Human detection in real scenes

Domonkos Varga[1], László Havasi[2], and Tamás Szirányi[3]

[1] MTA Sztaki, BME VIK HIT
varga.domonkos@sztaki.mta.hu
[2] MTA Sztaki
havasi.laszlo@sztaki.mta.hu
[3] MTA Sztaki, BME KJK ALRT
sziranyi.tamas@sztaki.mta.hu

Abstract. Detecting different categories of objects in an image and video content is one of the fundamental tasks in computer vision research. Human detection is a hot research topic, with several applications including robotics, surveillance and automotive safety. The progress of the past few years can be driven by the availability of challenging public datasets. Human detection is the first step for a number of applications such as smart video surveillance, driving assistance systems, and intelligent digital content management. It is a challenging problem due to the variance of illumination, color, scale, pose, and so forth. This article reviews various aspects of human detection in static images. We introduce our human detection system and compare it in detail with other state-of-the-art methods. We used a sliding-window approach for human detection. Sliding-window human detection systems scan the image all relevant positions and scales to detect a person. Our feature component encodes quickly the visual appearance of the human and the classifier component determines for each sliding-window independently whether it contains a human or not. We show in this article that our system is good trade-off among accuracy and speed. To get a feeling about the achievable performance of sliding-window based techniques we made a failure case analysis. We analyzed the missing recall and the false positive detections at equal error rate.

1 Introduction

Human detection is the first step for a number of applications such as smart video surveillance, driving assistance system, human-robot interaction, people-finding for military applications and intelligent digital management. It is a challenging problem due to the variance of illumination, color, scale, pose, clothing and so forth. Detecting humans in images is a task with a long history, in the past decade there has been a great interest in human detection. A robust human detector must deal with the above-mentioned problems. The detector must be capable of distinguishing the object from complex background regions.

The methods on human detection differ in three perspectives. First, they may use different feature such as edge features, Haar-like features, and gradient orientation features. Second, they use different learning methods such as NN, SVM or cascaded AdaBoost. Third they may treat the image region as a whole or detect each part first and then combine them by these parts' geometric configurations. Accurate detection is still a major interest in human detection, especially in terms of high detection rate with low FPPI (false positive per image).

The remainder of this paper is organized as follows. In Section 2, we give a brief introduction on public datasets. In Section 3 we describe sliding window techniques for human detection, including the applied features and machine learning methods. Our proposed human detection system is presented in Section 4, including the system architecture, the feature extraction and the process of training. In Section 5, we compare our human detection system with other state-of-the-art methods in respect of speed, accuracy and failure analysis. Brief conclusions and discussions on the proposed system can be found in Section 6.

2 Datasets

Multiple public pedestrian datasets have been collected over the years: INRIA [1], MIT [2], Caltech-USA [3], TUD-Brussels [4], KITTI [5], Daimler [6] and Daimler stereo [7] are the most commonly used ones. They all have different characteristics, weakness and strengths.

INRIA is the oldest among the above-mentioned databases. INRIA contains 1805 64*128 images of humans cropped from a varied set of personal photos. That is why it is commonly selected for training. ETH and TUD-Brussels are mid-sized video datasets. Daimler is not considered by all human detection methods because it lacks color channels. Daimler stereo, ETH, and KITTI provide stereo information. All datasets expected INRIA are obtained from video, and thus enable the use of optical flow as an additional cue.

Caltech-USA is outstanding for the large number of methods that have been evaluated. The test set of KITTI is slightly more diverse, but it is not used as frequently. A more detailed discussion of the darasets can be found in [8].

3 Sliding window techniques for human detection

Sliding window human or pedestrian detection systems scan the image all relevant positions and scales to detect a person. There are two major components of these systems: the *feature* component encodes the visual appearance of the person, the *classifier* component determines for each sliding window independently whether it contains a human or not. These techniques are computationally expensive because they scan many positions and scales. Fortunately, due to recent advances in GPUs, real-time people detection is possible as e.g. demonstrated by [9]. The rest of the section briefly describe some of the features and classifiers in more detail. The ultimate goal of feature extraction for object detection is to find one representation that yields high interclass variability and at the same time achieves low intraclass variability. Obviously, the choice of features is the most critical decision when designing a detector, and finding good features is still largely an empirical process with few theoretical guidelines. Machine learning techniques used to discriminate between these features.

