Recognition of Texting-While-Walking by Joint Features based on Arm and Head Poses

Fumito Shinmura[1,2], Yasutomo Kawanishi[3], Daisuke Deguchi[4],
Ichiro Ide[3], Hiroshi Murase[3], Hironobu Fujiyoshi[5]

[1] Institute of Innovation for Future Society (MIRAI), Nagoya University, Japan
[2] JST/COI, Nagoya, Japan
[3] Graduate School of Information Science, Nagoya University, Japan
[4] Information Strategy Office, Nagoya University, Japan
[5] Department of Robotics Science and Technology, Chubu University, Japan

Abstract. Pedestrians "texting-while-walking" increase the risk of traffic accidents, since they are often not paying attention to their surrounding environments and fails to notice approaching vehicles. Thus, the recognition of texting-while-walking from an in-vehicle camera should be helpful for safety driving assistance. In this paper, we propose a method to recognize a pedestrian texting-while-walking from in-vehicle camera images. The proposed approach focuses on the characteristic relationship between the arm and the head poses observed during a texting-while-walking behavior. In this paper, Pose-Dependent Joint HOG feature is proposed as a novel feature, which uses parts locations as prior knowledge and describes the cooccurrence of the arm and the head poses. To show the effectiveness of the proposed method, we constructed a dataset and evaluated it.

1 Introduction

Recently, pedestrian recognition methods using an in-vehicle camera have been studied widely to reduce traffic accidents with pedestrians. In the case of Advanced Emergency Braking System, it first detects obstacles in front of the vehicle and then actuates the brake to avoid collisions with pedestrians. However, this system provides limited support for avoiding collisions since it does not actuate the brake until a pedestrian runs out in the road. In order to realize safer driving without traffic accidents, predicting the probable risks of traffic accidents will be important. In the case of risk prediction against pedestrians, it relies on the prediction of their behavior. For prediction of pedestrians' behavior, not only the detection of their positions but also the recognition of their attributes (e.g. body orientation, etc.) becomes important. Although some methods for pedestrians' orientation estimation have been studied to predict their path directions [4], various attributes should be considered for risk prediction.

Pedestrian's carelessness is also an important issue. Pedestrians who are not paying attention to their surrounding environments often fail to notice approaching vehicles. Such pedestrians can be considered that they have a higher risk of

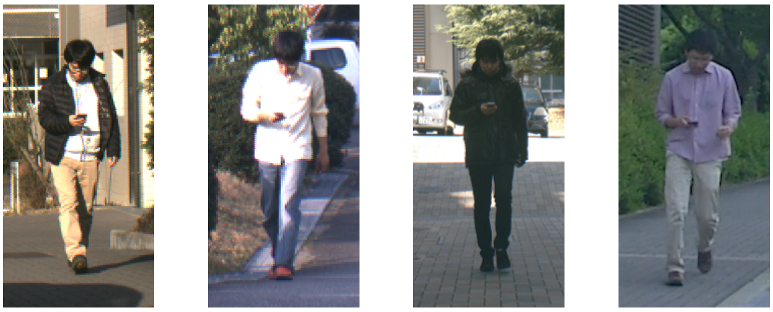


Fig. 1. Examples of pedestrians texting-while-walking.

running out into the road. Recently, dangerousness of pedestrian's "texting-while-walking" has been reported [5, 9]. Nintendo's Pokémon GO[1] player is a typical example of a pedestrian texting-while-walking. This is a behavior that a pedestrian's focus is on a hand-held electronic device such as a smartphone, while walking. Such a pedestrian tends to immerse him/herself into the operation of the device and is most likely not paying attention to his/her surrounding environment. Therefore, the recognition of a pedestrian's texting-while-walking behavior should contribute to the risk prediction of the pedestrian to be involved in an accident. In this paper, we propose a method to recognize whether a pedestrian is texting-while-walking or not from pedestrian images.

A pedestrian, who is texting-while-walking, holds a smartphone in his/her hand and looks down at the screen. Figure 1 shows examples of pedestrians texting-while-walking. Most of them take almost the same characteristic pose; bending the elbow and looking down.

However, there is a problem that poses similar to that of texting-while-walking may be observed while walking normally. For example, a pose merely bending the elbow, which occurs when a pedestrian waves his/her arms, appears similar to that of texting-while-walking. In order to prevent false recognitions on such cases, the proposed method focuses on the poses of both bending the elbow and looking down simultaneously.

