Proceedings of the IEEE ICRA 2009
Workshop on People Detection and Tracking
Kobe, Japan, May 2009

Visual People Detection – Different Models, Comparison and Discussion

Bernt Schiele, Mykhaylo Andriluka, Nikodem Majer, Stefan Roth and Christian Wojek
Department of Computer Science, TU Darmstadt

Abstract

Over the last few years, visual people detection has made impressive progress. The paper gives an overview of some of the most successful techniques for people detection and also summarizes a recent quantitative comparison of several state-of-the-art methods. As a proof-of-concept we show that the combination of visual and laser-based people detection can result in a significant increase in performance. We also briefly discuss future research directions for visual people detection.

1. Introduction

People detection is one of the most challenging problems in computer vision due to large variations caused by articulation, viewpoint and appearance. At the same time detecting and tracking people has a wide range of applications including robotics, image and video indexing, surveillance and automotive safety. Consequently visual people detection has been researched intensively with a rapid rate of innovation. Recently, several researchers have reported impressive results [23, 33, 6, 18, 1, 36] for this task.

The aim of this paper is threefold. First, we provide an overview of some of the most successful methods for visual people detection. Second, we summarize a comparative study of sliding-window techniques [35]. And third, we show the potential of combining visual people detection with other modalities such as laser.

Broadly speaking there are two major types of approaches for visual people detection. Sliding-window methods exhaustively scan the input images over positions and scales independently classifying each sliding window (e.g. [23, 33, 6]). Other methods generate hypotheses by evidence aggregation often using part-based human body models (e.g. [12, 9, 21, 18, 37, 28, 1]). After discussing some of the most successful sliding-window approaches in section 2 we summarize a comparative study of such methods in section 3. Section 4 briefly describes a part-based model that has shown to outperform sliding-window techniques in the presence of partial occlusion. Section 5 then describes an experiment to complement visual people detection with a laser-range finder thereby significantly reducing the number of false positives of the visual people detector. The last section 6 discusses promising research directions to improve the performance of today's visual people detection methods.

2. Sliding-window techniques

Sliding window detection systems scan the image at all relevant positions and scales to detect a person. Consequently there are two major components: the *feature* component encodes the visual appearance of the person, whereas the *classifier* determines for each sliding window independently whether it contains the person or not. As typically many positions and scales are scanned these techniques are inherently computationally expensive. Fortunately, due to recent advances in GPUs, real-time people detection is possible as e.g. demonstrated by [34]. In [35] we conducted a quantitative comparison that we briefly summarize in section 3.

As a complete review on people detection is beyond the scope of this work, we focus on most related work. An early approach [23] used Haar wavelets and a polynomial SVM while [33] used Haar-like wavelets and a cascade of AdaBoost classifiers. Gavrila [13] employs a hierarchical Chamfer matching strategy to detect people. Recent work often employs statistics on image gradients for people detection. [30] uses edge orientation histograms in conjunction with SVMs while [6] uses an object description based on overlapping histograms of gradients. [27] employs locally learned features in an AdaBoost framework and Tuzel [32] presents a system that exploits covariance statistics on gradients in a boosting classification setting. Interestingly, most approaches use discriminant classifiers such as AdaBoost or SVMs while the underlying object descriptors use a diverse set of features. Therefore the following section briefly describe some of these features in more detail.

Haar wavelets have first been proposed by Papageorgiou and Poggio [23]. They introduce a dense overcomplete representation using wavelets at the scale of 16 and 32 pixel with an overlap of 75%. Three different types are used, which allow to encode low frequency changes in contrast: vertical, horizontal and diagonal. Thus, the overall length of

the feature vector for a 64×128 pixel detection window is 1326 dimensions. In order to cope with lighting differences, for each color channel only the maximum response is kept and normalization is performed according to the window's mean response for each direction. Additionally, the original authors report that for the class of people the wavelet coefficient's sign is not carrying information due to the variety in clothing. Hence, only the absolute values for each coefficient is kept. During our experiments we found that an additional L_2 length normalization with regularization of the feature vector improves performance.

