Stereo-Based Framework for Pedestrian Detection with Partial Occlusion Handling

Anonymous AVSS submission for Double Blind Review

Paper ID 120

Abstract

The pedestrian detection literature has been recently renewed by the availability of large-scale multisensory datasets, able to capture complementary aspects of the objects of interest, namely, appearance, motion, and depth. In this paper, we exploit this multimodal scenario to propose a new set of composite descriptors dubbed CO^2, COvariances of visual features and CO-occurrences of depth fields. Covariances of visual features allow to integrate at low-level heterogeneous visual cues related to intensity and texture. Co-occurrences of depth fields are brand new descriptors, which use range information for characterizing the global shape of a pedestrian while being also able to identify its occluded parts. This paper illustrates how these descriptors can be instantiated and combined together for improving the detection capabilities, just taking benefit from the proper handling of occlusions. Experimental results show that CO^2, fed into a standard discriminative classification system, allow to set state-of-the-art performances on recent multimodal intensity- and stereo-based pedestrian datasets.

1. Introduction

Pedestrian detection is a very important and complex task for the computer vision community, with also significant implications in practical industrial applications, e.g., the surveillance and automotive sectors, to name a few. It also represents a hard benchmark for many classification theories and a testbed for the usage of novel image features, which should be discriminant and computationally light to cope with real-time requirements. Despite the impressive advances reported in the literature, state-of-the-art detectors seldom satisfy the strict specifications of such real applications and leave ample room for improvement. In particular, a recent survey on pedestrian detection classifiers [3] has revealed the importance of addressing two main problems in order to reach acceptable detection capabilities for real world applications, that is, the reduction of miss-detections at smaller scales and the robustness to partially occluded pedestrians.

Nowadays, the large release of cheap stereo/3D sensors poses new interesting challenges due to the possibility to exploit depth information for detecting people so as to improve the system efficiency. An important lesson from the recent literature is that combining complementary multimodal cues is vital to improve the state-of-the-art performance, and in the last few years some works addressed this issue. In general, earliest systems relied upon a stereo data pre-processing step aimed at restricting the detector usage in regions of well-defined depth, filtering out negative samples for both reducing the number of false positives and lightening the computational cost [6, 10]. More recently, Walk et al. [12] demonstrated good detection performance by using a new stereo-based feature in combination with a variant of HOG [2] adapted to disparity maps. Enzweiler at al. [4] proposed a Mixture of Experts approach, where each expert was trained with a single feature (HOG, LBP) extracted from three different modalities (intensity, depth, optical flow). The result of the detection was provided by fusing the output of each expert, thus implementing a fusion scheme at the classifier level. The same authors proposed a part-based model for human detection using depth information and motion for handling partial occlusions. To the best of our knowledge, this approach and the pioneering work of Wang et al. [13] are the only ones which tried to address the problem of partial occlusion handling for pedestrian detection.

On the same line, our work proposes a simple yet effective way to exploit the stereo information to tackle the problem of partial occlusions in pedestrian detection and classification. The idea is based on some assumptions that are valid in the detection task: i) fusing multiple cues at the raw data level and learning a single classifier on this composite feature is in general convenient; ii) visual features should be extracted locally in the image; iii) since depth information is strongly different from standard visual information it deserves an ad-hoc treatment. The first and second assumptions are witnessed by many recent detection strategies, in which local visual features are embedded in composite descriptors and fed into standard classifiers, showing good performances. For example, Tuzel et al. [11] proposed covariance matrices of basic cues (image derivatives, gradients' magnitudes and orientations) to encode the appearance of local sub-regions. Actually, covariance matrices naturally allow to encapsulate heterogeneous features, also encoding inter-feature correlations in a compact manner. Moreover, they are also robust to varying illumination and invariant to rotations.

