Fine-grained Walking Activity Recognition via Driving Recorder Dataset

Hirokatsu Kataoka (AIST), Yoshimitsu Aoki (Keio Univ.), Yutaka Satoh (AIST)
Shoko Oikawa (NTSEL), Yasuhiro Matsui (NTSEL)
Email: hirokatsu.kataoka@aist.go.jp
http://hirokatsukataoka.net/

***Abstract*—The paper presents a fine-grained walking activity recognition toward an inferring pedestrian intention which is an important topic to predict and avoid a pedestrian's dangerous activity. The fine-grained activity recognition is to distinguish different activities between subtle changes such as walking with different directions. We believe a change of pedestrian's activity is significant to grab a pedestrian intention. However, the task is challenging since a couple of reasons, namely (i) in-vehicle mounted camera is always moving (ii) a pedestrian area is too small to capture a motion and shape features (iii) change of pedestrian activity (e.g. walking straight into turning) has only small feature difference. To tackle these problems, we apply vision-based approach in order to classify pedestrian activities. The dense trajectories (DT) method is employed for high-level recognition to capture a detailed difference. Moreover, we additionally extract detection-based region-of-interest (ROI) for higher performance in fine-grained activity recognition. Here, we evaluated our proposed approach on "self-collected dataset" and "near-miss driving recorder (DR) dataset" by dividing several activities– *crossing*, *walking straight*, *turning*, *standing* and *riding a bicycle*. Our proposal achieved 93.7% on the self-collected NTSEL traffic dataset and 77.9% on the near-miss DR dataset.**

I. Introduction

According to the world health organization (WHO), the number of traffic deaths is approximately 1.24 million occurred worldwide in 2010 [1]. Specially, the task for pedestrian is on the alarming problem. The pedestrian deaths of traffic accident are likely to decrease further over the next several years. In an effort to decrease pedestrian deaths, examinations are being conducted in the area of intelligent transport systems (ITS) for pedestrian. Along these lines, we are currently trying to implement a system for pedestrian active safety, which is able to detection of pedestrian by means of a mounted camera. Up to date, pedestrians have been identified in outdoor scenes and from a cluttered background with various illuminations. In addition, pedestrians can be occluded by traffic elements, such as signs and vehicles. There are several surveys about pedestrian detection in traffic scenes. For example, Geronimo *et al.* [2], and Dollar *et al.* [3] have enumerated approaches of how to accurately detect pedestrian(s) on real-road. More recently, Benenson *et al.* [4] listed over 40 approaches and compared all of them on pedestrian detection datasets. According to their paper, pedestrian detection systems have been greatly improving in this decade. We have coped with a problem for pedestrian detection, however, it is preferable to infer a pedestrian intention in advance. Here, we believe a change of pedestrian's activity is important to infer a pedestrian intention. This is the reason why pedestrian activity should be recognized in traffic scenes. However, the recognition of pedestrian activity is extremely difficult from a couple of reasons. (i) In-vehicle mounted camera is always moving. Vision-based approach is weak in camera-motion because of it tends to be a cluttered background and a feature descriptor does not capture a stable motion feature. (ii) A pedestrian area is too small to capture a motion and shape features. In a small pedestrian area, a image-based feature such as HOG [5] cannot be derived clearly by using itself. We need to improve the framework for ITS domain. (iii) A change of pedestrian activity (e.g. "walking straight" changing into "turning") has only fine-grained feature difference. Here, Figure 1 shows the difference among pedestrian activities. Clearly, the feature description tends to be quite similar from the same person on the self-collected dataset. In particular, the three activities of "crossing", "walking straight" and "turning" are becoming fine-grained recognition from the similar shape, pedestrian's scale changing and these three are the same category as "walking".

Fig. 1. Four different pedestrian activities on the self-collected dataset: (a) crossing (b) walking straight (c) turning (d) riding a bicycle.

