FAST HUMAN DETECTION USING TEMPLATE MATCHING FOR GRADIENT IMAGES AND ASC DESCRIPTORS BASED ON SUBTRACTION STEREO

Makoto Arie[⋆†], *Masatoshi Shibata*[⋆†], *Kenji Terabayashi*[⋆], *Alessandro Moro*[†] *and Kazunori Umeda*[†]

[⋆†] Course of Precision Engineering, School of Science and Engineering, Chuo University
[⋆] Department of Mechanics, Faculty of Engineering, Shizuoka University
[†] Department of Precision Mechanics, Faculty of Science and Engineering, Chuo University

ABSTRACT

A fast human detection system using a stereo camera is constructed. "Subtraction stereo", that can measure distance information of foreground regions, is used to restrict regions for human detection and to adapt the detection window size. Two methods are introduced for human detection. One is a method based on template matching using gradient images, and the other is a method using approximated Shape Context (aSC) descriptors focusing on human upper bodies. High human detection performance better than the standard HOG-based method with low calculation cost is achieved by the combination of the two methods. The effectiveness of the proposed system is verified experimentally.

Index Terms— Human detection, subtraction stereo, template matching, gradient image, feature extraction

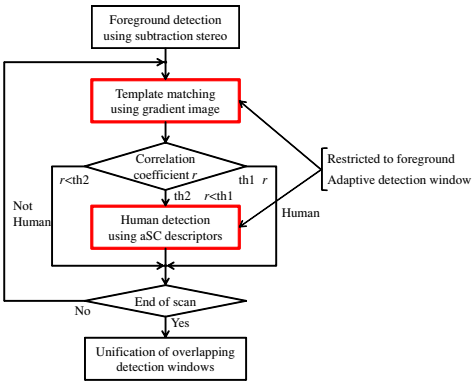


Fig. 1. Flow of the proposed scheme

1. INTRODUCTION

Human detection from images is an important issue for many applications, such as surveillance and marketing. Many studies have been presented for human detection. Most of them used a monocular vision [1, 2, 3, 4, 5, 6, 7, 8], while some adopted a stereo vision [9, 10, 11]. An overview of several approaches for human (pedestrian) detection is given in [12].

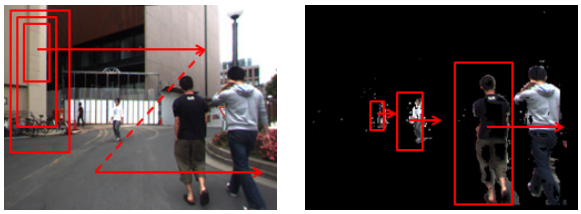
Various features for object detection from images have been proposed [1, 2, 3, 4]. Among them, the most standard one may be Histograms of Oriented Gradients (HOG) features proposed by Dalal and Triggs [1]. Human detection based on HOG features is known to be effective. However, HOG-based methods have an issue to require much calculation cost and are difficult to be executed in real time. The high calculation cost is due to the reasons that HOG features themselves require high calculation cost, and detection windows with different sizes have to be scanned repeatedly on a whole image to extract HOG features. Zhu et al. [13] extended the HOG descriptor and utilized a cascade classifier structure to fasten the detection speed. We have presented a fast human detection method [14] using "subtraction stereo" [15] with HOG features. Subtraction stereo gives distance information of foreground regions. Size of detection windows is determined based on the distance information. Additionally, calculation of HOG features and classification process are restricted to only foreground regions. Calculation cost and false detection are reduced by the usage of distance information and the restriction. However, the human detection method still relies on the time-consuming HOG features.

In this paper, we introduce a new scheme to further reduce the calculation cost without decreasing the capacity of human detection. We adopt a method based on template matching using gradient images and approximated Shape Context (aSC) descriptors [16]. Then we achieve fast and robust human detection by the combination of these two methods, with subtraction stereo.

2. FLOW OF THE PROPOSED SCHEME

Fig.1 shows the flow of the proposed scheme. The general structure is similar to the one in the previous method [14], except that template matching using gradient image and aSC descriptors are used instead of HOG features. We explain the overview of each module of Fig.1 in the following.

(1) Foreground detection using subtraction stereo In a human tracking scenario, humans are observed as foreground



(a) General (b) Proposed

Fig. 2. Adaptive scan of detection window with detection of foreground regions

regions in images. We extract foreground regions with distance information on each pixel using subtraction stereo [15]. Subtraction stereo is a modification of a standard stereo vision technique, in which background subtraction is applied to right and left camera images before stereo matching. Therefore, disparity is calculated only at foreground regions, which reduces both false stereo matching and calculation cost. The following procedures are applied only to the extracted foreground regions.

