Fast Human Detection Algorithm Based on Subtraction Stereo for Generic Environment

Alessandro Moro*, Makoto Arie†, Kenji Terabayashi† and Kazunori Umeda†

*University of Trieste / CREST, JST
†Chuo University / CREST, JST

Abstract. In this paper, we propose a fast and stable human detection based on subtraction stereo for generic environments. Scanning an input image by detection windows, the numbers and size is controlled by distance obtained from subtraction stereo. A preprocessing phase reduces the effect of noise and shadow, for a better segmentation. This control decrease the computation time, important in real-time applications, skipping a large number of detection windows. Experimental results show that the proposal is faster and less false detection.

Keywords: Human Detection, Subtraction Stereo, HOG, AdaBoost

INTRODUCTION

Human detection can be widely used in many applications, including people counting and security surveillance in public scenes. However, detection of humans is still a challenging task because of their various appearances, occlusion problem, and environmental conditions. In the recent years various methods for humans and objects detection have been proposed by Viola *et al.* [1], Levi *et al.* [2], and Wu *et al.* [3].

One of the most successful approaches of human detection is based on HOG (Histograms of Oriented Gradients) descriptor proposed by Dalal *et al.* [4], which is robust to illumination changes due to using edge information. On the other side, it requires a lot of computational time for the following three reasons: (i) computation of HOG feature, (ii) detection windows scanning whole area in an input image, (iii) multiple sizes of detection windows.

Additionally, false detection increases with the number of detection windows increasing rapidly because of the above points (ii) and (iii). Therefore, a large number of detection windows are inappropriate in terms of both computational cost and accuracy.

In this paper, we compare two features as descriptor of objects and human. We propose a flexible algorithm to reduce the computation time and increase the detection rate performance which is faster and less false detection than the method described in the reference [4]. It is based on subtraction stereo [5] which calculates distance information focusing on only foreground regions in an input image and it reduces the possible candidate regions. The size of detection window is determined appropriately and dynamically based on the distance information and paraperspective projection model. The detected region is then classified by a Real AdaBoost-HOG trained classifier.

The paper is organized as follow. A brief resume of other works. The proposed algorithm is described in details and results showed. Last, conclusions and future work.

FIGURE 1. Overview of the proposed human detection algorithm based on subtraction stereo. Distance information of foreground regions are obtained from the subtraction stereo through shadow removal (green pixel in the center image). From left to right. Input image, foreground candidate obtained by subtraction stereo, shadow detection, disparity image, detected humans.

RELATED WORKS

Recent research on human detection has used monocular vision [4, 6, 7], stereo vision [8, 9] and LIDAR sensing [10]. An overview of several approaches for pedestrian detection can be found in [11]. One of the most popular recent appearance based human detection algorithms is the HOG method proposed by Dalal and Triggs [4]. They characterized human regions in an image using HOG descriptor, which are a variant of the well-known SIFT descriptor [12]. Unlike SIFT, which is sparse, the HOG descriptor offers a denser representation of an image region by tessellating it into cells which are further grouped into overlapping blocks. Zhu et. al [13] extend the HOG descriptor and utilize a cascade classifier structure to increase detection speed.

Ess et al. [9] describe a stereo based system for three dimensional dynamic scene analysis from a moving vehicle, which integrates sparse three dimensional structure estimation with multi-cue image based descriptors to detect pedestrians. The authors show that the use of sparse three dimensional structures significantly improves of the performance of the pedestrian detector. Still, the best performance cited is 40% probability of detection at 1.65 false positive per image frame. While the structure estimation is done in real-time, the pedestrian detections is significantly slower.

Bajracharya et al. [8] describe a real-time stereo-based system that can detect human up to 40m in highly cluttered environments. The stereo range maps are projected into a polar-perspective map that is segmented to produce clusters of pixel corresponding to upright objects. Geometric features is computed for the resulted three dimensional point clouds and used to train pedestrian classifiers.

