Improving Human Detection by Long-Term Observation

Ikuhisa Mitsugami
Osaka University
Osaka, Japan
mitsugami@am.sanken.osaka-u.ac.jp

Hironori Hattori
Sony Corporation
Tokyo, Japan
Hironori.Hattori@jp.sony.com

Michihiko Minoh
Kyoto University
Kyoto, Japan
minoh@media.kyoto-u.ac.jp

***Abstract*—In this paper we propose a novel human detection method which is based on the existing learning-based method but designed so as to obtain the scene-specific knowledge and utilize it for improving the detection performance. The scene-specific knowledge contains two kinds of information. One of them is additional positive and negative samples that could not be detected by the initial detection method but extracted afterwards by tracking the initial detection results. The other is camera calibration using the size and direction of the detected people in the scene. By this calibration, we can drastically reducing the possibility to incidentally find a pattern which is not a human but looks similar to a human. Experimental results show the effectiveness of the proposed method.**

***Keywords*-Human detection, HOG, camera calibration**

I. Introduction

People's trajectories in facilities of a station or a shopping center are thought to be useful information for many fields; marketing, security, and so on. A system that can automatically acquire the trajectories is thus required. It is clear that a method to automatically detect people from the images is the fundamental and most important technique for this purpose. Various methods for the human detection have been proposed. Among those methods, the HOG-based method proposed by Dalal et al. [1] has become now the standard method supported by its good performance. There are currently a lot of extensions [2], [3], [4]. Those methods detect a human using a feature of his/her local shape learnt from a lot of positive and negative samples to generate a classifier. They, however, often fail when there is a pattern which is not a human but looks similar to a human from the viewpoint of the feature similarity. This is because the classifier is short of generalization capacity and because of the lack of the knowledge of the object scene. All the extensions are to struggle to this problem, but they still cannot overcome it completely. As long as they are based on the same framework, it is essentially impossible to avoid the problem.

In this paper, we propose a novel framework to overcome the problem that is designed for a fixed camera. As the fixed camera always captures the same scene, there should be some tendency of the human patterns in the images. If there is such tendency, it is expected that the classifier based on the human patterns in the images would works better than the classifier using the general human samples. Based on the notion, the proposed method uses additional information specific to each camera, which can be obtained by judging whether each of the detection results of the normal human detector is true or false. This judgement, of course, cannot be achieved only by the detected result in each frame, but can be achieved by analyzing time series of detected results. By this judgement, a new classifier specific to the scene can be automatically generated to increase the detection performance. In addition, an automatic camera calibration similar to [5], [6] is also executed to reduce the possibility to incidentally find a pattern which is not a human but looks similar to a human. This calibrated parameters can be also regarded as scene-specific knowledge.

In recent years, the similar framework for adaptive object detection has been proposed [7]. Compared with that study, the novelty of our method is the combination of the adaptive framework and the calibration. In fact, we experimentally show that the detection accuracy is much improved by combining these ideas while it is not so improved when each of these ideas is applied solely. Yamauchi et al [8] proposed another approach. They use a 3D human model to automatically generate training samples to improve the detection accuracy. It is an interesting idea, but the generated samples are not real ones. In our method, on the other hand, all the additional training samples are real and scene-specific, so that we expect good performance to the scene.

II. Learning-based Human Detection

A. Learning-based Human Detection Method

We briefly introduce the learning-based human detection method as represented by [1]. In this method, the distribution of intensity gradients of a human region is described as a feature vector. The feature vectors of the images included in positive and negative dataset are then trained by a binary classifier, such as SVM (Support Vector Machine)[9] or AdaBoost[10]. After the training, the classifier can be used for making decision about whether a certain region of an image is of a human or not. In the case that we would like to detect all people from an image, the detection window is scanned across the image and judged whether it is of a human or not at every position. Note that as this human detection method is designed for a general image so that it

