Robust Feature Descriptor and Vehicle Motion Model with Tracking-by-detection for Active Safety

Hirokatsu Kataoka, Kimimasa Tamura, Yoshimitsu Aoki
Keio University
Yokohama, Kanagawa, Japan
Email: kataoka@aoki-medialab.org

Yasuhiro Matsui
National Traffic Safety and Environment Laboratory
Chofu, Tokyo, Japan

Kenji Iwata, Yutaka Satoh
National Institute of Advanced Industrial Science and Technology (AIST)
Tsukuba, Ibaraki, Japan

***Abstract*—The percentage of pedestrian deaths in traffic accidents is on the rise in Japan. In recent years, there have been calls for measures to be introduced to protect vulnerable road users such as pedestrians and cyclists. In this study, a method to detect and track pedestrians using an in-vehicle camera is presented to perform braking controls, warn the driver, and develop improved safety systems for pedestrians. We improved the technology of detecting pedestrians using highly accurate images obtained with a monocular camera. We were able to predict pedestrian activity by monitoring the images, and developed an algorithm with which to recognize pedestrians and their movements more accurately. The effectiveness of the algorithm was tested using images taken on real roads. For the feature descriptor, we used an extended co-occurrence histogram of oriented gradients (ECoHOG) that accumulated the integration of gradient intensities. In the tracking step, we applied an effective motion model using optical flow and the proposed feature descriptor ECoHOG in a tracking-by-detection framework. These techniques were verified using images captured on the real road.**

I. Introduction

In Japan, the number of pedestrian deaths fell to 4411 in 2012. Over the next several years, the number is likely to decrease further. However, the percentage of pedestrian deaths among all deaths in traffic accidents is increasing [1]. In an effort to reduce pedestrian deaths, investigations are being conducted in the area of pedestrian intelligent transport systems.

Along these lines, we are currently studying a pedestrian active safety (collision avoidance) system, which is able to detect and track pedestrians by means of an in-vehicle sensor and employ automatic braking. There are great expectations for the system. The use of in-vehicle cameras is efficient in detecting obstacles, and many studies have been devoted to pedestrian detection by means of cameras.

Pedestrians must be identified in outdoor scenes. We must detect pedestrians in the context of a cluttered background and various illuminations. In addition, pedestrians can be occluded by traffic elements, such as signs and vehicles. Researchers have studied active safety technologies for traffic safety [2][3].Geronimo *et al*. used Real AdaBoost to select effective features from Haar-like and Edge Orientation Histograms [4]. Dalal *et al*. presented a human classification feature, called the Histograms of Oriented Gradients (HOG), and a linear support vector machine as a learning classifier. The HOG feature describes the shape of an object from a divided image, called the block and cell. The HOG can roughly represent a human edge feature employing statistical learning. Recently, the HOG has been widely used in the field of computer vision. Additionally, many HOG-based methods have been proposed, and the HOG has recently been improved for highly accurate detection. The HOG method has been improved by combining the HOG [6] with the Co-occurrence Probability Feature (CPF) [7]. These HOG feature and CPF provide the co-occurrence feature using the returning value of classifiers. Haar-like, Haar-wavelet and Edge Orientation Histograms are also used to detect pedestrians on real-world roads. Co-occurrence of Histograms of Oriented Gradients (CoHOG) is known as a state-of-the-art method of pedestrian detection. The CoHOG feature descriptor represents an edge-pair as a histogram for an image patch.

Recently, object detection using stereo cameras has been investigated. The stereo-based method can easily refine a background area and separate it from a pedestrian area. Many methods have adopted a stereo-based approach, e.g. Broggi *et al*[9], Krotosky *et al*[10].

However, because of the high costs of stereo technology, a monocular camera technique is preferable. The use of monocular cameras as driving recorders is on the rise, and the cameras are frequently used, for example, in taxis around the world. Accordingly, we devoted our efforts to developing a pedestrian active safety technique using a monocular camera.

