Multi-Cue Pedestrian Classification With Partial Occlusion Handling

Markus Enzweiler[1] Angela Eigenstetter[2] Bernt Schiele[2,3] Dariu M. Gavrila[4,5]

[1] Image & Pattern Analysis Group, Univ. of Heidelberg, Germany
[2] Computer Science Department, TU Darmstadt, Germany
[3] MPI Informatics, Saarbrücken, Germany
[4] Environment Perception, Group Research, Daimler AG, Ulm, Germany
[5] Intelligent Autonomous Systems Group, Univ. of Amsterdam, The Netherlands

Abstract

This paper presents a novel mixture-of-experts framework for pedestrian classification with partial occlusion handling. The framework involves a set of component-based expert classifiers trained on features derived from intensity, depth and motion. To handle partial occlusion, we compute expert weights that are related to the degree of visibility of the associated component. This degree of visibility is determined by examining occlusion boundaries, i.e. discontinuities in depth and motion. Occlusion-dependent component weights allow to focus the combined decision of the mixture-of-experts classifier on the unoccluded body parts.

In experiments on extensive real-world data sets, with both partially occluded and non-occluded pedestrians, we obtain significant performance boosts over state-of-the-art approaches by up to a factor of four in reduction of false positives at constant detection rates. The dataset is made public for benchmarking purposes.

1. Introduction

The ability to visually recognize pedestrians is key for a number of application domains such as surveillance or intelligent vehicles. Still, it is a particularly difficult problem, as pedestrians vary significantly in pose and appearance and may appear at low resolution. In case of a moving camera in a dynamic environment, ever-changing backgrounds and partial occlusions pose additional problems.

Most of the previous efforts in pedestrian classification assume full visibility of pedestrians in the scene. In a real environment however, significant amounts of partial occlusion occur as pedestrians move in the proximity of other (static or moving) objects. Pedestrian classifiers designed for non-occluded pedestrians do typically not respond well to partially occluded pedestrians. If some body parts of a pedestrian are occluded, the classification results often do not degrade gracefully.

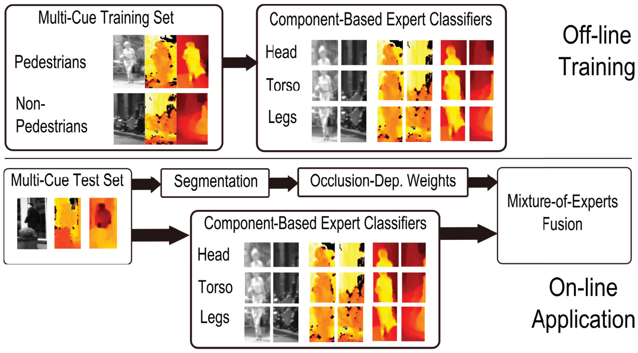


Figure 1. Framework overview. Multi-cue component-based expert classifiers are trained off-line on features derived from intensity, depth and motion. On-line, multi-cue segmentation is applied to determine occlusion-dependent component weights for expert fusion. Data samples are shown in terms of intensity images, dense depth maps and dense optical flow (left to right).

Component-based approaches which represent a pedestrian as an ensemble of parts, cf. [6], can only alleviate this problem to some extent without prior knowledge. The key to successful detection of partially occluded pedestrians is additional information about which body parts are occluded. Classification can then rely on the unoccluded pedestrian components to obtain a robust decision.

In this paper, we present a multi-cue component-based mixture-of-experts framework for pedestrian classification with partial occlusion handling. At the core of our framework is a set of component-based expert classifiers trained on intensity, depth and motion features. Occlusions of individual body parts manifest in local depth- and motion-discontinuities. In the application phase, a segmentation algorithm is applied to extract areas of coherent depth and motion. Based on the segmentation result, we determine

The authors acknowledge the support of the *Studienstiftung des deutschen Volkes* and of the *Bundesministerium für Wirtschaft und Technologie, BMWi* in the context of the *AKTIV-SFR* initiative.

occlusion-dependent weights for our component-based expert classifiers to focus the combined decision on the visible parts of the pedestrian. See Figure 1.

We are not concerned with establishing the best *absolute* pedestrian classification performance given state-of-the-art features and classifiers, cf. [6]. Instead, we explicitly turn to the problem of detecting partially occluded pedestrians and demonstrate the *relative* benefits obtained from the proposed mixture-of-experts framework.

2. Previous Work

Pedestrian classification has become an increasingly popular research topic recently. Most state-of-the art systems, cf. [4, 6, 12], derive a set of features from the available image data and apply pattern classification techniques. Popular features include Haar wavelets [18, 20, 25], adaptive local receptive fields [10, 28] or gradient histograms (HOG) [2, 3, 26, 29, 31]. These features are combined with a variety of classifiers, such as neural networks [10, 28], support vector machines (SVMs) [2, 3, 18, 20, 26, 31] or AdaBoost cascades [16, 24, 25, 30, 31]. Besides operating in the image intensity domain only, some authors have proposed multi-cue approaches combining information from different modalities, e.g. intensity, depth and motion [7, 10, 29]. We do not consider work in the domain of 3D human pose estimation [17], but focus on discriminative 2D approaches for pedestrian classification. See [6] for a current survey.