Histograms of oriented gradients have been proposed by Dalal and Triggs [6]. Image derivatives are computed by centered differences in x- and y direction. The gradient magnitude is then inserted into cell histograms (8×8 pixels), interpolating in x, y and orientation. Blocks are groups of 2×2 cells with an overlap of one cell in each direction. Blocks are L_2 length normalized with an additional hysteresis step to avoid one gradient entry to dominate the feature vector. The final vector is constituted of all normalized block histograms with a total dimension of 3780 for a 64×128 detection window.

Shape Context has originally been proposed as a feature point descriptor [4] and has shown excellent results for people detection in the generative ISM framework [18, 28]. The descriptor is based on edges which are extracted with a Canny detector. Those are stored in a log-polar histogram with location being quantized in nine bins. For the radius 9, 16 and 23 pixels are used, while orientation is quantized into four bins. For sliding window search we densly sampled on a regular lattice with a support of 32 pixels (other scales in the range from 16 to 48 pixels performed worse). For our implementation we used the version of Mikolajczyk [20] which additionally applies PCA to reduce the feature dimensionality to 36 dimensions. The overall length of all descriptors concatenated for one test window is 3024.

Classifiers. The second major component for sliding-window approaches is the deployed classifier. For the classification of single windows two popular choices are SVMs and decision tree stumps in conjunction with the AdaBoost framework. SVMs optimize a hyperplane to separate positive and negative training samples based on the *global* feature vector. Different kernels map the classification problem to a higher dimensional feature space. For our experiments we used the implementation *SVM Light* [16]. 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.

3. Comparison of sliding-window techniques

In [35] we conducted a systematic evaluation of different feature/classifier combinations. For this we reimplemented the respective features and classifiers. Comparisons with published binaries (whenever available) verified that our reimplementations perform at least as good as the originally proposed feature/classifier combinations. In the following 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 use the established INRIA Person dataset [1]. This data set contains images of humans taken from several viewpoints under varying lighting conditions in indoor and outdoor scenes. Unlike the original authors [6] we test the trained detectors on the full images. We do so, in order not only to evaluate the detector in terms of false positive detections per window (FPPW) but with respect to their frequency and spatial distribution. This gives a more realistic assessment on how well a detector performs for real image statistics. For further details see [35]

Due to space constraints we cannot report all the quantitative results from [35]. However, we still report the major results and figure 1 contains the results for four different settings.

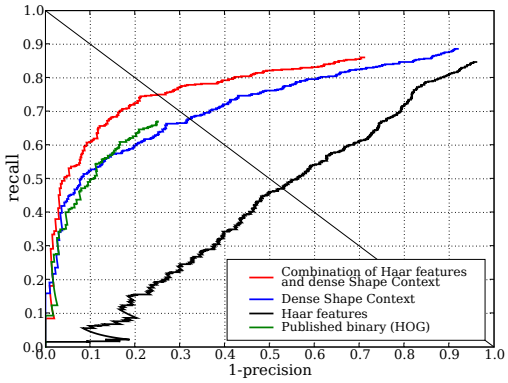


Figure 1. Recall-Precision detector performances for different features (Haar, HOG, Dense Shape Context, combination of Dense Shape Context and Haar) and linear SVM-classifier

Single feature detection. We start by summarizing the evaluation of using all features individually in combination with the three classifiers AdaBoost, linear SVM and RBF kernel SVM. First of all, the HOG descriptor and the similar Shape Context descriptor consistently outperform the other features (e.g. Haar-like features) independent of the learning algorithm. Overall, RBF kernel SVMs together with the gradient-based features HOG and Shape Context show the best results. All features except shapelets show better performance with the RBF kernel SVM compared to the linear SVM. AdaBoost achieves a similarly good performance in comparison with RBF kernel SVMs in particular for the Haar-like wavelet, the HOG feature and for shapelets. It does slightly worse for the dense Shape Context descriptor.

[1] http://pascal.inrialpes.fr/data/human

Multi-cue detection. A closer look on the single detectors' complementarity reveals that different features in combination with different classifiers have a varying performance on the individual instances. This can be explained by the fact, that the features encode different information. While gradients encode high frequency changes in the images, Haar wavelets as they are proposed by [23] also encode much lower frequencies. Figure 1 shows the combination of dense Shape Context features with Haar wavelets. In particular figure 1 shows, that in fact both features on their own cannot reach the performance that is reached with their combination. Compared to the state-of-the-art HOG object detector we improve recall considerably about 10% at 80% precision. Figure 2 shows sample detections of this multi-cue detector.