In [7], a pedestrian detection method based on the covariance matrix descriptor [14] is proposed and shows better performance on the INRIA dataset [4] than the HOG descriptor, but an experimental study conducted by Paisikriangkarai et al. [15] shows that the covariance matrix descriptor is slightly inferior to the HOG descriptor on the DaimlerChrysler pedestrian benchmark dataset created in [16].

PROPOSED ALGORITHM

We proposed an algorithm which performs fast and low false alarm. From a sequence of images, foreground pixels are extracted from an input stereo image by subtraction stereo. Then, the candidate moving pixels are used to perform human detection, by

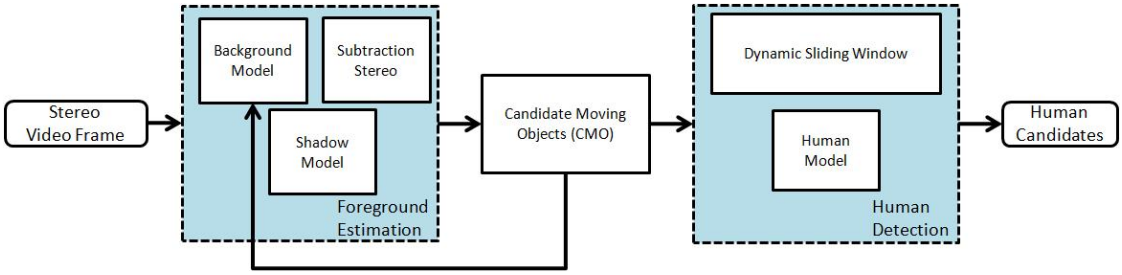

FIGURE 2. Overview of the proposed algorithm.

scanning the input image with a scrolling window only to foreground regions. Therefore, since a background region is not scanned, computation time and false alarm decreased. Gathered regions are in the end classified. A schematic model of the proposed algorithm is depicted in Fig. 2.

Foreground Detection

Foreground is important in human detection. People generally move in an environment and, due to their movements, it is possible to extract information about dynamic by analysing of temporary information. In literature several works have been proposed about change detection. Some of these works have good performance in specific field but we aim to a flexible system which can be easily adopted in different indoor or outdoor environments. We apply the method proposed in [17] which showed good performance and it is reliable in several environments. The image segmentation is divided in four steps: subtraction stereo, shadow detection, removal of periodic changes, background maintenance.

Subtraction Stereo

The edge information includes much information for human detection. Therefore we remove background information by the subtraction stereo. We show the algorithm of the subtraction stereo [5]. The subtraction stereo extracts foreground regions in a scene by background subtraction method and a disparity image is obtained by the stereo matching usable to measure actual heights and widths of the foreground regions. An example of disparity image is shown in Fig. 1

Shadow Detection

Shadow detection is used to refine the foreground. The image obtained using subtraction stereo includes noise affected by the shadow. This noise seriously affects the human detection. We have improved the shadow detection described in [18] combined

with the stereo information. When we define I(x,y) as the intensity of the pixel located in the two-dimensional image position (x,y) and I'(x,y) as the intensity of the background pixel, and d the distance obtained by the stereo system, the equation for the evaluation of shadow is described as

$$\Theta_{(t,x,y)} = \begin{cases} \alpha\Psi_{(x,y)} + \beta\Lambda_{(x,y)} + \frac{\gamma}{d^2}\Phi_{(x,y)} + \left(1-\alpha-\beta-\frac{\gamma}{d^2}\right)\cdot\Theta_{(t-1,x,y)} & \text{if } \frac{I_{x,y}}{\eta} < I'x,y \\ \infty & \text{otherwise} \end{cases} \tag{1}$$

where Θ corresponds to a shadow value. This value will be applied a threshold to determine if a pixel is a shadow. A small shadow value corresponds to a shadow point. The functions Ψ, Λ, Φ and show color constancy between pixels, color constancy within pixel and a distance function. α , β, γ, and η are constant weights of textures, colours, distance, and intensity, which are determined empirically in our experiments. The details of this method are explained in [17]. An example is shown in Fig. 1.