(2) Adaptation of Detection Window Size We adapt the window size for human detection using the distance information obtained using subtraction stereo. In a general human detection scenario, scan of multiple detection windows with different sizes is necessary, which is time-consuming. Using the distances of foreground regions (and assuming standard human dimensions such as 1.7m height and 0.7m width), we can adapt the detection window size.

Fig.2 illustrates the scan of the detection window in this study. In general, multiple detection windows with different sizes need to be scanned repeatedly on the whole image (see Fig.2(a)). On the other hand, a detection window with appropriate size is scanned once on only the extracted foreground regions in the proposed method as shown in Fig.2(b).

(3) Human Detection Using Template Matching and aSC Descriptors Human detection is carried out using the detection window with the adaptive size. First, template matching using gradient image is applied. If the correlation coefficient obtained by the template matching is high enough, i.e., larger than a threshold value $th1$, the target region is classified as a human region.

When a human image is partially occluded or distorted, the correlation coefficient tends to become smaller. Therefore, we apply a human detection using aSC descriptors to a target region where the template matching produces not large enough but still large correlation coefficient, i.e., larger than another threshold value $th2$. The human detection using aSC descriptors is based on local features and thus is robust to occlusion or distortion and works when a whole body of a human is not observed.

Details of the two methods are given in section 3 and 4.

(4) Unification of Overlapping Detection Windows After the human detection procedures are applied on the whole image, the human detection is finalized. Many overlapping detection windows that are classified as human regions tend to be extracted for each person. Therefore, we apply the mean-shift clustering technique [18] and unify overlapping detection windows for each person.

3. HUMAN DETECTION BASED ON TEMPLATE MATCHING

We introduce a template matching method for human detection. We do not use an image itself but a gradient image so that the edge information, which represents human contour and thus is effective for human detection, can be used. HOG features also use intensity gradient in nine directions. To reduce the calculation cost, we use intensity gradient only in one direction, i.e., horizontal direction, which corresponds to vertical edges. Furthermore, we produce two one-dimensional (1D) template images from a two-dimensional (2D) template image and use them for template matching.

3.1. Template Image Using Gradient Image

The orientation and magnitude of the intensity gradient at each point of an image are defined as

$$orientation(i,j) = \tan^{-1}\{I_j(i,j)/I_i(i,j)\} \quad (1)$$

$$magnitude(i,j) = \sqrt{I_i^2(i,j) + I_j^2(i,j)} \quad (2)$$

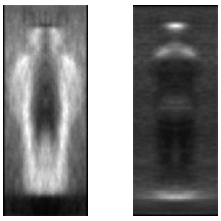
where $I_i(i,j)$ and $I_j(i,j)$ are horizontal and vertical intensity gradients respectively. The gradients can be obtained approximately using the subtraction of left and right pixel values for $I_i(i,j)$ and upper and lower pixel values for $I_j(i,j)$.

We adopt the magnitude value given in (2) when the orientation is between $\pm\pi/4$. Magnitude values are averaged for training images, and then a template image like Fig.3(a) is constructed. On the contrary, Fig.3(b) represents the gradient image obtained from the magnitude values when the orientation is from $\pi/4$ to $3\pi/4$. It is shown that horizontal gradient image represents the approximate contour shape better and is suitable for the template image.

We obtain three template images from the human images in the database for three poses: (a) front or back view, (b) side view, (c) walking. The reason of dividing in three is that the typical contour shape of a pedestrian corresponds to the three.

3.2. Template Matching Using Accumulated Gradient Features

A standard template matching using a 2D template image is still a time-consuming procedure especially when the number of pixels is large. Additionally, template matching using



(a) Horizontal (b) Vertical

Fig. 3. Gradient image obtained by averaging training images

a 2D template image tends to be too sensitive to the variation of gradient image. Therefore, we use two 1D template images obtained from a 2D template image as illustrated in Fig.4. We refer to the 1D template image as an (horizontal / vertical) accumulated gradient feature. This is constructed by accumulating the intensity values in a same column (horizontal) or row (vertical). The horizontal axis of the accumulated gradient feature represents the horizontal or vertical coordinates, and the vertical axis represents the accumulated value in a column or a row that is normalized from 0 to 1.

Template matching is carried out using the two accumulated gradient features. We adopt the normalized cross-correlation (NCC) method for template matching. The similarity between the template image and the target region in an image is evaluated using the correlation coefficient that is obtained by the following equation.

$$R_{NCC} = (R_{NCC_H} + R_{NCC_V})/2 \tag{3}$$

where R_{NCC_H} and R_{NCC_V} are the correlation coefficients for horizontal and vertical gradient features respectively.

R_{NCC} becomes from -1 to 1, and when similarity of the template image and the target region is high, it becomes close to 1. We prepare three templates as explained above. When R_{NCC} becomes larger than a threshold for one or more templates, the target region is classified as a human region.