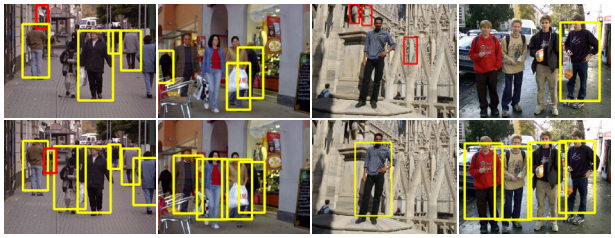


Figure 2. Sample detections at a precision of 80%. Red bounding boxes denote false detections, while yellow bounding boxes denote true positives. First row shows detection by the publically available HOG detector[6]; second row depicts sample detections for our combination of dense Shape Context with Haar wavelets in a linear SVM

Failure analysis. To get a feeling about the achievable performance of sliding-window based techniques we complete our brief summary with a failure case analysis. In particular, we analyzed the missing recall and the false positive detections at equal error rate (149 missing detections / 149 false positives) for the feature combination of Shape Context and Haar wavelets in combination with a linear SVM. Missing recall mainly occurred due to unusual articulations (37 cases), difficult background or contrast (44 cases), occlusion or carried bags (43 cases), under- or overexposure (18 cases) and due to detection at too large or too small scales (7). There were also 3 cases which were detected with the correct height but could not be matched to the annotation according to the PASCAL criterion due to the very narrow annotation.

False positive detections can be categorized as follows: Vertical structures like poles or street signs (54 cases), cluttered background (31 cases), too large scale detections with people in lower part (24 cases), too low scale on body parts (28 cases). There were also a couple of "false" detections (12 cases) on people which were not annotated in the database (mostly due to occlusion or at small scales). Some samples of missed people and false positives are shown in figure 3.

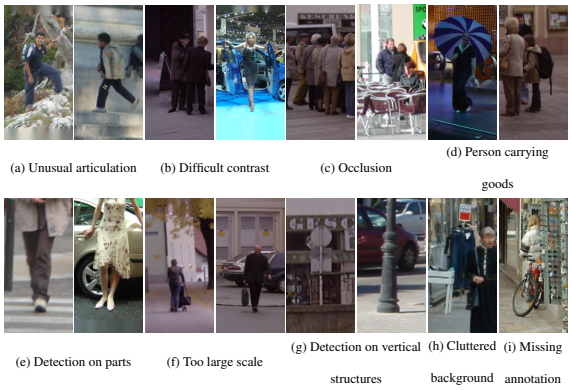


Figure 3. Missed recall (upper row) and false positive detections (lower row) at equal error rate

4. Part-based models for people detection

Part-based models have a long history in computer vision for object detection in general and for people detection in particular (e.g. [12, 9, 21, 18, 37, 28, 1]). There are two major components of these models. The first uses low-level features or classifiers to model individual parts or limbs of a person. The second component models the topology of the human body to enable the accumulation of part evidence.

A wide range of models have been proposed e.g. for upright people detection in traffic scenes [18], to estimate the pose of highly articulated people (e.g. in sports scenes [25]), or for upper body detection and pose estimation [11], e.g. for movie indexing. In this section we briefly summarize one of our own models [1] that builds upon and extends a number of previous approaches. The model is inspired by the pictorial structures model proposed by [10, 15], but uses more powerful part representations and detections, and as we will show outperforms recent pedestrian detectors [6, 28].