Removal of Periodic Changes

Generally speaking it is impossible to establish a priori if an environment contains only objects of interest or undesired objects. Outdoor environments are usually surrounded by trees and leaves which can be cause of troubles in the image segmentation. We used a method previously described in [17] which takes the advantage of compression based on codebook but considers only the periodic changes in a certain interval of time.

Background maintenance

Image segmentation proposed in this paper is based on change detection. It is important to manage the background in order to segment the image with the highest accuracy. We proposed a method of dynamic background maintenance pixel-wise based. If a pixel of coordinate *x,y* represent the background B and F is foreground, the background at time t is:

$$B_{x,y,t} = \begin{cases} F_{x,y} & \text{if} \Delta T_{x,y} \geq \vartheta_{x,y} \\ B_{x,y} & \text{otherwise.} \end{cases} \tag{2}$$

where ΔT consider the elapsed time between the current instant and last change detected, related to the total number of changes in a period of time, and $\vartheta_{x,y}$ is a dynamic threshold which depends on the changes detected. For further details please see [17].

Human Detection

Image segmentation has an important role in image processing. However with the evolution of the machines and applications in environment with human presence, it is necessary to be able to recognize humans in order to perform more complex tasks.

The human detection procedure based on the HOG feature at subtraction image: (1) extracts foreground and computes distance of its regions using the subtraction stereo described in the previous section; (2) dynamically changing detection window size by the distance; (3) compute the HOG feature of foreground regions for each scanning detection window; (4) discriminate whether human or not by an AdaBoost classifier.

Features

To properly describe an object or human is an important task. Local characteristics of an image are generally used for that purpose. However, because many different features exists, we compare two types of features in order to establish which could be used to describe human models. We compared HOG with Discrete Cosine Transform (DCT) function which are both widely used to describe local information.

DCT Features

DCT is an important function to describe the variation of information between two areas neglecting repetitions. It expresses the representation of a region of the image in forms of sum of cosine functions. DCT found large applications in lossy compression, and it can be used as feature in order to recognize objects or faces as in [19]. For the experiment we used DCT-II, and according to our results, we found best performance setting the parameters as follow. Region size 25 x 25 pixel, DCT expressed by 16 parameters and sliding window moving of 3 x 3 pixel.

HOG Features

The HOG feature has shown success in object detection [4] and they are accepted as one of the best features to capture gradient information. However, it cannot compute its information quickly. The histogram of the gradient orientation is used for analysis of the edge orientation and its magnitude. It is created in constant number, which is called cell. Since the size of the detection window changes dynamically, it also changes the size of cell according to it. We predefine the number of cell 6x12 and in a square region. The number of the bins of the histogram is decided by the number of partitions of the gradient orientation. We predefine the number of orientation bin is 9. Then the histogram is normalized in predefined regions, which is called block. We predefine the block size 3x3 cell and in a square region. Eq. 3 is used for normalization.

$$f = \frac{V}{\sqrt{||V||^2 + \varepsilon^2}} \tag{3}$$

where V is HOG feature vector, ε is a small regularization constant.

Features Performance Evaluation

We initially tested a RBF kernel Support Vector Machine trained with the descriptors from DCT or HOG. We tested the performance in a subset of the Caltech101 database [20]. We are interested in exploring which feature could be in use to describe objects models, then we removed the borders by a Cross Correlation algorithm

Given a template image (i.e. Fig. 5) and input image we searched the area which maximized the similarity scaling and rotating the input template. We manually discarded the results which does not contain images. The remaining images were divided in training set and classification set.