Haar wavelets are first proposed by Oren, et al. [10] as overcomplete Haar wavelets for pedestrian detection. Later, Papageorgiou and Poggio [11] made a study of the overcomplete Haar wavelets for the detection of face, car, and pedestrian. They used three different types Haar wavelets: vertical, horizontal and diagonal. The length of the feature vector for a 64*128 detection window was 1326. The authors reported that for the class of pedestrian the wavelet coefficients do not carry information due to the variety in clothing. Viola and Jones [12] proposed the Haar-like rectangle features and the fast evaluation method. Lienhart and Maydt [13] added rotated rectangle features to the feature set, which is called extended Haar-like features. The Haar-like features can be computed quickly using integral images, together with cascaded AdaBoost algorithm, forming a flexible object detection framework.

Histograms of oriented gradients have been proposed by Dalal and Triggs [14]. HOG is probably the most popular feature in human detection. We believe that contour is the most useful information for pedestrian detection. Our hypothesis is that for pedestrian detection the most important thing is to encode the contour, and HOG is mostly focusing on this information. To verify our hypothesis we used the original HOG detector for the Sobel version of our test images. Surprisingly, the detection accuracy is 68 % at 1 FPPI, which value is higher than the performance of Haar-AdaBoost-Cascade in Figure 6.

Shape Context has originally been proposed as a feature point descriptor [15]. Shape context describes the coarse distribution of the rest of the shape with respect to a given point on the shape. Finding correspondences between two shapes is then equivalent to finding for each sample point on one shape the sample point on the other shape that has the most similar shape context. The authors of [16] and [17] has shown good results for pedestrian detection using the generative ISM framework.

Local Binary Patterns (LBP) are a simple but efficient texture operator which labels the pixels of an image by thresholding the neighborhood of each pixel and considers the result as a binary number. LBP was described by Ojala et al. in 1994 [18]. The original LBP was later extended to many new variations. Recently, variants of LBP combined with other features i.e. HOG also show high potentials. While LBP encodes excellent the micro-structure of a human, the HOG focuses on the gradients.

Classifiers. The rest of this section focuses on discriminative methods that build a decision boundary directly from the input samples instead of density estimation methods. For the classification of single windows two popular choices are SVMs and decision tree stumps in conjuction with the AdaBoost framework. A SVM constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier. For our experiments and programs we used the implementation of LIBSVM [19]. In contrast, boosting is picking single entries of the feature vector with the highest discriminative power in order to minimize the classification error in each round.

4 Our system architecture

In this section we describe our human detection system. The key components are the Haar dictionary and the SVM classifier. Our applied architecture is shown in Figure 1.

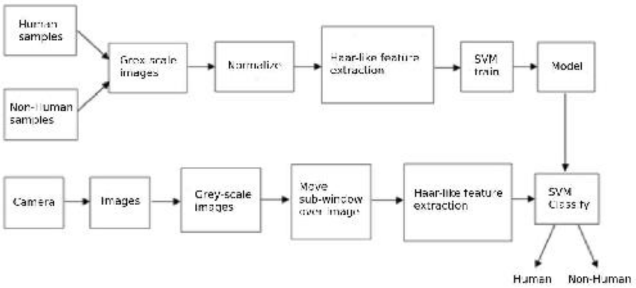

Fig. 1. Architecture of proposed human detection system.

To train our system, we gathered a set of 2,500 grey-scale sample images of pedestrians that have been aligned and scaled to the dimensions 128*64. The images were taken from MIT pedestrian dataset and INRIA person dataset. We made a database of negative samples too, which consists of 4,000 non-human images. In order to improve the performance we put 1,000 vertical structures like poles, trees or street signs to the negative samples.

The feature extraction method of our system uses rectangular Haar features. It is a feature extraction method for image and object identification proposed by Viola and Jones [12]. We used six simple rectangle features only to learn and classify images shown in Figure 2.

