Investigating Pedestrian Detection Performance Using Gradient-based Features in Outdoor Scenes

Masamichi Shimosaka, Atsunori Moteki, Taketoshi Mori, and Tomomasa Sato
Graduate School of Information Science and Technology, The University of Tokyo
7-3-1 Hongo Bunkyo-ku, Tokyo, JAPAN
Email: {simosaka, moteki, tmori, tsato}@ics.t.u-tokyo.ac.jp

Abstract—**This paper presents an effectiveness and problems of gradient-based pedestrian detection method in outdoor scenes. Taking into account implementation of pedestrian detection framework to mobility systems, equipment should be simple and systems should be feasible against dynamic changes of background. We apply a method based on Histogram of Oriented Gradients (HOG) to outdoor crowded situations by using a wheelchair equipped with a monocular video camera. We show experimental results in which mistaken detection is caused by diverse situations such as person's scale, direction and occlusion.**

I. Introduction

This paper considers problems of pedestrian detection method in outdoor scenes. Personal mobility system is expected to serve as one of the future transportation system and pedestrian detection is indispensable to the system for the purpose of realizing autonomous locomotion. Recent years, many pedestrian detection methods ([1], [2], [3], [4], [5]) have reported. Taking into account installation of pedestrian detection framework to the system, requirement which system must fulfill is twofold. (1) Personal mobility system is space-saving, so a sensor of the system should be as simple as possible. (2) An introduction of moving camera indicates that dynamic changes of background occurs, so background subtraction method is not available. To satisfy such requirements, we adopt single view camera as a sensor equipment and appearance model classification method in detecting human.

Dalal & Triggs [2] proposed a human detection algorithm which leverages appearance information and operates on a single color image. Their method uses a dense grid of Histograms of Oriented Gradients (HOG) as features extracted from a region of detection window, and classifies human by using a linear Support Vector Machine (SVM). However, in their experiment utilizing *INRIA Person Dataset* [2], due to the usage of per-window measures in evaluation, the accuracy of localizing pedestrian is not clear.

Therefore, we introduce a pedestrian detection dataset which is taken with single camera mounted on a wheelchair and has difficulty in occluded human and complex background. Then we make certification of feasibility of Dalal & Triggs algorithm under such situation.

The paper is structured as follows. The next section will introduce previous work of pedestrian detection. Section III will show the method in detail. Section IV will describe experimental results and discussion about problems.

II. Previous Work

There is a great deal of literature about human detection. According to the survey of Schiele *et al.* [6], human detection method is divided into two main classes: sliding window approach and part-based model approach.

In sliding window approach, detection systems scan an image at all possible positions and scales and classify whether the window contains human or not. Papageorgiou & Poggio [7] used Haar wavelets and a polynomial SVM. Viola & Jones [8] used Haar-like wavelets and a cascade of AdaBoost classifiers. Recently, Sabzmeydani *et al.* [5] employed locally learned features in an AdaBoost framework. In those approaches, a variety of features are used as object descriptors such as Haar wavelets ([1], [7]), HOG ([2], [3]), Shape Context ([4], [9]). Mainly used classifier is SVM or boosting.

On the other hand, part-based model approach uses low-level descriptors to model individual parts or limbs of person and models the topology of the human body to enable the accumulation of part evidence. Felzenszwalb *et al.* [3] proposed the pictorial structures model, and Andriluka *et al.* [10] advanced this idea to an implicit model of a-priori knowledge about possible body configurations. The drawback of part-based approach is that it costs more time to classify an object than sliding window approach.

In our study, the gradient-based method proposed by Dalal & Triggs is investigated. According to [11], HOG and linear SVM framework works well in their study. As HOG descriptor uses gradient information, it is insusceptible to variation of illumination or color and robust against rotations or translations. Additionally, some real-time feature extraction method such as [12] can be adopted if needed.

Studies in regard to investigation of human detection algorithms also exist. Dollár *et al.* [11] benchmarked seven detection algorithms with dataset constructed by theirs and highlighted situations in which existing methods fails to detect. Their dataset is taken with a camera mounted on a car and camera's height is taller than ours.