The classification results can be seen in Table 1. SVM trained with DCT with a subset of images showed better performance. However, even if DCT works well with a large size of categories, HOG performed better looking for human category.

Dynamic Sliding Window

To calculate HOG feature scanning the detection window end to end of an image, or multiple detection windows in different size, has high computation time. We propose a method of dynamically changing window size using subtraction stereo which reduce the computation time, and false detection. First, the scanning detection window size is computed from the distance of the foreground regions. From the foreground image, a graph-based algorithm is used to group the pixels belong to the same blob. Second, the height of detection window size is rectified by the position of the blob in the image. Because we assume the paraperspective projection, we rectify the height of detection window size. As a reason for selecting the paraperspective projection while there is various perspective projection, the weak perspective projection rectifies only the height to the distance, but the paraperspective projection rectifies it in consideration both the different in vision occurs with the position of blob in image and elevation angle and height of a camera. From these, the detection window of different size is scanned at once.

For example, when the height of a camera is 1.6m and elevation angle is 0 degree, the size (height and width of pixel) in each distance from a human to a camera is shown in Fig. 4. Since the distance and the size from a human to a camera have the relation of an inverse proportion from (4), the constant of proportion of height and width are computed. In addition, how to see a human in distance differs on a scene. Therefore, the height and width of the scanning detection window size are rectified by eq. 4, eq. 5, which assumed paraperspective projection.

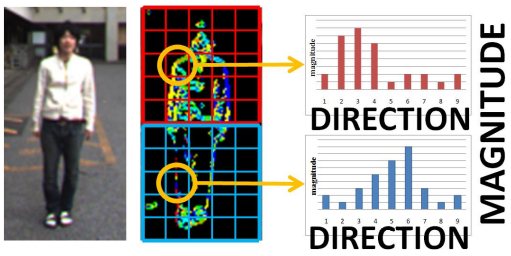

FIGURE 3. Example of estimation of HOG in a full body human figure.

$$height = \frac{k_h}{Y_w}(\cos\theta - y\sin\theta) \tag{4}$$

$$width = \frac{k_w}{Z_c} \tag{5}$$

Where, k_h,k_w are constant, Y_W is distance of world coordinate system, Z_C is distance of axis direction, θ is elevation angle of a camera, y is image coordinates which normalized the length to 1.

Classification and Clustering

Real-time applications are important for a large number of task. If SVM can cluster the class of a set of objects, the training and classification time make it not easily adaptable for this task. We opted for an AdaBoost algorithm which offer high classification rate and can be applied in real-time.

Real AdaBoost Training

Where our purpose is to have a flexible human detection algorithm, we gather a large dataset of images. For this goal we propose the use of Real AdaBoost [1] to learn the classification function. This is because Real AdaBoost is an effective and efficient learning algorithm for training on high-dimensional large dataset. Compared with other statistical learning approaches (e.g., SVM), which try to learn a single powerful discriminant function from all the specified features extracted from training samples, the AdaBoost algorithm combines a collection of simple weak classifiers on a small set of critical features to form a strong classifier using the weighted majority vote. This means the AdaBoost classifier can work very fast in the testing stage. Furthermore, AdaBoost is not prone to over fitting and provides strong bounds on generalization which guarantees the comparable performance with SVM. Gathering a representative set of negative samples is very difficult. To overcome the problem of defining this extremely large negative class, bootstrapping training is adopted. A preliminary classifier is trained on an initial training set, and then used to predict the class categories of a large set of patches randomly sampled from many added to the negative training set for the next iteration of training.