Recently there have been efforts to break down the complexity of pedestrian appearance into components usually related to body parts [5, 9, 14, 15, 16, 18, 22, 23, 26, 30]. After detecting the individual body parts, detection results are fused using statistical models [15, 16, 30], learning or voting schemes [5, 14, 18, 22] or heuristics [26].

In view of detecting partially occluded pedestrians, component-based classification seems an obvious choice. Yet, only a few approaches have used techniques to infer a measure of (partial) occlusion from the image data [23, 26, 30]. Sigal and Black proposed a technique for articulated 3D body pose estimation which is able to handle self-occlusion of body parts [23]. In our application however, we are not interested in (self-)occlusion handling of articulated 3D pose but focus on partial occlusions observed in 2D images of pedestrians. Particularly relevant to current work are the approaches of Wu and Nevatia [30] and Wang et al. [26]. They explicitly incorporate a model of partial occlusion into their 2D classification framework. However, both approaches make some restrictive assumptions.

The method of Wu and Nevatia, [30], requires a particular camera set-up, where the camera looks down on the ground-plane. Consequently, they assume that the head of a pedestrian in the scene is always visible. They further apply a binary threshold to ignore occluded components in their component-fusion algorithm.

Wang et al., [26], use a monolithic (full-body) HOG/SVM classifier to determine occlusion maps from the responses of the underlying block-wise feature set. Based on the spatial configuration of the recovered occlusion maps, they either apply a full-body classifier or activate part-based classifiers in non-occluded regions or heuristically combine both full-body and part-based classifiers. Since their method depends on the block-wise responses of HOG features combined with linear SVMs, it is unclear how to extend their approach to other popular features or classifiers, cf. [6].

The main contribution of our paper is a mixture-of-experts framework for pedestrian classification with partial occlusion handling. In contrast to [30], we do neither require a particular camera set-up nor assume constant visibility of a certain body part. Our method is independent of the employed feature/classifier combination and the pedestrian component layout, unlike [26]. A secondary contribution involves the integration of intensity, depth and motion cues throughout our approach. Off-line, we train multi-cue component-based expert classifiers involving feature spaces derived from gray-level images, depth maps (dense stereo vision) and motion (dense optical flow), cf. [3, 7, 21, 29]. On-line, we apply multi-cue (depth and motion) mean-shift segmentation to each test sample to recover occlusion-dependent component weights which are used to fuse the component-based expert classifiers to a joint decision, see Figure 1.

3. Pedestrian Classification

Input to our framework is a training set $\mathcal{D}$ of pedestrian (ω_0) and non-pedestrian (ω_1) samples $\mathbf{x}_i \in \mathcal{D}$. Each sample $\mathbf{x}_i = [\mathbf{x}_i^i; \mathbf{x}_i^d; \mathbf{x}_i^f]$ consists of three different modalities, i.e. gray-level image intensity ($\mathbf{x}_i^i$), dense depth information via stereo vision ($\mathbf{x}_i^d$) [11] and dense optical flow ($\mathbf{x}_i^f$) [27]. We treat $\mathbf{x}_i^d$ and $\mathbf{x}_i^f$ similarly to gray-level intensity images $\mathbf{x}_i^i$, in that both depth and motion cues are represented as images, where pixel values encode distance from the camera and magnitude of optical flow vectors between two temporally aligned images, respectively. In case of optical flow, we only consider the horizontal component of flow vectors, to alleviate effects introduced from a moving camera with a significant amount of changes in pitch, e.g. a vehicle-mounted camera. Longitudinal camera motion also induces optical flow. We do not compensate for the ego-motion of the camera, since we are only interested in local differences in flow between a pedestrian and the environment. As a positive side-effect, static pedestrians do not pose a problem in combination with a moving camera. See Figure 5.

3.1. Component-Based Classification

For pedestrian classification, our goal is to determine a class label ω_i for an unseen example $\mathbf{x}_i$. We consider a two-class problem with classes ω_0 (pedestrian) and ω_1 (non-pedestrian). Since $P(\omega_1|\mathbf{x}_i) = 1 - P(\omega_0|\mathbf{x}_i)$, it is sufficient to compute the posterior probability $P(\omega_0|\mathbf{x}_i)$ that an unseen sample $\mathbf{x}_i$ is a pedestrian. The final decision, i.e. ω_i, then results from selecting the object class with the highest posterior probability:

$$\omega_i = \underset{\omega_j}{\operatorname{argmax}} P(\omega_j|\mathbf{x}_i) \tag{1}$$

The posterior probability $P(\omega_0|\mathbf{x}_i)$ is approximated using a component-based mixture-of-experts model. A sample $\mathbf{x}_i$ is composed out of K components which are usually related to body parts. In the mixture-of-experts framework, [13], the final decision results from a weighted linear combination of so-called local expert classifiers which are specialized in a particular area of the feature space. With $\mathbf{F}_k(\mathbf{x}_i)$ representing a local expert classifier for the k-th component of $\mathbf{x}_i$ and $w_k(\mathbf{x}_i)$ denoting its weight, we approximate $P(\omega_0|\mathbf{x}_i)$ using:

$$P(\omega_0|\mathbf{x}_i) \approx \sum_{k=1}^{K} w_k(\mathbf{x}_i)\mathbf{F}_k(\mathbf{x}_i) \tag{2}$$

Note that the weight $w_k(\mathbf{x}_i)$ for each component expert classifier is not a fixed component prior, but depends on the sample $\mathbf{x}_i$ itself. These component weights allow to incorporate a model of partial occlusion into our framework, as shown in Sec. 3.3.

