Enhancing Real-time Human Detection based on Histograms of Oriented Gradients

Marco Pedersoli[1], Jordi Gonzàlez[2], Bhaskar Chakraborty[1], and Juan J. Villanueva[1]

[1] Computer Vision Center and Departament d'Informàtica. Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain marcopede@cvc.uab.es
[2] Institut de Robòtica i Informàtica Industrial(UPC-CSIC), Edifici U Parc Tecnològic de Barcelona. 08028, Spain.

Summary. In this paper we propose a human detection framework based on an enhanced version of Histogram of Oriented Gradients (HOG) features. These feature descriptors are computed with the help of a precalculated histogram of square-blocks. This novel method outperforms the integral of oriented histograms allowing the calculation of a single feature four times faster. Using Adaboost for HOG feature selection and Support Vector Machine as weak classifier, we build up a real-time human classifier with an excellent detection rate.

1 Introduction

Human detection is the task of finding presence and position of human beings in images. Many applications take advantage of it, mainly in the videosurvelliance and human-computer iteration domains. Thus, human detection is the first step of the full process of Human Sequence Evaluation [5].

A well known method for detecting and tracking humans in video sequences is the *segmentation of foreground* motion blobs. Modeling static background allows detecting the presence of new objects, as humans, in the scene. This solution is effective when the camera is stationary and background has gradual changes in illumination. However, human detection can be required when using a moving camera (i.e. pedestrian detection in car camera) or in image databases (i.e. human image retrieval). To tackle these problems, an approach based on single image human body detection must be used. This leads to detect humans using *appearance-based* image features. In practice, a feature pattern learnt by a classifier is exhaustively searched in the full image.

Detecting human bodies based on appearance is more difficult than detecting other objects such as cars or faces. Human bodies are non-rigid, and highly articulated. This implies that besides illumination changes, occlusions and perspective distortion, we have to deal with a high range of different

poses and postures. Additionally, in human detection it is not possible to take advantage of specific textures and color information due to the variability of worn cloths.

This paper is organized as follows: section 2 explains previous work on human detection, section 3 gives an overview of the human detection framework that we built and section 4 shows the results we obtained in experimental tests. Finally section 5 discusses the conclusions.

2 Previous Work

In literature, two main approaches for human detection based on classifiers have been investigated [3]. The first is the detection of human parts (i.e. head, torso, legs) while the second is detecting humans as a whole. So far, the latter method has shown better detection results. On the other hand, the part-based approach can deal with occlusions and a broader variety of poses.

Regarding body-part based approach, Felzenswalb and Huttenlocher [2] model the human body as a set of human parts joint with springs, where every part is detected using Gaussian Derivative filters. Ioffe and Forsyth [6] represent each body component as a projection of straight cylinders which are then assembled into a full body. Mikolajczyk et al. [8] use orientation-position histograms as representative feature of every body part. Once the parts have been detected, these are assembled with a probabilistic model. Wu and Nevatia [13] introduce edgelets, a new type of silhouette-oriented features for learning body parts within a boosting framework.

In the full body detection approach, Papageorgiou et al. [9] use a feature descriptor based on Haar wavelets and a polynomial Support Vector Machine (SVM) as classifier. Gavrilla and Philomen [4] calculate edge images and compare them with a set of learnt exemplar using Charmfer distance. Viola et al. [12] combine Haar-like wavelets and space-time differences features into an AdaBoost machine to exploit the additional information given by motion. An effective human detector is proposed by Dalal and Triggs [1], who describe humans as a dense grid of Histogram of Oriented Gradient (HOG) and detect them with the help of a SVM. Zhu at al. [14] enhanced Dalal and Triggs results using integral of histograms for a fast HOG computation and Adaboost for feature selection.

In this article, following the same approach of Zhu's, we present a new technique for feature extraction that improves the calculation speed. Then, we use the gained speed to add to the HOG features a gaussian weighting mask which is improving the final detection rate.

3 Human Detection Framework

Human detection can be seen as a classification problem. In practice, an image is densely scanned by a detection window that moves to any possible position

and with any possible scale, looking for humans. The task is then to construct a good classifier that can discriminate humans and not humans in a really short time and with very low false positive (because an image can contain from 10000 to 100000 detection windows depending on the density of the search).

Two main subjects are fundamental for human classification: *features extraction* and *classification learning*. The first part consists on extracting the most relevant information from the data available. In our case, we want to find out an optimal image pixel representation that can underline differences between human and not human images. The choice of the feature is done throughout an analysis of the semantic significance of the feature. We choose to use HOG feature because (i) they provide a good representation of silhouettes and borders, (ii) are invariant to light contrast changes and small image movements and (iii) can be computed in a constant time independently of their size.