In addition, pedestrians tend to walk in the same pose while texting-while-walking, in which case, the features of poses are almost constant for several seconds. On the other hand, when a pedestrian takes a similar pose momentarily, the features should vary widely during the same period. Thus, the proposed method sequentially observes the poses of pedestrian images during several seconds.

In summary, for accurate recognition of texting-while-walking, the proposed method takes the following approach:

– Focusing on the cooccurrence of poses of the arm and the head, we propose Pose-Dependent Joint HOG features, which is a variant of the Joint HOG

[1] http://pokemongo.nianticlabs.com/

features [7]. The proposed method selects features based on prior knowledge of parts locations.
- The proposed method prevents false recognition by observing the continuousness of the same pose.

2 Related Works

Various methods have been proposed for pedestrian detection and pedestrian attributes recognition. Many works that combine image feature extraction and supervised learning have been reported. Dalal et al. proposed a method that combined Histogram of Oriented Gradients (HOG) features and the Support Vector Machine (SVM) classifier [1], and Dollar et al. proposed a method that combined the Aggregate Channel Features (ACF) and the Boosted Trees [2] for pedestrian detection. For pedestrian attributes recognition, Shimizu and Poggio proposed a method for pedestrian's orientation estimation that combined the Haar-wavelet features and the SVM classifier [8]. Gandhi and Trivedi proposed a method that combined the HOG features and the SVM classifier [4].

Several methods for pedestrian detection and attributes recognition based on body parts features also have been proposed and proved to perform well. These methods extract image features not from the whole body but from body parts. Felzenszwalb et al. proposed a method using Deformable Part Models (DPM) [3] for pedestrian detection. Their models consist of a set of parts filters, and a pedestrian is detected based on the shape of parts and their positional relations. Tao and Klette proposed a method using Random Forest as a classifier and used body parts selectively for training [10] for pedestrian's orientation estimation. Their method extracted image features from part areas selected randomly, and estimated their orientation using a Decision Tree for each part.

In order to recognize texting-while-walking, the proposed method captures the features of the arm and the head poses. The methods introduced above based on image feature extraction can capture features of the arm and the head, but they may misrecognize a pedestrian with a similar pose to that of texting-while-walking.

To overcome this problem, the arm and the head poses need to be focused simultaneously. Hence, observing the cooccurrence between the arm and the head should be effective for the recognition of texting-while-walking. Mitsui and Fujiyoshi proposed joint features based on HOG (Joint HOG) [7] in order to represent the cooccurrence of appearances. Their joint features are obtained by combining the HOG features for several different local areas by means of AdaBoost, which allows to capture shape symmetry and edge continuity. In their method, more effective local areas for recognition are selected automatically by AdaBoost, although areas of the arm and the head may not always be selected. In the case of the recognition of texting-while-walking, selecting local areas from the arm and the head areas is more effective than selecting them from other areas. Therefore, the proposed method is improved to preferentially select local areas from the arm and the head areas.

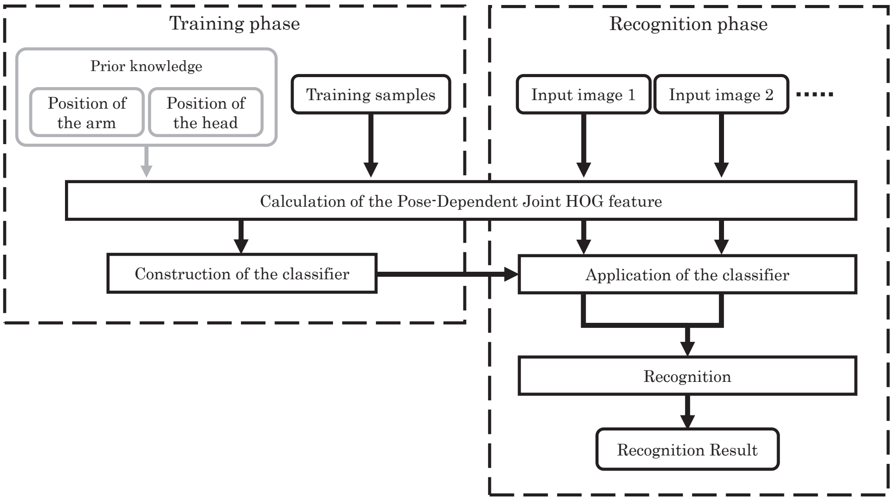


Fig. 2. Process flow of the proposed method.