In this paper, we present highly accurate pedestrian detection and tracking techniques using a monocular camera in a tracking-by-detection framework. An improved version of CoHOG [8] is used for detection, and an optical-flow-based motion model is implemented for tracking. An in-vehicle video is captured using a camera attached to a rearview mirror. The system detects any pedestrians that appear in the vehicle's path. After detection and tracking, the system sends a warning signal to the driver or it automatically engages the brake.

A signal is emitted by the vehicle to warn the pedestrian. The system is able to judge the situation and the human

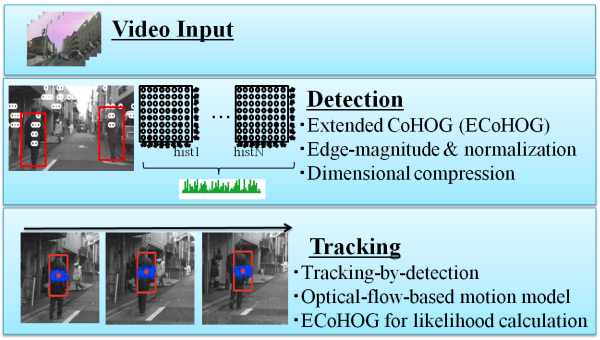

Fig. 1. The process flow of the pedestrian active safety system

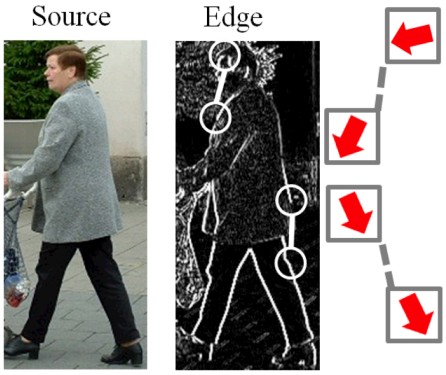

Fig. 2. How to acquire an edge pair from a pedestrian shape

reaction, and if it determines that there is the possibility of collision, it engages the brake after warning the driver. By avoiding collisions, the system is able to reduce the incidence of traffic accidents. Owing to the recent increase in the use of monocular in-vehicle cameras, such as in taxis or buses, it would be beneficial to use such cameras to detect pedestrians.

II. Proposed Method

Figure 1 shows the sequence of events for our proposed method. A video is made of the area in front of the vehicle by means of the in-vehicle camera. The current method for capturing a pedestrian's shape employs co-occurrence histograms of oriented gradients (CoHOG) [8]. In this paper, however, we propose the use of extended CoHOG (ECoHOG). We implemented tracking-by-detection with a particle filter [12] for pedestrian tracking. Optical flow [13] is used for the vehicle motion model that acquires the movement of a vehicle. The ECoHOG is applied to calculate likelihood from particles. To test the applicability of ECoHOG, we compared it with CoHOG using commonly used datasets. In preprocessing, we applied symmetrical judgment proposed by Kataoka *et al.* [14], who proposed the method for fast processing using the degree of symmetry between left and right images of a human, considering that a human body shape has symmetry. The thresholding value is set using a large number of positive and negative learning images.

A. Extended Co-occurrence Histograms of Oriented Gradients (ECoHOG)

We used ECoHOG as an improvement of CoHOG [8] for pedestrian detection.

For pedestrian-shape extraction, CoHOG investigates the pedestrian's edge direction minutely and uses two different pixels in an extraction window as a direction pair. From the edge direction pair, CoHOG can describe such features as the head and shoulders and the back and legs (shown in Figure 2) and reduce misdetection by means of pair extraction. Unlike CoHOG, ECoHOG acquires not only the edge-direction pair, but also an edge-magnitude pair. ECoHOG can represent the edge-magnitude characteristics of the pedestrian shape.