III. Pedestrian Detection Method

This section describes the pedestrian detection method leveraged in our study. Figure 1 shows the flow of overall detection process. In each step, many parameters have to be selected, and we adopted the values described in the

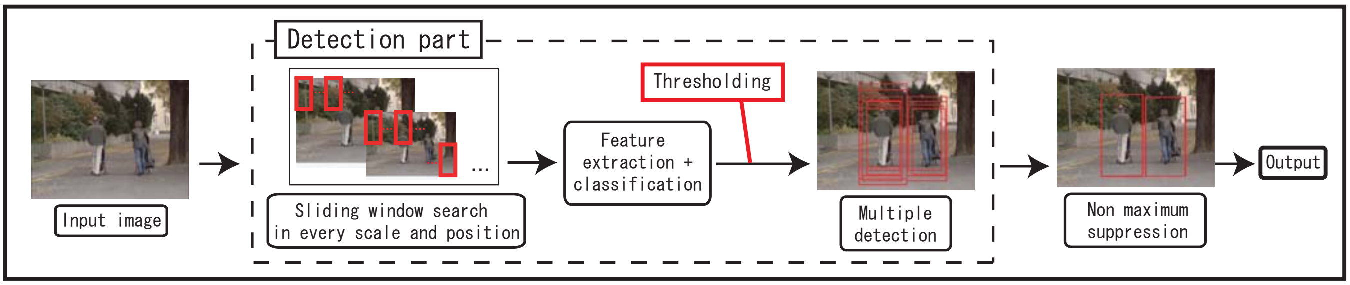
Detection part
Input image
Sliding window search in every scale and position
Thresholding
Feature extraction + classification
Multiple detection
Non maximum suppression
Output

Fig. 1. Overall algorithm configuration

following section because previous work [13] presented high performance with these values.

A. Features and Classifiers

To address a situation of cluttered background and diversity of person's appearance, Histogram of Oriented Gradients (HOG) features are employed. The system chooses the following parameters, described below: grayscale space with no gamma correction; [-1,0,1] gradient filter with no smoothing; linear gradient voting into 9 orientation bins in 0° - 180°; 16×16 pixel blocks of four 8×8 pixel cells; L1-norm block normalization; linear SVM classifier. LIBLINEAR [14] is used in our implementation. To reduce computational cost, a Gaussian mask and a tri-linear interpolation process are omitted.

B. Learning

Utilizing *INRIA Person Dataset* [2], a binary object classifier is constructed. The dataset contains 2478 positive training examples and 1218 non-human images. 12180 patches from 1218 non-human images are randomly sampled and initial negative set is constructed. Then, HOG features from each positive and negative example are extracted, and human / non-human classifier is trained in advance. To reduce false detection, re-training process is applied. False positive mistakes in 1218 images ('hard examples') are exhaustively searched for augmented dataset (initial 12180 + hard examples) in re-training process.

C. Detection

Sliding window approach is applied as a detection framework. First, the system scans each image at all scales and locations and runs the classifier in each window using learned binary classifier. After a thresholding process with the score of SVM, multiple detection per one person occur. Therefore, non maximum suppression (NMS) algorithm is needed to suppress the multiple detection. The system fuses detection with mean shift mode seeking algorithm [15] and yields final detection result.

Figure 2 shows an outline of the algorithm (details in [13]). This problem is replaced by one of kernel-density estimation. By iteratively computing a variable bandwidth mean shift vector for each data point until it converges, modes are obtained. The following parameter is employed, described below: a scale ratio of 1.2; a stride of 8 pixels; smoothing parameters $\sigma_x = 8$, $\sigma_y = 16$, $\sigma_s = \log(1.3)$; transformation function with hard clipping of negative scores to zero.

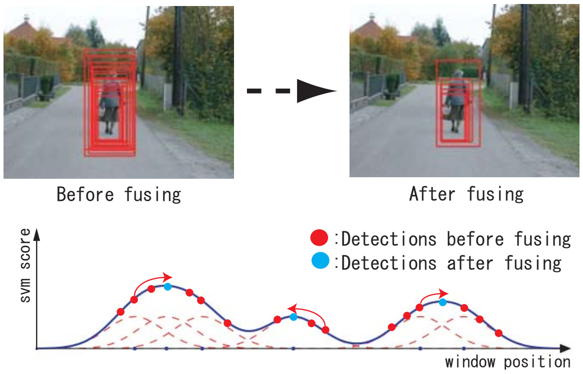

Fig. 2. Non maximum suppression outline

Sometimes multiple detection at the same person remains due to irregular density distribution. In that case, a window having the highest SVM score is regarded as final detection result.