In this paper, we propose fine-grained pedestrian activity recognition which is categorization task of subtle pedestrian's feature differences toward walking intent inferring. The paper mainly focuses on the fine-grained activity recognition, however, we need to get a location of pedestrian in an image sequence. Therefore, the process flow of fine-grained activity recognition is consist of two parts including "localization" and "activity recognition". In a localization part, we implement pedestrian detection based on Kataoka *et al.* [6]. The detection system allows us to put spatio-temporal pedestrian locations. In activity recognition part, the framework of dense trajectories (DT) [7] is employed to capture a detailed pedestrian motion. The basic idea of DT is to extract a large number of trajectories and execute feature description on the trajectories. Feature types include trajectory- motion- and shape-based descriptors.

The features are captured on the trajectoires inside of a pedestrian region given by the detection system. Although the activity recognition approach is based on DT [7], we employ improved DT [12] for noise canceling with camera motion estimation.

The contributions are two folds: (i) to the best of our knowledge, this is the first trial toward fine-grained pedestrian activity recognition for walking intent inferring; (ii) self-collecting fine-grained pedestrian activity datasets for validation and real-road recognition.

II. Related Works

The related works contain pedestrian detection and human activity recognition. The works are enumerated here;

Pedestrian detection. Since Dalal *et al*. presented the histograms of oriented gradients (HOG) [5], the pedestrian detection have been an active topic in computer vision. The HOG method has been improved into the joint HOG [14] and the Co-occurrence Probability Feature (CPF) [15]. As shown in the paper [3], Walk *et al.* [17] proposed state-of-the-art approach in pedestrian detection for active safety. In [17], they proposed improvement of edge-based feature (e.g. Haar-like[16], shapelets[18], and shape context[19]) by combining color self-similarity and the motion features from optic flow. However, this approach cannot be realized in real-time whose properties are obviously required for traffic safety systems. At this point, Co-occurrence of Histograms of Oriented Gradients (CoHOG) is known as a high-standard detection approach for pedestrian detection, by representing edge pair [20]. Moreover, CoHOG achieved a real-time detection on road-scene dataset due to individual processing at each block, which is an effective approach in multi-core systems such as in-vehicle hardware.

Human activity recognition. Several feature descriptors have been proposed to understand difficult activities; Laptev *et al.* [8] proposed space-time interest points (STIP) as an improvement of Harris corner detector. The STIP represents 3D (XYT) interested motion points to extract the same feature in an activity. This framework is a well-organized recognition methodology with temporal localization. Klaser *et al.* [9] proposed 3D-HOG and Marszalek *et al.* [10] described feature combination by using STIP framework. Chaudhry *et al.* implemented a remarkable feature descriptor on a benchmark using a time series optical flow based method [11].

Recently, the method of "dense trajectories (DT)" is proposed by Wang *et al.* [7], which is feature description on dense sampling feature points in an image sequence. DT approaches have also been proposed in [25], [26], [27], [13], [12] after proposal of the original one [7]. Raptis *et al.* tried to generate a middle-level representation yielding simple posture with location clustering [25]. To highly eliminate extra optical flows, Jain *et al.* applied affine matrix [26] and Peng *et al.* proposed dense optical flows capture on a motion boundary space [27]. Kataoka *et al.* improved DT feature by adding co-occurrence feature descriptor and dimensional compression. Wang *et al.* realized improved DT [12] by adding camera motion estimation, detection-based noise canceling, and Fisher vector [28].

Pedestrian dataset. Moreover, we here list several pedestrian datasets. At the beginning, Dalal *et al.* published INRIA person dataset [5] which includes pedestrian images in the wild. However, the dataset contains only reasonable scenes captured by digital cameras. Traffic-specified dataset is Daimler pedestrian benchmark [21]. Especially, the dataset employs small images of 18×36 pixels by assuming pedestrians far from in-vehicle camera. More natural traffic scenes are in the Caltech pedestrian detection benchmark [22]. INRIA person dataset and Daimler pedestrian benchmark are only including still images, however, Caltech pedestrian detection benchmark is consist of 10 hours of videos in real-road. The most advanced benchmark is the KITTI vision benchmark suite which focuses on a autonomous driving system. Several sensors are mounted on a vehicle such as GPS, Velodyne laser sensor not only stereo vision system. Although the datasets have natural traffic scenes, near-miss scenes are necessary to analyze and recognize a dangerous activity.