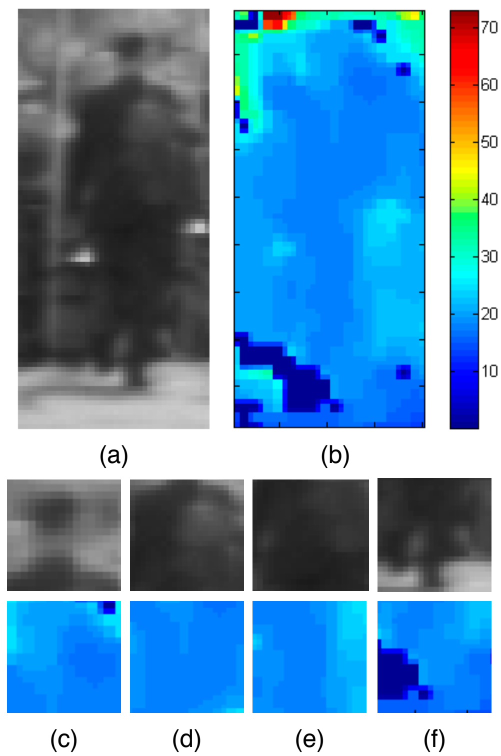

Figure 1. (a) The intensity map showing a pedestrian. (b) Corresponding depth map. (c)-(f) Sub-windows of intensity (up) and depth (down) maps: from left to right, head, torso, hipbone, and legs.

The third assumption comes out after a statistical investigation of many stereo images of pedestrians: as visible in Fig.1a-b, stereo data are able to codify the global human shape as a well-defined silhouette over the background clutter, while they contains far less discriminant information internally (i.e., to characterize the pedestrian). In particular, local patches of depth data are less descriptive than local patches of visual information (see Fig.1c-f), and this discourages a pure local analysis of stereo data. As an example, Fig.1c-f show portions of the human body described by intensity and depth fields, where the depth sub-windows look very similar, whereas intensity sub-windows are characterized by different intensity textural patterns, for example see the head (c) and torso (d). These considerations guided us to design our proposed technique, which is based on local covariance features for describing the visual human aspect, and co-occurrences of depth information for encoding the structure of the body and highlighting possible occlusions. We dubbed the ensemble of features CO^2, i.e., COvariances of visual features and CO-occurrences of depth fields. The features proved to be quite expressive and compact, as well as computationally light, being very fast to compute and oriented to embedded implementations. As for the effectiveness, we fed CO^2 into off-the-shelf classifiers, setting state-of-the-art performances on all the very recently proposed datasets dealing with stereo data, considering occluded and not-occluded situations without tailoring special solutions for one case or the other. In fact, this is a first effort towards the design of detection systems working in real environments, tailored to cope with pedestrians of any structure and shape (i.e., occluded or not), not customized for a single specific pedestrian class, as many of the works published to date [2, 11, 12].

In the rest of paper, we first detail the structure of the new composite feature in Sect. 2. In Sect. 3, the pedestrian detector approach with the explicit management of occlusions is described, and experimental results on the multimodal dataset are reported in Sect. 4, showing the effectiveness of the approach when dealing with both cases of occluded and non-occluded pedestrians. Finally, conclusions are drawn in Sect. 5.

2. The CO^2 feature set

2.1. Covariances of visual features

Let us assume that the image in Fig. 1 (a) contains the object of interest. We define 9 overlapped regions, corresponding to the left, center and right part in horizontal direction, and corresponding to head, torso and legs in vertical direction. More details will be given in Sec. 4.

For every region, we sample a uniform set of overlapped squared patches of size $S = 12 \times 12$ pixels, called *blocks* B. Given the set of $\{N_r\}_{r=1,\dots,9}$ patches, we calculate the corresponding set of covariance matrices denoted as $\{C_i\}_{i=1,\dots,N_r} \in Sym_d^+$ (the space of symmetric positive definite $d \times d$ matrices), where d is the number of features involved to build the matrices. In contrast to [11], we fed the covariance matrix with both gradient- and texture-based features. For each pixel (x, y) inside the patch, we extract $d = 8$ features, that are:

$$[\ x\ y\ |I_h|\ |I_v|\ \sqrt{I_h^2 + I_v^2}\ |I_{hh}|\ |I_{vv}|\ LBP\ ]^T, \quad (1)$$

where I_h, I_{hh}, etc. are grey-level intensity derivatives, and the last term represents the local binary patterns (LBP) feature (8-digit binary number [8]). From the features vector in Eq. (1), a $d \times d$ covariance matrix can be estimated. The space of covariance matrices can be equipped with a Riemannian metric (i.e., Euclidean distances cannot be computed), as in [11], turning it into a Riemannian manifold that we denote as $\mathcal{M}$ [9].

