A Performance Evaluation of Single and Multi-Feature People Detection

Christian Wojek, Bernt Schiele
{wojek, schiele}@cs.tu-darmstadt.de

Computer Science Department
TU Darmstadt

Abstract. Over the years a number of powerful people detectors have been proposed. While it is standard to test complete detectors on publicly available datasets, it is often unclear how the different components (e.g. features and classifiers) of the respective detectors compare. Therefore, this paper contributes a systematic comparison of the most prominent and successful people detectors. Based on this evaluation we also propose a new detector that outperforms the state-of-art on the INRIA person dataset by combining multiple features.

1 Introduction

People are one of the most challenging classes for object detection, mainly due to large variations caused by articulation and appearance. Recently, several researchers have reported impressive results [1–3] for this task. Broadly speaking there are two types of approaches. Sliding-window methods exhaustively scan the input images over position and scale independently classifying each sliding window, while other methods generate hypotheses by evidence aggregation (e.g. [3–7]). To the best of our knowledge there exist only two comparative studies on people detection methods. [8] compares local features and interest point detectors and [9] compares various sliding window techniques. However, [9] is focused on automotive applications and their database only consists of cropped gray scale image windows. While the evaluation on single image windows is interesting, it does not allow to assess the detection performance in real-world scenes where many false positive detections may arise from body parts or at wrong scales. This paper therefore contributes a systematic evaluation of various features and classifiers proposed for sliding-window approaches where we assess the performance of the different components and the overall detectors on entire real-world images rather than on cropped image windows.

As a complete review on people detection is beyond the scope of this work, we focus on most related work. An early approach [1] used Haar wavelets and a polynomial SVM while [10] used Haar-like wavelets and a cascade of AdaBoost classifiers. Gavrila [11] employs a hierarchical Chamfer matching strategy

G. Rigoll (Ed.): DAGM 2008, LNCS 5096, pp. 82–91, 2008.

to detect people. Recent work often employs statistics on image gradients for people detection. [12] uses edge orientation histograms in conjunction with SVMs while [2] uses an object description based on overlapping histograms of gradients. [13] employs locally learned features in an AdaBoost framework and Tuzel [14] 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.

This work contributes a systematic evaluation of different feature representations for general people detection in combination with discriminant classifiers on full size images. We also introduce a new feature based on dense sampling of the Shape Context [15]. Additionally, several feature combination schemes are evaluated and show an improvement to state-of-the-art [2] people detection.

The remainder of this paper is structured as follows. Section 2 reviews the evaluated features and classifiers. Section 3 introduces the experimental protocol, and section 4.1 gives results for single cue detection. Results for cue combination are discussed in section 4.2 and section 4.3 analyzes failure cases.

2 Features and Classifiers

Sliding window object detection systems for static images usually consist of two major components which we evaluate separately in this work. The *feature* component encodes the visual appearance of the object to be detected, whereas the *classifier* determines for each sliding window independently whether it contains the object or not.

Table 1 gives an overview of the feature/classifier combinations proposed in the literature. As can be seen from this table, many possible feature/classifier combinations are left unexplored therefore making it difficult to assess the respective contribution of different features and classifiers to the overall detector performance. To enable a comprehensive evaluation using all possible feature/classifier combinations, we reimplemented the respective methods. Comparisons with published binaries (whenever available) verifies that our reimplementations perform at least as good as the originally proposed feature/classifier combinations (cf. Figure 1(h)). The remainder of this section reviews the evaluated features and classifiers.

	Linear SVM	Kernel SVM	Ada-Boost	Other	Criterion
Haar wavelet [1]		polynomial			ROC
Haar-like wavelet [10]			cascaded		ROC
HOG [2]	✓	RBF			FPPW
Shapelets [13]			✓		FPPW
Shape Context [8]				ISM	RPC

Table 1. Original combination of features and classifiers

2.1 Features

Haar wavelets have first been proposed by Papageorgiou and Poggio [1]. 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.