On one hand, the society of automotive engineering of Japan (JSAE) released the near-miss incidents dataset [24]. We focus on the pedestrian's near-miss scenes in order to create fine-grained pedestrian activity dataset in a crucial situation.

III. Proposal

In this section we present our recognition framework toward fine-grained pedestrian activity recognition. Figure 2 indicates the framework of our proposal. The basic idea is to accumulate trajectory-based feature descriptor inside of a localized pedestrian. The localization and activity analysis approaches are *extended CoHOG + AdaBoost* and *dense trajectories*, respectively.

A. Pedestrian localization with co-occurrence feature [6].

The extended CoHOG (ECoHOG) is implemented as an improvement of CoHOG [20] 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-shoulder pair and the back-leg pair, which reduces miss detection. Unlike CoHOG, ECoHOG acquires not only the edge-direction pair, but also an edge-magnitude pair. Extended co-occurrence histogram of oriented gradient(ECoHOG) enable to be a more detailed description in a human shape. ECoHOG is an improvement descriptor from CoHOG with edge-magnitude to extract a human-like shape and its magnitude. CoHOG takes features with count of co-occurrence edge-pair. However, the description may be confusing if a human is in a cluttered scene. Against the counting co-occurrence edge-pair, ECoHOG represents co-occurrence features by weighting of edge-magnitude as below:

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

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

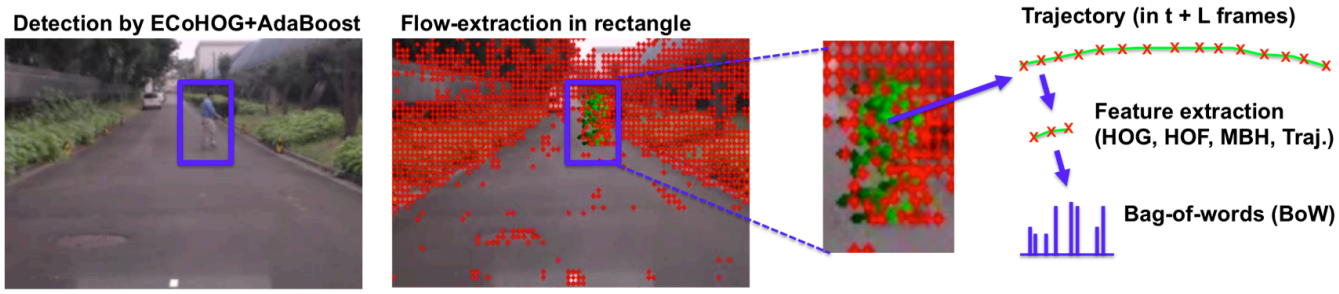

Fig. 2. Framework of our proposal.

$$C_{x,y}(i,j) = \sum_{p=1}^{n}\sum_{q=1}^{m} \begin{cases} m_1(x_1,y_1)+m_2(x_2,y_2) \\ (if\ d(p,q)=i\ and \\ d(p+x,q+y)=j) \\ \\ 0 \\ (otherwise) \end{cases} \quad (3)$$

$m_1(x,y)$ 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, respectively.

Moreover, classification is based on real-AdaBoost classifier and the proposed feature descriptor ECoHOG in the localization step. Real-AdaBoost outputs an evaluated value as a real number with a learning sample. A learning sample is distinguished either a positive (pedestrian) or negative (background) image. We prepare a large number of learning images for pedestrian recognition.

In this situation, we captured DR information which include low resolution videos. The computer vision approaches is not always well done in the environment. In the case of occurring miss-detection, we adjust pedestrian positions by hand. At the first step of fine-grained pedestrian activity recognition, we need to get a fixed location of pedestrian at each frame.