In order to disregard the expensive computation and the complex management of geodesic distances, it is recommended to project the covariance matrices in an Euclidean space [9]. The projection has to be carried out by selecting a projection point, over which the tangent plane of the projection is defined.

The most convenient projection point from the computational perspective is the $d \times d$ identity matrix $I_d \in \mathcal{M}$. More precisely, this projection is called *logarithmic mapping* and it is a standard Riemannian geometry operator which provide a linearized version of $\mathcal{M}$. See [9] for more details. Since C_i is a symmetric matrix, vectorization is applied to extract its upper triangular part and to linearize the content. Hence, the projection and the vectorization translate the covariances into $\{c_i\}_{i=1,\dots,N_r}$ vector descriptors, such that $c_i \in \mathbb{R}^{d \cdot (d+1)/2}$. For every region, $\{c_i\}_{i=1,\dots,N_r}$ are concatenated and organized as a single vector $\mathbf{c}_r$, the multi-feature covariance object descriptor COV.

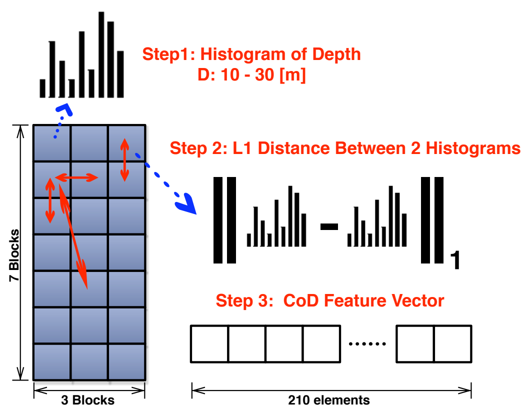

Figure 2. Co-occurrence Depth (CoD) feature vector.

2.2. Co-occurrences of depth

In real-world crowded scenes, pedestrians appear in a continue range of poses. This variability represents a hard issue for classical appearance-based human detection, because the appearance significantly changes in different views. On the contrary, depth information is similar for humans standing in an upright position, and irrespective of the point of view. These assumptions have been supported by the statistical analysis of depth maps extracted from about 50000 un-occluded pedestrians, selected from a public dataset [4]. Statistical evidence showed that the head, shoulders and torso are usually more correlated in terms of depth than legs or arms. Furthermore, the regions around unoccluded pedestrians are also usually correlated, corresponding to a flat background.

The idea underlying CoD features is to encode this depth coherence of the different body parts. Given a depth map, as preliminary operation, we apply a quantization procedure to discretize the depth range. We define a minimum and maximum depth value respectively equal to 10 and 30 meters. This range corresponds to the a priori defined search area in the 3D camera set-up configuration. Out-of-range depth data are saturated on the first and last histogram's bins. The histogram is calculated using a bin resolution of 0.5 meters.

In other words, CoD features are built through pairwise comparisons of histograms of depth, calculated on regions (blocks) inside the detection window. Fig. 2 illustrates the CoD feature building process in details. In short, CoD features are calculated in three steps. The first one is the histogram calculation. Given a depth map, we extract a region D which is equal in size to the detection window of the COV descriptor. We define a regular grid of square regions of size S, of the same size of the $blocks$ defined in 2.1. In each block, $B(m, n)$, we compute a local histogram of depth, $H(m, n)$, where m and n are respectively vertical and horizontal block indexes.

In the second step, we compare every possible pair of block descriptors (histograms). Each comparison is encoded as the distance between the two histograms. Experimental results, not reported here, revealed that L1 distance between histograms performs better than other distances such as L2, Bhattacharyya [1].