ECoHOG: this approach is a possible edge-magnitude accumulation method that can be used to generate the pedestrian shape model in minute detail. The edge-magnitude-accumulation method can describe detail more comprehensively than the CoHOG feature descriptor. In other words, ECoHOG considers not only the pedestrian shape features but also the degree of edge magnitude that can describe "shape strength". We believe the addition of the edge-magnitude pair is an effective way to classify the pedestrian and background with high accuracy. Equations relating to ECoHOG are

$$m_1(x_1,y_1) = \sqrt{f_{x1}(x_1,y_1)^2 + f_{y1}(x_1,y_1)^2} \quad (1)$$

$$m_2(x_2,y_2) = \sqrt{f_{x2}(x_2,y_2)^2 + f_{y2}(x_2,y_2)^2} \quad (2)$$

$$f_x(x,y) = I(x+1,y) - I(x-1,y) \quad (3)$$

$$f_y(x,y) = I(x,y+1) - I(x,y-1) \quad (4)$$

$$C_{x,y}(i,j) = \sum_{p=1}^{n}\sum_{q=1}^{m} \begin{cases} m_1(x_1,y_1) + m_2(x_2,y_2) \\ (if\ d(p,q) = i\ and \\ d(p+x,q+y) = j) \\ \\ 0 \\ (otherwise) \end{cases} \quad (5)$$

where $m_1(x,y)$ and $m_2(x,y)$ are the magnitude of the pixel of interest (at the center of the window) and the magnitude in the objective window, and $f_x(x,y)$ and $f_y(x,y)$ are the differences between two pixels in the x and y directions. $C(i,j)$ is the co-occurrence histogram that accumulates pairs of the pixel of interest and an objective pixel. Each $C(i,j)$ has 64 dimensions because the direction is quantized into eight possibilities and ECoHOG combines two directions. Coordinates (p,q) indicate the pixel of interest (center of window) and coordinates $(p+x,p+y)$ indicate the objective pixel. m and n are the width and height of the feature extraction window. $d(p,q)$ is a function that quantizes the edge direction as an integer from 0 to 7 at pixel (p,q).

ECoHOG normalizes the co-occurrence histogram to set off the difference in edge-magnitude due to image brightness. The

co-occurrence histogram is divided by its norm:

$$C'_{x,y}(i,j) = \frac{C_{x,y}(i,j)}{\sum_{i'=0}^{7}\sum_{j'=0}^{7} C_{x,y}(i',j')} \quad (6)$$

Dimensional Compression with Principal Component Analysis (PCA-ECoHOG): PCA is used for dimensional compression. We believe that a low-dimensional feature is easy to divide into positive and negative classes. In related work, CoHOG needed about 35,000 dimensions for pedestrian detection [8]; however, we propose a low-dimension feature with PCA. Employing this method, we compressed the feature dimensions from 4600 to a few hundred. The PCA comparison experiment shows the effective number of dimensions for pedestrian detection. In the field of pattern recognition, high-dimensional feature spaces are referred as "the curse of dimensionality" [15]. The problem is to increase feature dimensions according to information that can be taken from objectives. We believe that data reduction is the best way to solve this problem and improve the accuracy of recognition.

B. Tracking-by-detection in Particle Filter Algorithm

Recently, the tracking-by-detection framework has been found to be effective in the field of computer vision. In this section, we describe the latest works on likelihood calculations using ECoHOG and a vehicle motion model. The searching and optimization are performed using a particle filter [12]. A particle filter is an iterative algorithm including prediction, update and resampling, and is a type of Bayesian estimator that applies the Monte Carlo framework. The current state is estimated as the expectation of the likelihood distribution. The likelihood calculation evaluates the condition of an object by comparing texture between the model image and an image around a particle.

Step1 Initialization

We execute initialization after specifying a pedestrian's position using a detector. The system puts the tracker on a detected point. We arrange particles around the pedestrian and capture the pedestrian's texture as a tracking model. A particle has a motion model and observation model. The motion model is based on vehicle motion obtained from the front video, and the observation model is calculated by the ECoHOG detector. These are shown in the step2 state prediction and step3 likelihood calculation.

