Human Detection from Ground Truth Cameras through Combined Use of Histogram of Oriented Gradients and Body Part Models

Tian-Rui Liu, Valentine Copin and Tania Stathaki
Department of Electrical and Electronic Engineering, Imperial College London, London, U.K.

Keywords: Human Detection, Body Part-based Models, Histogram of Oriented Gradients.

Abstract: Vision based human detection continuously attracts research interest since it is a topic of practical significance. The well-established Histogram of Oriented Gradients (HOG) human detector, though regarded as a reference for human detection, still suffers from the typical problem of the trade-off between precision and recall, relying on the threshold of its classifiers. In this paper, we propose a human detection system which can provide both good precision and recall without the need for adjusting the classification thresholds. Our strategy is to combine the HOG detector with a body part model in order to eliminate the false detections that do not match the human silhouette (body) model. For this purpose, a probabilistic model of the human body is learned to describe the relative position between the distinctive body parts. A HOG detection would be retained if the body parts can be detected in the confidence areas provided by the learned body model. Moreover, the body parts detectors are boosted cascade classifier learned with the Haar, HOG or LBP features. The multi-modal feature representation of the different human body parts is more robust against variations in human appearances. Experiment results on the INRIA data sets show that our human detector achieves a precision of 70% at a recall of 50%, which cannot be achieved by the HOG detector under any parameter settings.

1 INTRODUCTION

Human detection has been at the forefront of current research in machine vision with many applications such as video surveillance, car safety, robotics, biometrics and others. The human detection problem is often hindered by difficulties such as various types of occlusions and changes in human pose and/or appearance. A substantial number of methods have been developed over the years and much progress has been done in terms of detection rate and accuracy and also computation time.

Many of the previous human detection approaches attempt to represent the entire human as a single object. What follows is a brief literature review on the problem of human detection. In (Papageorgiou and Poggio, 1999), the SVM classifier was learned to be applied on the entire human body for pedestrian detection. A shape model for human body has been proposed in (Felzenszwalb, 2001), where human positions are inferred via template matching based on the Chamfer distance. Viola and Jones used their Haar cascade detector for pedestrian detection in (Viola et al., 2003). The Haar detector was developed originally for real-time face detection (Viola and Jones, 2001). The basic idea of this method is to select weak classifiers with the AdaBoost algorithm (Freund and Schapire, 1996). However, direct utilization of the Haar features for human detection does not work well and therefore, the researchers mentioned above improved their detection system by using additional motion information, which achieved much better performance. In 2005, Dalal and Triggs introduced the well-established HOG-SVM detector, based on the Histogram of Oriented Gradient (HOG) descriptors (Dalal and Triggs, 2005). Following the work of (Viola et al., 2003), a boosted cascade classifier based on the HOG features was proposed in (Zhu et al., 2006) to speed up the HOG-SVM algorithm.

Another promising line of research commenced recently, exploring body part-based models to deal with occlusion and handle with multiple body poses. Mohan et al. (Mohan et al., 2001) divided human body into head-shoulder, legs, left and right arm and they trained SVM classifiers to learn each body part

Liu, T-R., Copin, V. and Stathaki, T.
Human Detection from Ground Truth Cameras through Combined Use of Histogram of Oriented Gradients and Body Part Models.
DOI: 10.5220/0005853407350740
In *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2016) - Volume 4: VISAPP*, pages 735-740
ISBN: 978-989-758-175-5

using Haar wavelet features. Mikolajczyk et al. (Mikolajczyk et al., 2004) modelled human body by employing seven parts. For each part, a detector was learned by using orientation-based features similar to those of the SIFT descriptor (Lowe, 1999) and aprior Gaussian mixture model for upper body was used to calculate a pose likelihood and handle various body poses. A similar method based on associating a probabilistic assemble of body parts into a full body configuration was presented in (Micilotta et al., 2005). Additional skin colour information was also taken into account to calculate the overall joint likelihood required for the final body configuration. In (Felzenszwalb et al., 2008) and (Felzenszwalb et al., 2010) the HOG-SVM (Dalal and Triggs, 2005) was used as a building block for a proposed so called deformable part-based models.