In the final step, all the comparisons (k=210) are collected in the CoD descriptor. These feature are employed to estimate the partial occlusions, and drives the detector accordingly, as described in the next section.

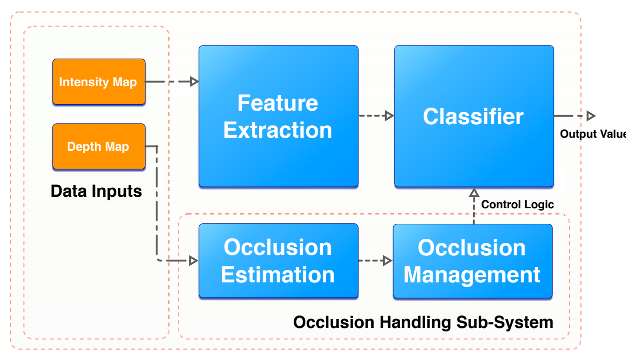

Figure 3. Architecture of the proposed system.

It is worth noting that both COV and CoD are fast to compute and suitable for an embedded implementation. Actually, covariances take advantage of the integral image representation for a rapid calculation (see [11]), and CoD do not require resource-heavy operations such as multiplications, divisions or trigonometric functions. Furthermore, there are no particular issues in terms of concurrent memory accesses because they encode information extracted from local patches (blocks) only.

3. The pedestrian detector

The inability of handling partial occlusions is one of the main limitations of current pedestrian classifiers as demonstrated in a recent survey [3]. In our opinion, the weakness of the current systems is that they rely only on intensity, without taking advantage from different cues such as depth and motion. Few authors have proposed effective solutions to address the partial occlusion problem. One of the most recent approaches [13] uses an heuristic to determine occlusion maps looking at the responses of a monolithic (full-body) SVM classifier. Based on the spatial configuration of the estimated occlusions, they recompute the weights of the linear SVM in order to give more importance to un-occluded regions. Ensweiler and Gavrila [4] detect occlusions by searching discontinuities on depth and optical flow images, showing better performances than [13]. The system is based on a component-based mixture of expert classifiers. They adopt the mean shift clustering algorithm to extract areas of coherent depth and motion. Based on the segmentation result, they determine occlusion-dependent weights for the component-based expert classifiers to focus the combined decision on the visible parts of the pedestrian.

Our system is composed by four modules (see Fig.3). The main core is the $feature\ extraction$ module, that builds the object descriptor, one for each of the 9 regions, as described in Sec.2.1. This module is fed exclusively with the intensity image, and, in turns, its output is fed to the $classifier$ module. The range map is fed into the $occlusion\ handling$ module, which is a pipeline of two stages designed to find the partial occlusions and to drive the $classifier$ module. The first stage is the $occlusion\ estimation$ module, that estimates the occluded regions inside the detection window and produces an occlusion map. The second stage is the $occlusion\ management$ module, that analyzes the occlusion map, and calculates the binary control signals that drive the classifier module. Here follow the details of the $occlusion\ handling$ module and the classifier.

3.1. Occlusion handling

Our assumption is that un-occluded pedestrians have similar CoD features, representing the correlations of depth among body parts. In contrast, partially occluded pedestrians generate different CoD configurations. In an off-line fashion, we compute the CoD statistics of un-occluded pedestrians, i.e., the mean and standard deviation $\mathbf{m}$ an σ of the histogram distances calculated between each block pair. During the test, first we compute CoD on the test image, and than we compare it with our parameters, producing a binary label vector $\mathbf{L}$:

$$\mathbf{L}(k) = \begin{cases} 0 \;\; if \;\; \mathbf{m}(k) - \sigma(k) < CoD(k) < \mathbf{m}(k) + \sigma(k) \\ 1 \;\; else \end{cases} \tag{2}$$

where $k = 1, 2, ...C$ (C is the CoD vector's lenght). If $\mathbf{L}(k) = 1$, we estimate an occlusion between the corresponding pair of blocks. As a consequence, it is possible to estimate how many elements of CoD, that refer to a particular block $B(m,n)$, are occluded ($B(m,n)$ is defined in Sec. 2.2). For each block $B(m,n)$ we count the number of corrupted CoD features W(m,n):

$$W(m,n) = \sum_{k \in B(m,n)} \mathbf{L}(k). \tag{3}$$

The whole matrix W can be reinterpreted as an occlusion map. Fig. 4 is a qualitative evaluation of the proposed occlusion estimation technique. It is evident that occluded blocks have a much higher W score, usually localized on the lowest part of the image.