B. Pedestrian activity analysis with dense trajectories (DT) [7].

We employ Wang's dense trajectories [7] (DT) to create bag-of-words (BoW) vectors [29] for activity recognition. The idea of DT includes dense sampling of an image and extraction of spatio-temporal features from the trajectories. Feature points at each grid cell are computed and tracked with Farneback optical flow in the other images of the video. To take care of scale changes, the DT extracts dense flows in multiple image scales, where the image size increases by a scale factor $1/\sqrt{2}$. In DT flow among the frames concatenates corresponding parts of the images. This setup allows us to grab a detailed motion at the specified patch. The length of trajectory is set as 15 frames, therefore, we records 0.5 seconds activities at 30 fps video.

In activity analysis, trajectory feature (traj.) [7], histograms of oriented gradients (HOG) [5], histograms of optical flow (HOF) [31] and motion boundary histograms (MBH) [30] are extracted at each image patch. The size of patch is 32×32 pixels which is divided into 2×2 blocks. The number of traj., HOG, HOF and MBH dimension are 14, 96, 108, and 96, respectively. HOG and MBH are consist of 2 (x) × 2 (y) × 3 (t) × 8 (directions) for the final description size. HOF is described by 2 (x) × 2 (y) × 3 (t) × 9 (directions) quantization. Here, trajectory feature (T) is calculated as below:

$$T = \frac{(P_t,...,P_{t+L-1})}{\Sigma_{j=t}^{t+L-1}||P_j||} \quad (4)$$

$$P_t = (P_{t+1}-P_t) = (x_{t+1}-x_t, y_{t+1}-y_t) \quad (5)$$

where L is the number of trajectory length.

Basically, the DT features are divided into the visual words representing BoW [29] by means of k-means clustering. In our implementation, the DT features are clustered into visual words with Gaussian mixture model (GMM). Here, k-means clustering has the equation as below:

$$\Sigma_{j=1}^{N_c}\Sigma_{i=1}^{N_v} min_{\mu\in C}(||x_j-\mu_i||^2) \quad (6)$$

where N_c and N_v are the number of cluster and feature vector. μ is the centroid of a cluster and x indicates the feature vector. GMM-based clustering shows:

$$\Sigma_{i=1}^{N_c}\pi_i\mathcal{N}(x|\mu_i,\Sigma_i) \quad (7)$$

where π means the parameter of Gaussian mixing coefficient. The difference between k-means and GMM is considering variance of feature vectors at a cluster. The GMM model has mean and variance of cluster, however, the k-means model is dividing clusters with centroid Voronoi tessellation. In this vectorization, activities of dataset is represented with the GMM-based BoW vector (GMM-BoW), containing the frequency of the visual words in activity videos.

IV. Experiment

In this section, we carry out an experiment of fine-grained activity recognition on NTSEL dataset (one-person example is in Figure 1) and near-miss DR dataset [24]. The section is consist of four steps, dataset description, experimental settings, results and feature vector analysis.

A. Datasets

NTSEL dataset. We have experimentally collected 100 videos with four activities in traffic scenes. The activities include *walking*, *crossing*, *turning*, and *riding a bicycle*, where

TABLE I. THE PERFORMANCE RATE WITH DENSE TRAJECTORIES ON THE NTSEL DATASET.

Descriptor	Rate (%)
Trajectory	76.5
HOF(Histograms of Optical Flow)	**93.7**
HOG(Histograms of Oriented Gradients)	85.6
MBHx(Motion Boundary Histograms– x dir.)	87.7
MBHy(Motion Boundary Histograms– y dir.)	86.7

TABLE II. THE PERFORMANCE RATE WITH DENSE TRAJECTORIES ON THE NEAR-MISS DR DATASET (AT 3-FRAME FEATURE ACCUMULATION).

Descriptor	Rate (%)
Trajectory	**77.9**
HOF(Histograms of Optical Flow)	75.9
HOG(Histograms of Oriented Gradients)	76.4
MBHx(Motion Boundary Histograms– x dir.)	59.3
MBHy(Motion Boundary Histograms– y dir.)	60.8

all of the activities indicate fine-grained pedestrian motion with three persons. The dataset contains cluttered background in small areas, therefore, the dataset has difficulties to capture flows. Presented activities are also fine grained since there is small variation between *walking*, *crossing* and *turning* which happen at almost similar motions of the pedestrians. We evaluated this dataset with 5-fold cross validation.