Several types of features have been applied to capture the key characteristics of humans. Various types of local features, such as SIFT, Haar wavelets, and HOG, have been compared in (Dalal and Triggs, 2005). Their experiments show that the HOG detector, which employs the HOG descriptors and a trained SVM classifier, outperforms the other types of features for the human detection task. Following this comparative study, the HOG detector has become a reference for human detection. However, this detector still manifests a trade-off between the detection precision and recall. Obviously, the false alarm and missdetection rates follow opposite trends and their values are related to the HOG descriptor parameters (block size, cell size, and block stride) and the classification threshold of the HOG-SVM classifier. As it has been verified in various experiments, a smaller block stride in the HOG detector, yields lower achieved missdetection rate (higher recall) but higher resulting false alarm rate (lower precision) (Dalal and Triggs, 2005). Besides, increasing the SVM classification threshold makes the classification more stringent so that the number of false detections are reduced (better precision) at the expense of higher number ofmissed detections (lower recall).

Motivated by the behaviour of the HOG detector with respect to the threshold of its corresponding SVM classifier, we aim to propose a human detection system providing both good precision and recall without adjusting the classification thresholds. For this purpose, we combine the HOG detection with a learned human body model so that the detections which do not match the body model (and they are hopefully false detections) are removed. The HOG detector is first employed with the goal of detecting as many human candidates as possible in order to ensure a high recall. The detection precision is thereafter increased relying on additional individual body part detections. A body model is learned, based on Gaussian distributions, to describe the one-to-one geometric relations between the body centroid and each body part. A candidate human detected by the HOG detector would be retained if at least one of the body parts is found in the confidence areas with respect to the learned Gaussian body model. To better capture the characteristics of different body parts, the Haar, HOG, and LBP feature descriptors are incorporated while we building the boosted classifiers for body parts detectors.

The rest of the paper is constructed as follows. Section 2 presents the proposed method. Experimental results are addressed in Section 3. Finally, conclusions are given in Section 4.

2 PROPOSED METHOD

2.1 Overview of the Proposed Human Detection Method

The proposed human detection method explores the part-based representation of human body to verify the detections of the HOG detector. The detection framework consists of two main phases.

Firstly, potential human bodies and body parts are detected at all locations and scales using the HOG pedestrian detector and body part detectors respectively. In particular, the HOG detector with a classification threshold $\Delta = 0$ is applied to select as many potential human candidate regions as possible. The detectors for the five human body parts are boosted cascade classifiers learned with different features.

In the second phase, the false HOG detections are removed based on information provided from additional body part detections. A probabilistic body model for upright humans is learned to describe the geometric relations between each body part and the body centroid. The learned body model provides high confidence neighbourhoods for searching for the head, upper body, and lower body. A candidate human detected with the HOG detector would be retained if at least one body parts lies in the corresponding neighbourhood.

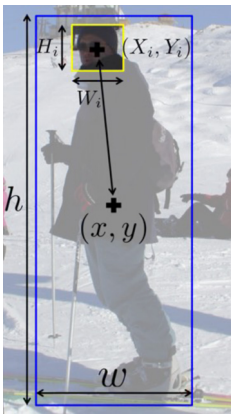

Figure 1: Illustration of the parameters to be used in the probablistic human body model.

2.2 Body Models

2.2.1 Body Parts

We use five body parts to represent a human, namely, frontal face, profile face, upper body, profile upper body, and lower body. In particular, the face is a very distinctive part of a human body due to specific features like eyes, nose and mouth, thus, it provides clear indication for discriminating between true and false human detections. Detecting both frontal and profile faces makes it possible to handle frontal and profile body poses. However, faces may become hardly detectable at low resolution. Including frontal and profile upper bodies in our set of body parts could help to overcome this limitation. The lower body is added to provide a complete coverage of the body space.

2.2.2 Learn the Relations between the Body Parts and Body Centroid

Motivated by (Mikolajczyk et al., 2004), a probability distribution model is learned with annotated humans to describe the relative location of the body parts. In our method, we model the relative distances between the body centroid and the body parts by bivariate Gaussian distributions. The whole body model consists of a set of one-to-one probabilistic relations. To reduce the complexity of the model, we assume that frontal and profile faces lie approximately at the same distance from the body centroid. The same hypothesis is made for frontal and profile upper bodies. Thus, three different Gaussian distributions are learned, one for face (frontal and profile faces) $\mathsf{N}(\mu_1, \Sigma_1)$, one for upper bodies (frontal and profile) $\mathsf{N}(\mu_2, \Sigma_2)$, and one for lower bodies $\mathsf{N}(\mu_3, \Sigma_3)$.

Let P be the set of body parts which include *upper body, lower body* and *face*. The relative distance Z_i between the centroid of the body, B, and a body part $p_i \in \mathsf{P}$ is calculated as:

$$Z_i = \left[\frac{X_i - x}{w}, \frac{Y_i - y}{h} \right]^{\mathsf{T}}, \quad (1)$$

where w, h and (x, y) define the width, height, and centroid coordinates of the body, while the pair (X_i, Y_i) is the centroid coordinates of the body part p_i (See Figure 1). The normalization with respect to the body size in Equation (1) makes it easy to deal with height and width variability among the annotated humans used for training.