3 Recognition of Texting-While-Walking by Joint Features

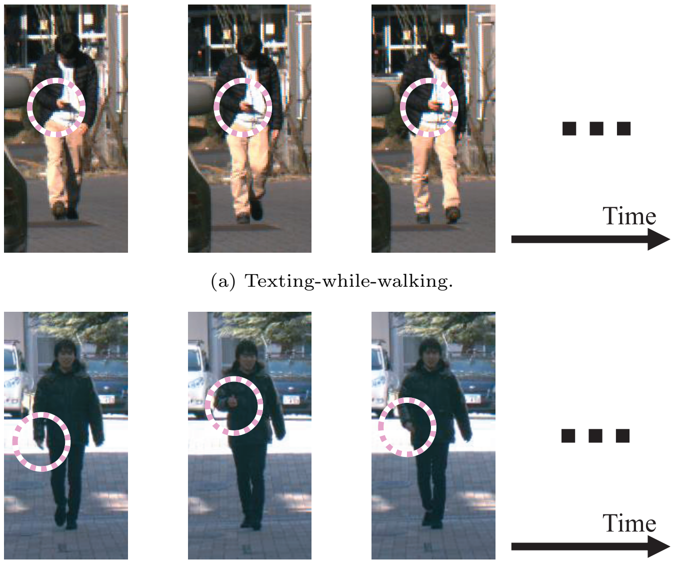
This paper proposes a method to recognize texting-while-walking from pedestrian images. Pedestrians are assumed to be detected beforehand by an arbitrary detection method. We have observed that a pedestrian's pose when he/she uses a smartphone tends to be as follows:

- Bending the elbow to hold a smartphone.
- Looking down at the screen of a smartphone.

Thus, the proposed method extracts image features that represent both the arm and the head poses, and constructs a classifier to recognize texting-while-walking using the extracted features. The process flow of the proposed method is shown in Fig. 2.

The pose of pedestrians simply bending their elbow or looking down are similar to the pose of texting-while-walking as shown in Fig. 3, even if they are not texting-while-walking. To distinguish between these poses, the proposed method needs to capture both the arm and the head poses simultaneously, and thus focuses on the cooccurrence of the arm and the head poses; It uses a joint feature of the arm and the head poses.

This method uses the Joint HOG features [7] to describe the cooccurrence of the arm and the head poses. It uses AdaBoost in two different contexts; 1) to calculate the joint features, and 2) to construct the classifier to recognize texting-while-walking. Since the conventional Joint HOG features select feature pairs by means of the second AdaBoost, the feature pairs are not always selected from the arm and the head areas. Thus, we propose the Pose-Dependent Joint



(a) Texting-while-walking.

(b) Walking normally.

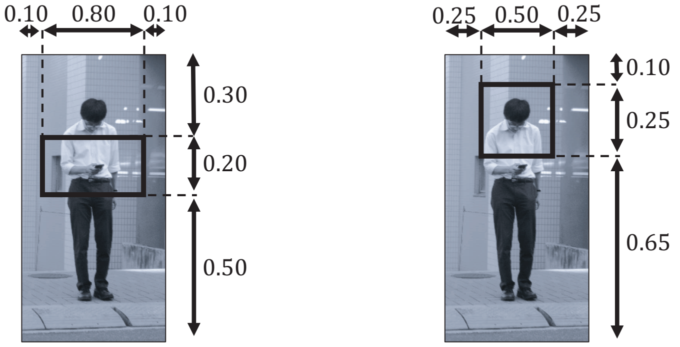
Fig. 3. Examples of a pedestrian texting-while-walking and that walking normally.

HOG feature, which is a variant of the Joint HOG feature, that is given a prior knowledge of parts locations to select feature pairs. This joint feature combines the HOG feature extracted from each of the arm and the head areas, which can describe the cooccurrence of their shapes.

Additionally, the proposed method sequentially observes the poses of pedestrians to reduce false recognition. Pedestrians texting-while-walking keep the same pose, although those walking normally varies their poses by waving their arms. Figure 3 shows examples of a pedestrian texting-while-walking and that walking normally during several frames. The red circles in the figure indicate the arm positions. A pedestrian texting-while-walking keeps his/her arm still, although that walking normally moves his/her arm. Thus, we consider that pedestrians recognized as in a texting-while-walking state for several seconds are actually texting-while-walking. On the contrary, pedestrians momentarily recognized as in a texting-while-walking state are not actually texting-while-walking. Thus, the proposed method assesses by majority voting of the recognition results in a two seconds period.