3.2. Multi-Cue Component Expert Classifiers

Given our component-based mixture-of-experts model, cf. Eq. (2), we model the component expert classifiers $\mathbf{F}_k(\mathbf{x}_i)$ in terms of our multi-cue (intensity, depth, flow) dataset. We extend the mixture-of-experts formulation by introducing individual component-based classifiers for each cue:

$$\mathbf{F}_k(\mathbf{x}_i) = \sum_{m \in (i,d,f)} v_k^m \; \mathbf{f}_k^m(\mathbf{x}_i^m) \tag{3}$$

In this formulation, $\mathbf{f}_k^m(\mathbf{x}_i^m)$ denotes a local expert classifier for the k-th component of $\mathbf{x}_i$, which is represented in terms of the m-th cue. As expert classifiers, we use feature-based pattern classifiers which are learned on the training set using data from the corresponding component and cue only. Each component/cue classifier is trained to discriminate between the pedestrian and non-pedestrian class in its local area of the feature space. We consider our framework to be independent from the actual type of feature/classifier combination used, given that the models are complex enough to handle our large and complex pedestrian and non-pedestrian datasets, cf. [6].

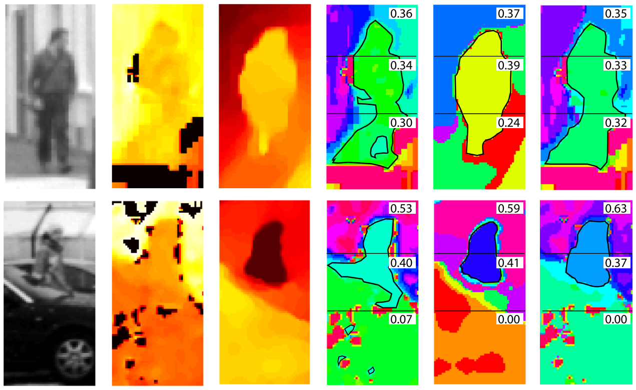


Figure 2. Segmentation results for a non-occluded (first row) and partially occluded pedestrian (second row). From left to right, the columns show: intensity image, stereo image, flow image, segmentation on stereo, segmentation on flow, combined segmentation on stereo and flow. The cluster chosen as pedestrian cluster $\vec{\phi}_{ped}$, cf. Eq. (8), is outlined in black. The computed occlusion-dependent component weights $w_k(\mathbf{x}_i)$, cf. Eq. (9), are also shown.

Weights v_k^m to each component/cue classifier are used to model the contribution of the individual classifiers according to their discriminative power. Some component classifiers have a better absolute performance than others, cf. lower body vs. upper body classifier in [18], similarly for different cues. Hence, we derive v_k^m using a validation dataset, by comparing the absolute classification performances (ROC performance at the same detection rates) of all component/cue classifiers. Weights v_k^m are then set to be proportional to the individual performance levels and normalized to sum to one.

3.3. Occlusion-Dependent Component Weights

Weights $w_k(\mathbf{x}_i)$ for component classifiers were introduced in Sec. 3.1. We derive $w_k(\mathbf{x}_i)$ from each example $\mathbf{x}_i$ to incorporate a measure of occlusion of certain pedestrian components into our model. Expert classifier outputs, related to occluded components, should have a low weight in the combined decision of the expert classifiers, cf. Eq. (2). We propose to extract visibility information from each sample $\mathbf{x}_i$ using the depth (stereo vision) and motion (optical flow) cues. Partially occluded pedestrians, e.g. a walking pedestrian behind a static object, exhibit significant depth and motion discontinuities at the occlusion boundary, as shown in Figures 2 and 5. Visible parts of a pedestrian are assumed to be in approximately the same distance from the camera (pedestrian standing upright on the ground) and move uniformly.

We employ a three-step procedure to derive component weights $w_k(\mathbf{x}_i)$ from an unseen sample $\mathbf{x}_i$: First, we apply a segmentation algorithm, cf. [8], to the dense stereo

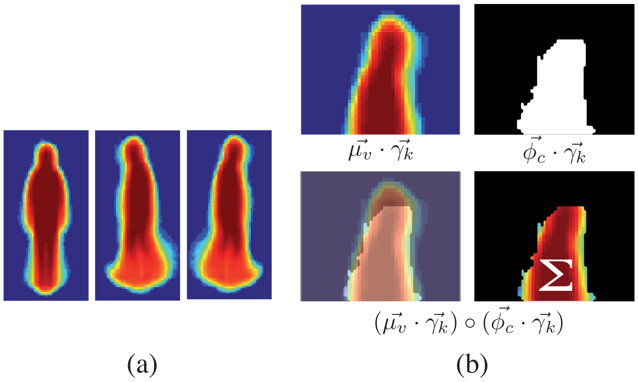

$\vec{\mu}_v \cdot \vec{\gamma}_k$
$\vec{\phi}_c \cdot \vec{\gamma}_k$
Σ
$(\vec{\mu}_v \cdot \vec{\gamma}_k) \circ (\vec{\phi}_c \cdot \vec{\gamma}_k)$
(a)
(b)

Figure 3. (a) Probability masks for front/back, left and right view. The values of the probability masks are in the range of zero (dark blue) to one (dark red). The values specify the probability of a certain pixel to be part of a pedestrian with the corresponding view. (b) Visualization of the correlation-based similarity measure $\Psi_{in}(\vec{\phi}_c, \vec{\gamma}_k, \vec{\mu}_v)$ for the head component, see text.

and optical flow images of $\mathbf{x}_i$. Second, we select the segmented cluster which likely corresponds to the visible area of a pedestrian. For this, a measure of similarity of a cluster to a generic model of pedestrian geometry in terms of pedestrian shape, size and location is utilized. Third, we estimate the degree of visibility of each component given the selected cluster.