Haar-like features have been proposed by Viola and Jones [10] as a generalization of Haar wavelets with arbitrary dimensions and different orientations (efficiently computed by integral images). They suggest to exhaustively use all possible features that can be sampled from a sliding window and let AdaBoost select the most discriminative ones. Thus, their approach is computationally limited to rather small detection window sizes. For our evaluation we us the OpenCV [1] implementation of their algorithm to select the relevant features and only use those appropriately scaled to our detection window's size of 64×128 pixels. Similarly to [1] we found that for the class of people the coefficient's sign is irrelevant due to different clothing and surroundings and therefore used absolute values. Moreover, we found that the applied illumination variance normalization performs worse than simple L_2 length normalization on the selected features.

Histograms of oriented gradients have been proposed by Dalal and Triggs [2]. 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.

Shapelets [13] are another type of gradient-based feature obtained by selecting salient gradient information. They employ discrete Adaboost on densly sampled gradient image patches of multiple orientations $(0°, 90°, 180°, 270°)$ at the scales of 5 to 15 pixels to classify those locally into people and non-people based on the local shape of the object. As preprocessing step, gradient images are smoothed to account for inaccuracies of the person's position within the annotation. Moreover, the underlying gradient image is normalized shapelet-wise to achieve illumination invariance. Compared to the published source code[2] we use stronger regularization for the normalization step, in order not to amplify noise. This improves the results considerably.

Shape Context has originally been proposed as a feature point descriptor [15] and has shown excellent results for people detection in the generative ISM

[1] `http://sourceforge.net/projects/opencvlibrary`
[2] `http://www.cs.sfu.ca/~mori/research/shapelet_detect`

framework [16, 3]. 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 [17] 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.

2.2 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* [18]. 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 Dataset and Methodology

To evaluate the performance for the introduced features and their combination with different classifiers we use the established INRIA Person dataset [3]. This data set contains images of humans taken from several viewpoints under varying lighting conditions in indoor and outdoor scenes. For training and testing the dataset is split into three subsets: the full size positive images, the scale-normalized crops of humans and full size negative images. Table 2 gives an overview of the number of images and the number of overall depicted people.

For training we use all 2416 positive images and for the negative training instances we randomly cropped a fixed set of 10 negative windows from every negative image. Unlike the original authors [2] 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. To allow this evaluation in terms of recall and precision,

	Positive set/ # instances	Normalized crops set	Negative set
Training	615 / 1208	2416	1218
Testing	288 / 566	1132	453

Table 2. Number of images and instances for the INRIA Person dataset

[3] `http://pascal.inrialpes.fr/data/human`

the nearby initial detections in scale and space need to be merged to a single final hypothesis. To achieve this, a mode seeking adaptive-bandwidth mean shift algorithm [19] is used. The width of the smoothing kernel was kept fixed for all experiments and no further postprocessing was applied. Ground truth and final detections are matched using the PASCAL criterion [20], which demands a minimum overlap of 50% for two matching bounding boxes.