To learn a specific pattern like a human silhouette, it is necessary to associate its corresponding feature pattern throughout a classification process. In our framework, the classification learning is obtained using Adaboost, a boosting technique based on the construction of a strong classifier as a linear combination of weak classifiers. In practice, from human images a set of HOG features with different sizes and positions are extracted and then the best ones are selected for detection. In the next subsections, a detailed explanation of every step of the human detection framework is discussed.

3.1 Feature Extraction

Most human detection systems are based on the original idea of Viola and Jones [11]. They, using an overcomplete set of wavelets features (Haar-like features), were the first to get a well performing real-time face detection system. The main contribution of that work is on the ability to build up, throughout AdaBoost, a cascade of classifiers that give at the same time real-time performance and good detection rate. Furthermore, they took advantage of the integral of the image to be able to speed up the Haar-like features calculation. Unfortunately this solution has been proved to be not good enough for human detection. Haar-like features has a too little discriminative power, so that in case of high appearance variability classes, like humans, they can not capture enough information to allow pattern learning. Alternatively recent works [1, 14] showed that HOG (Histogram of Gradients) features can give excellent results in human detection and also that can be calculated in a constant time using the integral histogram [10]. In this paper we propose a novel method for HOG extraction that relies on the recursive calculation of square-block histograms. It can be four times faster than the integral of the gradient.

Figure 1 represents the preprocessing necessary for the feature calculation. First the image is converted to gray-scale. Then image gradient module and

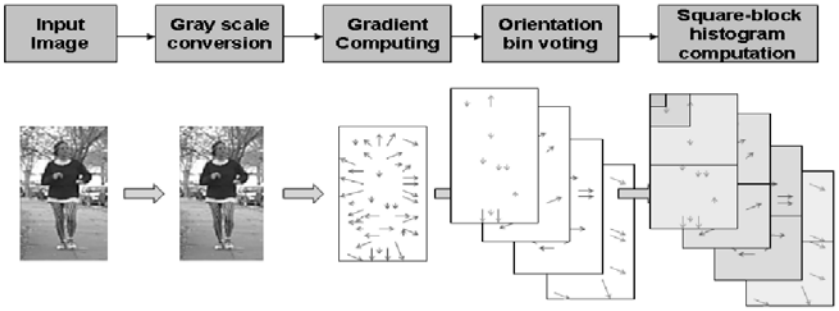

Fig. 1. Image Preprocessing: before calculating HOG features is necessary to preprocess the image to obtain square-blocks of histograms which are used for a fast HOG calculation.

phase are calculated applying an horizontal and vertical convolution with [-1 0 1] and [-1 0 1]' masks. At the next step, a clustering process based on the gradient module orientation is achieved. In practice the gradient image is split into 8 image bin, each one representing an orientation from 0° to 180°. For every pixel, the corresponding module of the gradient is stored in the appropriate orientation image bin. At this point, we calculate recursively the histogram of square-blocks. These blocks are the basic components for the following HOG construction. As it is possible to see in Figure 2, the blocks start to be relevant (in the image it is possible to distinguish a human silhouette) from a size of 4 pixels. Then, this is the basic block we compute. Then recursively, square-blocks of $2^2, 2^3, 2^4, 2^5$ size are calculated at steps of 4 pixels. As showed in Figure 3, each block can be calculated as sum of 4 blocks of the previous scale level or 16 blocks of 2 previous scale levels. The second case is useful when we want to apply a gaussian mask to the feature. This will be applied in the experiments.

After this image preprocessing, the feature calculation can be executed really fast, and in a time that does not depend on the feature size, but constant. In fact, the procedure to calculate a feature consists of recollecting the precalculated histogram blocks belonging to the feature and normalize them. The main difference of this method with the integral of histogram is that each histogram can be obtained with only one memory access instead of four (because the value is directly read from the memory), making the feature extraction process around four times faster.

