Combining Human Perception and Geometric Restrictions for Automatic Pedestrian Detection

M. Castrillón-Santana[1] and Q. C. Vuong[2]

[1] IUSIANI
Edificio Central del Parque Científico-Tecnológico
Campus Universitario de Tafira
Universidad de Las Palmas de Gran Canaria
35017 Las Palmas - Spain
mcastrillon@mcastrillon.ulpgc.es
[2] Max Planck Institute for Biological Cybernetics
Cognitive & Computational Psychophysics
Spemannstrasse 38
72076 Tübingen, Germany
quoc.vuong@tuebingen.mpg.de

Abstract. Automatic detection systems do not perform as well as human observers, even on simple detection tasks. A potential solution to this problem is training vision systems on appropriate regions of interests (ROIs), in contrast to training on predefined and arbitrarily selected regions. Here we focus on detecting pedestrians in static scenes. Our aim is to answer the following question: *Can automatic vision systems for pedestrian detection be improved by training them on perceptually-defined ROIs?*

1 Introduction

The present study investigates the detection of pedestrians by humans and by computer vision systems. This simple task is accomplished easily and quickly by human observers but still poses a challenge for current vision systems.

In the Computer Vision community, different automatic detection systems have been designed in the past using simple features for people detection based on the detection of different body elements: the face [4,17], the head [1,2], the entire body [15] or just the legs [11], as well as the human skin [6].

These systems make use of some selected regions where representations based on local features [9,14], sometimes combined with global cues [7], are employed for detection. Such systems perform fairly well but still have high miss rates. In order to overcome this problem, more recently a combination of body parts have been used to improve the performance as the false positive for an individual detector is higher than for several detectors [10].

However, the criteria to select the different body parts or regions have not been the focus in these earlier works. Rather, the parts or regions have been chosen *ad hoc* or arbitrarily. That said, we have observed a small correlation between

R. Marín et al. (Eds.): CAEPIA 2005, LNAI 4177, pp. 163–170, 2006.

the performance of these systems and human observers. This finding motivated us to systematically analyze human performance on a pedestrian detection task that tests whether these regions are the most semantically useful and whether other regions can also provide useful information. This study therefore allows us to determine which body parts should be included in an automatic detection system.

For that purpose, we used a psychophysical "bubbles" technique [3], described in Section 2, to isolate those regions used by humans for pedestrian detection–what we call perceptually-defined regions of interest (p-ROI). Section 3 describes the general object detection framework designed by Viola and Jones [14] used to train a p-ROI detector. Section 4 presents the geometric restrictions employed to make use of the global configuration of these parts, with the aim to reduce false detections. Results are presented in Section 5, and some conclusions are outlined in Section 6.

2 The Bubbles Technique

To investigate which ROIs are used more by humans, we used a psychophysical "bubbles" technique [3] to isolate the regions which help human observers determine the presence of pedestrians in an image. The technique was originally used in [3] to identify internal facial features that provided diagnostic information for gender and expression classification. For example, with high-resolution face images, the gender was correctly determined using just the eyes and mouth.

In the current study, images containing aligned pedestrians were revealed through a mask of small randomly distributed Gaussian windows ("bubbles"). That is, the presence of a bubble over a region showed that region, as shown in Figures 1 and 2. Eight subjects were shown stimuli masked by Gaussian bubbles and had to judge if a human was present. Half the trials contained a human. Across observers, masks leading to correct responses are summed and normalized to reveal image regions that were useful for this task. This procedure is illustrated in Figure 1. The result of the bubbles paradigm is a diagnostic image, which is presented in Figure 2.

The diagnostic image (Figure 2) indicates that observers relied predominantly on head and leg regions, and to a lesser extent on arm regions. These results confirm some of the regions already considered by automatic pedestrian detectors.