4 Experiments

4.1 Single feature detection

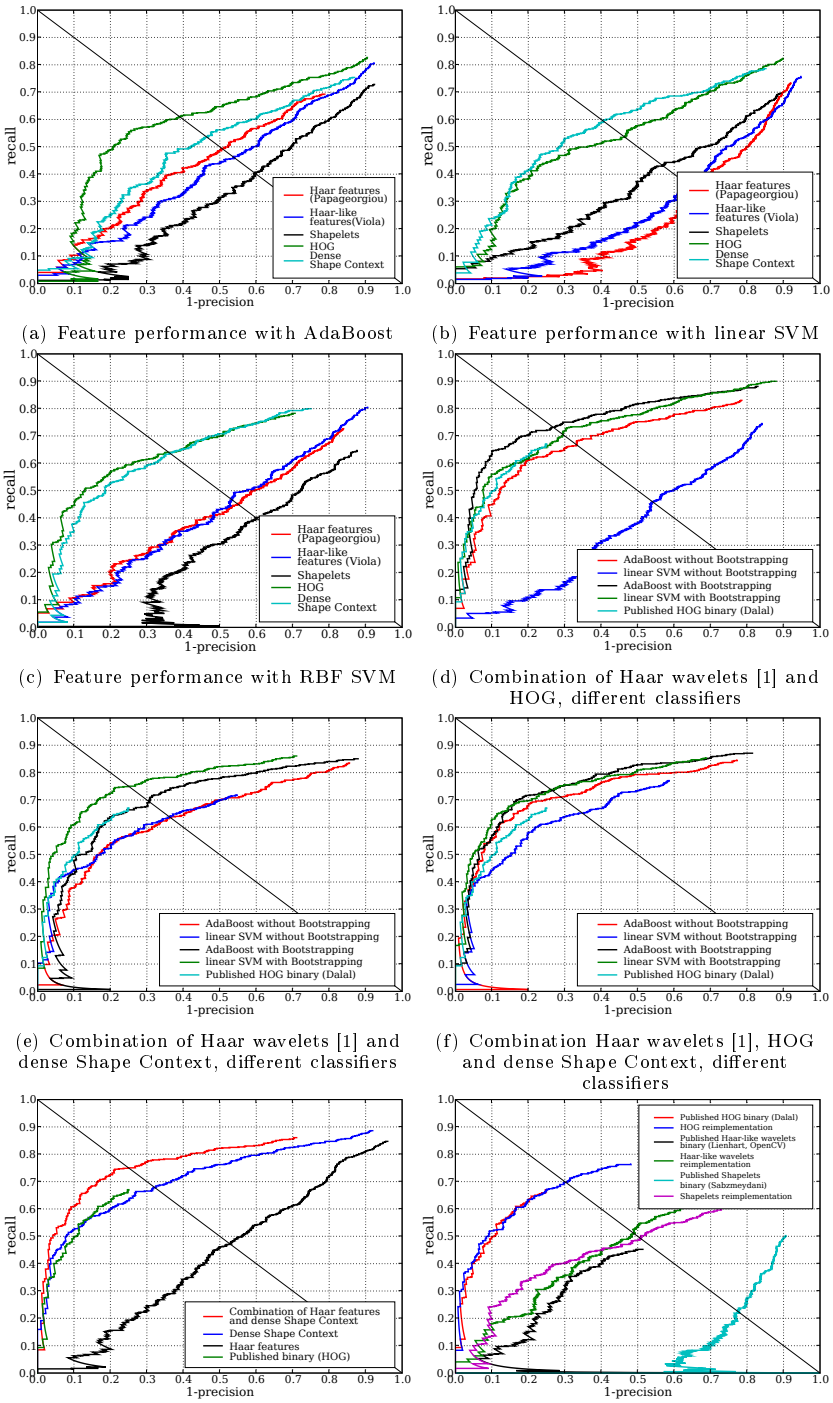
We start by evaluating all features individually in combination with the three classifiers AdaBoost, linear SVM and RBF kernel SVM. In order not to introduce bias by the selection of negative samples a fixed set was used and no bootstrap learning was employed. Figures 1 (a)-(c) show the results we have obtained.

First of all, the HOG descriptor and the similar Shape Context descriptor consistently outperform the other features independent of the learning algorithm. They are able to achieve around 60% equal error rate. The two Haar-like wavelet-based approaches perform similar, while the Haar features by [1] perform slightly better in combination with AdaBoost and the Haar-like features by [10] show better results when combined with a linear SVM. Shapelets are not performing as well as suggested by the reported FPPWs in the original paper. Only in combination with a linear SVM they do better than the wavelet features.