Near-miss DR dataset [24]. The society of automotive engineering of Japan (JSAE) is providing the Hiyari-Hatto database which includes ten-thousands order near-miss incidents. The database contains visual sources, GPS (global positioning system) and CAN (controller area network). We focused on pedestrian to analyze its fine-grained activities in real-road information. The 50 movies are collected in this experiments. The driving recorders are attackhed on a moving vehicle, therefore we set four activities– *walking*, *crossing*, *standing* and *riding a bicycle*. Moving camera makes the computer vision problem difficult such as motion blur, relative motion between vehicle and pedestrian. We collected 15 (walking), 43 (crossing), 13 (standing) and 11 (riding a bicycle) videos, respectively. We calculated performance rate with 5-fold cross validation.

B. Experimental settings

Pedestrian localization. We adjust the parameters of detection descriptor same as [6]. The window size (7×7 pixels), block division (2×2 blocks), histogram division (8 directions) are set for extended feature. Although the number of dimension is 4608, a low-dimensional feature space is preferable to divide two classes between positive (pedestrian) and negative (background). To create a real-AdaBoost classifier, we input 11,000 positive and 75,000 negative samples. The 100 weak classifiers are applied. Currently, vision-based detection is not perfect, therefore, we manually complement the pedestrian positions.

Pedestrian activity analysis. Here, dense trajectory features (traj., HOF, HOG, MBH) are clustered into $4,000$ visual words with GMM. In MBH feature, we divided into x- and y-direction to analyze detailed horizontal and vertical motions. The number of length (L in equation (1)) is adjusted 15 frames on the NTSEL dataset. However, we cannot accumulate a long-term feature in a real-scene. Then we set 2 – 20 (2, 3, 5, 7, 10, 15, 20) frames on the near-miss DR dataset.

C. Results

Table I shows the performance rate of DT on the NTSEL dataset. The 5 trajectory-based features are listed in the table. In the NTSEL dataset, four activities are defined for fine-grained pedestrian activity recognition, namely *walking*, *crossing*, *turning* and *riding*. The HOF feature outperformed the other four features in the classification. The HOF feature recorded 93.7% since the feature has the characteristics of flow feature extraction at each block. A large amount of flow features are captured on an image patch and inserted into feature vector. A detailed motion is expressed to distinguish a fine-grained motion on a pedestrian area. The MBHx, MBHy, HOG, and traj. features are following to the HOF feature on the dataset. Dense trajectories significantly represent the difference among fine-grained activities, however, the feature performances are not the same. The MBH feature is described as a best feature on the Wang's paper [7], however, we divided x- and y-direction to analyze the characteristics. The MBHx output better performance rate (87.7%) than MBHy (86.7%) on the dataset since the walking activities are almost horizontal motion. The HOG and traj. features were not good performance rates (85.6% and 76.5%), since the HOG feature has no motion attribute, and traj. does not contain primitive features it is just a trajectory shape.

Figure 3 shows the relationship between performance rate and feature accumulating time (frame). The videos in near-miss DR dataset are captured on the in-vehicle cameras. The cameras are obviously moving, therefore we must consider relative pedestrian motions on the dataset. In this experiment, surprisingly, the fewer frames accumulation recorded the best performance rate. According to the Figure 3, traj. feature at 3-frame accumulation is the top performance rate on the dataset. The HOG and HOF features are following at the 3-frame accumulation. Basically the fewer accumulation is better than long-term accumulation at the near-miss situation since any near-miss incidents suddenly occur in real-road. The features of walking activities are included a few frames in near-miss DR dataset. The subtle changes should be accumulated with vision-based feature descriptors.