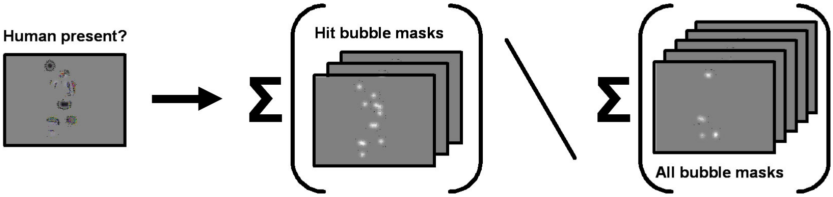

Fig. 1. Building the diagnostic image using bubbles

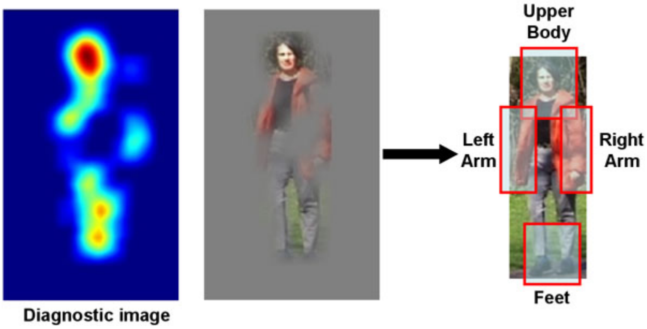

Fig. 2. The diagnostic image produced from eight human observers for the pedestrian detection problem

3 Viola-Jones General Object Detection Framework

The Viola-Jones [14] general object detection framework achieves fast and robust performance by means of a cascade of weak classifiers trained using boosting. This approach has been used to train the p-ROI detector in order to confirm if those regions are particularly discriminable for certain pattern matching problems, and to what extent an area is important. This framework has already been successfully applied to different objects categories, and it is well known in the face detection community because it provides real-time performance.

Recently, an implementation has been made available in OpenCV (Open Computer Vision Library) [5]. This framework, designed for rapid object detection, is based on the idea of a boosted cascade of weak classifiers [13] but extends the original feature set and provides different boosting variants for learning [8].

The cascade learning algorithm is similar to decision-tree learning. Essentially, a classifier cascade can be seen as a degenerated decision tree. For each stage in the cascade a separate subclassifier is trained to detect almost all target objects while rejecting a certain fraction of the non-object patterns. The resulting true detection rate, D, and the false detection rate, F, of the cascade is given by the combination of each single stage classifier rates:

$$D = \prod_{i=1}^{K} d_i \qquad\qquad F = \prod_{i=1}^{K} f_i \tag{1}$$

For example, given a 20-stage cascade of weak classifiers designed to reject 50% of non-object patterns (false detection rate) while accepting 99.9% of object patterns (detection rate) at each stage, the overall detection rate will be $0.999^{20} \approx 0.98$ with a false detection rate of $0.5^{20} \approx 0.9 * 10^{-6}$. This scheme allows for a high image processing rate, due to the fact that background regions of the image are quickly discarded. Consequently, more processing time can be dedicated to promising object-like regions. Thus, to achieve a desired trade-off between accuracy and speed for the resulting classifier, the designer can choose the desired number of stages, the target false detection rate and the target true detection rate per stage.

4 Geometric Restrictions

As stated in recent works, the probability of a false detection for an individual detector is higher than for several detectors [10]. Using this assumption, the coocurrence of coherently located detections can provide the system with evidence to reject detections which are inconsistent with the majority of detections. This fact is evident in Figure 3. The left column reflects the hits obtained for a frame using the p-ROI detector, while the right column indicates the resulting filtered detection image after applying some heuristic geometric restrictions for typical standing pedestrians. The basic rules applied to determine if a detection is coherent in the image are applied only when at least two different detectors of different nature reported a hit. These rules are summarized as:

- The feet must be below the head, to the right of the left arm and to the left of the right arm.
- The head must be above the feet, to the right of the left arm and to the left of the right arm.
- The centroid of the left arm must be to the left of head, feet and right arm.
- The centroid of the right arm must be to the right of head, feet and left arm.

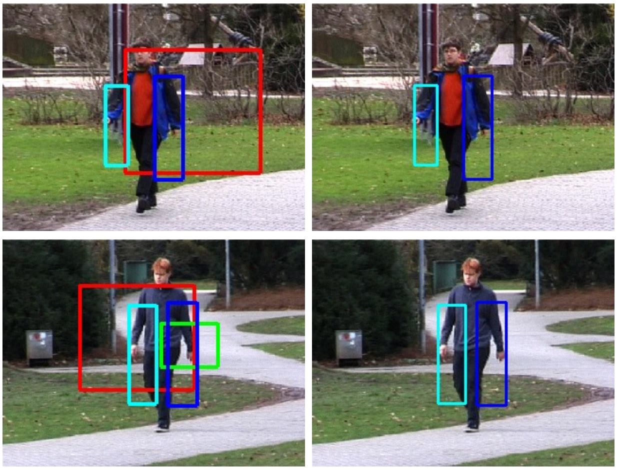

Fig. 3. Left column shows detection samples using the p-ROIs detector. Right column shows hits accepted after geometric filtering. Colors are related to the detector: red for heads and shoulders, green for feet, blue for right (in the image) arm and cyan for left (in the image).