978-1-4799-2190-4/13 $26.00 © 2013 IEEE
DOI 10.1109/ACPR.2013.40

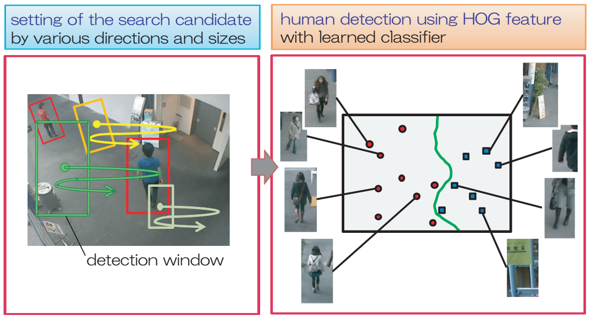

setting of the search candidate by various directions and sizes
human detection using HOG feature with learned classifier
detection window

Figure 1. Outline of human detection using HOG feature.

cannot preliminarily obtain the knowledge about the image, the scanning has to be executed with various scales and orientations of windows. The abstract of the method is described in Fig.1.

B. Problems of Learning-based Method

Fig.2 shows the detection results by the method mentioned in Section II-A. You see some false detections, while in most cases the method works well. These false detections can be classified into the following three classes:

FP1: False positives that appear at fixed background regions which incidentally look similar to people as found in Fig.2(ii).

FP2: False positives that randomly appear in the image as found in Fig.2(iii); they may appear at the fixed background, or may appear partially containing a human.

FN: False negatives that occur when a human in the image is too small he/she is partial occluded by other things, as found in Fig.2(iv).

There have been several studies for reducing these false detections. Dalal improved the method by combining the HOG feature and the optical flow[11]. It was reported that using co-occurrence of the features improved the performance in [12], [13]. These studies are in fact helpful for reducing the errors. However, even when we apply these extensions, it is still impossible to remove all the false detections. This is because of the limitation of generalization capacity of the machine learning methods.

III. Scene Adaptation Using Time Series of Detected Results

As long as the human detection method is designed so as to be utilized for arbitrary images, the issue mentioned in Section II-B cannot be overcome; even when a huge amount of samples are used for training, or whichever training method is used, the issue cannot be solved fundamentally so that the detection failure would not be reduced drastically.

If we focus on a certain fixed camera and the human detection is executed for sequential images of the camera, another approach to overcome the issue should exist. Although it is impossible to judge whether a detection result is true or not just by a single image, which is apparent because the result is given after the judgement, it is possible if sequences of the results are analyzed; true detections can be found as ones that move smoothly (following the human's motion) while false detections are fixed or appear/vanish/move randomly as they are not corresponding with the people in fact.

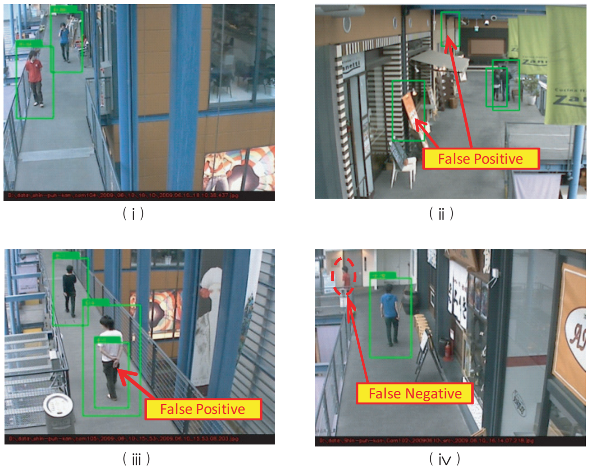


Figure 2. Example of human detection.

Based on this notion, the detection performance can be slightly increased by adding the sequential analysis after the normal detection process. However, this is just a post-processing, so that the detection method itself essentially does not get better.