We assume that Z_i follows a bi-variate Gaussian distribution, i.e., $Z_i \sim \mathsf{N}(\mu, \Sigma)$. The mean μ and covariance matrix Σ of $\mathsf{N}(\mu, \Sigma)$ can thus be estimated using Maximum-likelihood:

$$\begin{aligned} \hat{\mu} &= \textstyle\sum_i z_i \\ \hat{\Sigma} &= \frac{1}{N} \textstyle\sum_{i=1}^{N} (z_i - \mu)(z_i - \mu)^{\mathrm{T}} \end{aligned} \quad (2)$$

where z_i is a realization of Z_i provided by the annotated training set, i.e. $z_i = \left(\frac{x_i - x}{w}, \frac{y_i - y}{h} \right)$.

2.2.3 Confidence Areas for Body Parts Locations

The normalized distance Z_i between the body centre and body parts has been modelled as Gaussian distribution, i.e., $Z_i \sim \mathsf{N}(\mu, \Sigma)$. The distribution could provide us a high confidence neighbourhood for a particular body part with respect to the body's centroid. The covariance matrix Σ describes the spread of the distribution around the mean. If we set the confidence level as 0.95 when selecting the confidence region, then the region that contains 95% of all samples that can be drawn from the Gaussian distribution forms a confidence region for the distribution. The Mahalanobis distance is used to measure the distance from the test point z to the bivariate Gaussian:

$$m_d^2(z) = (z - \mu)\Sigma^{-1}(z - \mu) \quad (3)$$

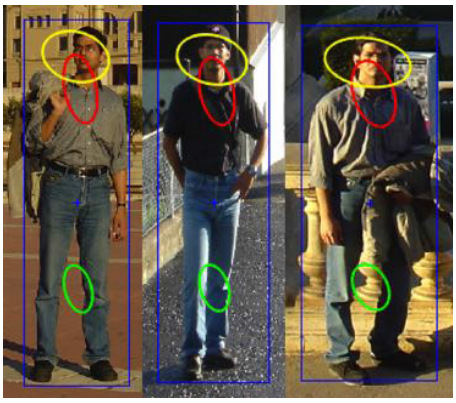

Figure 2: Confidence areas for body parts (yellow: face, red: upper body, green: lower body).

The eigenvectors v_1 , v_2 of Σ define a new coordinate system R in which the Mahalanobis distance between the test point z' and distribution $\mathsf{N}(0,D)$ where $D = diag(\lambda_1,\lambda_2)$, is:

$$m_d^2(z') = z'^T D^{-1} z' = (\frac{z_1'}{\sqrt{\lambda_1}})^2 + (\frac{z_2'}{\sqrt{\lambda_2}})^2 \quad (4)$$

In the coordinate system R' , the 95% confidence area for $Z' = [v_1, v_2]^T (Z - \mu)$ is:

$$P(m_d^2(z) \le s^*) = 0.95\,, \quad (5)$$

where s^* is a scalar equal to 5.991 according to the probability table of the chi-squared distribution. The solution of Equation (5) is a confidence ellipse ε' in the coordinate system R'. The confidence ellipse ε can be obtained by rotation and translation of ε' to provide us the confidence area for searching the body parts within the range of image coordinates. Figure 2 shows an example of the confidence areas for a few pedestrians, in which the three learned confidence areas for face, upper body and lower body are annotated with yellow, red and green ellipses, respectively.

2.3 Body Part Detectors

Table 1: The body parts detectors.

Body part detector	Feature type
Frontal face	LBP
Profile face	LBP
Upper body	HAAR
Upper body profile	HOG
Lower body	HAAR

Five boosted cascade classifiers are trained with the labelled training set to detect the five body parts in our body model. The detectors are employed at a range of locations and scales by applying a multi-scale pyramid and a sliding window. We use different feature pools for describing different body parts. For the frontal and profile face classifiers we utilised Local Binary Patterns (LBP) (Ojala et al., 2002) because they are efficient for texture description, a characteristic which has been found to be effective for face detection (Ahonen et al., 2004). For the frontal upper body and lower body, we applied the Haar features (Viola and Jones, 2001) which perform well even at low resolution. Finally, the profile upper body classifier is learned with the HOG features. These features provide a better performance and also reduced training time compared to the Haar features. Table 1 provides more details about the body parts classifiers in terms of feature types employed.