A part-based person model [1]. Following the general pictorial structures idea, a person is represented as a joint configuration of her body parts. In such a model the problem of locating a person in a test image is formulated as search for the modes of the posterior probability distribution $p(L|E)$ of the body part configuration L given the image evidence E and (implicit) class-dependent model parameters θ. In our model, the configuration is described as $L = \{\mathbf{x}^o, \mathbf{x}^1, \ldots, \mathbf{x}^N\}$, where $\mathbf{x}^o$ is the position of the body center and its scale, and $\mathbf{x}^i$ is the position and scale of body part i. The image evidence, which here is defined as a set of local features observed in the test image, will be denoted as $E = \{\mathbf{e}_k^{app}, \mathbf{e}_k^{pos} | k = 1, \ldots, K\}$, where $\mathbf{e}_k^{app}$ is an appearance descriptor, and $\mathbf{e}_k^{pos}$ is the position and scale of the local image feature with index k.

An important component of the pictorial structures model is an implicit model of a-priori knowledge about possible body configurations, which must be expressive enough

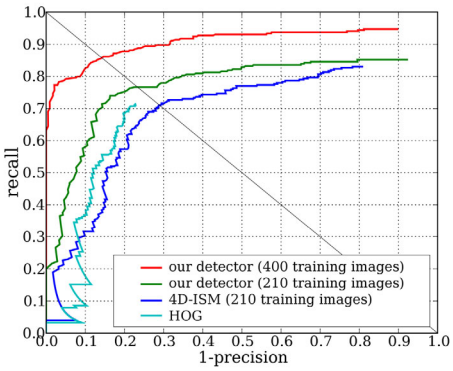


Figure 4. Comparison of our pedestrian detector with 4D-ISM detector [29] and HOG [6].

to capture all important dependencies between parts. For particular object categories, such as walking people, we can introduce auxiliary state variables that represent the *articulation state* or an *aspect* of the object, such as different phases in the walking cycle of a person [17], and make the parts conditionally independent. As we are not interested in knowing the articulation state, but only the object and limb positions, the articulation state a can be marginalized out: $p(L|E) = \sum_a p(L|a, E)p(a)$.

From decomposing $p(L|a, E) \propto p(E|L, a)p(L|a)$, assuming that the configuration likelihood can be approximated with product of individual part likelihoods [10] $p(E|L, a) \approx \prod_i p(E|\mathbf{x}^i, a)$, and assuming uniform $p(\mathbf{x}^i|a)$, it follows that

$$p(L|a, E) \approx p(\mathbf{x}^o) \prod_i p(\mathbf{x}^i|a, E)p(\mathbf{x}^i|\mathbf{x}^o, a). \quad (1)$$

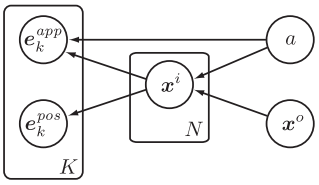


Figure 5. Graphical model structure describing the relation between articulation, parts, and features.

Please refer to [1] for the details concerning model training and inference. In the experiment (as presented in detail in [1]) we use shape context feature descriptors [3] and the Hessian-Laplace interest point operator [19] as detector. Ee compare the above detector on a challenging dataset of street scenes containing 311 side-view pedestrians with significant variation in clothing and articulation[2]. Fig. 4 shows the comparison of our detector with two state-of-the-art detectors. Using the same training set as [28] our detector outperforms the 4D-ISM approach [28] as well as the HOG-detector [6]. Increasing the size of the training set further improves performance significantly.

[2] Available at www.mis.informatik.tu-darmstadt.de.

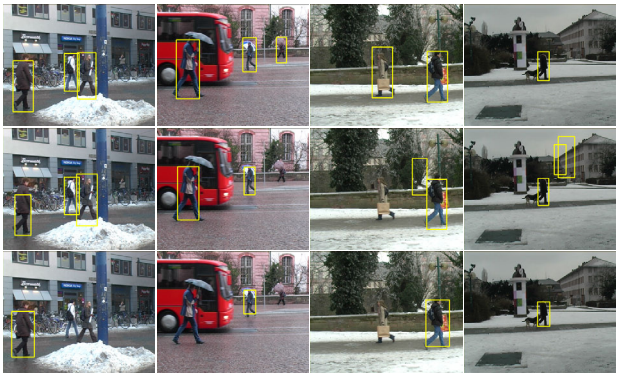


Figure 6. Example detections at equal error rate of our detector (top), 4D-ISM (middle) and HOG (bottom) on the "TUD-Pedestrians" dataset.