Considering the above discussion, this paper proposes a novel method that works as the general human detection at the beginning but additionally obtains the scene-specific knowledge so as to perform better as time passes. The scene-specific knowledge consists of the classifier of the human pattern of that scene and the relation of the size and orientation of a human in the image. The additional classifier is generated by true and false detections that are detected by the general human detector and judged by the time series analysis. The size-orientation relation is obtained also by the true detections judged by the time series analysis. Fig.3 describes the comparison between the normal human detector and the proposed method. By using the two kinds of scene-specific knowledge at the searching and classifying steps, the proposed method runs more accurately than the normal one. The following sections discuss the two steps in order.

A. Additional Learning of Feature

In the proposed method, the detection results of the normal detector are collected for a long time. These results are then classified into true and false detections by the difference of their time series. Fig.4 shows the examples of the time series of the detections. The positions of true detections are expected to move smoothly as they correspond

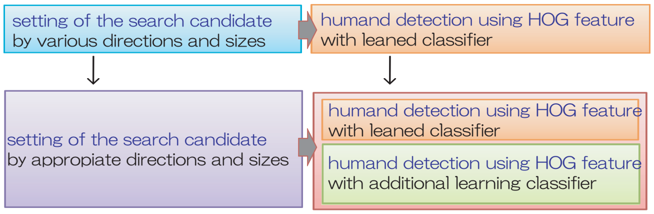

setting of the search candidate by various directions and sizes
humand detection using HOG feature with leaned classifier
setting of the search candidate by appropiate directions and sizes
humand detection using HOG feature with leaned classifier
humand detection using HOG feature with additional learning classifier

Figure 3. Comparison between the general human detector and the proposed method.

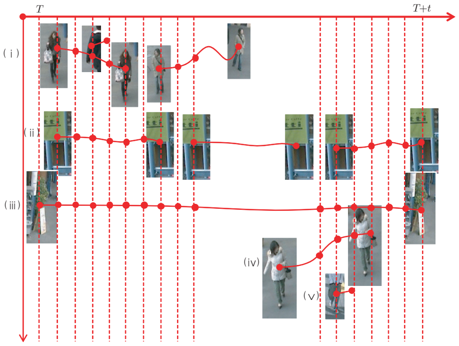


Figure 4. Analysis time series of detected results.

to the real movements of people, who move smoothly and never vanish and appear suddenly. On the other hand, the false positives often continue to be located at a fixed position as shown in Fig.4 (ii) when they are in the background, or appear randomly at a random position in the image as shown in Fig.4 (iii). Considering these characteristics, the proposed method judges the true and false detections according to the following criteria:

- If a detection sequence does not move for more than k_1 frames, it is regarded as a false detection against a fixed background.
- If a detection sequence moves smoothly so as to be tracked for more than k_1 frames, it is regarded as a positive detection corresponding to a human in fact.
- If only one or two frames are lacked from the smooth sequence, the lacked ones are used for positive samples.
- If a detection sequence is alive less than k_1 frames, it is judged as a a false positives that appears randomly.

Based on the above judgement, a lot of positive and negative samples specific to the scene can be collected. Their HOG features are thus trained by new binary classifiers, as shown in Fig.5.

B. Camera Calibration Using Pedestrian

Since the feature is calculated from a pattern in a detection window, it is not invariant against the rotation and the scale. It is thus necessary to scan the image several times with different directions and sizes of detection windows in order to detect a human of an arbitrary size and direction. However, as the variation of size and direction gets increased, the number of false detections would also increase, because the possibility that the pattern in the window looks incidentally similar to a human would increase. This means that it is necessary to set the direction and size appropriately to improve the accuracy of human detection. In the proposed method, therefore, an automatic camera calibration is executed using the true detections so as to get the relation between the position, size and direction of a human specific to the scene.

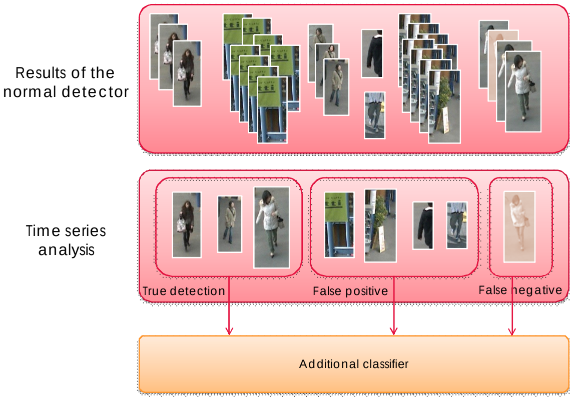


Figure 5. Time series analysis to generate additional classifier.