2.4 Combine HOG Detections with Body Part Detections

In this subsection, the body part detections are combined with the HOG detections in order to eliminate the false detections. For each body B detected by the HOG detector, the confidence areas are searched sequentially for potential detection of body parts. When a part p_i is located in the correct confidence ellipse, the assembly $\{p_i, B\}$ is formed. The candidates for the part p_i are searched within the 95% confidence neighbourhoods provided by the Gaussians distributions $Z \sim \mathsf{N}(\mu_i, \Sigma_i)$. As the relative distance z_i between p_i and B has been modelled to follow the distribution $\mathsf{N}(\mu_i, \Sigma_i)$, the likelihood of the assembly can be calculated as:

$$L_i = L(z_i \mid \mu_i, \Sigma_i) = \frac{1}{2\pi\sqrt{|\Sigma_i|}} \exp\left(-\frac{1}{2}(z_i - \mu_i)^T \Sigma_i^{-1} (z_i - \mu_i)\right). \quad (6)$$

When more than one body parts candidates are detected inside the same confidence ellipse, the most credible body part is selected based on the log-likelihood criterion and used only once before being removed from the set of detected body parts:

$$l_i = \log(L_i) = -\log(2\pi) - \frac{1}{2}\log|\Sigma_i| - \frac{1}{2}(z_i - \mu_i)^T \Sigma_i^{-1} (z_i - \mu_i) \quad (7)$$

We retain the HOG detection which has at least one body part located in a correct confidence area.

Otherwise, the HOG detections will be considered as false detections and will be removed.

3 EXPERIMENTS

To learn the Gaussian distributions of the body parts, we use the INRIA person positive training set (Guillaumin et al., 2009) which is composed of 260 positive and 1218 negative images. Since only entire body annotations are provided in the dataset, we annotate the human body parts manually.

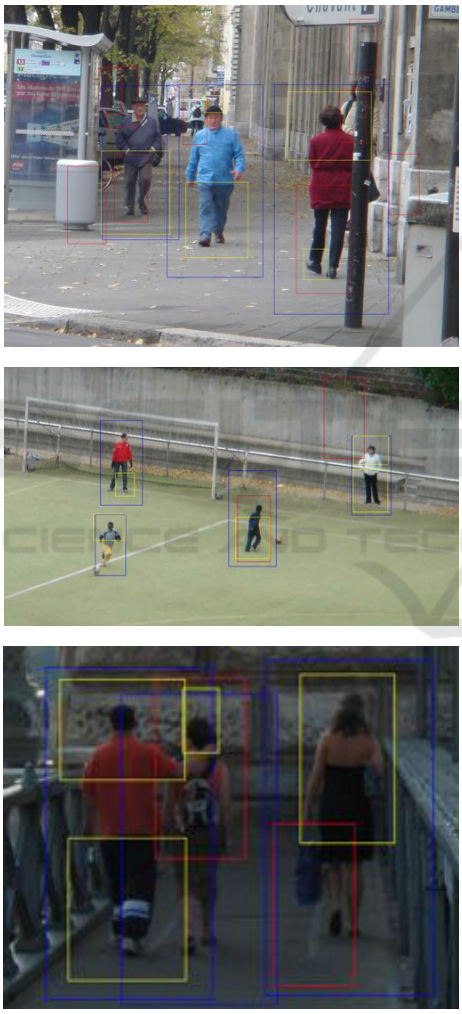

Figure 3: Examples of human detections with the proposed method. Blue rectangles indicate those HOG detections kept after post processing with body part detectors, while red rectangles indicates those discarded. Yellow boxes are used for frontal and profile face, purple boxes for frontal and profile upper body, and green boxes for lower body.

We test the performance of our detector on the INRIA person positive test. In our experiments, the detection of humans and human body parts takes on average 1.6 seconds per testing image, with a 2.6 GHz Intel Core i5 processor. The post-processing (searching body part candidates within the confidence areas in order to eliminate false HOG detections) takes on average 0.03 seconds per image. The detected human regions are compared with the annotated humans of the ground truth images with which exhibit a minimum overlap of 35%. Figure 3 shows some indicative examples of human detection results obtained with the proposed detector. The blue rectangles and the red rectangles indicate respectively those HOG detections retained and removed after our body part model-based post processing technique is applied. The detected body parts are also represented in Figure 3, with yellow boxes for frontal and profile face, purple boxes for frontal and profile upper body, and green boxes for lower body.

