, and many jobs are related to two types of problems [22]. It can make

visible the pattern of the high-dimensional feature space. Additionally, it can identify the optimal bandwidth.

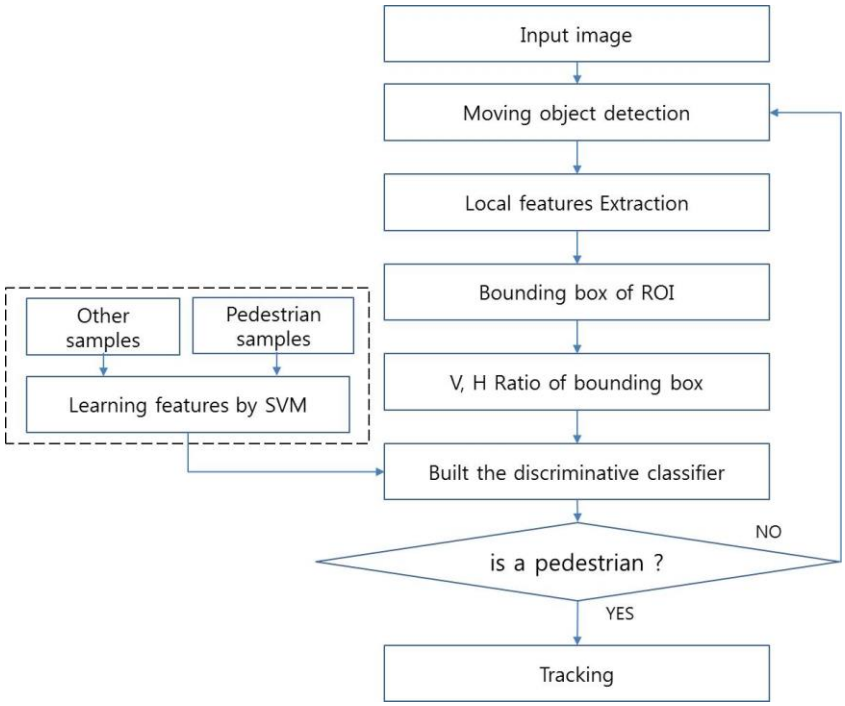


Figure 4. Process of Discriminative Classifier by SVM

The optimal separating hyperplane (OSH) of the SVM is the linear classifier with the maximum margin for a given finite set of learning patterns [23]. The points at the arrow indicate the support vector that is able to get the OSH. The space between the support vectors is called the maximum margin. The maximum margin reduces the information and false alarm rate. Only then can values W and b be obtained by the OSH. There is a method to obtain the optimal W and b—the saddle point of the Lagrangian function. Using the Lagrangian function, the maximum margin and OSH are calculated by the kernel function. According to the purpose of the SVM, the discriminative classifier was built for distinguishing whether the detectable object is a pedestrian. The feature points are all extracted from the minimum bounding boxes, which are found by the SURF algorithm, as shown in Figure 3, of the human blobs detected by the method depicted above. For the learning data of SVM, the horizontal and vertical ratio of the bounding box is found. There are two merits to selecting the ratio of the bounding box as a feature source: First, dynamic information extracted from the bounding boxes by SURF is easy to process; second, it is generally easier to analyze the ratio of humans compared with other approaches. For example, the ratio of humans who are pedestrians is greater than other things, such as sitting people, low people, and Tables. Figure 4 shows the learning process of the discriminative classifier by SVM. The features of pedestrians were that height was about twice the length of the width. If the ratio of size satisfied the threshold value, the discriminative classifier places it into the category of pedestrians. For pedestrians and others, for learning to build the discriminative classifier, 2000 samples were used.

Table 1. The Information of Pedestrians

Sample number	Width (mm)	Height (mm)	Proportion
1	12.6	27.2	1:2.1587
2	18.1	40.9	1:2.2597

3	9.6	23.7	1:2.4687
4	17.2	32.7	1:1.9011
.	.	.	.
.	.	.	.
.	.	.	.
1001	24.0	60.2	1:2.508
1002	15.0	44.4	1:2.9600
1003	15.4	39.7	1:2.5702
.	.	.	.
.	.	.	.
.	.	.	.
1999	20.8	44.9	1:2.1586
2000	17.4	49.2	1:2.8275
Average	19.5	45.2	1:2.3179

Table 1 and Table 2 show the information on pedestrian size and proportion and others. The average proportion is 1:2.4327. For the discriminative classifier, the OSH of the SVM can be expressed as Eq. 7.