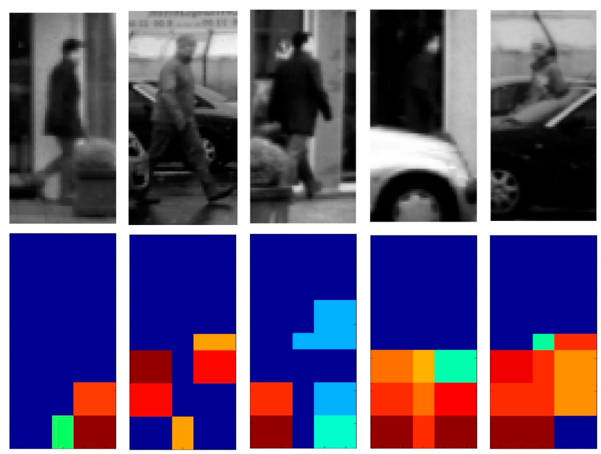

Figure 4. First row: Pedestrian examples from Daimler Multi-Cue, Occluded Pedestrian Classification Benchmark [4]. Second row: occlusion maps W estimated using the CoD features.

We apply a thresholding to W as noise removal filtering. If $W(m,n) < T$, $W(m,n) = 0$, otherwise the block is labeled as occluded. Once the filtered occlusion map has been built, we can generate the control signals in order to activate/deactivate the region classifiers. A region classifier is activated only if *all* the blocks belonging to the region are labelled as not occluded. In practice, a control vector of 9 binary signals $\mathbf{J}$ is generated, one signal for each region classifier.

3.2. The classifier

Given the region descriptors, we learn a set of binary classifiers $\{F_r\}_{r=1,...,9}$, one for each region, adopting a $linearSVM$. When the 9 region classifiers are learnt, we combine their strong responses into a unique classification response as follows:

$$\mathcal{F} = \sum_{r=1}^{9} \frac{\mathbf{w}(r) \cdot \mathbf{J}(r) \cdot F_r(\mathbf{c}_r)}{\sum_{r=1}^{9} \mathbf{w}(r) \cdot \mathbf{J}(r)}, \tag{4}$$

where $\mathbf{w}$ is a vector of region-dependent weights, heuristically estimated during the training phase, and kept fixed for all the experimental phase, and $\mathbf{J}$ is the binary vector sent by the occlusion management unit. The denominator acts as a normalization term, taking into account the different number of region classifiers that can be active. The output of the system is $\mathcal{F}$, the classification confidence value.

4. Experimental results

The recently introduced Daimler Multi-Cue, Occluded Pedestrian Classification Benchmark [4][1] is the only benchmark that incorporates stereo information of occluded and non-occluded pedestrians, representing thus the most valid testbed for multimodal, real-world detectors. It is composed by a single training set of 52112 positives (non-occluded human images) and 32465 samples for the background. The benchmark is equipped with two test sets, one where the pedestrians are partially occluded (11160 samples), the other containing non-occluded pedestrians (25608 samples). Both share the same set of background images (16235 samples). All the images have size 72×24 with a 12-pixel border around each sample, and have been captured from a vehicle-mounted calibrated stereo camera rig in an urban environment.Intensity, depth and optical flow maps are provided for each sample. Dense stereo is computed using the semi-global matching algorithm [7]. At the moment, the best systems benchmarked on the adopted dataset are those of Enzweiler and Gavrila [4] and a modified version of Wang et al. [13], whose detection performance are extracted from [4]. For all these approaches, SVM with linear kernel is adopted as baseline classifier.