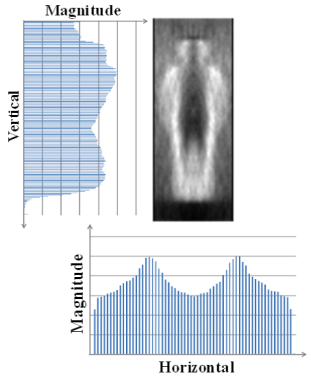
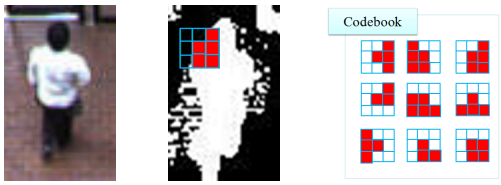


Fig. 4. Accumulated gradient features obtained from template gradient image



(a) Color image (b) aSC descriptor (c) Codebook

Fig. 5. Outline of the aSC descriptors

4. HUMAN DETECTION USING ASC DESCRIPTORS

The human detection using template matching works well if the whole body of a person is observed. However, it is often the case that the lower body of the person is not observed well or occluded especially when a camera is set at a high position and tilted. To deal with such cases, we extract local features from an upper body of a person and use them for classification.

We adopt aSC descriptors [16] as the local features for representing an upper body. This feature is easy to extract with low calculation cost and thus combination with the human detection using template matching is possible. Fig.5 illustrates the aSC descriptors. They consist of 3×3 cells and are extracted from a binary image as shown in Fig.5(b). Each cell has a binary value.

aSC descriptors are used to represent a human upper body. A human upper body is represented using a set of aSC descriptors, which is called a Codebook. Fig.5(c) shows examples of the aSC descriptors of a Codebook. A head or shoulders of a human, which are discriminative parts to represent a human, tend to have horizontal edges and have large vertical gradients (see Fig.3(b)). Therefore, we extract aSC descriptors from pixels having large vertical gradients. We use 15 kinds of aSC descriptors for constructing a Codebook. In human detection, we extract 15 kinds of aSC descriptors from a target region corresponding to a detection window. If more than half of the features correspond to the ones in the Codebook, then the region is classified as a human region.

5. EXPERIMENTS

We show experimental results to evaluate the proposed human detection system. Dalal and Triggs' method using the HOG features [1] was compared as a reference.

5.1. Experimental Conditions

We used a stereo camera Point Grey Research Bumblebee2 (color, f=3.8mm, 48fps) and implemented the proposed methods using a laptop PC Lenovo W700 (CPU: Intel Core2 Duo 3.06GHz, RAM 6GB). Range images and color images of 320×240 pixels are obtained simultaneously with the stereo

camera. In the reference method, the size of detection window was set to 30×60 pixels and the detection window was scanned on the whole image. The threshold values $th1$ and $th2$ in Fig.1 were set empirically to 0.3 and 0 respectively. We used NICTA pedestrian dataset [19] to make three template images for the template matching. 1000 pedestrian images were used. To make a codebook of aSC descriptors, we prepared our own dataset. 100 binary images were used.

5.2. Experimental Results

Table 1 and Fig.6 show the experimental results. The height of the camera position is 5.5m, and camera's tilt angle is 45° in the experiments. TPrate, FDrate, Precision, and Time represent True Positive rate: rate of the detected human out of every human, False Detection rate: rate of falsely detected non-human against every human, the rate that the detected human is truly a human, and the processing time per a image respectively. We used 1000 frames for each experiment.

Table 1 shows the following results. The proposed scan based on subtraction stereo works well, i.e., calculation cost is reduced much and the indexes of false detection rate and precision are improved. Each of the two proposed methods is worse than the reference method based on HOG features, which is quite natural. However, each of them produces good performance comparable to the reference method when combined with the proposed scan. The combination of the template-matching-based and aSC-descriptors-based methods gives good results thanks to their complementary characteristics. Even in full scan case, the performance is comparable to the reference method, and when the proposed scan is applied, every index is better than the reference method.

And we obtain the following results from Fig.6. In (a), we can see that human detection at the distorted human regions fails. In (b), we can see some falsely detected regions. And in (c), the best human detection result is obtained by the combination of the two methods.

Table 1. Performance comparison between the proposed methods and the standard method using HOG features [1]

	TPrate (%)	FDrate (%)	Precision (%)	Time (ms)
HOG(full scan)	79.4	9.6	89.2	158.3
Template(full scan)	73.6	15.2	82.8	32.8
Template(proposed)	72.3	8.2	89.8	8.2
aSC(full scan)	82.9	37.9	68.6	51.2
aSC(proposed)	81.4	12.8	86.4	11.8
Both(full scan)	82.9	15.0	84.7	64.8
Both(proposed)	80.5	7.9	91.1	16.2

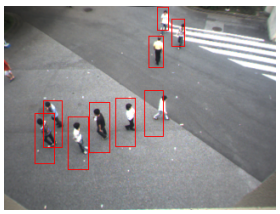
Fig.7 and Table 2 show human detection results for other scenes. We used 100 frames for Scene 1 and 2. Scene 0 is the same one in Table 1. From these results and Table 1, we can say that the proposed scheme works at various scenes.