5 Experiments

The detectors used in these experiments were trained using the OpenCV implementation of the Viola-Jones framework. The training set consisted of the CBCL pedestrian images [12] augmented with additional positive and negative samples

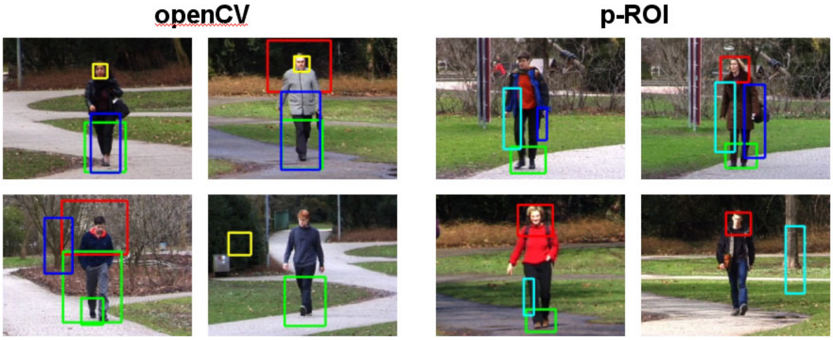

Fig. 4. Results achieved using the detectors included in the OpenCV release, and results achieved using the basic p-ROI approach. Colors meaning: left) yellow for faces, red for heads and shoulders, green for lower body, blue for full body , and right) red for heads and shoulders, green for feet, blue for right (in the image) arm and cyan for left (in the image).

Table 1. Default OpenCV detectors results. Double detections refers to overlapping detections of the same target, true detections consider multiple correct detections of the same target as a single one.

	Total Detections	Double Detections	True Detections	False Detections	d'
Face	248	1	170	77	0.42
Upper Body	136	0	134	2	1.66
Lower Body	711	132	505	74	1.23
Full Body	213	3	69	41	0.24
Total			14.5%	3.5%	0.89

Table 2. p-ROI detection results

	Total Detections	Double Detections	True Detections	False Detections	d'
Upper Body	1166	56	984 (66%)	126 (8%)	1.79
Feet	909	89	636 (43%)	184 (12%)	0.97
Left Arm	616	81	407 (27%)	128 (9%)	0.76
Right Arm	601	70	353 (24%)	178 (12%)	0.46
Total			40%	10%	

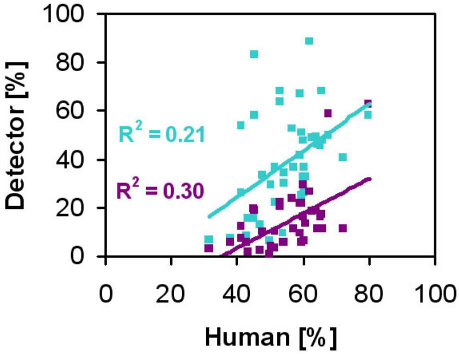

Fig. 5. For each of the 41 pedestrian test video sequences, we observed a correlation between how well human observers performed in a pedestrian detection task and how well the openCV (purple) and p-ROI (cyan) detectors on a detection task. Each dot represents one of the pedestrians. The human data are from [16].

of upper bodies and faces for the default openCV detectors and (a different set of) upper bodies for the p-ROI detectors. The test consisted of image sequences of 41 pedestrians walking through a park [16]. In total there were 1448 images.

Figure 4 illustrates different detection results achieved with the two detectors. Tables 1 and 2 compared the results achieved using the default detectors included in the OpenCV release, and the results achieved with the basic p-ROI detector. As evident in the results tables, the true detection rate is increased with the new approach, however, the false detection rate is also increased. As raised above, we also found a small correlation between both detectors' true detection rate with human detection performance, as shown in Figure 5.