For segmentation, we chose the mean-shift algorithm, [1], out of many possible choices. As shown in [8], mean-shift provides a good balance between segmentation accuracy and processing efficiency. The result of the mean-shift segmentation is a set of C clusters ϕ_c with $c = 1, \dots, C$, as shown in Figure 2. The actual number of clusters C is optimized during mean-shift itself [1]. Note that we evaluate both single-cue segmentation using depth or motion and simultaneous multi-cue segmentation using both cues in our experiments, as shown in Sec. 4.

Let $\vec{\phi}_c$ and $\vec{\gamma}_k$ denote binary vectors defining the membership of pixel-locations of the sample $\mathbf{x}_i$ to the c-th cluster ϕ_c and k-th component γ_k, respectively. Note that $\vec{\phi}_c$ results from the segmentation algorithm, whereas $\vec{\gamma}_k$ is given by the geometric component layout. Further, we utilize a two-dimensional probability mass function $\mu_v(\mathbf{p})$ which represents the probability that a given pixel $\mathbf{p} \in \mathbf{x}_i$ corresponds to a pedestrian, solely based on its location within $\mathbf{x}_i$. $\mu_v(\mathbf{p})$ is obtained from the normalized superposition of a set of S aligned binary pedestrian foreground masks $m_s(\mathbf{p})$, obtained from manually labeled pedestrian shapes:

$$\mu_v(\mathbf{p}) \propto \sum_{s=1}^{S} m_s(\mathbf{p}), \quad 0 \leq \mu_v(\mathbf{p}) \leq 1 \tag{4}$$

To increase specificity, we use view-dependent probability masks $\mu_v(\mathbf{p})$ in terms of separate masks for front/back, left and right views. Those probability masks represent a view-dependent model of pedestrian geometry in terms of shape, size and location. See Figure 3(a). Again, a vectorized representation of μ_v is denoted as $\vec{\mu}_v$.

To select the segmented cluster, which corresponds to the visible area of a pedestrian, we utilize a correlation-based similarity measure Ψ, as defined in Eq. (5). Our similarity measure employs the cluster information and the probability masks to assess the likelihood that a cluster ϕ_c corresponds to the visible parts of a pedestrian. We model Ψ as the sum of two terms, Ψ_{in} and Ψ_{out}:

$$\Psi(\vec{\phi}_c, \vec{\gamma}_k, \vec{\mu}_v) = \Psi_{in}(\vec{\phi}_c, \vec{\gamma}_k, \vec{\mu}_v) + \Psi_{out}(\vec{\phi}_c, \vec{\gamma}_k, \vec{\mu}_v) \tag{5}$$

The first measure $\Psi_{in}(\vec{\phi}_c, \vec{\gamma}_k, \vec{\mu}_v)$ is designed to evaluate how well a cluster ϕ_c matches typical pedestrian geometry, represented by a view-dependent pedestrian probability mask μ_v, in a certain component γ_k. To compute $\Psi_{in}(\vec{\phi}_c, \vec{\gamma}_k, \vec{\mu}_v)$, we correlate the cluster $\vec{\phi}_c$ and the probability mask $\vec{\mu}_v$ within the component given by $\vec{\gamma}_k$ and normalize:

$$\Psi_{in}(\vec{\phi}_c, \vec{\gamma}_k, \vec{\mu}_v) = \frac{(\vec{\mu}_v \cdot \vec{\gamma}_k) \circ (\vec{\phi}_c \cdot \vec{\gamma}_k)}{\vec{\mu}_v \circ \vec{\gamma}_k} \tag{6}$$

Here, $\cdot$ denotes point-wise multiplication of vectors, while $\circ$ denotes a dot product. Note that the main purpose of $\vec{\gamma}_k$ in this formulation is to restrict computation to a local body component γ_k. See Figure 3(b).

The second measure $\Psi_{out}(\vec{\phi}_c, \vec{\gamma}_k, \vec{\mu}_v)$ relates to the specificity of the cluster ϕ_c. The idea is to penalize clusters which extend too far beyond a typical pedestrian shape. For that we perform similar correlation using an “inverse” probability mask $\vec{\nu}_v = 1 - \vec{\mu}_v$:

$$\Psi_{out}(\vec{\phi}_c, \vec{\gamma}_k, \vec{\mu}_v) = 1 - \frac{(\vec{\nu}_v \cdot \vec{\gamma}_k) \circ (\vec{\phi}_c \cdot \vec{\gamma}_k)}{\vec{\nu}_v \circ \vec{\gamma}_k} \tag{7}$$

The cluster similarity measure $\Psi(\vec{\phi}_c, \vec{\gamma}_k, \vec{\mu}_v)$, see Eq. (5), is computed per cluster, component and view-dependent probability mask. To choose the cluster $\vec{\phi}_{ped}$ which most likely corresponds to visible parts of the pedestrian, we apply a maximum operation over components and views:

$$\vec{\phi}_{ped} = \underset{\vec{\phi}_c}{\operatorname{argmax}} \left(\max_{\vec{\gamma}_k \vec{\mu}_v} \left(\Psi(\vec{\phi}_c, \vec{\gamma}_k, \vec{\mu}_v) \right) \right) \tag{8}$$

From our experiments we observed that the visible parts of a pedestrian do not significantly disintegrate in the mean-shift segmentation results, see Figure 2. Hence, we only consider single clusters ϕ_c and pairs of clusters merged together as possible candidates.