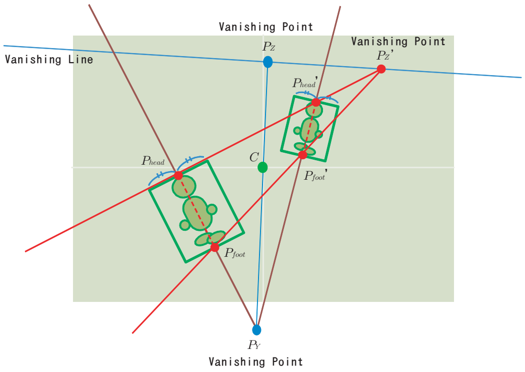


Figure 6. The vanishing point and line.

In this approach, it is assumed that a human walks on a plain floor and the whole body of the human is not occluded. As shown in Fig.6, when a human is detected in two different frames, the line $P_{head}P_{foot}$ and $P'_{head}P'_{foot}$ are parallel because he/she is standing perpendicular to the floor. On the other hand, the line $P_{head}P'_{head}$ and $P'_{foot}P'_{foot}$ are also parallel because his/her height is constant in the two frames. We can thus estimate vanishing point $P_Y(u_Y, v_Y)$ and $P_Z(u_Z, v_Z)$ by using these two pairs of lines. Note that the proposed method uses the RANSAC[14] using many sample head position P^i_{head} and foot position P^i_{foot} for making the estimation of the vanishing point more robust.

Here the coordinate system is shown in Fig.7. The projection of a 3D scene point $\boldsymbol{M} = (X, Y, Z)^T$ onto a point in the image plane $\boldsymbol{m} = (u, v)^T$ can be modeled by the equation:

$$s\hat{\boldsymbol{m}} = KA\hat{\boldsymbol{M}}, \tag{1}$$

where s is the scale factor. K denotes the upper triangular

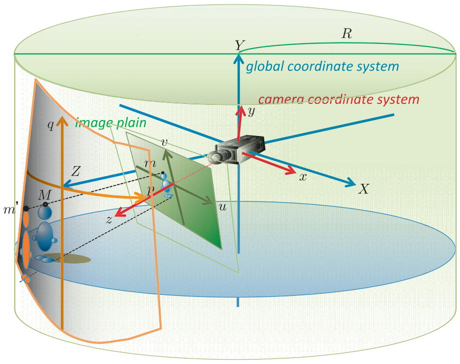


Figure 7. Coordinate System.

3×3 matrix and is expressed by

$$K = \begin{pmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{pmatrix}, \tag{2}$$

where f is the focal length, and the center of the image is $(0, 0)$. The extrinsic matrix A is defined as

$$A = (R|\boldsymbol{T}), \tag{3}$$

where R is the rotation matrix defined by

$$R = \mathbf{Rot}(z, \gamma)\ \mathbf{Rot}(y, \beta)\ \mathbf{Rot}(x, \alpha). \tag{4}$$

and $\boldsymbol{T}$ is the translation vector defined by $T = (T_X, T_Y, T_Z)^T$. Here, we determine $T_X = 0$, $T_Y = 0$, $T_Z = 0$ and $\beta = 0$, which do not loose generality.