(a) Template matching only

(b) aSC descriptors only

(c) Combination of template matching and aSC descriptors

Fig. 6. Comparison of human detection results



(a) Scene 1

(b) Scene 2

Fig. 7. Human detection examples for other scenes

6. CONCLUSIONS

We have proposed a fast human detection system using a stereo camera. Regions to detect a human are restricted and the detection window size is adapted using "subtraction stereo". We introduced two human detection methods: a method based on template matching using the gradient images and a method using aSC descriptors extracted from upper bodies. High human detection performance better than the standard HOG-based method with low calculation cost is achieved by the combination of the two methods with subtraction stereo. Future work includes improvement of the method for occlusion, etc., and application to practical problems.

Table 2. Human detection results for various scenes

	TPrate (%)	FDrate (%)	Precision (%)	Time (ms)
Scene 0	80.5	7.9	91.1	16.2
Scene 1	83.6	6.1	93.1	14.2
Scene 2	75.2	10.3	87.9	17.8

7. REFERENCES

[1] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, Proc. of CVPR, Vol.1, pp.886–893, June 2005.

[2] P. Viola, J. Jones, Rapid object detection using a boosted cascade of simple features, Proc. of CVPR, pp.511–518, 2001.

[3] K. Levi, Y. Weiss, Learning object detection from a small number of example: The importance of good feature, Proc. of CVPR, Vol.2, pp.53–60, 2004.

[4] B. Wu, R. Nevatia, Detection of multiple, partially occluded human in a single image by bayesian combination of edgelet part detectors, Proc. of ICCV, vol.1, pp.90–97, 2005.

[5] A. Shashua, Y. Gbalyahu, and G. Hayun, Pedestrian detection for driver assistance systems: Single-frame classification and system level performance, Proc. of the IEEE Intelligent Vehicle Symposium, 2004.

[6] P. Viola, M. Jones, and D. Snow, Detection pedestrian using patterns of motion and appearance, Proc. of ICCV, pp.734–741, 2003.

[7] P. Sabzmeydani, G. Mori, Detection pedestrians by learning shapelet features, Proc. of CVPR, 2007.

[8] O. Tuzel, F. Porinki, and P. Meer, Human detection via classification on riemannian manifolds, Proc. of CVPR, 2007.

[9] D. M. Gavrila, S. Munder, Multi-cue pedestrian detection and tracking from a moving vehicle, IJCV, vol.73, pp.41–59, 2007.

[10] M. Bajracharya, B. Moghaddam, A. Howard, S. Brennan, and L. H. Matthies, Results from a real-time stereo-based pedestrian detection system on a moving vehicle, IEEE Workshop on People Detection and Tracking at ICRA, 2009.

[11] A. Ess, B. Leibe, K. Schindler, and L. Van. Gool, Moving obstacle detection in highly dynamic scenes, Proc. of ICRA, pp.56–63, 2009.

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

[13] Q. Zhu, S. Avidan, M. Yeh, and K. Cheng, Fast human detection using a cascade of histograms of oriented gradients, Proc. of CVPR, pp.1491–1498, 2006.

[14] M. Arie, A. Moro, Y. Hoshikawa, T. Ubukata, K. Terabayashi, K. Umeda, Fast and Stable Human Detection Using Multiple Classifiers Based on Subtraction Stereo with HOG Features, Proc. of 2011 IEEE International Conference on Robotics and Automation, pp.868–873, May 2011.

[15] K. Umeda, et al., Subtraction Stereo - A Stereo Camera System That Focuses On Moving Regions -, Proc. of SPIE-IS&T Electronic Imaging, Vol. 7239 Three-Dimensional Imaging Metrology, 723908, 2009.

[16] C. Beleznai, Fast Human Detection in Crowded Scene by Contour Integration and Local Shape Estimation, Proc. of CVPR, pp.2246–2253, June 2009.

[17] A. Moro, et al., " Auto-adaptive threshold and shadow detection approaches for pedestrian detection, Proc. of AWSVCI, pp.9–12, 2009.

[18] D. Comaniciu, P. Meer, Mean Shift Analysis and Applications, IEEE International Conference on Computer Vision, pp.1197-1203, 1999.

[19] G. Overett, L. Petersson, N. Brewer, L. Andersson and N. Pettersson, A New Pedestrian Dataset for Supervised Learning, Proc. IEEE Intelligent Vehicles Symposium, pp.373–378, 2008.