Fig. 6 shows sample detections of the 3 methods on test images. The 4D-ISM detector is specifically designed to detect people in cluttered scenes with partial occlusions. Its drawback is that it tends to produce hypotheses even when little image evidence is available (image 3 and 4), which results in increased number of false positives. The HOG detector seems to have difficulties with the high variety in articulations and appearance present in out dataset. However, we should note that it is a multi-view detector designed to solve a more general problem than we consider here.

Summary. From these experiment we can conclude that part-based people model can outperform sliding-window based methods (such as HOG) in the presence of partial occlusion and significant articulations. It should be noted however, that part-based models tend to require a higher resolution of the person in the image than most sliding-window based approaches.

5. Combining vision and laser to improve people detection

Cameras are not the only sensor that can be used for people detection. In robotics laser range scanners are widely used for tasks like localization and position estimation but have been also used for people detection [2] and place classification [22]. Recent approaches to fuse visual and laser information for classification and object detection tasks show promising results [26, 24, 38, 31]. This section explores a simple yet effective technique to combine vision and laser information for improved people detection. As visual people detection is never perfect laser range information is used to constrain the search space of plausible hypotheses.

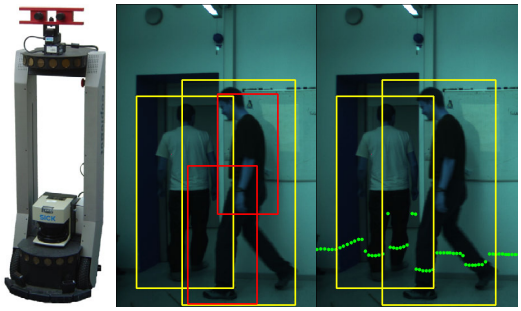
Setting. The platform used for data acquisition is a PeopleBot that runs a distributed component architecture devel-

oped during the CoSy project. The robot is equipped with a SICK LMS (180° fov, 1° angular resolution) mounted approximately 30 cm above the floor and a color camera stereo head located 97cm above the LMS. Only one camera is used for visual people detection. The camera is calibrated using the CALIB toolbox [5] while the transformation parameters between the camera and the LMS coordinate system are set by measuring the robots geometry.

Approach. In this section we use a sliding-window approach for people detection where we choose the HOG descriptor as feature and a linear SVM as classifier (see section 2). To achieve good generalization performance in various environments we decided to train the classifier on the INRIA people data set (see section 3). As expected the visual people detector already achieves good results. Figure 7 shows sample detections as well as typical false positive detections e.g. on partial people or vertical edge structures.

Many false positive detections do not fulfill simple constraints assuming that people usually walk on the floor and therefore the object scale is proportional to distance. This assumption can be formulated with the following two constraints to prune the space of valid hypotheses obtained from the HOG detection stage; (1): laser range measurements projected onto the image plane should hit the lower third (legs) of the detection window h_i. We denote the set of the associated range values that meet this condition with R_i. And (2): detection scale s_i of h_i is bounded by a factor proportional to the largest/smallest distance measurement found in R_i : $s^*/min(R_i) + c > s_i > s^*/max(R_i) - c$ where s^* is a scale estimate at 1m distance and c is a small constant accounting for errors in scale estimations. Both parameters are in pixel units and dependent on camera parameters. Since all detection hypotheses have the same aspect ratio we set s_i to the detection window width.We initially set s^* to 550 and c to 25. These values are estimated from a subset of the recorded data.

If a visual person hypothesis does not meet these constraints it is rejected. Figure 7(c) shows the effect of rejecting hypotheses that do not match these two constraints.



(a) (b) (c)

Figure 7. The PeopleBot Robot (a). Typical false positives from visual people detection (b). Rejection by simple range based constraints (c).