At the same time, Table II shows the performance rate at 3-frame accumulation on the near-miss DR dataset. The traj. feature achieved the best performance rate on the near-miss DR dataset. Other features are all quantized features in the case of vectorization. The traj. feature is creating vectors based on a shape of trajectory from captured optical flows. In real-road dataset, the detailed shape feature was good at short-time classification, 3 frames in this situation. HOF and HOG are near performance rate to traj. feature since these representations are kind of quantized shape of optical flow (HOF) and image edge (HOG). In this case, the raw shape feature is better than the quantized one. Although the HOG, HOF and MBH are outperforming the traj. feature on a basic vision-based activity recognition problem [7], here we got better results with the traj. feature on the near-miss DR dataset.

D. Feature vector analysis

Figure 4 shows a feature vector analysis on the near-miss DR dataset. The figure includes average feature vectors of traj. feature, HOG and MBHx. Traj. feature and HOF

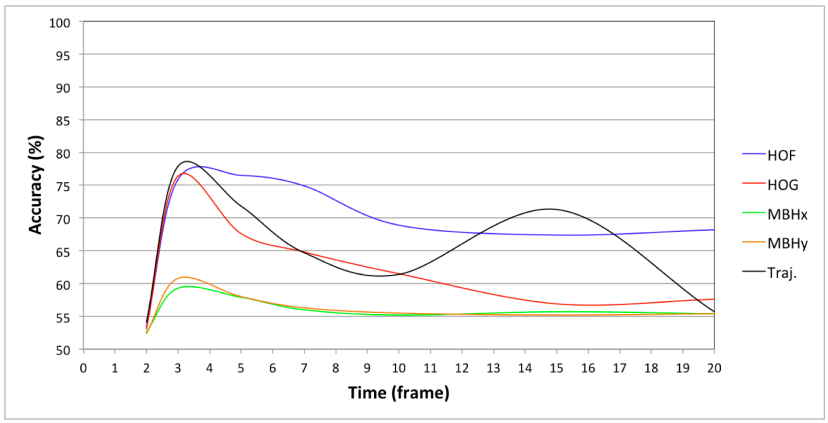

Fig. 3. The performance rate on near-miss DR dataset with number of frame comparison.

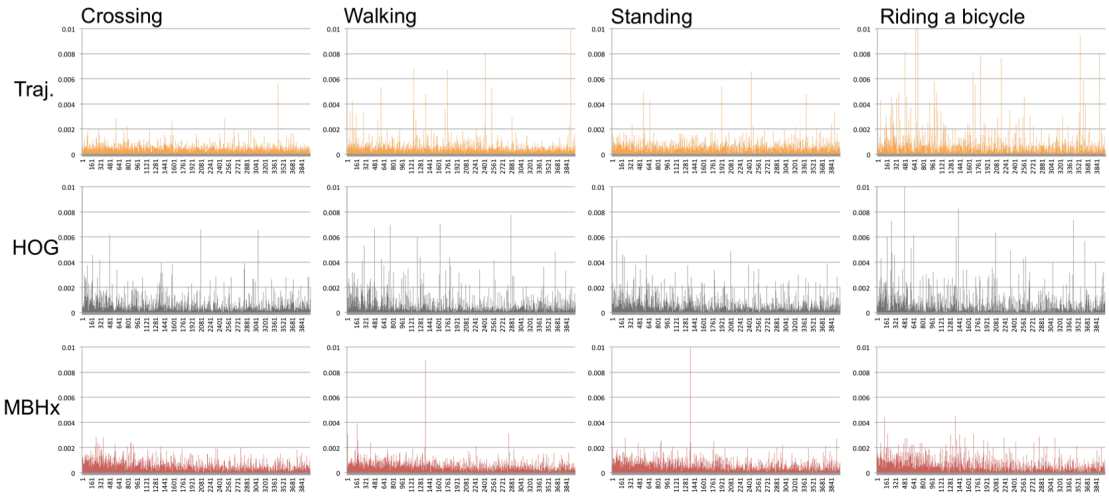

Fig. 4. Average feature vectors at each activity.