Here $P_Y(u_Y, v_Y)$ and $P_Z(u_Z, v_Z)$ are at infinity point along Y-axis and Z-axis. So $Y \to \infty$ corresponds with (u_Y, v_Y) and $Z \to \infty$ corresponds with (u_Z, v_Z). Therefore u_Y, v_Y and u_Z, v_Z is calculated as follows:

$$\begin{cases} u_Y = -f\frac{\sin\gamma\cos\alpha}{\sin\alpha} \\ v_Y = f\frac{\cos\gamma\cos\alpha}{\sin\alpha} \end{cases} \tag{5}$$

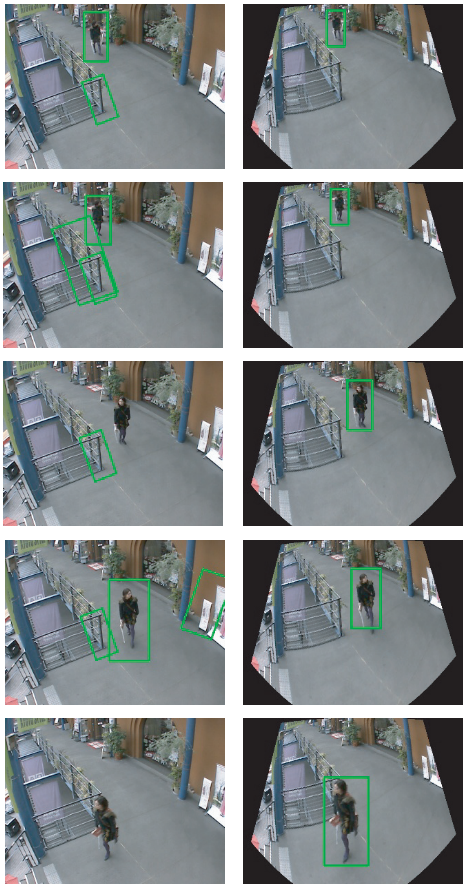
$$\begin{cases} u_Z = f\frac{\sin\gamma\sin\alpha}{\cos\alpha} \\ v_Z = -f\frac{\cos\gamma\sin\alpha}{\cos\alpha} \end{cases} \tag{6}$$

Therefore, f,α,γ are estimated as follows:

$$f = \sqrt{-u_Y u_Z - v_Y v_Z} \tag{7}$$

$$\alpha = \tan^{-1}\sqrt{-\frac{u_Z}{u_Y}}, \qquad \gamma = \tan^{-1}\left(-\frac{u_Y}{v_Y}\right) \tag{8}$$

Once the above calibration is finished, the size and direction of a human at arbitrary position in the image can be calculated, so that the window for the detection is arranged only with the appropriate size and direction according to the position. Note that the 2D image is transformed by the projection onto a cylinder surface as shown in figure 7 for implementation. By this transformation, the size and direction of the detection window can be constant on the transformed image.



(i) Simple HOG detector (ii) Proposed method

Figure 8. Comparison between simple HOG detector and proposed method.

IV. Implementation and Evaluation

A. Environment

We used an outdoor scene where the illumination changes by sunshine. The frame rate was 2.5 fps and the frame size was VGA (320×240 pixel). We used the simple HOG-based method for the initial human detection. We used images captured in three days for the additional learning step, and a hundred images randomly selected from other three days were used to evaluate accuracy of the detection.

B. Comparison between Proposed Method and Previous Method

We compared the results of the proposed method with those of the simple HOG human detector[1]. The direction of the detection window was controlled from -40 to +40 degree and the size was controlled from 32×64 pixels to 96×192 pixels.

The result of these two methods was shown in Fig.8. It was visually confirmed that the proposed method worked much better than the simple one thanks to the additional knowledge specific to the scene.

Table I
QUANTITATIVE EVALUATION

Method	Precision Rate	Recall Rate
Simple HOG detector	60.1	79.7
Simple HOG detector with additional learning	85.2	75.8
Simple HOG detector with calibration	64.9	85.2
Proposed method (containing both additional learning and calibration)	89.8	91.8

In addition, Table I shows the result of quantitative evaluation. By applying only the additional learning, the recall rate remained bad while the precision rate got increased. This is because the false negatives could not be reduced though the false positives were suppressed. On the other hand, when applying only the calibration, the false negatives were reduced but the false positives remained. It is considered that the calibration is helpful for correct the aspect ration of the human appearance but the corrected appearance may look unnatural because the correction is just 2D image warping. Compared with them, the proposed method, which combined the additional learning and the calibration, worked much better than the others. In this method, the false positives were suppressed by the additional learning, and the false negatives were also reduced because the aspect ratio of human was corrected and its unnaturalness was learnt additionally. By this consideration, it was confirmed that the idea to combine the additional learning and the calibration was very important.