The regions within these features have the same size, shape and horizontally or vertically adjacent (Fig. 1). Processing time of feature extraction is very important for real-time detection. With the help of integral images we can obtain very quickly the values of Haar-like feature at any position in an image. For given image $f(x, y)$, we obtain integral image $I(x, y)$ as follow:

$$I(x, y) = \sum_{m=1}^{x} \sum_{n=1}^{y} f(m, n) \tag{1}$$

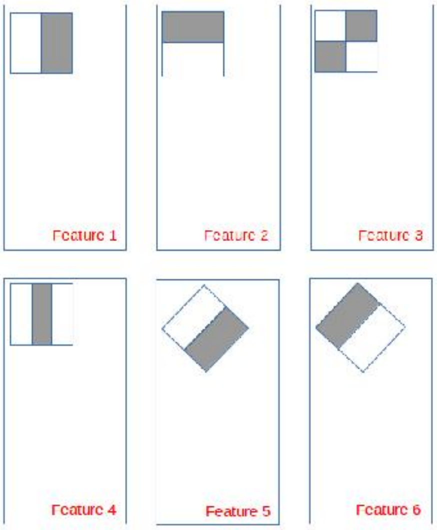

Fig. 2. Six kinds of feature are used to extract features of human.

Where $I(x, y)$ indicates the sum of the pixels above and to the left of x, y (Fig. 3a). Using the integral image $I(x, y)$, any rectangular sum can be computed in four array references (Fig. 3b).

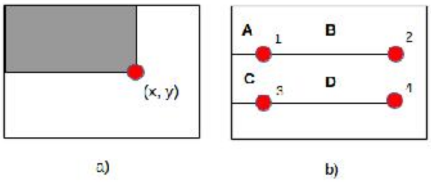

Fig. 3. (a) Integral image and (b) Sum of the pixels within D region is 1+4-(2+3).

All of human and non-human samples are converted to gray-scale images and resized to 128*64 pixels. A sub-window with size of 24*24 pixels is used to move 8 pixels for each step over a sample image. 432 feature values are obtained from this processing (each sub-window extracts 6 features). Next, a sub-window with size of 12*12 pixels is used. The number of pixels for each moving step is equal to 4. And 2352 feature values are obtained after this processing. Totally, with each sample image, a 2784-dimension vector is achieved.

To detect pedestrians in a new image, we shift the 128*64 detection window over all location in the image. This will only detect pedestrians at a single scale. To achieve multi-scale detection, we incrementally resize the image and run the detection window over each of these resized images. This brute force search over the image is time consuming, however several methods can be used to reduce the computation. We tested our method with many size of input images. The correct detection and processing time depend on size of input images. Several results can be seen in Figure 4 and in Figure 5.

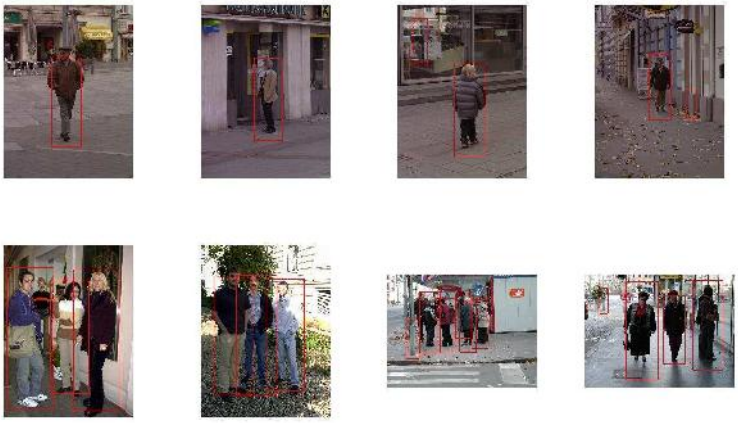

Fig. 4. Example detections of our human detector.

5 Comparison of sliding window techniques

In the followings we report on some of the results that illustrate the state-of-the-art in sliding window based detection techniques. To evaluate the performance for the introduced features and their combination with different classifiers we used the established INRIA Person dataset.