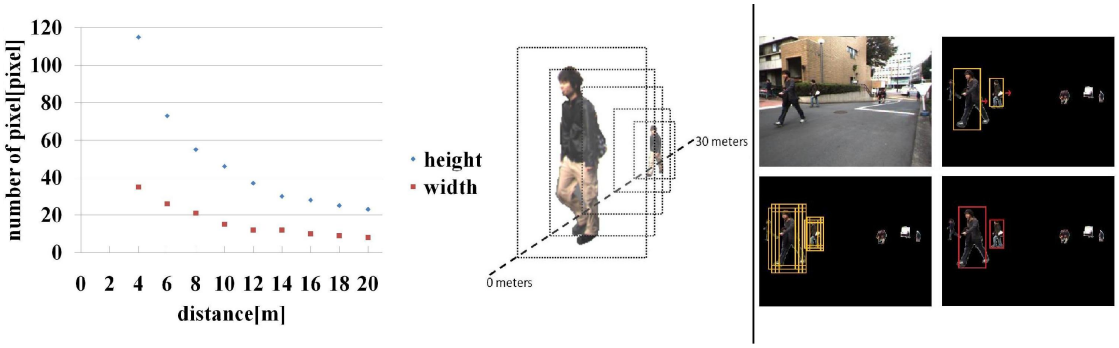

FIGURE 4. Detection windows recognized as human are unified using mean shift clustering. On the left side an example of how the size of the detection window is calculated. On the right, an example of foreground detection, multiple regions are classified as human, and final position.

Clustering

To obtain the number of targets and the exact location of each target from these detection windows, we scan sliding windows in the subtraction image. As a result, as shown in Fig. 4, there will be many detection around each target. The detection window recognized as a human is unified using mean shift clustering [21].

EXPERIMENTAL RESULTS

In this section we compare the performance of two different features (HOG and DCT) with a database of multiple objects and with a single class of human to detect. Our multiple classifiers using two-step boosting with that of the HOG-SVM detector [4], one state-of-the-art human pedestrian detector by the data set built uniquely. We also compare the human detection results between normal image and subtraction image. The results have been obtained using an Intel Core2 Duo CPU, 3.00 GHz with 4GB RAM.

We verify the validity of our method by the dataset built uniquely. The positive sample of the data set is constituted in consideration of elevation angle from 0 to 50 degrees. The number of the images of positive sample is each 3500 images per 10 degrees in total 17,500 images. The number of negative sample images is 20,000, and the elevation angle is not related. The size of all sample images is 64x128. Since it aims at applying to a surveillance camera, it is verifying on the scene of various elevation angle. The images in Fig. 5 are an example of images used as training data.

In first we compared the features performance to evaluate which would better describe objects and human shape. Results in table 1 shows that DCT with SVM can classify better in a large set of images. HOG however better perform in human classification because enhance the gradients and then less suffer of light conditions.

We compared the performances of human detection between normal image and subtraction image on an original data set using HOG feature which better describe humans. The classifier used in this experiment is a one-step boosting. The human detection results are evaluated by four measures, True Positive rate (TP), False Positive rate (FP),

FIGURE 5. Training images: Caltech [20] samples with example of template used for the cross-correlation (left), and the database used for human detection (right) with positive samples (human) and negative samples.

TABLE 1. SVM Features comparison

	HOG(%)	**DCT(%)**
Caltech101 [20]	14.1	32.9
Only Human - INRIA database [4]	85.4	76.3

Precision (TP/(TP+FP)) and Processing Speed (PS). There are two verification methods by the normal image: (1) scanning detection window size is fixed by 30x60; (2) its size is fixed by 60x120. Then paraperspective projection is assumed and the size of detection window recognized to be a human is rectified. As Table 2 shows, the human detection with subtraction image improved FP and PS than with normal image [4].

CONCLUSION

In this paper, we presented a fast and stable human detection algorithm using the subtraction stereo with HOG feature. We compared the classification performance of two different features and proposed the results. Since the scanning region was putted down to the foreground regions by using the subtraction stereo, the accuracy and processing speed of human detection have been improved. In the future we will aim to improve the detection results by sensor-fusion.

REFERENCES

1. P. Viola, and J. Jones, *Rapid object detection using a boosted cascade of simple features,* **IEEE Computer Society Conference on Computer Vision and Pattern Recognition**, pp. 511–518 (2001).