3.2 Classification Learning

Our classification algorithm is based on Adaboost. This is a technique of combining a sequence of weak classifiers (classifier with a classification rate

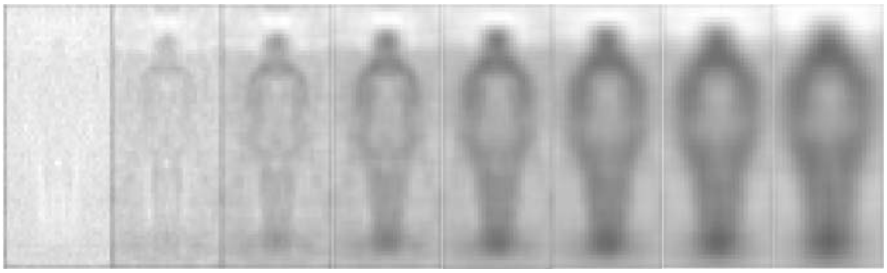

Fig. 2. Histograms of square-block variance calculated from 100 images of the MIT human database. From left to right are represented the results using a constant moving step of 2 and different block sizes: {2x2,4x4,8x8,12x12,16x16,20x20,24x24,32x32}.

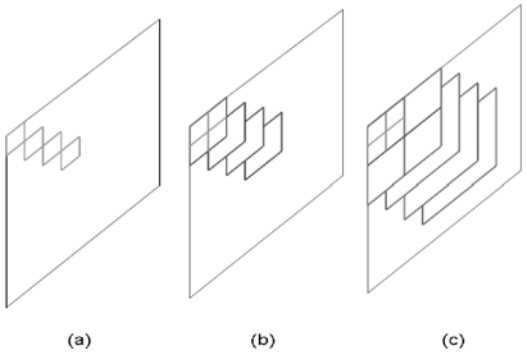

Fig. 3. Distribution of features: (a) *4x4 square-blocks*: each block is calculated directly form the gradient image and they are not overlapping (b) *8x8 square-blocks*: each block is calculated from 4 blocks of the previous level and they are overlapping each other of 50% (c) *16x16 square-blocks*: each block can be calculated either with 16 blocks of size 4x4 or with 4 of size 8x8. They are overlapping each other of 75%

slightly better than 0.5) into a strong classifier. Given a set of samples $X = \{x_1, x_2, .., x_n\}$ and a set of weak classifier h_t which assigns every element to -1 or +1 whether the element belong to a class or not, Adaboost provides a strong classifier

$$f(x) = \sum_{t=1}^{T} \alpha_t h_t(x)$$

which is a linear combination of h_t. Boosting technique is obtained iteratively assigning importance weight to the samples used for the training session. At the start all the weight have the same value, but after the first weak classifier has been added, the weight value is updated. The weight of samples with a right classification keeps the same value, while the weight of samples with wrong classification is increased, to give them more importance in the next iteration. In the same way, at every iteration, the classifier co-

efficient values α_t are calculated based on the relative goodness of the weak classifier.

In our case, the weak classifiers are Support Vector Machine associated to a single HOG feature, and these are sorted by their own classification capability. When Adaboost is called from the first time for every feature a SVM is learnt. The SVM with the best classification score is selected. At this point Adaboost update the importance weight associated to every sample and calculate the α_1 coefficient. Then, again, for every feature the corresponding SVM is learnt, but this time with different importance weights and then also with different result. This cycle is repeated until a certain detection rate is reached or until the strong classifier has reached a certain complexity. We considered 50 weak classifiers as a good trade off between detection rate and complexity.

4 Experimental Results

We tested our detector with the MIT pedestrian database. This database is composed by 924 full body pedestrian images of size 128x64 pixels. We used 800 of them for the training and the rest for the test. The total training session is then composed by 800 positive images (humans) form the MIT database and 800 negatives (not humans) taken from random windows cropped from no human images. For every sample image we extracted 619 features which are the sum of:

- $29 * 13$ features of size 16x16 pixels
- $25 * 9$ features of size 32x32 pixels
- $17 * 1$ features of size 64x64 pixels

This approach is something like a trade-off between Dalal and Triggs approach which is using a low number (105 per human image), fixed size, computationally slow HOG features and Zhu et al. who are using a high number (5031), variable size and ratio faster HOG implementation.

We did two different kind of tests: the first using normal HOG features exactly as in Zho's, the second applying a gaussian weighting mask to the features. The idea of the second experiment is to use the increased speed obtained form our feature calculation technique to gain better detection performance. In fact, applying to the feature a gaussian mask of standard deviation equal to a half the feature size, reduces the effect of small perspective changes, making it more robust [7]. The gaussian weighted feature can be calculated using 16 square-blocks (instead of 4) of two inferior scale size features and multiply each of them for the corresponding coefficient.