$$W^T\mathrm{X} + \omega_0 = \sum_{i=0}^{N} \omega_i x_i + \omega_0, \;\; g(x,y) \leq W, X > \omega_0 \tag{7}$$

$$W^T\mathrm{X} + \omega_0 > 0 \;\; \rightarrow \;\; +1 \; class \tag{8}$$

$$W^T\mathrm{X} + \omega_0 < 0 \;\; \rightarrow \;\; -1 \; class \tag{9}$$

Where W is the normal vector of the hyperplane, X is the input data, and $\mathrm{g(x,y)}$ is the discriminant. The +1 class in Eq. (8) denotes the silhouette area to track a pedestrian, and the -1 class shown in Eq. (9) denotes that an object is not a pedestrian.

Table 2. The Information of Other Samples

Sample number	Width (mm)	Height (mm)	Proportion
1	27.3	22.6	1:0.8278
2	52.4	32.9	1:0.6278
3	51.7	20.0	1:0.3868
4	25.8	28.7	1:1.1124
1001	50.7	24.0	1:0.4733
1002	30.4	45.5	1:1.4967
1003	31.8	33.9	1:1.0660
1999	39.4	26.3	1:0.6675
2000	37.7	31.2	1:0.8275
Average	49.8	32.6	1:0.6546

3.2. Pedestrian Tracking

The Camshift algorithm is primarily intended to perform efficient head and face tracking in a perceptual user interface. It is an adaptation of the Meanshift algorithm. The primary difference between Camshift and the Meanshift algorithms is that Camshift uses continuously adaptive probability distributions, while Meanshift is based on static

distributions. Camshift works by tracking the hue of an object—in this case, the color of the ROI. To obtain the hue, the video frames were all converted to HSV space before individual analysis was conducted.

Camshift is implemented as follows:

Step 1: Set the ROI of the probability distribution image to the entire image.

Step 2: Select an initial location of the Meanshift search window. The location selected by the above method is the target distribution to be tracked.

Sept 3: Calculate a color probability distribution of the region centered on the Meanshift search window.

Step 4: Iterate the Meanshift algorithm to find the centroid of the probability image. Store the zero moment and centroid location.

Step 5: For the following frame, center the search window at the mean location found in Step 4 and set the window size to a function of the zeroth moment. Go to Step 3.

To evaluate our method, we used the two video sequences. Table 3 shows the detailed information of each video sequence used in the experiments. It shows the number of the total frames, the frame rates (frame/sec), frame size, and the number of objects in each video sequence.

Table 4 shows the comparison of the accuracy for each method. The accuracy of the proposed method is higher than that of the other algorithm, as shown in Table 4. Table 5 shows the comparison of the processing time for each method in the experimental videos. Figure 5 shows the experimental videos, which are "Pets2006 - video 1," "Pets2006 - video 2," [24] and "Video #." For this experiment, Figure 6 and Figure 7 show the tracking results of pedestrians using the proposed method.

Table 3. Detailed Information of Each Experimental Video Sequence

Videos	Information		
	Total frames	Frame rate (frame/sec)	Frame size
Pets2006	3020	25	320x240
Video #	1618	25	320x240

Table 4. The Comparison of the Accuracy for Each Method (%)

Videos	Method		
	Meanshift	Camshift	Our method
Pets2006	86.654	76.952	92.413
Video #	92.745	90.386	96.328

Table 5. The Comparison of the Processing Time for Each Method (Msec/Frame)

Videos	Method		
	Meanshift	Camshift	Our method
Pets2006	42.25	42.32	50.59
Video #	37.79	38.74	38.99

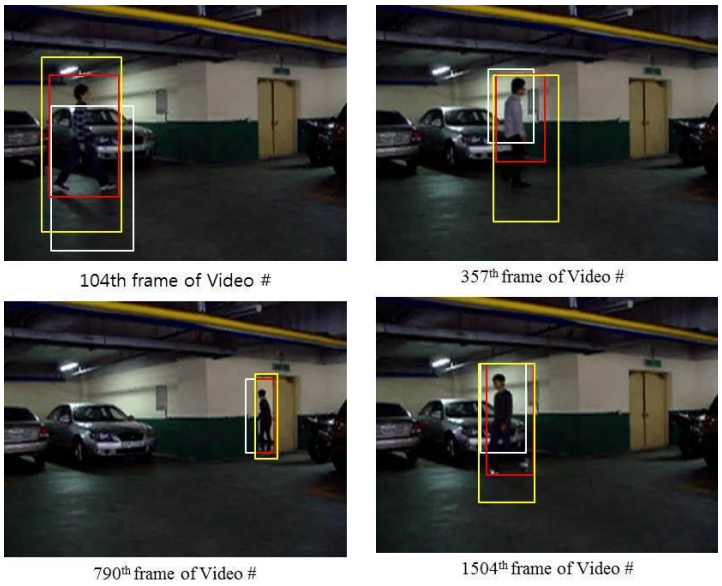


Figure 5. Samples of Tracking Results for Experimental Videos (1)