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. For the wavelet features, linear SVMs are not able to learn a good classifier with limited data. AdaBoost and RBF kernel SVMs are doing better in this case due to their ability to separate data non-linearly. Remarkably, linear SVMs show better performance in combination with Shape Context compared to HOG. This might be an effect of the log-polar sampling for the feature histograms which allows for a better linear separation.

4.2 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 [1] also encode much lower frequencies. Thus, it is worth to further investigate the combination of features. To this end, we conducted several experiments employing early integration with



(a) Feature performance with AdaBoost

(b) Feature performance with linear SVM

(c) Feature performance with RBF SVM

(d) Combination of Haar wavelets [1] and HOG, different classifiers

(e) Combination of Haar wavelets [1] and dense Shape Context, different classifiers

(f) Combination Haar wavelets [1], HOG and dense Shape Context, different classifiers

(g) Bootstrapped single feature detectors of (e) and their combination, linear SVM

(h) Performance of available detector binaries and our implementations; for Haar-like wavelets we improved regularization, for Shapelets we set the regularization properly

Fig. 1. Recall-Precision detector performances

linear SVMs and AdaBoost as classifier. RBF kernel SVMs have not further been employed for computational complexity reasons.

Before stacking feature vectors in an linear SVM classifier, each feature cue was L_2 length normalized to avoid a bias resulting from the features' scale range. In order to keep the comparison fair, we also used the same normalization for AdaBoost. We have combined all possible subsets of HOG, Shape Context and Haar wavelet-based features [1]. Combinations with shapelets have not been tried due to the poor performance of the feature. In the following we will focus on the combinations which yielded the best results.

Additionally we also employed a bootstrapping method, which has shown to improve performance [2, 9]. For this an initial classifier is trained with all available positive training data and random negative samples. Then "hard examples" are collected by scanning the negative training images. The final classifier is then trained on the set of the initial and hard samples.

Our most successful experiments yielded results as depicted in plots 1(d)-(f). For easier comparison the curve of the best performing published[4] binary ([2], bootstrapped SVM classifier) is also shown. Figure 1(d) shows the performance of Haar wavelets [1] and HOG features. Even without boostrapping, the combined features with the AdaBoost classifier almost reach the performance of the published HOG binary. This is due to local optimization of AdaBoost that concentrates on the most discriminative feature in each round. An analysis shows that 67.5% are HOG features while 32.5% are Haar features. The performance of SVM with this feature combination is in between the performance of the two original features. This result can be explained by the global optimization strategy of SVMs, which needs more data to obtain a good fit. Obviously, the bootstrapping method provides more data and consequently performance increases substantially to little above the performance of the bootstrapped HOG features. However, it does not reach the performance of the bootstrapped AdaBoost classifier. As already discussed in section 4.1, AdaBoost is doing better in separating HOG features and Haar wavelets when used individually. Thus, it is only little surprising for the combination to perform also well.

Figure 1(e) shows the combination of dense Shape Context features with Haar wavelets. Without boostrapping AdaBoost and linear SVM perform similar and better than for the single features alone. Adding bootstrapping the SVM classifier again gains a significant improvement. This is due to the same fact we have pointed out in section 4.1. Shape Context features show good linear separability and thus linear SVMs are able to achieve a high classification performance. Again we reviewed the features chosen by AdaBoost. Those were 66.25% Shape Context features and 33.75% Haar wavelet features. We also analyzed the performance of the individual features in a linear SVM when learned with a bootstrapping strategy. Figure 1(g) 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.

[4] `http://pascal.inrialpes.fr/soft/olt`

Finally, figure 1(f) shows results of the combination of HOG, Shape Context and Haar features. For this combination AdaBoost already outperforms the HOG object detector by Dalal [2] even without bootstrapping. The linear SVM classifier again profits from the boostrapping step and performs similarly to the bootstrapped AdaBoost classifier. Interestingly, the performance obtained by the combination of HOG, Shape Context and Haar features is highly similar to the pairwise combinations of Haar features with either HOG or Shape Context. Here the analysis on the chosen features yields the following distribution: 45.25% HOG, 34.0% Shape Context, 20.75% Haar. Additionally adding Haar-like features [10] resulted in almost unchanged detections.