Step2 State prediction

We apply the vehicle movement model as a motion model. Vehicle movement is calculated using Lucas-Kanade optical flow [13] that tracks a feature point between spatio-temporal frames as seen in Figure 3. The average of the optical flow direction and magnitude in a frame is the movement model. Furthermore, we add Gaussian noise to a particle's velocity and location. We estimate from the linear uniform motion if there is no vehicle movement:

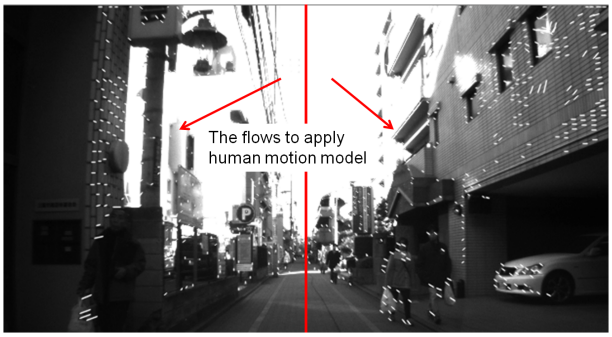

Fig. 3. Flow vector calculation from driver's view

$$X_{t+1} = FX_t + w_t \quad (7)$$

$$X_t = (x, y, v_x, v_y)^{\mathrm{T}} \quad (8)$$

$$F = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad (9)$$

where x and y are the location of the particle, and v_x and v_y are the velocity between frames. w_t is the system noise added to the particle location and velocity.

Step3 Likelihood calculation

The pedestrian model is important for pedestrian tracking. We believe that a pedestrian model based on machine learning can be a universal pedestrian model in the likelihood calculation.

We apply a Real AdaBoost classifier by learning the proposed feature descriptor ECoHOG in the tracking step. Real AdaBoost outputs the evaluated value as a real number from the learning sample. In this situation, the learning sample is distinguished as a positive (pedestrian) and negative (background) image. We prepare a large number of learning images for pedestrian recognition. Here we show the value of the weak classifier and the value of the strong classifier accumulating weak classifiers in the Real AdaBoost algorithm:

$$h(x) = \frac{1}{2}\ln\frac{W_{positive}+\epsilon}{W_{negative}+\epsilon} \quad (10)$$

$$H(x) = \sum_{t=1}^{T} h_t(x) \quad (11)$$

where h(x) and H(x) are respectively the returning values of the weak and strong classifier. $W_{positive}$ and $W_{negative}$ are the weight values of Real AdaBoost. ϵ is set as $\frac{1}{N}$ using the number of learning samples N.

A rectangle area around a particle is input to calculate the likelihood. We can obtain the likelihood with a classifier in the rectangle area. The likelihood is applied to capture a pedestrian.

Step 4 Likelihood evaluation

The center of gravity is calculated in this step:

$$(g_x, g_y) \quad = \quad (\sum_{i=1}^{n} L_i x_i, \sum_{i=1}^{n} L_i y_i) \tag{12}$$

where g is the center of gravity of the pedestrian. L is each particle's likelihood and x and y are the position of the particle.

III. Experiment

We carried out experiments to validate the detection and tracking approaches. Experiment 1 was conducted to choose the feature descriptor, and experiment 2 to verify the tracking methods. We applied the INRIA person dataset as a benchmark for experiment 1, and real-road captured videos for experiment 2.

First, we describe how to implement the feature descriptor and tracking-by-detection.

A. Implementation

The parameters and detailed implementations are given below. We describe ECoHOG and tracking-by-detection.

ECoHOG: In this method, we set the parameters the same as for the CoHOG feature descriptor. The window size, image block division, number of offsets, histogram division, and dimensions should be given to detect pedestrians. The parameters of ECoHOG are given in Table 1.