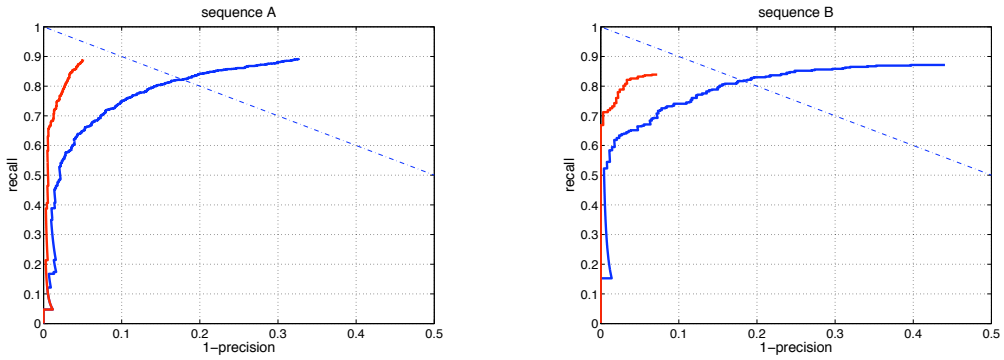
Evaluation. We evaluate this simple procedure on two sequences. Sequence (A) (samples shown in figures 7(b) and 9(c)) is recorded in an office sized room and sequence (B) (samples shown in figures 9(a) and 10) in a large foyer of a university building. Ground truth is annotated manually in form of bounding boxes and is quite complete in the sense that occluded people or people entering the visual field are also annotated if approximately half of the person is visible. As a consequence we cannnot expect to reach full recall. For both sequences the robot was placed to have a good visual view of the scene. Due to the nature of the concurrent distributed component architecture the sampling process for each sensor is asynchronous and tends to vary slightly with the overall system load. We aligned the laser and the vision sensor in a semi-automatic fashion such that each image frame is associated to the laser scan with the smallest temporal difference.

Sequence A. This sequence consists of 1023 image frames sampled at 2.5 Hz on average while laser recordings reach 20.8 Hz. The environment is an office sized room with people entering and leaving the room through two doorways. People might occlude each other and be occluded by the wall. Figure 8(a) shows the detection performance for this sequence.The HOG detector reaches a maximal recall of 89.2% with a precision of 67.3% The equal error rate (EER) is 82.2%. The use of laser range information clearly improves precision to 95% with a loss smaller than 0.25% in maximally achievable recall.

Sequence B. The second sequence consists of 124 images sampled at 0.7 Hz on average while laser recordings reach 37.5 Hz. In this sequence more people appear also at large scales so that they are not fully visible. The HOG detector reaches a maximal recall of 87.1% with a precision of 56% The EER is 81.7%. Laser range information improves precision to 92.8% at 83.9% recall (i.e. 3.2% loss in recall).

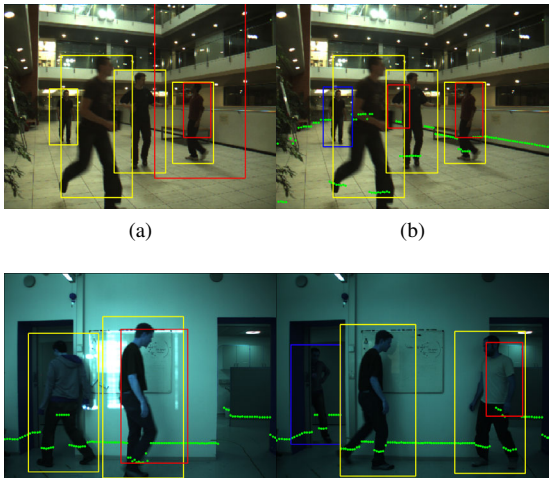
Loss of recall is a sign that the posed constrains are not necessarily true for all ground truth instances. This happens in cases where a true positive hypothesis occludes the laser which leads to rejection of a true positive detection at a smaller scale due to the missing laser readings. Not achieving full precision means that cases occur where false positive detections fulfill the laser constraints. This happens if multiple hypotheses at similar scales are found as true positive hypotheses or if a false positive hypothesis and laser range readings fulfill the constraints by chance. See Figure 9 for failure cases and figure 10 for sample detections.

Conclusion. Overall, in terms of EER, the proposed combination of camera and laser information improves precision/recall by 12.8% / 7 % on sequence A and 11.1% / 2.2 % on sequence B. This improvement is clearly significant and highly encouraging given the simplicity of the de-



(a) (b)

Figure 8. Detection performance for test sequences A and B. HOG detection in blue. Laser constrained HOG in red.