TABLE 2. Performance comparison between the proposal and HOG-SVM

	TP(%)	**FP(%)**	**Precision(%)**	**PS(ms)**
Reference (30 x 60) [4]	77.2	10.5	88.0	423.4
Reference (60 x 120) [4]	71.2	7.8	90.1	223.8
Proposed	78.5	3.2	96.1	58.3

FIGURE 6. Example of human detection in some indoor and outdoor environments.

2. K. Levi, and Y. Weiss, *Learning object detection from a small number of example: The importance of good feature,* **IEEE Computer Vision and Pattern Recognition, vol. 2**, pp. 53–60 (2004).
3. B. Wu, and R. Nevatia, *Detection of multiple, partially occluded human in a single image by a bayesian combination of edgelet part detectors,* **ICCV, vol. 1**, pp. 90–97 (2004).
4. N. Dalal, and B. Triggs, *Histogram of oriented gradients for human detection,* **Int. Conf. on Computer Vision and Pattern Recognition, vol. 2**, pp. 886–893 (2005).
5. K. Umeda, et al., *Subtraction Stereo - A Stereo Camera System that focuses on Moving Regions,* **Proc. Of SPIE-IS&T Electronic Imaging, 7239 Three-Dimensional Imaging Metrology** (2009).
6. P. Sabzmeydani, and G. Mori, *Detection pedestrians by learning shapelet features,* **CVPR** (2007).
7. O. Tuzel, F. Porinki, and P. Meer, *Human detection via classification on Remannian manifolds,* **CVPR** (2007).
8. 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).
9. A. Ess, B. Leibe, K. Schindler, and L. Van. Gool, *Moving obstacle detection in highly dynamic scenes,* **ICRA** (2009).
10. K. Fuerstenberg, K. Dietmayer, and V. Willhoeft, *Pedestrian recognition in urban traffic using a vehicle based multilayer laserscan,* **IEEE Intelligent Vehicle Symposium, vol. 1** (2002).
11. P. Dollar, C. Wojek, B. Schiele, and P. Perona, *Pedestrian detection: A benchmark,* **CVPR** (2009).
12. D. G. Lowe, *Distinctive image feature from scale-invariant key-points,* **IJCV, vol. 60**, pp. 91–110 (2004).
13. Q. Zhu, S. Avidan, M. Yeh, and K. Cheng, *Fast human detection using a cascade of histograms of oriented gradients,* **CVPR** (2006).
14. O. Tuzel, F. Porinki, and P. Meer, *Region covariance: A fast descriptor for detection and classification,* **CVPR** (2006).
15. S. Paisitkriangkrail, C. Shen, and J. Zhang, *An experimental study on pedestrian classification using local features* **IEEE Inter. Symp. on Circuit and System (ISCAS)** (2008).
16. S. Munder, and D. Gavrila, *An experimental study on pedestrian classification,* **PAMI, 28(11)**, pp. 53–60 (2006).
17. A. Moro, K. Terabayashi, and K. Umeda, *Detection of moving objects with removal cast shadow and periodic changes using stereo vision,* **ICPR**, pp. 328–331 (2010).
18. A. Moro, et al., *Auto-adaptive threshold and shadow detection approaches for pedestrian detection,* **In Proc. AWSVCI, 2**, pp. 9–12 (2009).
19. A. Nefian, *A Hidden Markov Model-Based Approach for Face Detection and Recognition,* **PhD Thesis**, Georgia Institute of Technology (1999).
20. L. Fei-Fei, R. Fergus, and P. Perona, *Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories,* **IEEE. CVPR, Workshop on Generative-Model Based Vision.** (2004).
21. D. Comaniciu, and P. Meer, *Mean Shift Analysis and Applications,* **ICCV**, pp. 1197–1203 (1999).