In summary we can state that the combination of different features is successful to improve state-of-the-art people detection performance. We have shown, that a combination of HOG features and Haar wavelets in a AdaBoost classification framework as well as dense Shape Context features with Haar wavelets in a linear SVM framework are able to achieve about 10% better recall with a precision of 80% compared to a single feature HOG detector. Figure 2 shows the improvement on sample images. Similarly to [21] we also observe that a combination of features can achieve better detection performance than the standalone features when trained on the same amount of training data. Additionally, SVMs were able to benefit from a bootstrapping strategy during learning as noted by [9]. While AdaBoost also improves by bootstrapping, the effect is much weaker compared to SVMs.

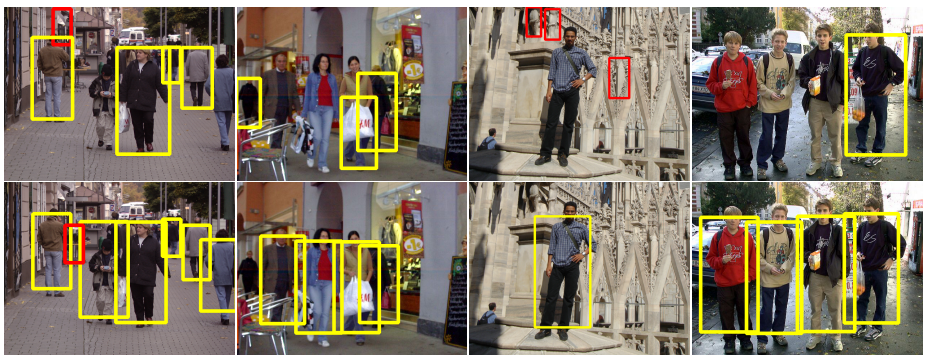


Fig. 2. Sample detections at a precision of 80%. Red bounding boxes denote false detections, while yellow bounding boxes denote true positives. First row shows detection by the publically available HOG detector[2]; second row depicts sample detections for our combination of dense Shape Context with Haar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

4.3 Failure analysis

To complete our experimental evaluation we also conducted a failure case analysis. In particular, we have 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.

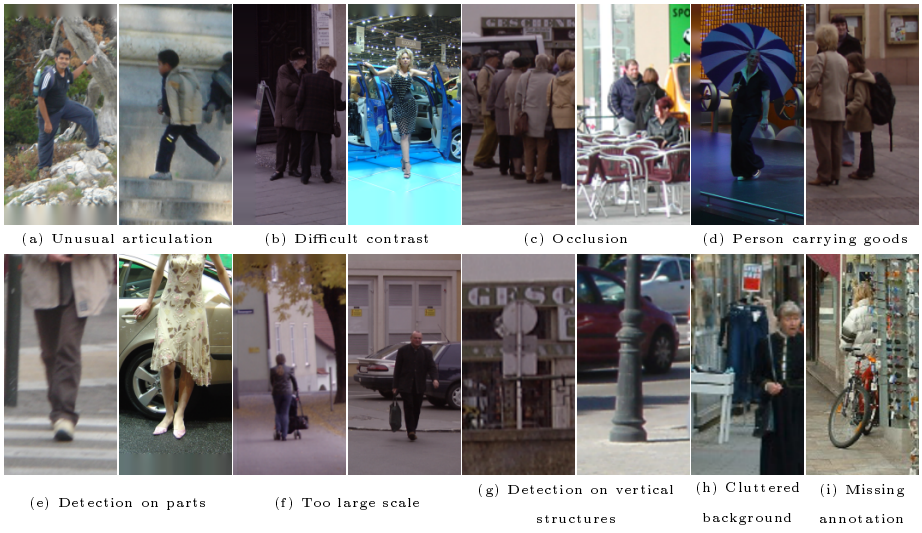


Fig. 3. Missed recall (upper row) and false positive detections (lower row) at equal error rate