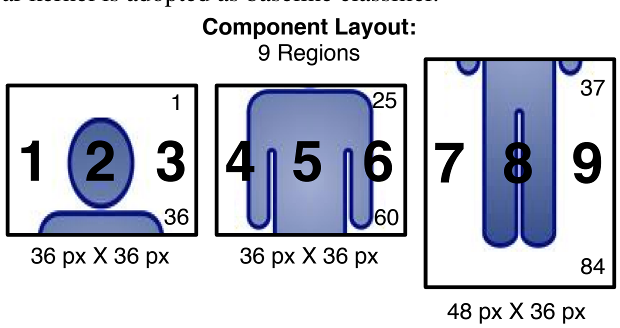

Figure 5. Component layout as used in our experiments.

4.1. Our system

As object model, we pick the central region of 84×36 pixels inside the pedestrian detection window (corresponding to the 72×24 actual region where the pedestrian is enclosed, with a 6-pixel border around each sample). We divide the region in 13×5 square blocks of size 12 pixels, overlapping half their size. A covariance matrix is calculated on each block. The set of covariance matrices is organized in 9 sub-sets, $\mathbf{c}_r$, one for each region (see Sec. 2.1). The component layout is illustrated in Fig. 5.

The CoD feature is calculated on a depth map, computing the histograms in square blocks of size 12 pixels. See Sec. 2.2 for details.

For classification we employ linear SVM, one for each region. The linear SVM classifiers have been trained using Liblinear SVM tool [5] running on off-the-shelf Intel(©) Xeon(©) CPU 2.33 GHz with 8 GB of RAM. To avoid memory overflow issues due to the large pool of positive and negatives training sets, bootstrapping is employed: an initial SVM classifier is trained with the positive images and 10000 background patches randomly selected from the image database. Afterwards, the SVM classifier is used to classify patches of non-pedestrian extracted from the 32465 non pedestrian samples. A set of false-positive are collected and added to the initial negative training set. The process has been repeated until no significant improvement of the performance of the classifier has been noted.

We explore the capabilities of CO^2 in detecting pedestrians, considering the two sets of test data: 1) Occluded pedestrians and 2) Non-occluded pedestrians. The performances are evaluated using the Receiver Operating Characteristic (ROC) curve, that expresses the proportion of false positives against the proportion of true positives. The curve is estimated by varying the confidence threshold τ in the range $[-5, 5]$.

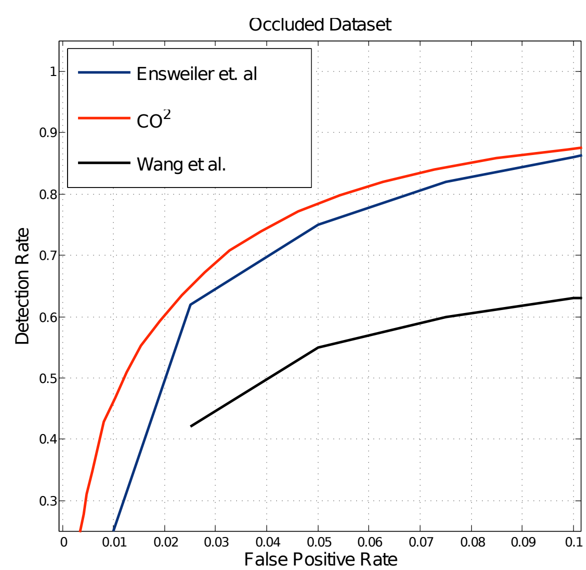

Figure 6. Classification performance on partially occluded testset (best viewed in colors).

[1] See http://www.science.uva.nl/research/isla/downloads/pedestrians/

4.2. Performance on partially occluded test data

In the first experiment we evaluate the performance of our CO^2-based occlusion-handling technique on the occluded dataset. As comparative methods, we consider the approaches proposed by Enzweiler et al. [4] and Wang et al. [13]. Detection performances are reported in Fig. 6. Our approach is superior to both classifiers, demonstrating the effectiveness of our covariance-based framework.The second best performance reported are those of [4] which are based on a mixture of experts of three independent classifiers respectively trained on head, torso and legs. Our technique provides a better performance using a much simpler algorithm to detect occlusion patterns, just by exploiting the CoD feature.