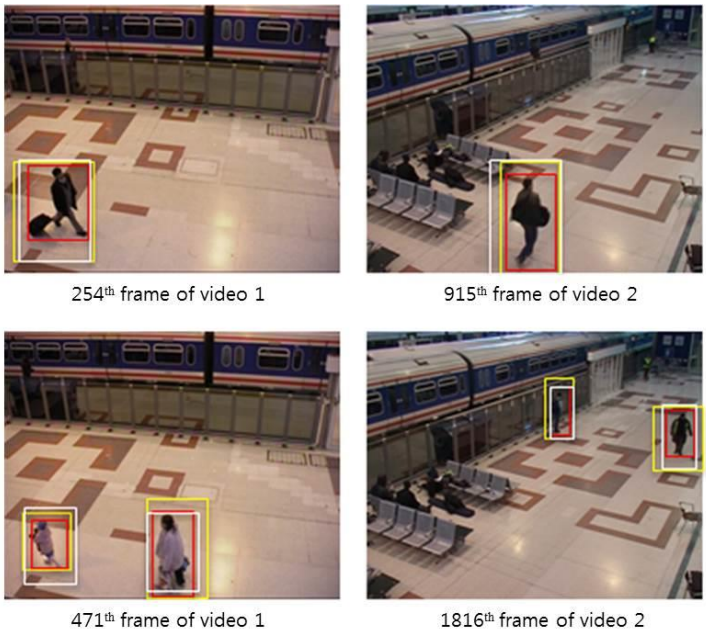


Figure 6. Samples of Tracking Results for Experimental Videos (2)

Figure 5 shows samples of the experimental results for the 104th, 357th, 790th, and 1504th frames in video#. Figure 6 shows samples of the experimental results for the 254th and 471th frames in video 1 and the 915th and 1816th frames in video 2.

In these figures, the region of red boxes shows the detected objects using our method; the white boxes are the Meanshift results; and the yellow boxes are the Camshift results. Through the image results, we can confirm that the minimum bounding box was found.

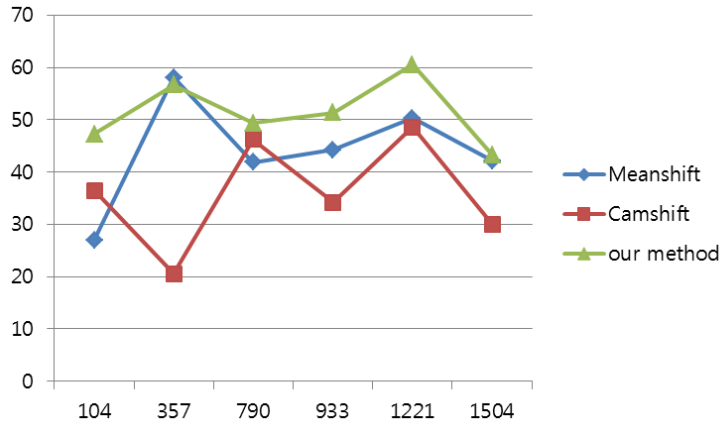


Figure 7. Comparison Results of Rate of Objects in Bounding Boxes (Video#)

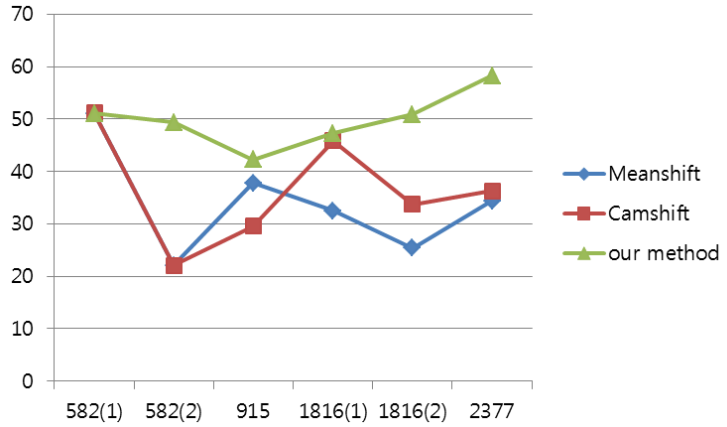


Figure 8. Comparison Results of Rate of Objects in Bounding Boxes (Video1)