Once the cluster $\vec{\phi}_{ped}$, corresponding to visible parts of the pedestrian, is selected, the degree of visibility of each

	Pedestrians (labeled)	Pedestrians (jittered)	Non-Pedestrians
Train Set	6514	52112	32465
Partially Occluded Test Set	620	11160	16235
Non-Occluded Test Set	3201	25608	16235

Table 1. Training and test set statistics.

component is approximated. For each component $\vec{\gamma}_k$, we chose to relate the spatial extent of $\vec{\phi}_{ped}$ against clusters corresponding to occluding objects. The set of all clusters $\vec{\phi}_j$, which are possible occluders of $\vec{\phi}_{ped}$, is denoted by Υ. Possible occluders of $\vec{\phi}_{ped}$ are clusters which are closer to the camera than $\vec{\phi}_{ped}$. If depth information is not available for segmentation, all clusters are regarded as possible occluders. With $n(\vec{v})$ denoting the number of non-zero elements in an arbitrary vector $\vec{v}$, occlusion-dependent component weights $w_k(\mathbf{x}_i)$, with $\sum_k w_k(\mathbf{x}_i) = 1$, are then given by:

$$w_k(\mathbf{x}_i) \propto \frac{n(\vec{\phi}_{ped} \cdot \vec{\gamma}_k)}{\sum_{\vec{\phi}_j \in \Upsilon} \left(n(\vec{\phi}_j \cdot \vec{\gamma}_k) \right) + n(\vec{\phi}_{ped} \cdot \vec{\gamma}_k)} \tag{9}$$

See Figure 2 for a visualization of the cluster $\vec{\phi}_{ped}$, corresponding to visible parts of the pedestrian, and the recovered occlusion-dependent component weights $w_k(\mathbf{x}_i)$.

4. Experiments

4.1. Experimental Setup

The proposed multi-cue component-based mixture-of-experts framework was tested in experiments on pedestrian classification. Since we require partially occluded multi-cue (intensity, dense stereo, dense optical flow) training and test samples, we cannot use established datasets for benchmarking, e.g. [2, 4, 6, 19]. To our knowledge, our dataset is the first to comprise "real" partially occluded pedestrians in the field of pedestrian classification. Wang et al. only simulated partial occlusion by synthetically adding other objects as occluders to pedestrian samples [26]. Our multi-cue/occlusion dataset is made publicly available to non-commercial entities for research purposes.[1]

We chose to evaluate our approach in a pedestrian classification setting, where we assume that initial pedestrian location hypotheses already exist, e.g. using methods described in [6, 7, 10, 22] or non-vision sensors. In our experiments, we focus on the central part of a pedestrian detection system, the classifier, to eliminate auxiliary effects arising from various detector parameters such as grid granularity, non-maximum suppression, scene and processing constraints or tracking, cf. [6].

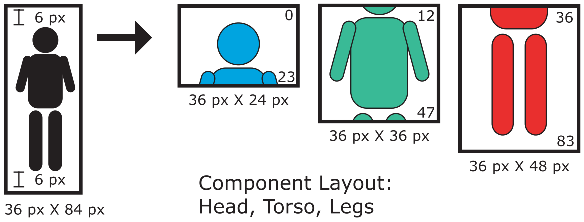


Figure 4. Component layout as used in our experiments. We employ three overlapping components, corresponding to head, torso and leg regions, see text.

Our training and test samples consist of manually labeled pedestrian and non-pedestrian bounding boxes in images captured from a vehicle-mounted calibrated stereo camera rig in an urban environment. For each manually labeled pedestrian, we created additional samples by geometric jittering. Non-pedestrian samples were the result of a shape detection pre-processing step with relaxed threshold setting, i.e. containing a bias towards more "difficult" patterns. Dense stereo is computed using the semi-global matching algorithm [11]. To compute dense optical flow, we use the method of [27].

Training and test samples have a resolution of 36 × 84 pixels with a 6-pixel border around the pedestrians. In our experiments, we use $K = 3$ components γ_k, corresponding to head/shoulder (36×24 pixels), torso (36×36 pixels) and leg (36×48 pixels) regions, see Figure 4. Note that our components vertically overlap by 12 pixels, i.e. each component has a 6-pixel border around the associated body part. In preliminary experiments, we determined this overlap to improve performance.

Regarding features for the component/cue expert classifiers $\mathbf{f}_k^m$, see Eq. (3), we chose histograms of oriented gradients (HOG) out of many possible feature sets, cf. [2, 4, 6, 19]. The motivation for this choice is two-fold: First, HOG features are still among the best performing feature sets available; second, we compare our framework to the approach of Wang et al. [26] which explicitly requires and operates on the block-wise structure of HOG features. We compute histograms of oriented gradients with 12 orientation bins and 6 × 6 pixel cells, accumulated to overlapping 12 × 12 pixel blocks with a spatial shift of 6 pixels. For classification, we employ linear support vector machines (SVMs). Note that the same HOG feature set is extracted from intensity, dense stereo and dense flow images, cf. [3, 21]. In our implementation of [26], we use the occlusion handling of Wang et al. together with the same component layout (head, torso, legs), features (HOG) and classifiers (linear SVMs) as in our approach, but only for the intensity cue.