Our detector performs better especially at low false positives rates; specifically, for a false positive rate of 0.01 the detection rate is increased by 15% with respect to [4]. We think this improvement is due to our partial occlusion handling technique and to the well-known capabilities of covariance matrices to encode inter and intra-feature correlation, which is very effective especially with images of medium/low resolution.

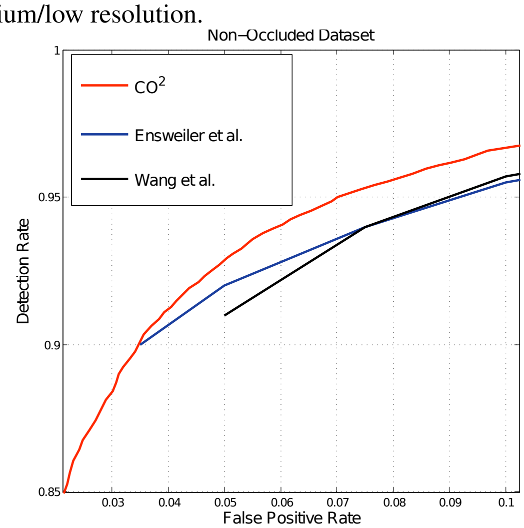

Figure 7. Classification performance on non-occluded testset (best viewed in colors).

4.3. Performance on non-occluded test data

In the second experiment, we evaluate the detection performances of all the algorithms on the non-occluded dataset. Detection performances are reported in Fig. 7. Even in this case, we outperform the two competitors, and this witnesses once again the capability of the covariance to capture robustly the human visual nature.

5. Conclusions

In this paper, we proposed a new set of features suited for pedestrian detection in stereo settings, i.e., when range information is also available. Visual features, pooled together under the form of covariances, characterize human body parts in a robust way. On the other side, depth information is organized as co-occurrence matrices encoding the human shape, so allowing to individuate possible pedestrian occlusions. Such features, fed into a simple classifier, give detection performances on recent multimodal datasets that are definitely superior to all the other competitors in the literature, while they show also the advantage of not being customised for a specific class of pedestrians (e.g., non occluded), as many of the works in the literature to date. In the end, they suggest a interesting recipe for designing real-world commercial detection systems, especially in applications where a pedestrian is immersed in cluttered, real, scenarios.

References

[1] A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc*, 35(99-109):4, 1943.

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. CVPR*, volume 1, pages 886–893. IEEE, 2005.

[3] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *Proc. CVPR*, pages 304–311. IEEE, 2009.

[4] M. Enzweiler, A. Eigenstetter, B. Schiele, and D. Gavrila. Multi-cue pedestrian classification with partial occlusion handling. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 990–997. IEEE, 2010.

[5] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[6] D. Gavrila and S. Munder. Multi-cue pedestrian detection and tracking from a moving vehicle. *International journal of computer vision*, 73(1):41–59, 2007.

[7] H. Hirschmuller. Stereo processing by semiglobal matching and mutual information. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(2):328–341, 2008.

[8] T. Ojala, M. Pietikäinen, and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, 29(1):51–59, 1996.

[9] X. Pennec, P. Fillard, and N. Ayache. A riemannian framework for tensor computing. *International Journal of Computer Vision*, 66(1):41–66, 2006.

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

[11] O. Tuzel, F. Porikli, and P. Meer. Pedestrian detection via classification on riemannian manifolds. *PAMI, IEEE Transactions on*, 30(10):1713–1727, 2008.

[12] S. Walk, K. Schindler, and B. Schiele. Disparity statistics for pedestrian detection: Combining appearance, motion and stereo. *Computer Vision–ECCV 2010*, pages 182–195, 2010.

[13] X. Wang, T. Han, and S. Yan. An hog-lbp human detector with partial occlusion handling. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 32–39. IEEE, 2009.