Tracking-by-detection: The tracking-by-detection and its machine learning are described here. In the particle filter algorithm, the number of particles, motion model, and likelihood calculation should be input to capture the objective movement. Additionally, the extraction size is input for calculating ECoHOG. To construct Real AdaBoost, we set the number of weak classifiers and learning samples. Table 2 gives the settings of tracking-by-detection and machine learning.

The tracking algorithm is implemented on a standard laptop (Windows 7, Intel Core i7-2620M, 2.70 GHz CPU, 4.0 GB RAM). We applied C++ and OpenCV2.3 programming for pedestrian tracking.

TABLE I
Parameters of ECoHOG for pedestrian detection

Window size	7×7 pixels
Block division	4 blocks (2×2)
Number of offsets	18 offsets
Histogram division	8 directions
Dimensions	4608 dimensions
Dimensions(after PCA)	100 dimensions

B. Experiment 1: Feature Descriptor Verification using the INRIA Person Dataset

We apply the INRIA person dataset [16], which includes pedestrian images and a Detection Error Tradeoff (DET) curve for the verification. The number of pixels for feature extraction is 18 pixels. ECoHOG has 4608 dimensions, and PCA-ECoHOG 100 dimensions.

TABLE II
Settings of tracking-by-detection and machine learning for pedestrian tracking

Number of particles	150
Extraction size	80×160 pixels
Learning samples	3,000 positive images 20,000 negative images
Motion model	Optical flow
Likelihood	ECoHOG classifier
Weak classifier	50 classifiers

First, we compare the proposed and previous frameworks in Figure 4. Figure 5 shows the comparison with PCA-ECoHOG and ECoHOG. Figure 6 shows the relationship between the number of dimensions and the detection rate. The vertical axis of the DET curve is miss rate, and the horizontal axis is the false positive rate; therefore, the bottom-left of the DET curve shows higher performance.

TABLE III
Processing time of feature descriptors (ECoHOG, PCA-ECoHOG)

Feature descriptor	Processing time
ECoHOG	11.2 ms
PCA-ECoHOG	12.9 ms

Figure 4 shows that the proposed framework provides the best performance; ECoHOG accumulates the edge magnitude and normalizes an image. ECoHOG is thus much more effective than CoHOG.

In the comparison of ECoHOG and PCA-ECoHOG (Figure 5), PCA-ECoHOG proves the best method for pedestrian detection. It is easy to divide the feature space into two classes because PCA-ECoHOG compresses the number of dimensions from 4608 to 100. Table 1 gives the processing times of ECoHOG and PCA-ECoHOG for each frame. Both methods realized fast processing for the INRIA person dataset.

Figure 6 shows the relationship between the dimension number and detection rate. The number of dimensions is compressed by a factor of 5-200 (i.e., 5, 7, 15, 20, 50, 100, 200 dimensions) in the figure. This experiment shows that the best number of dimensions is 100. Five dimensions provide little information, whereas 200 dimensions provide a large feature space.

C. Experiment 2: Effectiveness of Tracking in Real Scenes

We carried out a tracking experiment using our real-road dataset. The objective is to track a single pedestrian without occlusion or the crossing of other pedestrians. We applied three related works: (i) grayscale histogram matching + linear uniform motion, (ii) texture matching + linear uniform motion, and (iii) texture matching + vehicle motion model are compared as related approaches. Additionally, the proposed approach (iv) ECoHOG classifier + vehicle motion model was applied. The initial position was set manually and the histogram or texture was extracted from the initial rectangle in related works. Our proposed framework includes a pedestrian