The ROC curves of the two experiments are shown in Figure 4. Both detectors are well performing, with a detection rate at 10^{-4} false positive per window greater than 85%. As expected the features with gaussian weighting are performing better than the others. However, the increment of detection rate in the ROC curve is varying from 1% to 5%, which is probably a too small gain considering the loss of 4X of speed necessary for that.

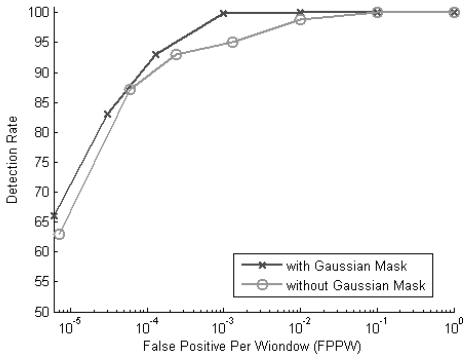

Fig. 4. ROC curve of two detectors with different HOG features.

5 Conclusions

In this paper we presented a human detector framework based on a HOG features, which are the actual state of art in the field of detection systems. We defined a new way of computing HOG features based on square-block of histograms which is four time faster than Zhu's implementation based on the integral of histograms. Finally, we exploited this increased speed to apply an approximate gaussian mask to the features in order to improve the feature quality and the corresponding detection rate.

The results described in the previous section show the quality of the approach. Nevertheless, we did not reach the performances of [1] and [14]. We believe that this is due to the reduced number of negative examples we could employ because of the limited amount of memory we could use in our Matlab implementation. Thus, it can be useful to convert at least part of the code in a more efficient language, like C or C++. This allows us to experiment more, and with bigger and more difficult examples. A good testing set could be the INRIA database which has more, and more difficult human position and situation. Finally, Another point to improve the framework is a cascade of classifiers. This can provide a twofold advantage. The speed of the final detector can be highly enhanced using an approach based on subsequent classification refinements. Moreover, in the learning phase of the cascade, difficult examples can be selected to feed the next cascade stage, avoiding the necessity of a high number of negative examples which slows down the learning process.

Acknowledgements

This work is supported by EC grants IST-027110 for the HERMES project and IST-045547 for the VIDI-video project, and by the Spanish MEC under

projects TIN2006-14606 and DPI-2004-5414. Jordi Gonzàlez also acknowledges the support of a Juan de la Cierva Postdoctoral fellowship from the Spanish MEC.

References

1. N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In Cordelia Schmid, Stefano Soatto, and Carlo Tomasi, editors, *International Conference on Computer Vision & Pattern Recognition*, volume 2, pages 886–893, INRIA Rhône-Alpes, ZIRST-655, av. de l'Europe, Montbonnot-38334, June 2005.
2. P. Felzenszwalb and D. Huttenlocher. Efficient matching of pictorial structures. In *CVPR*, pages 66–73, 2000.
3. D. Gavrila. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding: CVIU*, 73(1):82–98, 1999.
4. D. Gavrila and V. Philomin. Real-time object detection for "smart" vehicles. In *CVPR*, pages 87–93, Fort Collins, Colorado, USA, 1999.
5. J. Gonzàlez. *Human Sequence Evaluation: The Key-frame Approach*. PhD thesis, Universitat Autònoma de Barcelona, Spain, 2004, `http://www.cvc.uab.es/~poal/HSE/aThesis.zip`.
6. S. Ioffe and D. A. Forsyth. Probabilistic methods for finding people. *Int. J. Comput. Vision*, 43(1):45–68, 2001.
7. D. Lowe. Distinctive image features from scale-invariant keypoints. In *International Journal of Computer Vision*, volume 20, pages 91–110, 2003.
8. K. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust part detectors. In *European Conference on Computer Vision*, volume I, pages 69–81, 2004.
9. C. Papageorgiou and T. Poggio. A trainable system for object detection. *Int. J. Comput. Vision*, 38(1):15–33, 2000.
10. F. Porikli. Integral histogram: A fast way to extract histograms in cartesian spaces. In *CVPR*, 2005.
11. P. Viola and M. Jones. Robust real-time object detection. *International Journal of Computer Vision - to appear*, 2002.
12. P. Viola, M. J. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. *Int. J. Comput. Vision*, 63(2):153–161, 2005.
13. B. Wu and R. Nevatia. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In *ICCV '05: Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, pages 90–97, Washington, DC, USA, 2005. IEEE Computer Society.
14. Q. Zhu, M. C. Yeh, K. T. Cheng, and S. Avidan. Fast human detection using a cascade of histograms of oriented gradients. In *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1491–1498, Washington, DC, USA, 2006. IEEE Computer Society.