The precision-recall curve of our detector is shown in Figure 4. We also tested the HOG detector at classification thresholds Δ between 0 and 3 for references (refer also to Table 2). As can be seen from Table 2, the HOG detector with a classification threshold of $\Delta = 0$ provides 37% precision and 65% recall on the test set. Starting from that point, our detector removes a substantial number of the false HOG detections and exhibits only few false alarms, achieving a high precision of 70%. However, there are also some true HOG detections that are erroneously removed when body parts are not detectable, resulting a recall of 50%. This is nonetheless a better overall performance that cannot be achieved by the HOG detector solely, at any threshold settings.

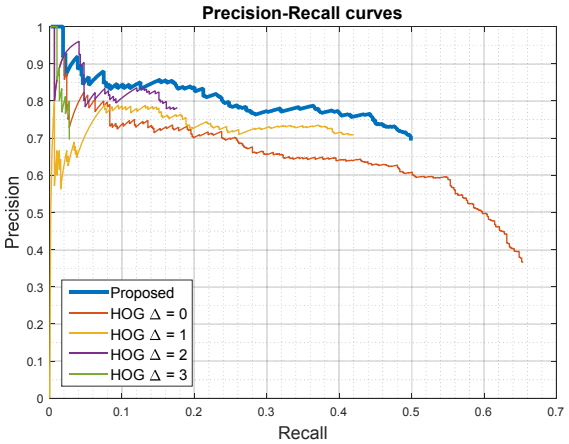

Figure 4: Precision-Recall curves of the detectors on the INRIA Person test set.

Table 2: Comparison of the overall precision and recall on the INRIA Person test set.

	Pro-posed	**HOG Δ=0**	**HOG Δ=1**	**HOG Δ=2**	**HOG Δ=3**
Preci-sion	70%	37%	71%	78%	70%
Re-call	50%	65%	42%	18%	3%

4 CONCLUSIONS

In this paper, by exploring an additional probabilistic human body model, we proposed an enhanced human detection method based on the HOG detector. Taking the HOG detector as a starting point, we use a body model to eliminate the false HOG detections and increase the precision. We demonstrate the efficiency of our human detection method on the INRIA person test set. Experimental results show that the proposed human detector can provide both good precision (70%) and recall (50%) with no need for adjusting the classification thresholds.

REFERENCES

Ahonen, T., Hadid, A., and Pietikäinen, M. (2004). Face recognition with local binary patterns *Computer vision-eccv 2004* (pp. 469-481): Springer.

Dalal, N., and Triggs, B. (2005). *Histograms of oriented gradients for human detection.* Paper presented at the Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on.

Felzenszwalb, P., McAllester, D., and Ramanan, D. (2008). *A discriminatively trained, multiscale, deformable part model.* Paper presented at the Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on.

Felzenszwalb, P. F. (2001). *Learning models for object recognition.* Paper presented at the Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 32*(9), 1627-1645.

Freund, Y., and Schapire, R. E. (1996). *Experiments with a new boosting algorithm.* Paper presented at the ICML.

Guillaumin, M., Mensink, T., Verbeek, J., and Schmid, C. (2009). *Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation.* Paper presented at the Computer Vision, 2009 IEEE 12th International Conference on.

Lowe, D. G. (1999). *Object recognition from local scale-invariant features.* Paper presented at the Computer vision, 1999. The proceedings of the seventh IEEE international conference on.

Micilotta, A. S., Ong, E.-J., and Bowden, R. (2005). *Detection and Tracking of Humans by Probabilistic Body Part Assembly.* Paper presented at the BMVC.

Mikolajczyk, K., Schmid, C., and Zisserman, A. (2004). Human detection based on a probabilistic assembly of robust part detectors *Computer Vision-ECCV 2004* (pp. 69-82): Springer.

Mohan, A., Papageorgiou, C., and Poggio, T. (2001). Example-based object detection in images by components. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 23*(4), 349-361.

Ojala, T., Pietikäinen, M., and Mäenpää, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 24*(7), 971-987.

Papageorgiou, C., and Poggio, T. (1999). *Trainable pedestrian detection.* Paper presented at the Image Processing, 1999. ICIP 99. Proceedings. 1999 International Conference on.

Viola, P., and Jones, M. (2001). *Rapid object detection using a boosted cascade of simple features.* Paper presented at the Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on.

Viola, P., Jones, M. J., and Snow, D. (2003). *Detecting pedestrians using patterns of motion and appearance.* Paper presented at the Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on.

Zhu, Q., Yeh, M.-C., Cheng, K.-T., and Avidan, S. (2006). *Fast human detection using a cascade of histograms of oriented gradients.* Paper presented at the Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on.