3.1 Extraction of the Pose-Dependent Joint HOG Features

In order to represent the pose of the arm and the head, this method uses the HOG feature proposed by Dalal et al. [1] as a low-level feature. First, an input



(a) Local area around the arm.

(b) Local area around the head.

Fig. 4. Pre-determined local areas used as prior knowledge.

pedestrian image is resized to 100×200 pixels, and then divided into local areas called cells with a size of 10×10 pixels. A histogram of gradient orientations containing nine orientation bins is created for each cell. The HOG features are calculated by normalizing these histograms in a block with a size of 3×3 cells. This is repeated by sliding the block one cell after another. One HOG feature is represented by a normalized histogram in a cell.

Next, two HOG features for different cells are selected to calculate the joint features. Since this method uses features of the arm and the head, the arm and the head positions are given as prior knowledge as shown in Fig. 4. One of the HOG features is selected from a cell in the arm area, and the other is selected from a cell in the head area.

From the two selected HOG features, the proposed method describes the cooccurrence of the two HOG features. The value of a binary symbol s is determined with the following equation:

$$s(\boldsymbol{V}) = \begin{cases} 1 & \text{if} \quad pv_o > p\theta \\ 0 & \text{otherwise} \end{cases}, \tag{1}$$

where $\boldsymbol{V} = [v_1, v_2, \cdots, v_9]$ is the HOG feature, θ is the threshold value, $v_o \in \boldsymbol{V}$ is the value of a histogram of orientation gradient, o is the orientation of gradient, p is a parity indicating the direction of the inequality sign and takes the values $p \in \{+1, -1\}$. The value of a certain histogram of orientation gradient v_o is selected from $\boldsymbol{V}$ and used for the calculation of $s(\boldsymbol{V})$. o, p and θ are determined when the AdaBoost learns as described below. The binary symbol s represents whether the target pedestrian is texting-while-walking or not. Two binary symbols that are calculated from cells around the arm and the head are obtained. The cooccurrence feature is generated by combining these two symbols. It takes four values based on the combination of the values of the symbols from the two areas.

The features effective for discrimination are selected from the cooccurrence features by the Real AdaBoost algorithm. When a set of N labeled training samples $(x_1, y_1), \ldots, (x_N, y_N)$, where $y_i \in \{+1, -1\}$ is the class label associated with a training sample x_i, is given, the strong classifier in the AdaBoost is constructed with the following equation:

$$H(x) = \sum_{t=i}^{T} h_t(x), \tag{2}$$

where $h_t(x)$ is the weak classifier in the AdaBoost. When a cooccurrence feature $J(x) = j$ is observed, $h_t(x)$ is expressed as follows:

$$h_t(x) = \frac{1}{2} \ln \frac{P_t(y = +1|j) + \varepsilon}{P_t(y = -1|j) + \varepsilon}, \tag{3}$$

where $t = 1, \ldots, T$ is the number of training rounds, $\varepsilon = 10^{-7}$ is a very small value to prevent division by zero. $P_t(y = +1|j)$ and $P_t(y = -1|j)$ are the respective conditional probability distributions calculated with the following equations:

$$P_t(y = +1|j) = \sum_{i}^{N} I(y = +1) D_t(i), \tag{4}$$

$$P_t(y = -1|j) = \sum_{i}^{N} I(y = -1) D_t(i), \tag{5}$$

where $I \in \{1, 0\}$ is an indicator that takes $I = 1$ when the condition is satisfied. $D_t(i)$ is a weight of the training sample x_i calculated with the following equation

$$D_{t+1}(i) = D_t(i) \exp(-y_i h_t(x_i)). \tag{6}$$

The weights are initialized by $D_1(i) = 1/N$.