IV. Experiments

A. Dataset Acquisition

To study an effect of various difficult situations in outdoor scenes, a series of video data is prepared. The dataset includes person's various condition. A direction of person is grouped into three classes: *front/back* (51 %), *right/left* (21 %) and *diagonal* (28 %). These video data are taken with mobile recording system (shown in Figure 3:left). We utilize a wheelchair equipped with a monocular video camera, and this camera is 0.62 meters high. The video resolution is 640×480 and the frame rate is 30 fps.

B. Experimental Preparation

At the beginning, the video data containing approximately 2000 frames were manually annotated. Occluded pedestrians are annotated as one ground truth is labeled per one pedestrian. To alleviate a burden on the annotation, every 10 frames of the video are only annotated and interpolation method were utilized so that intermediate frames are automatically labeled. An aspect ratio of a window is constantly (2:1) because an output window of the system is also constantly

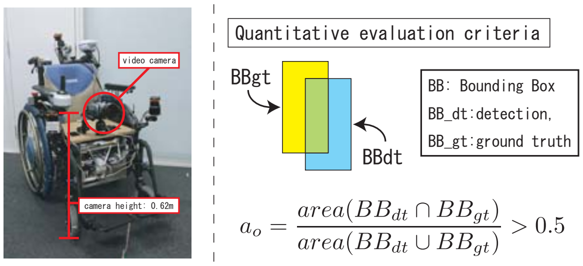
video camera
camera height: 0.62m
Quantitative evaluation criteria
BBgt
BBdt
BB: Bounding Box
BB_dt:detection,
BB_gt:ground truth
$a_o = \frac{area(BB_{dt} \cap BB_{gt})}{area(BB_{dt} \cup BB_{gt})} > 0.5$

Fig. 3. Experimental equipment and evaluation criteria

TABLE I
PERFORMANCE OF PEDESTRIAN DETECTION PER SEQUENCE

	frame	TP	FN	FP	pre.[%]	rec.[%]	FPPI
seq1	470	645	256	567	53.2	71.6	1.2
(only MS)	470	670	231	1439	31.8	74.4	3.1
seq2	460	659	208	676	49.4	76.0	1.5
(only MS)	460	688	179	1620	29.8	79.3	3.5
seq3	490	661	579	621	51.6	53.3	1.3
(only MS)	490	771	482	1229	38.5	61.5	2.5
seq4	610	503	567	563	47.2	47.0	0.92
(only MS)	610	560	532	1049	34.8	51.2	1.9

(2:1). Pedestrians are annotated in a little larger size than that of formal human (about 10 pixels larger), for images of INRIA training dataset have margins around human in order to classify human well using an environmental context.

As quantitative evaluation criterion, PASCAL measure [16] is employed. Detected bounding box is denoted as BB_{dt} and ground truth bounding box as BB_{gt}. If an area of overlap a_o exceeds 50 %, pedestrians are correctly detected (Figure 3:right).

This matching is done at every detection and ground truth combination, and at the end of the evaluation, unmatched BB_{dt} are counted as false positive (FP) and unmatched BB_{gt} as false negative (FN). Matched BB are counted as true positive (TP). Then, precision p, recall r and FPPI (false positive per image) are computed as an accuracy measure. BB_{gt} whose size is less than 50×100 is ignored because minimum size of a bounding box is 64×128.

$$p = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{1}$$

$$r = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{2}$$

C. Results and Discussion

First, we apply our implementation to the test dataset of *INRIA*. The result is that recall rate is 71.7 % and FPPI is 0.91. Next, we show the result of the experiment. Detection examples are shown in Figure 4. Video sequence number, frame number of each sequence, and performance rates are shown in Table I. A row presented as ’only MS’ means only mean shift mode seeking is conducted in NMS.