(a) (b)

(c) (d)

Figure 9. Visual detection only (a) & (c). Laser constrained detections (b) & (d). The laser range based constraints correct one false positive but also reject one true positive detection. True positive detections are marked yellow, missed objects blue and false detections are red.

scribed algorithm. We believe that this clearly demonstrates the potential to combine camera and laser information and that more elaborate algorithms should enable to improve performance further.

6. Conclusion and discussion

The primary aim of this paper was to give an overview of promising techniques for visual people detection (sections 2–4). In recent years the field has been moving rapidly thereby continuously improving detection performance. Given today's state-of-the-art in visual people detection it is clear however that the currently achievable performance is often neither sufficient nor satisfactory for many applications. In this last section we briefly discuss several research directions that have the potential to improve overall performance.

Motion cues. It is clear that human motion is an important cue for people detection. Quite surprisingly however, motion is seldom used for people detection. Notable exceptions are the work by Viola et. al [33], Dalal and Triggs [7] and Wojek et al. [36]. All three papers clearly demonstrate the potential gain when using motion information for visual people detection. However, we strongly believe that the current approaches still leave room for further improvement.

Integration of detection and tracking. Both detection and tracking people are challenging problems. People detectors have been shown to be able to locate pedestrians even in complex scenes, but false positives have remained frequent. Tracking methods are able to find a particular individual in image sequences, but are severely challenged by real-world scenarios such as crowded scenes. Therefore it is a promising research direction to combine the advantages

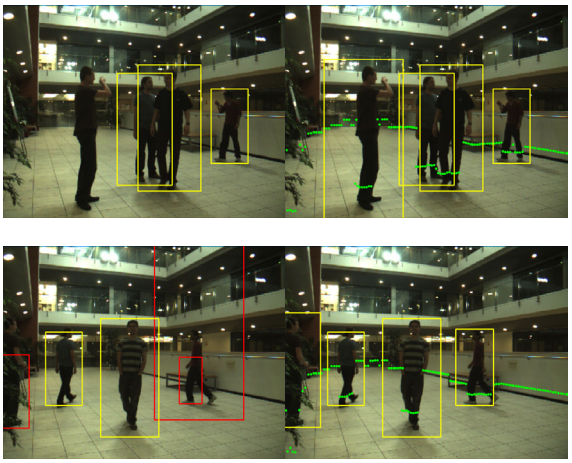


Figure 10. HOG hypotheses thresholded at EER on the left. Laser constrained hypotheses on the right.

of both detection and tracking in a single framework. In [1] we have proposed such an integrated framework that allows to detect and track multiple people in cluttered scenes with reoccurring occlusions. While this research direction is again largely under-explored we strongly believe that this is a highly promising route to pursue.

System integration. It seems clear that the integration of all of the above mentioned information into a single overall system has the potential to obtain an improved overall performance. Due to the complexity of this task however relatively view such systems exist. Probably the best known examples are the system by Gavrila and colleagues [14] and more recently the work by Ess and colleagues [8]. In these systems different components are integrated such as stereo and depth estimation, structure from motion, texture based classifiers and part-based people detectors.

Combination with other sensor modalities. Section 5 already demonstrated the potential of combining vision and laser information to improve overall detection performance. While this research direction has gained attention recently [26, 24, 38, 31] it is again under-explored and has the potential to enable robust people detection e.g. for robotics and automative applications.

References

[1] M. Andriluka, S. Roth, and B. Schiele. People-tracking-by-detection and people-detection-by-tracking. *CVPR 2008*.

[2] K. O. Arras, O. M. Mozos, and W. Burgard. Using boosted features for the detection of people in 2D range data. *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 3402–3407, 2007.

[3] S. Belongie, J. Malik, and J. Puzicha. Shape context: A new descriptor for shape matching and object recognition. *NIPS*2000*.

[4] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *PAMI*, 24(4):509–522, 2002.

[5] J.-Y. Bouguet. Camera calibration toolbox, 2008.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *CVPR 2005*.

[7] N. Dalal., B. Triggs., and C. Schmid. Human detection using oriented hist. of flow and appearance. *ECCV*, 2006.