The results of comparing the ratio of pixels per object in the bounding boxes for video#, 1, and 2 are shown in Figures 7, 8, and 9, respectively. Figure 7 shows the results of the average ratio of detected objects in the bounding boxes for video# using three methods (Meanshift, Camshift, and our method). The Meanshift result is 43.935%; the Camshift result is 36.006%; and the result of our method is 51.406%. Figure 8 shows the comparison graph of the average ratios of detected objects in the bounding box for video 1. The Meanshift result is 32.188%; the result of Camshift is 39.266%; and the result of our method is 48.373%. Figure 9 shows the results for video 2. The results of Meanshift, Camshift, and the proposed method are 25.655%, 32.531%, and 40.770%, respectively.

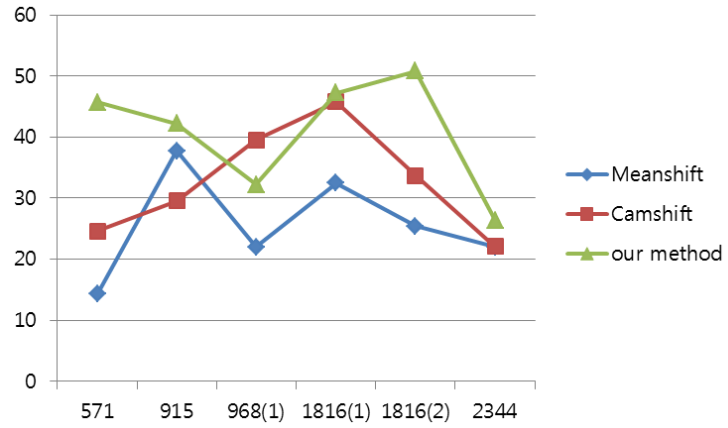


Figure 9. Comparison Results of Rate of Objects in Bounding Boxes (Video2)

4. Conclusion

In this paper, a pedestrian classifier and tracking method is presented, which involves feature fusion with the adaptive bounding box. Feature fusion is composed of color and feature points. Camshift is based on color and is robust to specific color objects and simple to implement. The SURF algorithm is based on feature points and can describe the object appropriately. To then discriminate the pedestrians, a discriminative classifier using SVM was built. For tracking, the Camshift algorithm based on color was applied. For the experiment, the tracking results of Meanshift, Camshift, and the proposed method were compared. In general, the accuracy of the proposed method is higher than that of other algorithms. The average accuracy for the proposed method is 94.3705% in the experiments of various videos. Therefore, the proposed method to classify and track the pedestrians was more robust in tracking applications.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2011-0011735)

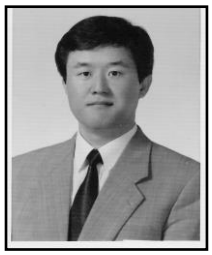
References

[1] Hu W. M., Tan T. N., Wang L. and Maybank S., "A Survey on Visual Surveillance of Object Motion and Behaviors. Systems, Man, and Cybernetics, Part C: Applications and Reviews", IEEE Transactions on, vol. 34, no.3, **(2004)**, pp. 334-352.
[2] Stauffer C. and Grimson W. E. L., "Adaptive background mixture models for real-time tracking. In: Computer Vision and Pattern Recognition", 1999. IEEE Computer Society Conference on, IEEE Press, Fort Collins, vol. 2, **(1999)**.
[3] Stauffer C. and Grimson W. E. L., "Learning patterns of activity using real-time tracking. Pattern Analysis and Machine Intelligence", IEEE Transactions on, vol. 22, no.8, **(2000)**, pp. 747-757.
[4] Heikkila M. and Pietikainen M., "A texture-based method for modeling the background and detecting moving objects", IEEE transactions on pattern analysis and machine intelligence, vol. 28, no.4, **(2006)**, pp. 657-662.
[5] Comaniciu D. and Meer P., "Mean shift analysis and applications In: Computer Vision", 1999. The Proceedings of the Seventh IEEE International Conference on, vol. 2, **(1999)**, pp. 1197-1203.
[6] Comaniciu D. and Meer P., "Mean shift: A robust approach toward feature space analysis. Pattern Analysis and Machine Intelligence", IEEE Transactions on, vol. 24, no. 5, **(2002)**, pp. 603-619.
[7] Fashing M. and Tomasi C., "Mean shift is a bound optimization", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 3, **(2005)**, pp. 471-474.