The process described above is applied to all combinations of cells. When two different cells are expressed as c_m and c_n, the strong classifier is expressed as follows:

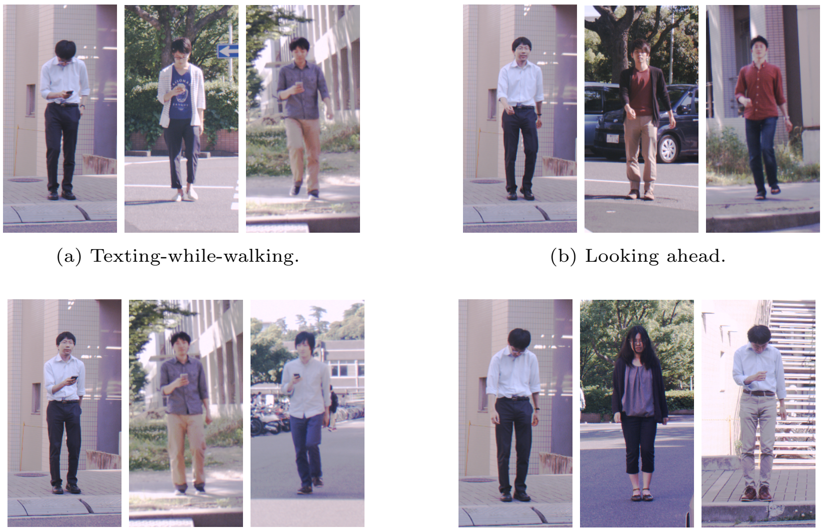
$$H^{c_m, c_n}(x) = \sum_{t=i}^{T} h_t^{c_m \in C_{\text{arm}}, c_n \in C_{\text{head}}}(x), \tag{7}$$

where C_{arm} and C_{head} are the cells in the arm and the head areas, respectively, and $H^{c_m, c_n}(x)$ is the joint feature. The joint features are generated for all combinations of cells. Here, since there were 108 cells for the arm and 81 cells for the head, 8,748 Pose-Dependent Joint HOG features were generated in total.

3.2 Construction of a Classifier to Recognize Texting-While-Walking

A classifier is constructed by Real AdaBoost using the calculated Pose-Dependent Joint HOG features. The features effective for discrimination are selected by training of the AdaBoost. This means that the classifier learns the positions and combinations of cells effective for discrimination.

Therefore, the classifier for recognition of texting-while-walking from an input pedestrian image is constructed.



(a) Texting-while-walking.

(b) Looking ahead.

(c) Bending the elbow.

(d) Looking down.

Fig. 5. Samples from the dataset used in the experiment.

3.3 Recognition of Texting-While-Walking

This method recognizes whether a pedestrian is texting-while-walking or not from an input pedestrian image sequence by using the constructed classifier.

First, the Pose-Dependent Joint HOG features are calculated from a pedestrian image detected from each frame. Next, the constructed classifier is applied, and the recognition result is obtained for each frame. Finally, whether the pedestrian is texting-while-walking or not is determined by majority voting of the recognition results of all the frames in a two seconds window.

4 Experiments

An experiment on recognition of texting-while-walking was conducted in order to evaluate the effectiveness of the proposed method.

4.1 Dataset

For the experiment, we prepared images captured in outdoor environments by a commercially available camera (Point Grey Grasshopper®3) in daytime. The bounding boxes of the pedestrians were manually annotated beforehand. The resolution of the images were 1,920×1,440 pixels, and the sizes of the cropped pedestrian images ranged from 275×550 to 409×818 pixels. Samples from the dataset are shown in Fig. 5.

The prepared dataset consists of 3,960 pedestrian images casted by eleven individuals. It contains four kinds of poses as follows:

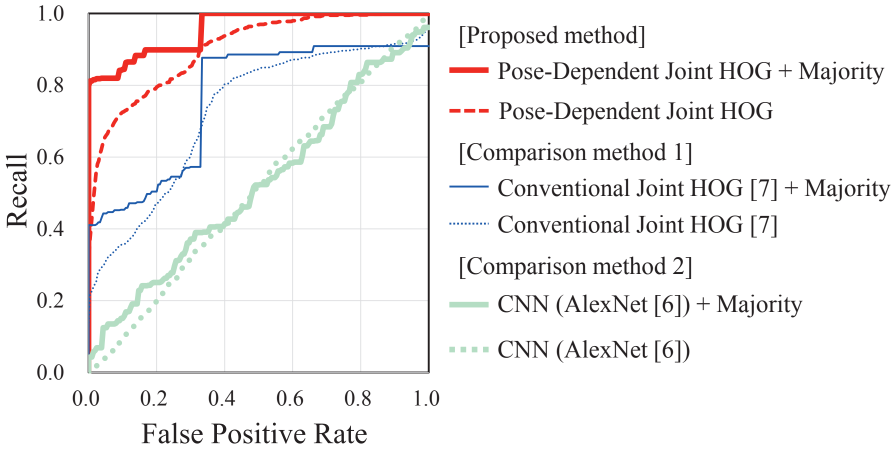


Fig. 6. Experimental results.

