Integrated Pedestrian Classification and Orientation Estimation

Markus Enzweiler[1] Dariu M. Gavrila[2,3]

[1] Image & Pattern Analysis Group, Univ. of Heidelberg, Germany
[2] Environment Perception, Group Research, Daimler AG, Ulm, Germany
[3] Intelligent Autonomous Systems Group, Univ. of Amsterdam, The Netherlands

Abstract

This paper presents a novel approach to single-frame pedestrian classification and orientation estimation. Unlike previous work which addressed classification and orientation separately with different models, our method involves a probabilistic framework to approach both in a unified fashion. We address both problems in terms of a set of view-related models which couple discriminative expert classifiers with sample-dependent priors, facilitating easy integration of other cues (e.g. motion, shape) in a Bayesian fashion. This mixture-of-experts formulation approximates the probability density of pedestrian orientation and scales-up to the use of multiple cameras.

Experiments on large real-world data show a significant performance improvement in both pedestrian classification and orientation estimation of up to 50%, compared to state-of-the-art, using identical data and evaluation techniques.

1. Introduction

Pedestrian recognition is a key problem for a number of application domains, e.g. surveillance, robotics and intelligent vehicles. Yet, it is a difficult task from machine vision perspective because of the wide range of possible pedestrian appearance, due to changing articulated pose, clothing, lighting, and background. The lack of explicit models has spawned the use of implicit representations, based on pattern classification techniques [18].

Beyond detecting a pedestrian in the scene, many application areas benefit from knowledge of body orientation of a pedestrian. In the domain of intelligent vehicles [13], known pedestrian orientation can enhance path prediction, to improve risk assessment. Other applications include perceptual interfaces [32], where body orientation can be used as a proxy for interaction.

Orientation could be inferred by trajectory information (tracking). Yet, trajectory-based techniques fail in case of pedestrians which are static or just about to move. Tracking approaches also require a certain amount of time to converge to a robust estimate. Quick adaptation to sudden changes in movement is often problematic. Particularly in the intelligent vehicle application, time is precious and fast reaction is necessary.

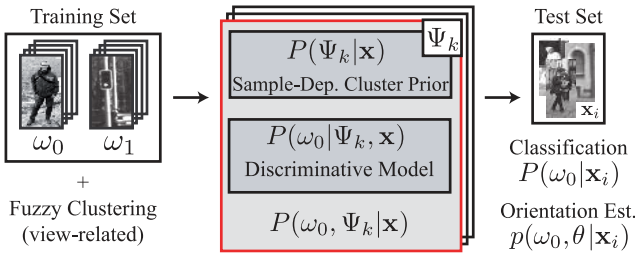


Figure 1. Framework overview. K view-related models specific to fuzzy clusters Ψ_k are used for pedestrian classification and orientation estimation. The models capture sample-dependent cluster priors and discriminative experts which are learned from pedestrian (class ω_0) and non-pedestrian (class ω_1) samples $\mathbf{x}$.

As a way out, methods to infer pedestrian orientation from a single image have been proposed [12, 26, 31]. Such approaches augment an existing pedestrian detector by a post-processing step to estimate orientation. Single-frame orientation estimation allows to recover pedestrian heading without integration over time; static pedestrians do not pose a problem.

In this paper, we present a novel integrated method for single-frame pedestrian classification and orientation estimation. Both problems are treated using the same underlying probabilistic framework, in terms of a set of view-related models which couple discriminative expert models with sample-dependent priors. Pedestrian classification involves a maximum-a-posteriori decision between the pedestrian class and non-pedestrian class. Orientation estimates are inferred by means of approximating the probability density of pedestrian body orientation. See Figure 1.

The general approach is independent from the actual type of discriminative models used and can be extended to other object classes. This paper is not concerned with establishing the best *absolute* performance given various state-of-the-art discriminative models [7, 9, 11, 14, 17, 33, 35, 36, 37, 39]. Rather, our aim is to demonstrate the *relative* performance gain resulting from the proposed integrated approach, exemplified using two state-of-the-art feature sets and classifiers in our experiments (see Sec. 4).

2. Previous Work

There is extensive literature on image-based classification of pedestrians. See [11, 14] for recent surveys. We do not consider 3D human pose estimation techniques [24], but focus on single camera 2D approaches suited for medium resolution pedestrian data (i.e. pedestrian height smaller than 80 pixels).