achieved the top 2 performance rate on the near-miss DR dataset. In classification step, a remarkable vector element is important to classify a specified class from the other classes. Traj. and HOG feature have several peaks in the average vector, however, MBHx has small peaks. On top of that, the top 2 features have a characteristics at each activity. MBHx tends to become flat histograms through the four pedestrian activities. The distinction is connecting to the classification accuracy (in Table II) between traj. feature (77.9%), HOG (76.4%) and MBHx (59.3%). Wang *et al.* proved the motion boundary feature highly recognizes human activities in stable camera scenes [7]. The feature captures minute changes on human area in a stable scene, therefore, a feature extraction on the near-miss DR dataset must be a troublesome situation to classify fine-grained activities. Against the motion boundary feature, traj. and HOG are consist of detailed shape descriptor. Especially traj. feature directly take a look at optical flow shape with a motion descriptor.

The correlations between two activities are shown in Figure 5. We calculated the correlations with Bhattacharyya coefficient as below:

$$L = \Sigma_{u=1}^{N} \sqrt{\bar{V}_u^1 \bar{V}_u^2} \tag{8}$$

where N is the number of dimension (=4,000), $\bar{V}$ is the average vector at each activity.

The lower value is indicating better performance in Figure 5 since the dissimilarity of vector induces that simple classification with kernel SVM. The Figure 5 describes the traj. feature and HOG are better approaches than the MBHx feature descriptor in terms of the similarity values. The average values are 0.727 (traj. feature), 0.778 (HOG) and 0.832 (MBHx), respectively. In detailed analysis, a similarity between *riding a bicycle* and other activities tends to become a lower value. The appearance of bicycle is dramatically different in vectorization step. The other three activities (*crossing*, *walking*, *standing*) are divided by means of a subtle changes such as walking direction and arm swinging motion.

V. Conclusion

The paper proposed a fine-grained walking activity recognition by means of pedestrian localization and activity analysis

Traj.

	Crossing	Walking	Standing	Riding
Crossing	--	.808	.821	.685
Walking	.808	--	.768	.654
Standing	.821	.768	--	.624
Riding	.685	.654	.624	--

Average: 0.727

HOG

	Crossing	Walking	Standing	Riding
Crossing	--	.781	.787	.790
Walking	.781	--	.803	.752
Standing	.787	.803	--	.755
Riding	.790	.752	.755	--

Average: 0.778

MBHx

	Crossing	Walking	Standing	Riding
Crossing	--	.857	.851	.823
Walking	.857	--	.846	.821
Standing	.851	.846	--	.793
Riding	.823	.821	.793	--

Average: 0.832

Fig. 5. Correlations of each features.

toward an inferring pedestrian intention. We implemented ECoHOG + Real AdaBoost based on Kataoka [6] to annotate pedestrian locations in an image sequence. In the pedestrian locations, the framework of dense trajectories extracts a large number of trajectory-based features. The traj. feature, HOG, HOF and MBH are prepared as feature descriptors on the trajectories.

As the results of the experiments, the proposed approach achieved 93.7% with HOF on the self-collected NTSEL traffic dataset and 77.9% with Traj. feature on the near-miss DR dataset. Especially in the near-miss DR dataset, we got a surprising result that short-term feature accumulation is better approach. We set 3 frame accumulation in the dataset, it equals to 0.1 second at a 30 fps video. The vector analysis allows us to induce a satisfactory knowledge from the data.

In the future, we would like to improve the recognition accuracy with more sophisticated feature vector. Moreover, we expand the near-miss dataset to allow further investigation into how much the algorithm is practical in fine-grained activity recognition.

REFERENCES

[1] World Health Organization, "Road traffic deaths Data by country", http://apps.who.int/gho/data/node.main.A997, 2010.

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

[3] P. Dollar, C. Wojek, B. Schiele, P. Perona, "Pedestrian Detection: An Evaluation of the State of the Art", IEEE Trans. Pattern Analysis and Machine Intelligence (TPAMI), vol.34, no.4, pp.743-761, 2012.

[4] R. Benenson, M. Omran, J. Hosang, B. Schiele, "Ten years of pedestrian detection, what have we learned?", ECCV Workshop CVRSUAD, 2014.