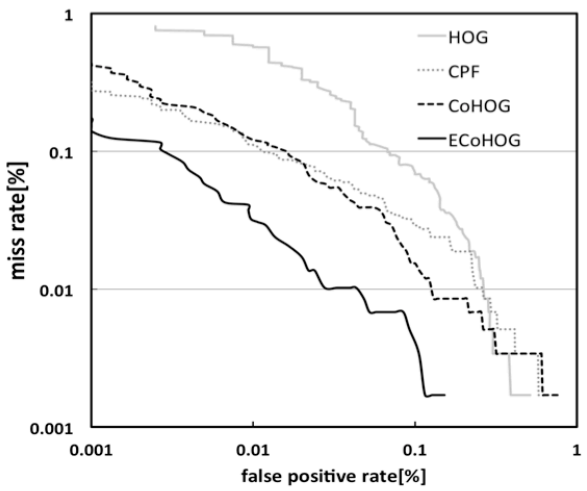

Fig. 4. Comparison of four feature descriptors applying the Detection Error Tradeoff (DET) curve : ECoHOG, CoHOG, CPF, HOG

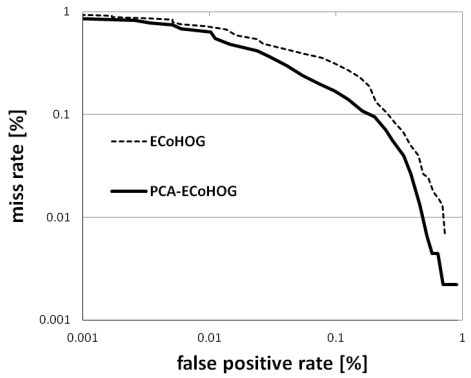

Fig. 5. Two proposed methods

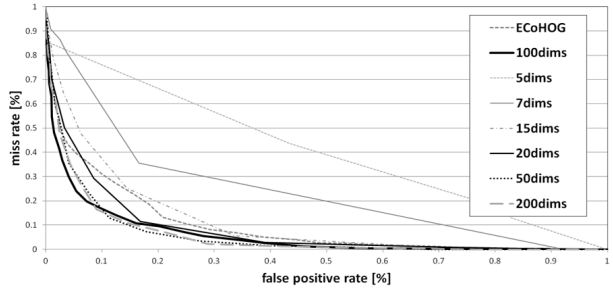

Fig. 6. PCA-ECoHOG

TABLE IV
RESULT OF TRACKING EXPERIMENT: (I) GRAYSCALE HISTOGRAM MATCHING + LINEAR UNIFORM MOTION, (II) TEXTURE MATCHING + LINEAR UNIFORM MOTION, AND (III) TEXTURE MATCHING + VEHICLE MOTION MODEL ARE COMPARED AS RELATED APPROACHES. (IV) ECOHOG CLASSIFIER + VEHICLE MOTION MODEL (PROPOSED APPROACH)

Method	Accuracy	Frame per seconds
(i)	18.7%	62.42
(ii)	72.3%	52.00
(iii)	87.2%	26.99
(iv)	99.2%	5.85

model obtained by learning positive and negative samples; therefore, we do not need to capture the pedestrian model in the initial frame. Table 4 shows the result of the tracking experiment and Figure 7 shows the tracking for the real-road dataset.

The grayscale histogram does not evaluate human likelihood effectively. This result arises because the grayscale histogram ignores the location of the human's texture of pose and close. The histogram feature is confusing unless it is color depth. Furthermore, the grayscale value changes dramatically in real time. We applied texture matching including location information from the pedestrian. In this case, the time-series texture of the pedestrian has few changes. Texture matching is a better way to capture a pedestrian from in-vehicle video. However, texture accumulates a gap due to a changing angle and scale if tracking continues for a long period. We need to fix the template using a point-of-interest tracker; for example, the SIFT (Scale Invariant Feature Transform) [17] or SURF (Speeded Up Robust Features) [18].

Our proposed approach had the best accuracy of 99.2% among the examined methods. The approach includes a general pedestrian model in the classifier obtained from positive and negative samples. According to calculation of the processing time, our proposed approach can track near to real time. Although the processing time of 5.85 fps is fast, we require program optimization with in-vehicle hardware to process in real time. Additionally, we should implement a fast and accurate recognition method on the software side.