Roughly, approaches to pedestrian classification can be distinguished by the type of models involved. Generative approaches model the appearance of the pedestrian class by approximating its class-conditional density. Combined with class priors, posterior probabilities for classification can be inferred using a Bayesian approach.

Discriminative approaches model the Bayesian decision by learning discriminative functions (decision boundaries) to separate object classes within a feature space. While generative models implicitly establish a feature-space specific to the object class, discriminative models combine a more generic lower-level feature set with pattern classification methods, see [34]. In on-line application, discriminative models are often faster, since they do not require estimation of a large set of model parameters (model fitting) from the data. Yet, generative models can handle partially labeled data and allow to generate virtual samples [10].

For generative models, shape cues are widely used because of their robustness to variations in pedestrian appearance due to lighting or clothing [6, 10, 15]. Other work considered joint shape and texture models [6, 10, 19], which require data normalization methods and involve a significantly increased feature space.

Discriminative models comprise a combination of feature extraction and classification. Non-adaptive Haar wavelet features have been popularized by [27] and used by many others [25, 31, 35]. The particular structure of local texture features has also been optimized, in terms of local receptive field features [36] which adapt to the underlying data during training. Other texture-based features are codebook patches, extracted around interest points in the image [1, 20, 29] and linked via geometric relations.

Gradient-based features have focused on discontinuities in image brightness. Normalized local histograms of oriented gradients have found wide use in both sparse (SIFT) [22] and dense representations (HOG) [7, 12, 38, 39]. Spatial variation and correlation of gradients have been encoded using covariance descriptors enhancing robustness towards brightness variations [33]. Yet others have proposed local shape filters exploiting characteristic patterns in the spatial configuration of salient edges [23, 37].

Regarding classifier architectures, support vector machines (SVM) have become increasingly popular in the domain of pedestrian classification, in both linear [7, 8, 26, 31, 38, 39] and non-linear variants [25, 27]. However, performance boosts resulting from the non-linear model are paid for with a significant increase in computational costs and memory. Other popular classifiers include neural networks [13, 18, 36] and AdaBoost cascades [23, 33, 35, 37, 38, 39].

To improve pedestrian classification performance, several approaches have attempted to break-down the complexity of the problem into sub-parts. Besides component-based approaches involving a representation of pedestrians as an ensemble of parts [23, 25, 37], mixture-of-experts strategies are particularly relevant to current work. Here, local pose-specific clusters are established, followed by the training of specialized classifiers for each subspace [13, 26, 30, 31, 37, 38]. The final decision of the classifier ensemble involves maximum-selection [26, 37], majority voting [31], trajectory-based data association [38] or shape-based selection [13]. Approaches performing object detection/classification in multiple cameras at different viewpoints are also relevant to current work [3, 16].

Besides work in the domain of 3D human pose estimation [24], few approaches have tried to recover an estimate of pedestrian orientation based on 2D lower-resolution images [12, 26, 31]. Existing approaches re-used popular features, i.e. Haar wavelets [31] or gradient histograms [12], and applied them in a different classification scheme. While pedestrian classification usually involves a two-class model (pedestrian vs. non-pedestrian), [12, 26, 31] have not used non-pedestrian training samples for orientation estimation. Instead, *one vs. one* [12] and *one vs. rest* [26, 31] multi-class schemes have been trained using pedestrian data only. Recovering the most likely discrete orientation class then involved maximum-selection over the associated multi-class model.

We consider the main contribution of our paper to be the integrated framework for pedestrian classification and orientation, see Figure 1. Previous approaches to orientation estimation, [12, 26, 31], assumed classification to be solved beforehand by some other approach or treated both problems separately with different models and different training data. In our approach, both problems are addressed in a unified fashion, using the same underlying mixture-of-experts model within a probabilistic framework. The integrated treatment improves the performance of both classification and orientation estimation. Unlike [12, 26, 31], we utilize readily available negative samples not only for classification

but also for orientation estimation, to better map out the feature space and stabilize the learned discriminative models. Our orientation estimate involves approximating the density function of pedestrian body orientation. This is quite unlike [12, 31], where pedestrian heading is only recovered in terms of pre-defined orientation classes, e.g. front, back, etc., using multi-class classification techniques. Such orientation classes are implicitly contained in our approach by integrating the density function. A secondary contribution is concerned with the integration of other cues, e.g. shape [13] or motion [2], as sample-dependent priors, by means of a Bayesian model.