In order to test possible benefits of the geometric restrictions, the hits achieved by the p-ROI detector have been filtered using the geometric coherence tests. The results summarized in Table 3 reflect an important reduction of false detection rates together with a slight reduction in true detection rates. Specifically, false detection are reduced by ∼ 50% (averaging across the individual detectors) whereas true detections are reduced by ∼ 8%.

Table 3. p-ROI with geometric restrictions detection results

	Total Detections	Double Detections	True Detections	False Detections	d'
Upper Body	1062	38	963 (65%)	61 (4%)	2.12
Feet	675	47	547 (37%)	81 (5%)	1.27
Left Arm	507	60	373 (25%)	74 (5%)	0.98
Right Arm	464	43	315 (21%)	106 (7%)	0.67
Total			37%	5%	

6 Conclusions

We have described a pedestrian detector which is based on body parts selected according to their perceptual importance. The results achieved with this set of features improved hits for automatic detection of pedestrian compared with the default detectors included in OpenCV. The method provides a means to select training features to improve automatic vision systems.

Occlusions are not considered explicitly, but the use of a body part approach allows occlusions up to a certain level due to the fact that the system requires only the presence of a single part.

In sum, we believe that this perceptually-based approach combined with simple rules (i.e., our geometric constraints) can be useful for assigning different weights to regions and points not only based on their discrimination features but also on their perceptual significance for the problem considered.

Future work will consider the integration of temporal coherence to improve the detection rate and reduce the false detections when a sequence is available. The application to crowded scenes will also be a challenging problem.

Acknowledgments

Work partially funded by research projects Univ. of Las Palmas de Gran Canaria UNI2003/06, UNI2004/10 and UNI2004/25, Canary Islands Autonomous Government PI2003/160 and PI2003/165 and the Spanish Ministry of Education and Science and FEDER funds (TIN2004-07087).

References

1. Stan Birchfield. Elliptical head tracking using intensity gradients and color histograms. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 232–237, June 1998.
2. Marco La Cascia and Stan Sclaroff. Fast, reliable head tracking under varying illumination: An approach based on registration of texture-mapped 3d models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(4):322–336, April 2000.
3. F. Gosselin and P. G. Schyns. Bubbles: a technique to reveal the use of information in recognition tasks. *Vision Research*, pages 2261–2271, 2001.
4. Erik Hjelmas and Boon Kee Low. Face detection: A survey. *Computer Vision and Image Understanding*, 83(3):236–274, 2001.
5. Intel. Intel Open Source Computer Vision Library, b4.0.
www.intel.com/research/mrl/research/opencv, August 2004.
6. Michael J. Jones and James M. Rehg. Statistical color models with application to skin detection. Technical Report Series CRL 98/11, Cambridge Research Laboratory, December 1998.
7. B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In *Proceedings of the CVPR'05)*, 2005.

8. Rainer Lienhart, Alexander Kuranov, and Vadim Pisarevsky. Empirical analysis of detection cascades of boosted classifiers for rapid object detection. In *DAGM'03*, pages 297–304, Magdeburg, Germany, September 2003.
9. D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Comnputer Vision*, 60(2):91–110, 2004.
10. Krystian Mikolajczyk, Cordelia Schmid, and Andrew Zisserman. Human detection based on a probabilistic assembly of robust part detectors. In *European Conference on Computer Vision*, volume I, pages 69–81, 2004.
11. C. Papageorgiou, M. Oren, and T. Poggio. A general framework for object detection. In *Proceedings of the International Conference on Computer Vision*, pages 555–562, 1998.
12. C. Papageorgiou and T. Poggio. A trainable system for object detection. *International Journal of Computer Vision*, 38(1), 2000.
13. Paul Viola and Michael J. Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition*, pages 511–518, 2001.
14. Paul Viola and Michael J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):151–173, May 2004.
15. Paul Viola, Michael J. Jones, and Daniel Snow. Detecting pedestrians using patterns of motion and appearance. In *Proc. of the International Conference on Computer Vision*, volume 2, pages 734–741, October 2003.
16. Q. C. Vuong, A. Hof, H. H. Bülthoff, and I. M. Thornton. An advantage for detecting human targets in dynamic versus static composite stimuli. In *4th annual meeting of the Vision Sciences Society*, 2004.
17. Ming-Hsuan Yang, David Kriegman, and Narendra Ahuja. Detecting faces in images: A survey. *Transactions on Pattern Analysis and Machine Intelligence*, 24(1):34–58, 2002.