The seven systems we compare include Dalal and Trigg's HOG-SVM-Bootstrapping system, Viola and Jones' Haar-AdaBoost-Cascade system, a HOG-AdaBoost-Cascade system, our Haar-SVM-Bootstrapping system, Wang et al.'s HOG-LBP-SVM-Bootstrapping system [20], Dollar et al.'s ChnFtrs human detector [21] and the C^4 human detector [22] which uses the so-called CENTRIST descriptor. The results can be seen in Figure 6, we tested the speed on the INRIA human database.

From the results, we can see that HOG features are more powerful than Haar-like features on classification; the SVM-Bootstrapping classifier structure outperforms the AdaBoost-Cascade structure in classification performance, but the AdaBoost-Cascade structure can increase the speed greatly.

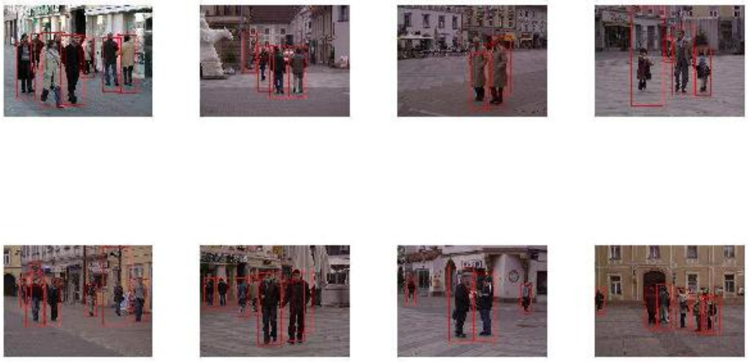

Fig. 5. Example detections of our human detector.

Method	VGA speed	Accuracy
HOG-SVM-Bootstrapping [14]	0.075 fps	74.4 %
Haar-AdaBoost-Cascade [11]	**70 fps**	64.5 %
HOG-AdaBoost-Cascade	15 fps	73.5 %
HOG-LBP-SVM-Bootstrapping [20]	4 fps	**87 %**
ChnFtrs [21]	0.5 fps	86 %
C^4 [22]	20 fps	83.5 %
Haar-SVM-Bootstrapping	40 fps	70.1 %

Fig. 6. Speed comparison of several vision-based human detection methods. VGA resolution is 640*480. Accuracy is at 1 FPPI (false positive per image).

To get a feeling about the achievable performance of sliding-window based techniques we made a failure case analysis. We analyzed the missing recall and the false positive detections at equal error rate (150 missing detections / 150 false positives). The cause of missing recalls can be categorized as follows: unusual articulations, difficult background or contrast, occlusion or carried bags, under- or overexposure. The results of the analysis are shown in Figure 7.

False positive detections can be categorized as follows: vertical structures like poles or street signs, cluttered background, too large scale detections with people in lower part, too low scale on body parts. The results can be seen in Figure 8.

The classification results of three systems on INRIA pedestrian dataset are shown in Figure 9. From the results, we can see that HOG features are more powerful than Haar-like features on classification at a great cost of speed. The SVM-Bootstrapping classifier structure obviously the AdaBoost-Cascade structure in classification performance, but the AdaBoost-Cascade structure can increase the

Method	Unusual articulations	Difficult background or contrast	Occlusion or carried bags	Under- or overexposure	Other
HOG-SVM-Bootstrapping [14]	17.33 %	14 %	50 %	10.66 %	8 %
Haar-AdaBoost-Cascade [11]	27.33 %	30 %	26.66 %	11.33 %	**4.66 %**
HOG-AdaBoost-Cascade	18.66 %	18 %	50 %	**6.66 %**	6.66 %
HOG-LBP-SVM-Bootstrapping [20]	10 %	**12 %**	40.66 %	26.66 %	10.66 %
ChnFtrs [21]	**2.66 %**	24 %	26 %	33.33 %	14 %
C^4 [22]	12.66 %	19.33 %	**22.66 %**	35.33 %	10 %
Haar-SVM-Bootstrapping	24.66 %	29.33 %	28.66 %	12 %	5.33 %

Fig. 7. Missing recall.