Figure 8 shows the likelihood transition under pedestrian tracking. The black line and gray line show the results of tracking-by-detection and texture matching. Both methods apply a vehicle motion model. Tracking-by-detection outputs a higher value in most cases. The pedestrian model obtained by machine learning allows us to keep a likelihood in tracking. The many variations of the image are included in the machine learning step; for example, occluded situations, complicated backgrounds, and dark or bright scenes. Machine learning allows tracking in human-human occluded scenes.

IV. CONCLUSION

This paper proposed a pedestrian tracking method for pedestrian active safety. The contributions of this approach are (i) a state-of-the-art detector that is a combination of ECoHOG and Real AdaBoost, and (ii) tracking-by-detection in the particle

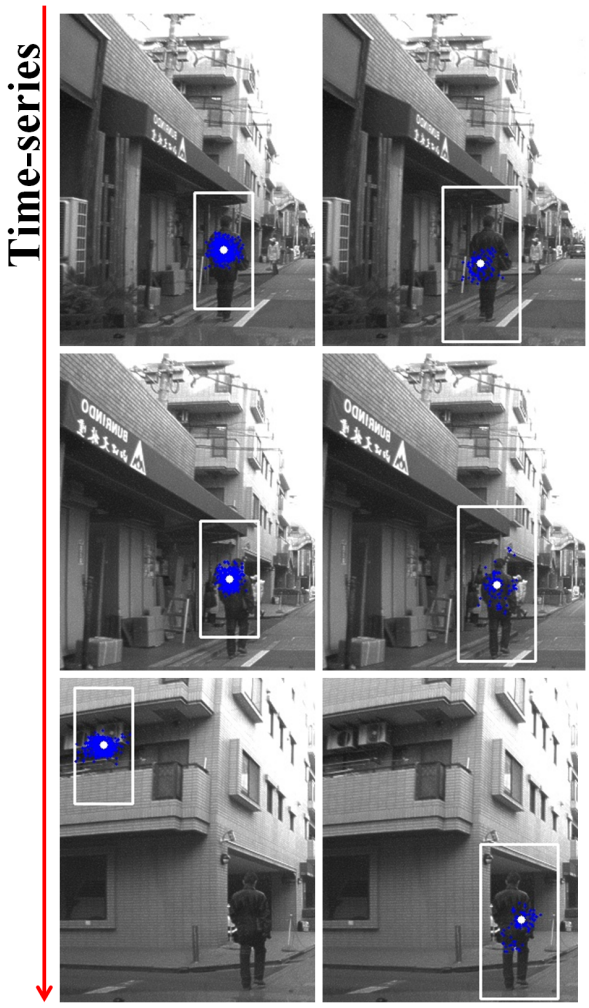

Fig. 7. Pedestrian tracking-by-detection: (left) texture matching and vehicle motion model (right) ECoHOG + Real AdaBoost classifier and vehicle motion model.

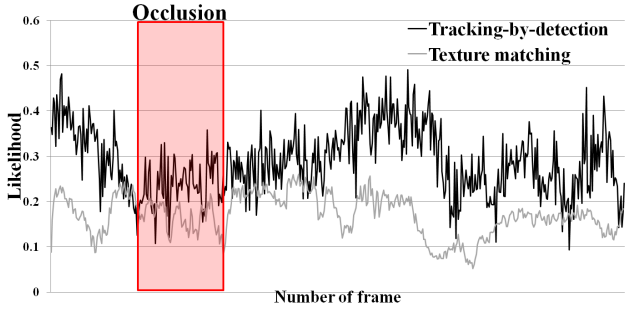

Fig. 8. Likelihood transition under pedestrian tracking.

filter framework using the ECoHOG classifier and a vehicle motion model.