V. CONCLUSIONS

This paper proposed a novel human detection method which is based on the existing learning-based method but designed so as to obtain the scene-specific knowledge and utilize it for increasing the detection performance. The scene-specific knowledge contains two kinds of information; a classifier generated by using positive and negative samples appeared in the images of the scene, and the camera calibration for reducing the possibility to incidentally find a pattern which is not a human but looks similar to a human. Experimental results showed the effectiveness of the proposed method.

Future work contains investigation of (i) the stability of the accuracy when the proposed method continues to run for a long time, and (ii) the relation between the initial and final performance in our method. Although we used the simple HOG method for the initial detector, it can be replaced by more improved methods, so that the final performance might be increased.

REFERENCES

[1] N.Dalal and B.Triggs, “Histograms of Oriented Gradients for Human Detection,” IEEE Computer Vision and Pattern Recognition, pp.886–893, 2005.

[2] B.Wu and R.Nevatia, “Detection of Multiple, Partially Occluded Humans in a Single Image by Bayesian Combination of Edgelet Part Detectors,” IEEE European Conference on Computer Vision, 2005.

[3] F.Suard and A.Broggi, “Pedestrian Detection using Infrared images and Histograms of Oriented Gradients,” IEEE Symposium on Intelligent Vehicle, pp.206–212, 2006.

[4] Q.ZhuS, Avidan, M.Yeh and K.Cheng, “Fast Human Detection Using a Cascade of Histograms of Oriented Gradients,” IEEE Computer Vision and Pattern Recognition, Vol. 2, pp.1491–1498, Jun, 2006.

[5] Imran Junejo, Hassan Foroosh, “Robust Auto-Calibration from Pedestrians,” IEEE International Conference on Advanced Video and Signal Based Surveillance, 2006.

[6] F.Lv, T.Zhao, R.Nevatia, “Camera calibration from video of a walking human,” IEEE Transaction on Pattern Analysis and Machine Intelligence, vol.28, no.9, pp.1513–1518, 2006.

[7] P. Roth, S. Sternig, H. Grabner, and H. Bischof, “Classifier Grids for Robust Adaptive Object Detection,” IEEE Conference on Computer Vision and Pattern Recognition, 2009.

[8] “Automatic Generation of Training Samples and a Learning Method Based on Advanced MILBoost for Human Detection,” Asian Conference on Pattern Recognition, 2011.

[9] J.C. Burges, “A Tutorial on Support Vector Machines for Pattern Recognition,” Data Mining and Knowledge Discovery, Vol.2, No.2, pp.121–167, 1998.

[10] Y. Freund, R. E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” In Computational Learning Theory: Eurocolt’95, pp.23–37, 1995.

[11] N.Dalal, B.Triggs and C.Schmid, “Human Detection Using Oriented Histograms of Flow and Appearance,” IEEE European Conference on Computer Vision, vol.2, pp.428–441, 2006.

[12] T. Watanabe, S. Ito, K. Yokoi, “Co-occurrence Histograms of Oriented Gradients for Human Detection,” Transactions on Computer Vision and Applications, Vol.2, pp.39–47, 2010.

[13] Y. Yamauchi, H. Fujiyoshi, Y. Iwahori, T. Kanade, “People Detection Based on Co-occurrence of Appearance and Spatiotemporal Features,” National Institute of Informatics Transaction on Progress in Informatics, No.7, pp.33–42, 2010.

[14] M. A. Fischler, R. C. Bolles, “Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography,” Comm. of the ACM, No.24, pp.381–395, 1981.