Method	Vertical structures	Cluttered background	Too large scale detections	Too low scale on body parts	Other
HOG-SVM-Bootstrapping [14]	42.66 %	18 %	9.33 %	18.66 %	11.33 %
Haar-AdaBoost-Cascade [11]	42 %	20.66 %	12.66 %	**8 %**	16.66 %
HOG-AdaBoost-Cascade	51.33 %	17.33 %	10 %	15.33 %	**6 %**
HOG-LBP-SVM-Bootstrapping [20]	37.33 %	16 %	12 %	24.66 %	10 %
ChnFtrs [21]	34 %	**12.66 %**	**6 %**	27.33 %	20 %
C^4 [22]	**26 %**	30 %	13.33 %	11.33 %	19.33 %
Haar-SVM-Bootstrapping	29.33 %	21.33 %	15.33 %	19.33 %	14.66 %

Fig. 8. False positive detections.

speed greatly. We can see that our system can detect human accurately, irrespective of the illumination.

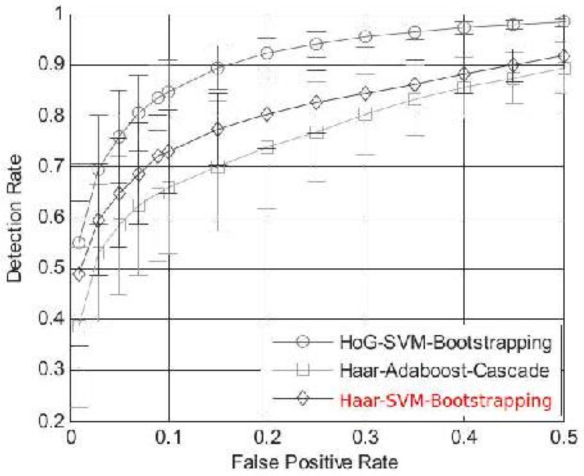

Fig. 9. Classification performance of three systems.

6 Conclusion

In this paper we proposed our human detection system. In the presented system, Haar-like features with SVM are used to detect humans in a static image. We gathered a set of 2,500 grey-scale sample images of humans and a set of 4,000 non-human sample images. We used this image database to train our system. We compared the presented system with other state-of-the-art human detection methods based on sliding-window approach. We have created a good trade-off among the accuracy and the computational time. Although our presented system is less accurate than the most reliable state-of-the-art sliding-window methods but our system outperforms in speed all the other slding-window techniques except Haar-AdaBoost-Cascade, which accuracy is much less than our results.

Acknowledgements

This work has been supported by the EU FP7 Programme (FP7-SEC-2011-1) No. 285320 (PROACTIVE project). The research was also partially supported by the Hungarian Scientific Research Fund (No. OTKA 106374).

References

1. http://pascal.inrialpes.fr/data/human
2. http://cbcl.mit.edu/software-datasets/PedestrianData.html
3. P. Dollar, C. Wojek, B. Schiele and P. Perona, "Pedestrian detection: A benchmark", *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 304-311, 2009
4. http://www.mpi-inf.mpg.de/departments/computer-vision-and-multimodal-computing/research/people-detection-pose-estimation-and-tracking/multi-cue-onboard-pedestrian-detection
5. http://www.cvlibs.net/datasets/kitti/
6. http://www.computervisiononline.com/dataset/daimler-pedestrian-benchmarks
7. C. G. Keller, M. Enzweiler and D. M. Gavrila, "A new benchmark for stereo-based pedestrian detection", *Intelligent Vehicle Symposium (IV), IEEE*, pp. 691-696, 2011
8. P. Dollar, C. Wojek, B. Schiele and P. Perona, "Pedestrian detection: An evaluation of the state of the art", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 34, pp. 743-761, 2012
9. C. Wojek, G. Dorko, A. Schulz and B. Schiele, "Sliding-windows for rapid object class localization: A parallel technique", *Pattern Recognition.* Springer Berlin Heidelberg, pp. 71-81, 2008
10. M. Oren, C. Papageorgiou, P. Sinha, E. Osuna and T. Poggio, "Pedestrian detection using wavelet templates", *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 193-199, 1997
11. C. Papageorgiou and T. Poggio, "A trainable system for object detection", *International Journal of Computer Vision*, vol. 38, pp. 15-33, 2000
12. P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features", *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 511-518, 2001