To train the component classifiers, only non-occluded pedestrians (and non-pedestrian samples) are used. For testing, we evaluate performance using two different test sets:

[1] See http://www.science.uva.nl/research/isla/downloads/pedestrians/index.html or contact the last author.

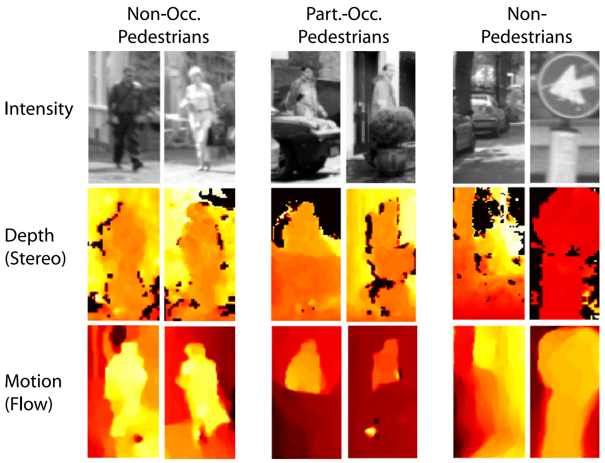


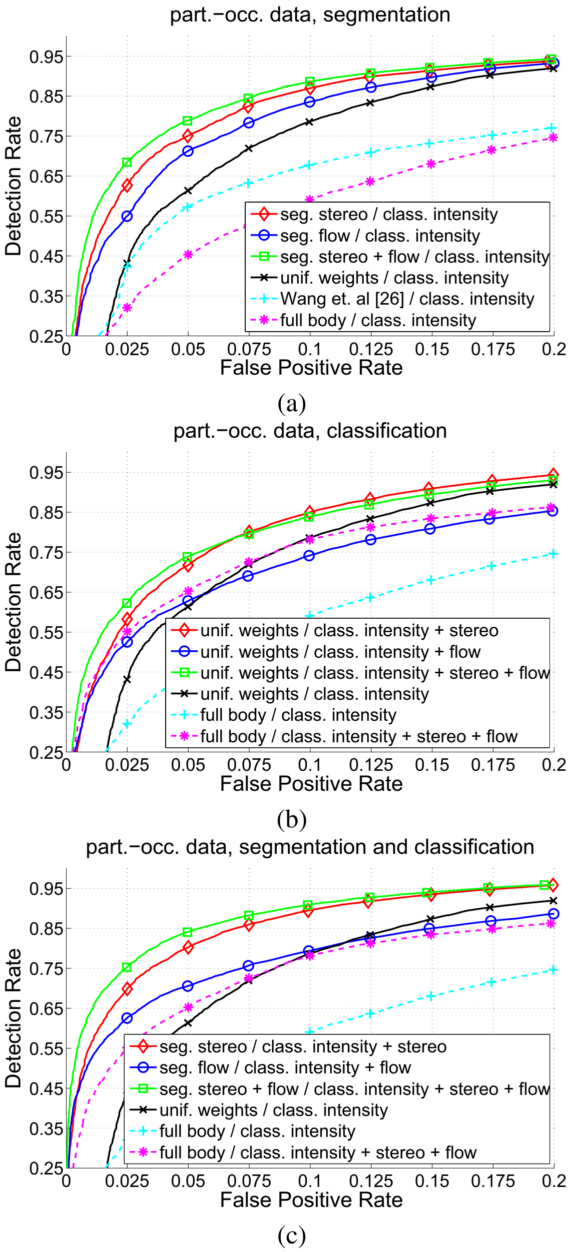
Figure 5. Non-occluded pedestrians, partially occluded pedestrians and non-pedestrians samples in our data. In depth (stereo) images, darker colors denote closer distances. Note that the background (large depth values) has been faded out for visibility. Optical flow images depict the magnitude of the horizontal component of flow vectors, with lighter colors indicating stronger motion.

one involving non-occluded pedestrians and one consisting of partially occluded pedestrians. The non-pedestrian samples are the same for both test sets. See Table 1 and Figure 5 for an overview of the dataset.

4.2. Performance on Partially Occluded Test Data

Partial Occlusion Handling In our first experiment, we evaluate the effect of different models of partial occlusion handling. We do not consider multi-cue classifiers yet. All expert component classifiers are trained on intensity images only. As baseline classifiers, we evaluate the full-body HOG approach of [2] (we use the code provided by the original authors) and the approach of [26], which uses an occlusion model based on the block-wise response of a full-body HOG classifier to activate part-based classifiers in areas corresponding to non-occluded pedestrian parts. Our framework is evaluated using four different strategies to compute occlusion-dependent component weights $w_k(\mathbf{x}_i)$ for $\mathbf{x}_i$, as defined in Sec. 3.3: We consider weights resulting from mean-shift segmentation using depth only, flow only and a combination of both depth and flow. Additionally, we consider uniform weights $w_k(\mathbf{x}_i)$, i.e. no segmentation. Note that weights v_k^m, as given in Eq. (3), are still in place. Results in terms of ROC performance are given in Figure 6(a).