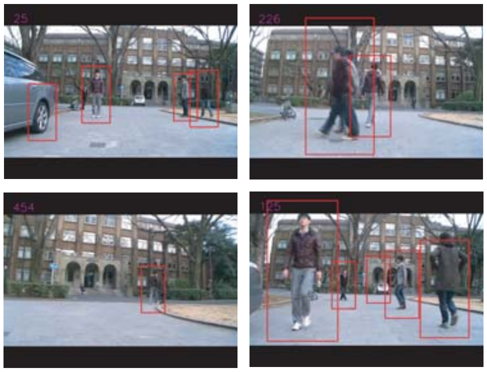

Fig. 4. Examples of final detection result with our system

TABLE II
DIFFERENCE OF RECALL RATE BY PEDESTRIAN’S DIRECTION

	right/left	front/back	diagonal
recall[%]	48.3	69.0	57.1

Sometimes mean shift mode seeking algorithm does not work sufficiently and in consequence precision rate descends (shown in Figure 5). Multiple detection is centered in leg portion, and this kind of mistaken detection is seen in other frames frequently. Table I shows that final result selection method which adopts the window having highest SVM score improves performance. However, this simple method does not work in the following situation: detecting pedestrians at different scales (Figure 6:left), or having the highest score at the wrong position (Figure 6:right).

In contrast, recall rates are varied. Roughly speaking, seq3 and seq4 have more occlusions of human (especially horizontally adjacent pairs) than seq1 and seq2. Although variability among sequences does exist, a relationship between location of occlusion and classify performance should be investigated closely. In addition, there is a possibility that the difference of dataset between training and test affects the performance of recall rate. At the same time, real world situation is full of uncertainty such as person’s cloth, physical attribute, environment and climate. A larger dataset in outdoor scenes is necessary to learn the classifier sufficiently.

The system localizes pedestrians of various sizes and directions. Table II shows performance at different pedestrian directions. This result indicates that recall rate of *front/back* is better than that of *right/left*. This is because INRIA dataset contains comparatively more pedestrians of *front/back* than that of *right/left*.

Unexpected false positive and false negative mistakes sometimes occur (Figure 7). Owing to the sudden movement of a camera, an image is blurred and the system cannot detect edges which are necessary in gradient feature calculation. The blurred image is captured when road surface condition is bad (sudden bump, ballast *etc.*). To deal with this blurring, the usage of information obtained from past several frames may improve the detection performance because these mistakes

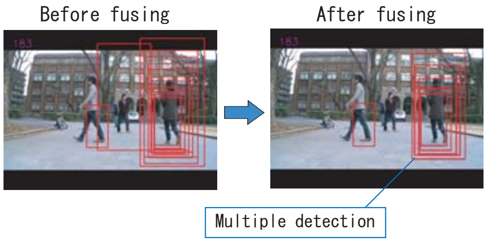

Fig. 5. An Example of insufficient mean shift mode seeking

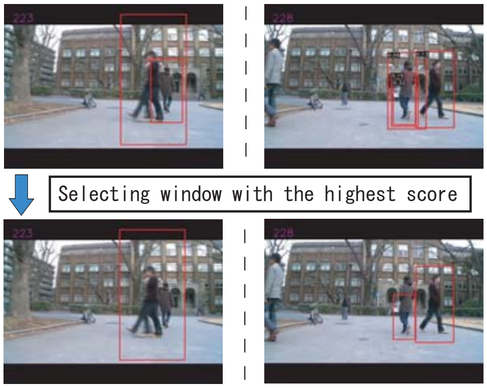

Fig. 6. Examples of the image simple NMS method does not work well

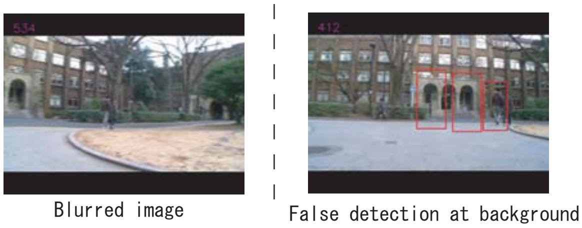

Fig. 7. An Example of mistake due to blurred image and cluttered background

occur suddenly and do not last several frames. In addition, due to cluttered background, false positive mistake occurs and system using appearance cues alone cannot always eliminate mistakes. Robust detection framework against false positive mistakes is necessary.