In future work, a tracking method for situations of occlusion and multiple people should be considered. Our current pedestrian model includes only a person model obtained from machine learning. We must consider occlusion and a multi-person model in pedestrian tracking. Furthermore, we will attempt "pedestrian motion prediction" based on an estimation algorithm. The prediction framework is important for the development of advanced active safety systems.

REFERENCES

[1] "Institute for Traffic Accident Research and Data Analysis of Japan (ITARDA)", Annual Traffic Accident Report in 2010, Tokyo, 2011.

[2] Tarak Gandhi, Mohan M. Trivedi, "Pedestrian Collision Avoidance Systems:A Survey of Computer Vision Based Recent Studies", IEEE Intelligent Transportation Systems Conference (ITSC2006), pp.976-981, 2006.

[3] David Geronimo, Antonio M. Lopez, Angel D. Sappa, Thorsten Graf, "Survey of Pedestrian Detection for Advanced Driver Assistance Systems", IEEE Trans. Pattern Analysis and Machine Intelligence, Vol.32, No.7, 2010.

[4] D. Geronimo, A. Sappa, A. Lopez, and D. Ponsa, "Adaptive Image Sampling and Windows Classification for On-Board Pedestrian Detection", Int' Conf. Computer Vision Systems, 2007.

[5] N. Dalal, B. Triggs, "Histograms of oriented gradients for human detection", IEEE Conference on Computer Vision and Pattern Recognition (CVPR2005), pp886-893, 2005.

[6] T. Mitsui, Y. Yamauchi, and H. Fujiyoshi, "Object Detection by Two-Stage Boosting with Joint Features", The Institute of Electronics, Information and Communication Engineers Transactions on Information and Systems,Vol. J92-D, No. 9, pp. 1591-1601, 2009.

[7] Y. Yamauchi, H. Fujiyoshi, Y. Iwahori, and T. Kanade, "People Detection based on Co-occurrence of Appearance and Spatio-Temporal Features", National Institute of Informatics Transactions on Progress in Informatics, No. 7, pp. 33-42, 2010.

[8] Tomoki Watanabe, Satoshi Ito, Kentaro Yokoi, "Co-occurrence Histograms of Oriented Gradients for Pedestrian Detection", PSIVT2009, LNCS vol.5414, pp37-47, 2009.

[9] A. Broggi, M. Bertozzi, A. Fascioli, and M. Sechi, "Shape-Based Pedestrian Detection", IEEE Intelligent Vehicles Symp., pp.215-220, 2000.

[10] S. Krotosky and M.M. Trivedi, "Multimodal Stereo Image Registration for Pedestrian Detection", IEEE Int Conf. Intelligent Transportation Systems, pp.109-114, 2007.

[11] A. Bhattacharyya,"On a measure of divergence between two statistical populations defined by their probability distributions", Bulletin of the Calcutta Mathematical Society, 35, pp99-109, 1943.

[12] Isard M, Blake A, "Condensation : conditional density propagation for visual tracking", International Journal of Computer Vision, vol.28, no.1, pp.5-28, 1998.

[13] B.D.Lucas, Takeo Kanade, "An iterative image registration technique with an application to stereo vision", Proceedings of Imaging Understanding Workshop, pp.121-130, 1981.

[14] Hirokatsu Kataoka, Yoshimitsu Aoki, Yasuhiro Matsui, "Symmetrical Judgment Area Reduction and ECoHOG Feature Descriptor for Pedestrian Detection", International Journal of Vehicle Safety, 2012.

[15] Charu C. Aggarwal, "On k-anonymity and the curse of dimensionality", VLDB, pp.901-909, 2005.

[16] INRIA Person Dataset, http://pascal.inrialpes.fr/ data/human

[17] David G. Lowe, "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, 60, 2 (2004), pp. 91-110

[18] Herbert Bay, Andreas Ess, Tinne Tuytelaars, Luc Van Gool "SURF: Speeded Up Robust Features", Computer Vision and Image Understanding (CVIU), Vol. 110, No. 3, pp. 346–359, 2008