All component-based approaches outperform the full-body HOG classifier (magenta *). The approach of Wang et al. [26] (cyan +) significantly improves performance over the full-body HOG classifier by a factor of two (reduction in false positives at constant detection rates). All variants of our framework in turn outperform the method of Wang et al. [26], with segmentation on combined depth and flow (green □) performing best. Compared to the use of uniform weights $w_k(\mathbf{x}_i)$ (black ×), the addition of multi-cue segmentation to compute component weights (green □) improves performance by approximately a factor of two.



(a)

(b)

(c)

Figure 6. Classification performance on partially occluded testset. (a) Evaluation of partial occlusion handling strategies. (b) Multi-cue classification in comparison to intensity-only classification. (c) Combined multi-cue partial occlusion handling and classification.

Multi-Cue Classification In our second experiment, we evaluate the performance of multi-cue component classifiers, as presented in Sec. 3.2, compared to intensity-only component classifiers. Uniform component weights $w_k(\mathbf{x}_i)$, i.e. no segmentation, were used throughout all approaches. Results are given in Figure 6(b) (solid lines). As baseline classifiers, we use a full-body intensity-only HOG classifier and a multi-cue full-body HOG classifier trained on intensity, stereo and flow data (dashed lines). Multi-cue classification significantly improves performance both for the full-body and for the component-based approach. The best performance (particularly at low false positive rates) is reached by the component-based approach involving intensity, stereo and flow (green □). The performance improvement over a corresponding component-based classifier using intensity-only (black ×) is up to a factor of two reduction in false positives.

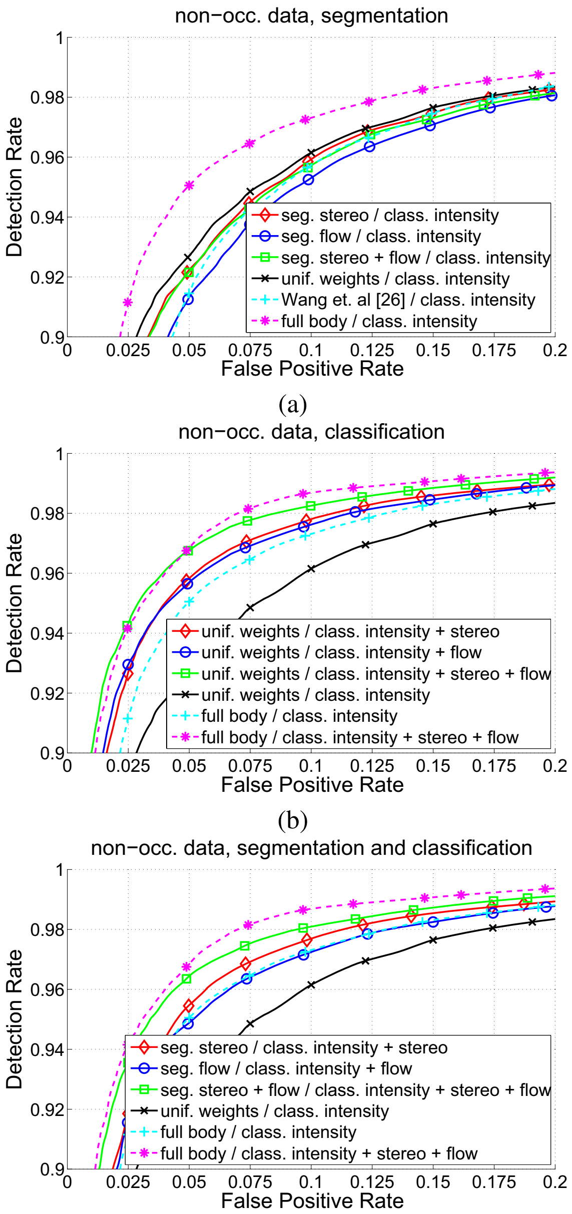
Multi-Cue Classification with Partial Occlusion Handling In the next experiment, we evaluate the proposed multi-cue framework involving occlusion-dependent component weights derived from mean-shift segmentation combined with multi-cue classification. Instead of presenting results for all possible combinations of cues for segmentation and classification, we chose to use the same cues for both segmentation and classification. We did evaluate all cue-combinations and found no better performing combination. Similar to the previous experiment, the baseline is given by full-body classifiers (cyan + and magenta *), as well as a component-based intensity-only classifier using uniform weights (black ×). See Figure 6(c).

The best performing system variant is the proposed component-based mixture-of-experts architecture using stereo and optical flow concurrently to determine occlusion-dependent weights $w_k(\mathbf{x}_i)$ and for multi-cue classification (green □). Compared to a corresponding multi-cue full-body classifier (magenta *), the performance boost is approximately a factor of four. A similar performance differences exists between our best approach (green □) and a component-based intensity-only classifier using uniform component weights (black ×).

4.3. Performance on Non-Occluded Test Data

After demonstrating significant performance boosts on partially occluded test data, we evaluate the performance of the proposed approach using non-occluded pedestrians (and non-pedestrians) as test set. Similar to our previous experiments, we evaluate the effect of partial occlusion handling independently from the use of multiple cues for segmentation and classification.