Time required to evaluate one image is shown in Table III. Window number per frame is depend on *scale ratio* S_r used to compute the scale-steps during multi-scale scan. Sparse scan means $S_r = 1.2$ and dense scan means $S_r = 1.05$ on a 2.40 GHz Intel (R) Core (TM) 2 Duo processor and 2 GB RAM. If tracking human are taking into account, processing speed should be more than 5 fps. To reduce computational cost, the usage of temporal and spatial contexts must be important. As for temporal context, an implicit assumption that the amount of movement in position and scale from the last frame is mostly not large can be applied. In addition, spatial context such as ground plane assumption [17] can be available in the situation of using a wheelchair.

TABLE III
TIME REQUIRED TO EVALUATE ONE IMAGE

Resolution	320x240	320x240	640x480	640x480
Scan density	sparse	dense	sparse	dense
Window number	764	2098	7581	23003
Time[sec]	0.14	0.43	1.0	3.1

V. CONCLUSION

We described in this paper an effectiveness and problems of gradient-based pedestrian detection method in outdoor situation. Gradient-based pedestrian detection method is investigated by using a wheelchair equipped with a monocular video camera. Experimental results show that person's scales and directions affect detection performance. Mistaken results are as follows: insufficiency of fusing multiple overlapping detections of adjacent, variable scale people, the loss of information due to blurred images and false detection in cluttered background. Processing time also need to be improved from the practical point of view. In future work we plan to analyze mistaken detection closely. The novel method will be taken into account which uses spatial or temporal context as a cue for robust human detection.

REFERENCES

[1] P. Viola, M. J. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. in *Proc. of CVPR*. Vol. 2. 2003. pp. 734–741.

[2] N. Dalal and B. Triggs. Histograms of Oriented Gradients for human detection. in *Proc. of CVPR*. Vol. 1. 2005. pp. 886–893.

[3] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. in *Proc. of CVPR*. 2008.

[4] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. in *Proc. of CVPR*. Vol. 1. 2005. pp. 878–885.

[5] P. Sabzmeydani and G. Mori. Detecting pedestrians by learning shapelet features. in *Proc. of CVPR*. 2007.

[6] B. Schiele, M. Andriluka, N. Majer, S. Roth, and C. Wojek. Visual people detection: Different models, comparison and discussion. in *Proc. of ICRA 2009 Workshop on People Detection and Tracking*. 2009.

[7] C. Papageorgiou and T. Poggio. A trainable system for object detection. *International Journal of Computer Vision*. Vol. 38. No. 1. pp. 15–33. 2000.

[8] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. in *Proc. of CVPR*. Vol. 1. 2001. pp. 511–518.

[9] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Trans. on PAMI*. Vol. 27. No. 10. pp. 1615–1630. 2005.

[10] M. Andriluka, S. Roth, and B. Schiele. People-tracking-by-detection and people-detection-by-tracking. in *Proc. of CVPR*. 2008.

[11] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. in *Proc. of CVPR*. 2009. pp. 304–311.

[12] T. P. Cao, G. Deng, and D. Mulligan. Implementation of real-time pedestrian detection on FPGA. in *Proc. of IVCNZ*. 2008.

[13] N. Dalal. Finding people in images and videos. Ph.D. dissertation. Institut National Polytechnique de Grenoble / INRIA Rhône-Alpes. Grenoble. 2006.

[14] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*. Vol. 9. pp. 1871–1874. 2008.

[15] D. Comaniciu. An algorithm for data-driven bandwidth selection. *IEEE Trans. on PAMI*. Vol. 25. No. 2. pp. 281–288. 2003.

[16] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*. Vol. 88. No. 2. pp. 303–308. 2009.

[17] D. Hoiem, A. A. Efros, and M. Hebert. Putting object in perspective. in *Proc. of CVPR*. Vol. 2. 2006. pp. 2137–2144.