3. Classification and Orientation Estimation

Input to our framework is a training set $\mathcal{D}$ of pedestrian and non-pedestrian samples $\mathbf{x}_i^* \in \mathcal{D}$. Associated with each sample is a class label ω_i, (ω_0 for the pedestrian and ω_1 for the non-pedestrian class), as well as a K-dimensional cluster membership vector $\mathbf{z}_i$, with $0 \leq z_i^k \leq 1$ and $\sum_k z_i^k = 1$. $\mathbf{z}_i$ defines the fuzzy membership to a set of K clusters Ψ_k, which relate to the similarity in appearance to a certain view of a pedestrian. Note that the same also applies to non-pedestrian training samples, where the image structure resembles a certain pedestrian view, see for example the first non-pedestrian sample in Figure 2. Our definition of cluster membership $\mathbf{z}_i$ is given in Sec. 4.1.

3.1. Pedestrian Classification

For pedestrian classification, our goal is determine the class label ω_i of a previously unseen sample $\mathbf{x}_i$. We make a Bayesian decision and assign $\mathbf{x}_i$ to the class with highest posterior probability:

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

We decompose $P(\omega_0|\mathbf{x}_i)$, the posterior probability that a given sample is a pedestrian, in terms of the K clusters Ψ_k as:

$$P(\omega_0|\mathbf{x}_i) = \sum_k P(\Psi_k|\mathbf{x}_i) P(\omega_0|\Psi_k, \mathbf{x}_i) \tag{2}$$

$$\approx \sum_k w_k(\mathbf{x}_i) f_k(\mathbf{x}_i) \tag{3}$$

In this formulation, $P(\Psi_k|\mathbf{x}_i)$ represents a sample-dependent cluster membership prior for $\mathbf{x}_i$. We approximate $P(\Psi_k|\mathbf{x}_i)$ using sample-dependent weights $w_k(\mathbf{x}_i)$, with $0 \leq w_k(\mathbf{x}_i) \leq 1$ and $\sum_k w_k(\mathbf{x}_i) = 1$, as defined in Eq. (5), Sec. 3.2.

$P(\omega_0|\Psi_k, \mathbf{x}_i)$ represents the cluster-specific probability that a given sample $\mathbf{x}_i$ is a pedestrian. Instead of explicitly computing $P(\omega_0|\Psi_k, \mathbf{x}_i)$, we utilize an approximation given by a set of discriminative models f_k, as follows. We train K texture-based classifiers f_k on the full training set $\mathcal{D}$ to discriminate between the pedestrian and the non-pedestrian class. For each training sample $\mathbf{x}_i^*$, the fuzzy cluster membership vector $\mathbf{z}_i$ is used as a sample-dependent weight during training. The classifier outputs $f_k(\mathbf{x}_i)$ can be seen as approximation of the cluster-specific posterior probabilities $P(\omega_0|\Psi_k, \mathbf{x}_i)$.

In principle, the proposed framework is independent from the actual discriminative model used. We only require example-dependent weights during training, e.g. [4], and that the classifier output (decision value) $f_k(\mathbf{x}_i)$ relates to an estimate of posterior probability. In the limit of infinite data, the outputs of many state-of-the-art classifiers, e.g. neural networks or support vector machines, can be converted to an estimate of posterior probabilities [18, 28]. We use this in our experiments.

3.2. Sample-Dependent Cluster Priors

Prior probabilities for membership to a certain cluster Ψ_k of an unseen sample $\mathbf{x}_i$, $P(\Psi_k|\mathbf{x}_i)$, are introduced in Eq. (2). Note, that this prior is not a fixed prior, but depends on the sample $\mathbf{x}_i$ itself. At this point, information from other cues besides texture (on which the discriminative models f_k are based) can be incorporated into our framework in a probabilistic manner. We propose to model cluster priors using a Bayesian approach as:

$$P(\Psi_k|\mathbf{x}_i) = \frac{p(\mathbf{x}_i|\Psi_k)P(\Psi_k)}{\sum_l p(\mathbf{x}_i|\Psi_l)P(\Psi_l)} \tag{4}$$

Cluster conditional-likelihoods $p(\mathbf{x}_i|\Psi_k)$ involve the representation of $\mathbf{x}_i$ in terms of a set of features, followed by likelihood estimation. Possible cues include motion-based features, i.e. optical flow [2, 8], or shape [13]. Likelihood estimation can be performed via histogramming on training data or fitting parametric models [13].