Figure 7(a) shows the effect of different models of partial occlusion handling combined with intensity-only component-based classifiers. The full-body HOG classifier (magenta *), as well as the approach of Wang et al. [26] (cyan +), serve as baselines. The best performance is reached by the full-body HOG classifier. All component-based approaches perform slightly worse. Of all component-based approaches, uniform component weights $w_k(\mathbf{x}_i)$, i.e. no occlusion handling, yields the best performance by a small margin. This is not surprising, since all components are visible to the same extent. On non-occluded test samples, our best approach with occlusion handling (green □) gives the same performance as Wang et al. [26] (cyan +).



(a)

(b)

(c)

Figure 7. Classification performance on non-occluded testset. (a) Evaluation of partial occlusion handling strategies. (b) Multi-cue classification in comparison to intensity-only classification. (c) Combined multi-cue partial occlusion handling and classification.

Multi-cue classification, as shown in Figure 7(b), yields similar performance boosts compared to intensity-only classification as observed for the test on partially occluded data, cf. Sec. 4.2. Figure 7(c) depicts results of our integrated multi-cue mixture-of-experts framework with partial occlusion handling. Compared to a full-body classifier involving intensity, stereo and flow (magenta *), our best performing mixture-of-experts approach gives only slightly worse performance, particularly at low false positive rates. In relation to intensity-only full-body classification (cyan +), i.e. the approach of [2], our multi-cue framework improves performance by up to a factor of two.

5. Conclusion

This paper presented a multi-cue mixture-of-experts framework for component-based pedestrian classification with partial occlusion handling. For the partially occluded dataset, we obtained in the case of depth- and motion-based occlusion handling an improvement of more than a factor of two versus the baseline (component-based, no occlusion handling) and state-of-the-art [26]. We obtained in the case of multi-cue (intensity, depth, motion) classification an additional improvement of a factor of two versus the baseline (intensity only). The full-body classifiers performed worse than the beforementioned baselines. For the non-occluded dataset, occlusion handling does not appreciably deteriorate results, while multi-cue classification improves performance by a factor of two. We take the results as evidence for the strength of our approach.

References

[1] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE PAMI*, 24(5):603–619, 2002.

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

[3] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *Proc. ECCV*, pages 428–441, 2006.

[4] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. *Proc. CVPR*, 2009.

[5] P. Dollar et al. Multiple component learning for object detection. *Proc. ECCV*, pages 211–224, 2008.

[6] M. Enzweiler and D. M. Gavrila. Monocular pedestrian detection: Survey and experiments. *IEEE PAMI*, 31(12):2179–2195, 2009.

[7] A. Ess, B. Leibe, and L. van Gool. Depth and appearance for mobile scene analysis. In *Proc. ICCV*, 2007.

[8] F. J. Estrada and A. D. Jepson. Benchmarking image segmentation algorithms. *IJCV*, 85(2):167–181, 2009.

[9] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61(1):55–79, 2005.

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

[11] H. Hirschmüller. Stereo processing by semi-global matching and mutual information. *IEEE PAMI*, 30(2):328–341, 2008.

[12] M. Hussein, F. Porikli, and L. Davis. A comprehensive evaluation framework and a comparative study for human detectors. *IEEE ITS*, 10(3):417–427, 2009.

[13] R. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

[14] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In *Proc. CVPR*, pages 878–885, 2005.

[15] A. S. Micilotta, E. J. Ong, and R. Bowden. Detection and tracking of humans by probabilistic body part assembly. In *Proc. BMVC*, pages 429–438, 2005.

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

[17] T. B. Moeslund and E. Granum. A survey of advances in vision-based human motion capture and analysis. *CVIU*, 103(2-3):90–126, 2006.

[18] A. Mohan, C. Papageorgiou, and T. Poggio. Example-based object detection in images by components. *IEEE PAMI*, 23(4):349–361, 2001.

[19] S. Munder and D. M. Gavrila. An experimental study on pedestrian classification. *IEEE PAMI*, 28(11):1863–1868, 2006.

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

[21] M. Rohrbach, M. Enzweiler, and D. M. Gavrila. High-level fusion of depth and intensity for pedestrian classification. In *Proc. DAGM*, pages 101–110, 2009.

[22] E. Seemann, M. Fritz, and B. Schiele. Towards robust pedestrian detection in crowded image sequences. In *Proc. CVPR*, 2007.

[23] L. Sigal and M. J. Black. Measure locally, reason globally: Occlusion-sensitive articulated pose estimation. *Proc. CVPR*, pages 2041–2048, 2006.

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

[25] P. Viola, M. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. *IJCV*, 63(2):153–161, 2005.

[26] X. Wang, T. Han, and S. Yan. A HOG-LBP human detector with partial occlusion handling. *Proc. ICCV*, 2009.

[27] A. Wedel, D. Cremers, T. Pock, and H. Bischof. Structure- and motion-adaptive regularization for high accuracy optic flow. *Proc. ICCV*, 2009.

[28] C. Wöhler and J. K. Anlauf. A time delay neural network algorithm for estimating image-pattern shape and motion. *IVC*, 17:281–294, 1999.

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

[30] B. Wu and R. Nevatia. Detection and tracking of multiple, partially occluded humans by Bayesian combination of edgelet based part detectors. *IJCV*, 75(2):247 – 266, 2007.

[31] Q. Zhu, S. Avidan, M. Ye, and K.-T. Cheng. Fast human detection using a cascade of histograms of oriented gradients. In *Proc. CVPR*, pages 1491–1498, 2006.