[5] N. Dalal, B. Triggs, "Histograms of Oriented Gradients for Human Detection", in CVPR, pp.886–893, 2005.

[6] H. Kataoka, K. Tamura, K. Iwata, Y. Satoh, Y. Matsui, Y. Aoki, "Extended Feature Descriptor and Vehicle Motion Model with Tracking-by-Detection for Pedestrian Active Safety", in IEICE Trans. on Information and Systems, Vol.97, No.2, pp.296–304, 2014.

[7] H. Wang, A. Klaser, C. Schmid, C. L. Liu, "Action Recognition by Dense Trajectories", in CVPR, pp.3169–3176, 2011.

[8] I. Laptev, "On Space-Time Interest Points", International Conference on Computer Vision (IJCV), No. 64, pp.107-123, 2005.

[9] A. Klaser, M. Marszalek, C. Schmid, "A spatio-temporal descriptor based on 3D-gradients", British Machine Vision Conference (BMVC), 2008.

[10] M. Marszalek, I. Laptev, C. Schmid, "Actions in context", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.2929-2936, 2009.

[11] R. Chaudhry, A. Ravichandran, G. Hager, R. Vidal, "Histograms of oriented optical flow and binet cauchy kernels on nonlinear dynamical systems for the recognition of human actions", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1932-1939, 2009.

[12] H. Wang, C. Schmid, "Action Recognition with Improved Trajectories", in ICCV, pp.3551–3558, 2013.

[13] H. Kataoka, K. Hashimoto, K. Iwata, Y. Satoh, N. Navab, S. Ilic, Y. Aoki, "Extended Co-occurrence HOG with Dense Trajectories for Fine-grained Activity Recognition", in ACCV, 2014.

[14] 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.

[15] 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.

[16] P. Viola, M. J. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2001.

[17] S. Walk, N. Majer, K. Schindler, B. Schiele, "New Features and Insights for Pedestrian Detection", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010.

[18] P. Sabzmeydani, G. Mori, "Detecting Pedestrians by Learning Shapelet Features", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2007.

[19] G. Mori, S. Belongie, J. Malik, "Efficient Shape Matching Using Shape Contexts", IEEE Trans. Pattern Analysis and Machine Intelligence (TPAMI), vol.27, no.11, pp.1832–1837, 2005.

[20] T. Watanabe, S. Ito, K. Yokoi, "Co-occurrence Histograms of Oriented Gradients for Pedestrian Detection", in PSIVT2009, pp.37–47, 2009.

[21] S. Munder, D. M. Gavrila, "An Experimental Study on Pedestrian Classification", IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI), vol.28, no.11, pp.1863-1868, 2006.

[22] P. Dollar, C. Wojek, B. Schiele and P. Perona, "Pedestrian Detection: A Benchmark", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009.

[23] A. Geiger, P. Lenz, R. Urtasun, "Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012.

[24] http://www.jsae.or.jp/hiyari/

[25] M. Raptis, I. Kokkinos, S. Soatto, "Discovering discriminative action parts from mid-level video representation", in CVPR, pp.1242–1249, 2013.

[26] M. Jain, H. Jegou, P. Bouthemy, "Better exploiting motion for better action recognition", in CVPR, pp.2555–2562, 2013.

[27] X. Peng, Y. Qiao, Q. Peng, X. Qi, "Exploring Motion Boundary based Sampling and Spatial Temporal Context Descriptors for Action Recognition", int BMVC, 2013.

[28] F. Perronnin, J. Sanchez, T. Mensink, "Improving the Fisher Kernel for Large-scale image classification", in ECCV, pp.143–156, 2010.

[29] G. Csurka, C. Dance, L. X. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints, ECCV Workshop, 2004.

[30] N. Dalal, B. Triggs, C. Schmid, "Human Detection using Oriented Histograms of Flow and Appearance", in ECCV, pp.428–441, 2006.

[31] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, "Learning realistic human actions from movies", in CVPR, pp.1-8, 2008.