- Texting-while-walking, namely, using a smartphone.
- Looking ahead.
- Bending the elbow without using a smartphone.
- Looking down without using a smartphone.

All pedestrians were captured from the frontal side.

4.2 Experiments and Results

The prepared dataset was used for an experiment on recognition of texting-while-walking. In the experiment, data for ten subjects (3,600 images) were used for training and those for one other subject (360 images) were used for testing. The evaluation was repeated eleven times with a different subject for testing, and then the accuracy for each evaluation was averaged.

The experimental result is shown in Fig. 6 in a ROC curve. We compared the proposed method with two methods. Comparison method 1 was a method using the conventional Joint HOG features [7]. Comparison method 2 was a method using AlexNet that is a variety of the Convolutional Neural Network (CNN) [6], where we modified the output of its final layer to binary output for texting-while-walking recognition. As shown in Fig. 6, the proposed method using the cooccurrence features and majority voting achieved the best performance.

4.3 Discussion

Comparison method 1 also combined two HOG features and should have selected the effective pairs for recognition. It could also describe the cooccurrence features, but its accuracy was lower than that of the proposed method. Thus, we considered that the feature pairs truly effective for the recognition were not selected. In order to select truly effective feature pairs for recognition by this method, a training data with large variety is required. Likewise, comparison method 2 also requires a large variety of training data to train the CNN. In this experiment, the size of training data was insufficient to sufficiently train a CNN. If there were huge amount of training data, the results of CNN would be better. However, collecting many pedestrian images of texting-while-walking is even more difficult than collecting many pedestrian images for pedestrian detection, so it is better if we can cope with a small number of training data.

In the proposed method, by giving prior knowledge of parts location, we succeeded to select effective feature pairs for the recognition of texting-while-walking without a large training dataset.

5 Conclusions

This paper proposed a method to recognize a pedestrian texting-while-walking by focusing on the arm and the head poses. The characteristic feature for recognition of texting-while-walking is simultaneously bending the arm and looking down. The proposed method improved the Joint HOG feature to capture the cooccurrence of the two poses.

Our future work will include improvement to deal with pedestrian's orientation. Since the appearance changes according to his/her orientation, the recognition performance will vary according to the pedestrian's orientation. We also plan to conduct an experiment with a larger dataset including various pedestrian orientations.

Acknowledgement. This research is partially supported by the Center of Innovation Program (Nagoya-COI) from Japan Science and Technology Agency and Grant-in-Aid for Scientific Research from The Ministry of Education, Culture, Sports, Science and Technology.

References

1. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proc. 2005 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition. (2005) 886–893
2. Dollar, P., Appel, R., Belongie, S., Perona, P.: Fast feature pyramids for object detection. IEEE Trans. on Pattern Analysis and Machine Intelligence **36** (2014) 1532–1545
3. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. IEEE Trans. on Pattern Analysis and Machine Intelligence **32** (2010) 1627–1645
4. Gandhi, T., Trivedi, M.: Image based estimation of pedestrian orientation for improving path prediction. In: Proc. 2008 IEEE Intelligent Vehicles Symposium. (2008) 506–511

5. Haga, S., Sano, A., Sekine, Y., Sato, H., Yamaguchi, S., Masuda, K.: Effects of using a smart phone on pedestrians' attention and walking. Procedia Manufacturing **3** (2015) 2574–2580
6. Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. In: Proc. 26th Annual Conf. on Neural Information Processing Systems. (2012) 1097–1105
7. Mitsui, T., Fujiyoshi, H.: Object detection by joint features based on two-stage boosting. In: Proc. 12th IEEE Int. Conf. on Computer Vision Workshops. (2009) 1169–1176
8. Shimizu, H., Poggio, T.: Direction estimation of pedestrian from multiple still images. In: Proc. 2004 IEEE Intelligent Vehicles Symposium. (2004) 596–600
9. Stavrinos, D., Byington, K., Schwebel, D.: Distracted walking: Cell phones increase injury risk for college pedestrians. Journal of Safety Research **42** (2011) 101–107
10. Tao, J., Klette, R.: Part-based RDF for direction classification of pedestrians, and a benchmark. In: Workshop on Intelligent Vehicles with Vision Technology in the 12th Asian Conf. on Computer Vision. (2014) w11-p2