5 Conclusion

We have presented a systematic performance evaluation of state-of-the-art features and classification algorithms for people detection. Experiments on the challenging INRIA Person dataset showed that both HOG and dense Shape Context perform better than other features independent of the deployed classifier. Moreover, we have shown that a combination of multiple features is able to improve the performance of the individual detectors considerably. Clearly, there are several open issues which cannot be solved easily with single image classification. Thus, additional motion features and the integration across multiple frames are necessary to further improve performance. Motion for instance can help to resolve false detections due to vertical structures while multiple frame integration is likely to yield better results with cluttered background.

Acknowledgements: We gratefully acknowledge support by the Frankfurt Center for Scientific Computing. This work has been funded, in part, by the EU project CoSy (IST-2002- 004250).

References

1. Papageorgiou, C., Poggio, T.: A trainable system for object detection. IJCV **38**(1) (2000) 15–33
2. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR. (2005) 886–893
3. Seemann, E., Leibe, B., Schiele, B.: Multi-aspect detection of articulated objects. In: CVPR. (2006) 1582–1588
4. Forsyth, D., Fleck, M.: Body plans. In: CVPR. (1997)
5. Wu, B., Nevatia, R.: Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In: ICCV. (2005)
6. Felzenszwalb, P., Huttenlocher, D.: Efficient matching of pictorial structures. In: CVPR. (2000)
7. Mikolajczyk, K., Schmid, C., Zisserman, A.: Human detection based on a probabilistic assembly of robust part detectors. In: ECCV. (2004) 69–81
8. Seemann, E., Leibe, B., Mikolajczyk, K., Schiele, B.: An evaluation of local shape-based features for pedestrian detection. In: BMVC. (2005)
9. Munder, S., Gavrila, D.M.: An experimental study on pedestrian classification. PAMI **28**(11) (2006) 1863–1868
10. Viola, P.A., Jones, M.J.: Robust real-time face detection. IJCV **57**(2) (2004) 137–154
11. Gavrila, D.: Multi-feature hierarchical template matching using distance transforms. In: Proceedings of the International Conference on Pattern Recognition. Volume 1. (1998) 439–444
12. Shashua, A., Gdalyahu, Y., Hayun, G.: Pedestrian detection for driving assistance systems: Single-frame classification and system level performance. In: International Symposium on Intelligent Vehicles. (2004) 1–6
13. Sabzmeydani, P., Mori, G.: Detecting pedestrians by learning shapelet features. In: CVPR. (2007)
14. Tuzel, O., Porikli, F., Meer, P.: Human detection via classification on Riemannian manifolds. In: CVPR. (2007)
15. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. PAMI **24**(4) (2002) 509–522
16. Leibe, B., Seemann, E., Schiele, B.: Pedestrian detection in crowded scenes. In: CVPR. (2005) 878–885
17. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. PAMI **27**(10) (2005) 1615–1630
18. Joachims, T.: Making large–scale SVM learning practical. In Schölkopf, B., Burges, C.J.C., Smola, A.J., eds.: Advances in Kernel Methods — Support Vector Learning, Cambridge, MA, MIT Press (1999) 169–184
19. Comaniciu, D.: An algorithm for data-driven bandwidth selection. PAMI **25**(2) (2003) 281–288
20. Everingham, M., Zisserman, A., Williams, C., van Gool, L.: The PASCAL visual object classes challenge 2006 (VOC2006) results. Technical report
21. Levi, K., Weiss, Y.: Learning object detection from a small number of examples: The importance of good features. In: CVPR. (2004) II: 53–60