[8] A. Ess, K. Schindler, B. Leibe, and L. van Gool. Robust mulit-person tracking from a moving platform. *CVPR*, 2008.

[9] P. Felzenszwalb and D. Huttenlocher. Efficient matching of pictorial structures. *CVPR*, 2000.

[10] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61:55–79, 2007.

[11] V. Ferrari, M. Marin, and A. Zisserman. Progressive search space reduction for human pose estimation. *CVPR 2008*.

[12] D. Forsyth and M. Fleck. Body plans. *CVPR*, 1997.

[13] D. Gavrila. Multi-feature hierarchical template matching using distance transforms. *Proceedings of the International Conference on Pattern Recognition*, v. 1, pp. 439–444, 1998.

[14] D. Gavrila and S. Munder. Multi-cue pedestrian detection and tracking from a moving vehicle. *IJCV*, 73(1):41–59, 2007.

[15] S. Ioffe and D. Forsyth. Human tracking with mixtures of trees. *ICCV 2001*.

[16] T. Joachims. Making large–scale SVM learning practical. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods — Support Vector Learning*, pp. 169–184, Cambridge, MA, 1999. MIT Press.

[17] X. Lan and D. P. Huttenlocher. Beyond trees: Common-factor models for 2d human pose recovery. *ICCV 2005*.

[18] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. *CVPR*, pp. 878–885, 2005.

[19] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *IJCV*, 60:63–86, 2004.

[20] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *PAMI*, 27(10):1615–1630, 2005.

[21] K. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust part detectors. *ECCV*, pp. 69–81, 2004.

[22] O. M. Mozos, C. Stachniss, and W. Burgard. Supervised learning of places from range data using AdaBoost. *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 1742–1747, Barcelona, Spain, 2005.

[23] C. Papageorgiou and T. Poggio. A trainable system for object detection. *IJCV*, 38(1):15–33, 2000.

[24] A. Pronobis, O. Martínez Mozos, and B. Caputo. SVM-based discriminative accumulation scheme for place recognition. *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA08)*, Pasadena, CA, USA, 2008.

[25] D. Ramanan. Learning to parse images of articulated objects. *NIPS*2006*.

[26] A. Rottmann, O. M. Mozos, C. Stachniss, and W. Burgard. Semantic place classification of indoor environments with mobile robots using boosting. *Proceedings of the National Conference on Artificial Intelligence*, pp. 1306–1311, Pittsburgh, PA, USA, 2005.

[27] P. Sabzmeydani and G. Mori. Detecting pedestrians by learning shapelet features. *CVPR*, 2007.

[28] E. Seemann, B. Leibe, and B. Schiele. Multi-aspect detection of articulated objects. *CVPR*, pp. 1582–1588, 2006.

[29] E. Seemann and B. Schiele. Cross-articulation learning for robust detection of pedestrians. *DAGM*, 2006.

[30] A. Shashua, Y. Gdalyahu, and G. Hayun. Pedestrian detection for driving assistance systems: Single-frame classification and system level performance. *International Symposium on Intelligent Vehicles*, pp. 1–6, 2004.

[31] L. Spinello, R. Triebel, and R. Siegwart. Multimodal people detection and tracking in crowded scenes. *Proc. of The AAAI Conference on Artificial Intelligence (Physically Grounded AI Track)*, 2008.

[32] O. Tuzel, F. Porikli, and P. Meer. Human detection via classification on Riemannian manifolds. *CVPR*, 2007.

[33] P. Viola, M. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. *CVPR*, 2003.

[34] C. Wojek, G. Dorko, A. Schulz, and B. Schiele. Sliding-windows for rapid object class localization: a parallel technique. *Pattern Recognition (DAGM) 2008*.

[35] C. Wojek and B. Schiele. A performance evaluation of single and multi-feature people detection. *Pattern Recognition (DAGM) 2008*.

[36] C. Wojek, S. Walk, and B. Schiele. Multi-cue onboard pedestrian detection. *CVPR 2009*.

[37] B. Wu and R. Nevatia. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. *ICCV*, 2005.

[38] Z. Zivkovic and B. Kröse. Part based people detection using 2d range data and images. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2007.