[8] Allen J. G., Xu R. Y. and Jin J. S., "Object tracking using camshift algorithm and multiple quantized feature spaces," In: Proceedings of the Pan-Sydney area workshop on Visual information processing, Australian Computer Society, Inc., Darlinghurst, **(2004)**, pp. 3-7.
[9] Wang Z., Yang X., Xu Y. and Yu S., "CamShift guided particle filter for visual tracking", Pattern Recognition Letters, vol. 30, no. 4, **(2009)**, pp. 407-413.
[10] Isard M. and Blake A., "Condensation-conditional density propagation for visual tracking", International journal of computer vision, vol. 29, no. 1, **(1998)**, pp. 5-28.
[11] Jung D.W. and Lee C.S., "Exploiting Adaptive Background Image and Dynamic Search Window for Fast Object Tracking", International Journal of Multimedia and Ubiquitous Engineering, vol. 9, no. 1, **(2014)**, pp. 123-132.
[12] Lowe D. G., "Object recognition from local scale-invariant features, In: Computer vision", 1999. The proceedings of the seventh IEEE international conference on, vol. 2, **(1999)**, pp. 1150-1157.
[13] Kisku D. R., Phalguni G. and Jamuna K. S., "Face recognition using SIFT descriptor under multiple paradigms of graph similarity constraints," International Journal of Multimedia and Ubiquitous Engineering, vol. 5, no. 4, **(2010)**, pp. 1-18.
[14] Lim H. Y. and Kang D. S., "Object tracking system using a VSW algorithm based on color and point features", EURASIP Journal on Advances in Signal Processing, vol. 2011, no. 1, **(2011)**, pp. 1-12.
[15] Bay H., Tuytelaars T. and Van G. L., "Surf: Speeded up robust features. In: Computer Vision–ECCV 2006", Springer Berlin Heidelberg, Graz, **(2006)**, pp. 404-417.
[16] Bay H., Ess A., Tuytelaars T. and Van G. L., "Speeded-up robust features (SURF)", Computer vision and image understanding, vol. 110, no. 3, **(2008)**, pp. 346-359.
[17] Rublee E., Rabaud V., Konolige K. and Bradski G., "ORB: an efficient alternative to SIFT or SURF", In: 13th International Conference on Computer Vision (ICCV2011), IEEE Press, Barcelona, **(2011)**, pp. 2564-2571.
[18] Philli P. J., "Support vector machines applied to face recognition", U.S. Department of Commerce, Technology Administration, National Institute of Standards and Technology, vol. 285, **(1998)**.
[19] Vishwanathan S. V. M. and Murty M. N., "SSVM: a simple SVM algorithm. In: Neural Networks, 2002", IJCNN '02. Proceedings of the 2002 International Joint Conference on, Hawaii, vol. 3, **(2002)**, pp. 2393-2398.
[20] Kim M. J., Lim H. Y. and Kang D. S., "Human Motion Tracking using the Image Subtraction and SVM with the Physical Characteristics", ICGHIT2014. International Conference on Green and Human Information Technology, Ho Chi Minh, **(2014)**, pp. 220-223.
[21] Natalya D. and Aizhan T., "Algorithm for detection of image consistent features using surf method", The Kazakh-American Free University Academic Journal, no. 4, **(2012)**, pp. 257-262.
[22] Li W. W., Xing X. X., Liu F. and Zhang Y., "Application of Improved Grid Search Algorithm on SVM for Classification of Tumor Gene", International Journal of Multimedia and Ubiquitous Engineering, vol. 9, no. 11, **(2014)**, pp. 181-188.
[23] http://www.support-vector-machines.org/SVM_osh.html
[24] Pets 2006: http://www.cvg.rdg.ac.uk/PETS2006/data.html

Authors

Ming-Shou An, He received a B.S. degree from Yanbian University, China, in 2007, an M.S. degree from Dong-A University, in 2009. He now is a Ph.D candidate of the Department of Electronic Engineering, Dona-A University, Busan, Korea. His research interests are signal processing and pattern recognition.



Dae-Seong Kang, He received a B.S. degree from Kyungpook National University, Daegu, Korea, in 1984, M.S. degree and D.Sc. degree in electrical engineering from Texas A&M University, in 1991 and 1994, respectively. He is currently full professor of the Department of Electronic Engineering, Dona-A University, Busan, Korea. His research interests are image processing and compression.