In our experiments, we consider both uniform priors, as well as shape-based priors based on [13]: Within each cluster Ψ_k, a discrete set of shape templates specific to Ψ_k is matched to the sample $\mathbf{x}_i$. Shape matching involves correlation of the shape templates with a distance-transformed version of $\mathbf{x}_i$. Let $D_k(\mathbf{x}_i)$ denote the residual shape distance between the best matching shape in cluster Ψ_k and sample $\mathbf{x}_i$. By representing $\mathbf{x}_i$ in terms of $D_k(\mathbf{x}_i)$ and using Eq. (4), sample-dependent shape-based priors for cluster Ψ_k are approximated as:

$$w_k(\mathbf{x}_i) = \frac{p(D_k(\mathbf{x}_i)|\Psi_k)P(\Psi_k)}{\sum_l p(D_l(\mathbf{x}_i)|\Psi_l)P(\Psi_l)} \approx P\left(\Psi_k|\mathbf{x}_i\right) \tag{5}$$

Priors $P(\Psi_k)$ are assumed equal and cluster-conditionals are modeled as exponential distributions of $D_k(\mathbf{x}_i)$:

$$p(D_k(\mathbf{x}_i)|\Psi_k) \propto \lambda_k e^{-\lambda_k D_k(\mathbf{x}_i)} \tag{6}$$

Parameters λ_k of the exponential distributions are learned via maximum-likelihood on the training set.

3.3. Pedestrian Orientation Estimation

Instead of simply assigning a test sample to one of the K view-related clusters Ψ_k used for training (i.e. a maximum a-posteriori decision over the expert classifiers), we aim to estimate the actual body orientation θ of a pedestrian ω_0. For this, we use a mixed discrete-continuous distribution $p(\omega_0, \theta|\mathbf{x}_i)$ which is approximated by a Gaussian mixture model:

$$p(\omega_0, \theta|\mathbf{x}_i) \approx \sum_k \alpha_{k,i}\ g_k(\theta|\mathbf{x}_i) \tag{7}$$

In each cluster Ψ_k, a Gaussian with mean μ_k and standard deviation σ_k is used to approximate the component density $g_k(\theta|\mathbf{x}_i)$ of pedestrian body orientation associated with cluster Ψ_k. For mixture weights $\alpha_{k,i}$, we re-use $w_k(\mathbf{x}_i)f_k(\mathbf{x}_i)$, the weighted classifier outputs, as defined in Eq. (3):

$$g_k(\theta|\mathbf{x}_i) = \mathcal{N}(\theta|\mu_k, \sigma_k^2)\ \ ;\ \ \alpha_{k,i} = w_k(\mathbf{x}_i)f_k(\mathbf{x}_i) \tag{8}$$

The most likely pedestrian orientation $\hat{\theta}_i$ can be recovered by finding the mode of the density in Eq. (7), e.g. [5]:

$$\hat{\theta}_i = \underset{\theta}{\operatorname{argmax}}\left(p\left(\omega_0, \theta|\mathbf{x}_i\right)\right) \tag{9}$$

Besides estimating $p(\omega_0, \theta|\mathbf{x}_i)$, our framework allows to recover so-called orientation classes, similar to [12, 26, 31]: The probability that a sample $\mathbf{x}_i$ is a pedestrian with orientation in a range of $[\tilde{\theta}_a, \tilde{\theta}_b]$ is given by:

$$P(\omega_0, \theta \in [\tilde{\theta}_a, \tilde{\theta}_b] \mid \mathbf{x}_i) = \int_{\tilde{\theta}_a}^{\tilde{\theta}_b} p(\omega_0, \theta|\mathbf{x}_i)d\theta \tag{10}$$

We do not use *one vs. one* [12, 26] or *one vs. rest* [26, 31] multi-class models for orientation estimation. Given the similarity of front/back or left/right views in low-resolution scenarios, such schemes would require highly similar training samples (often of the same physical pedestrians) to appear in both positive and negative training data, see Figure 2. As a result, the training procedure might become unstable and the recovered decision boundaries error-prone.

Instead, we tightly integrate orientation estimation and pedestrian classification by means of re-using our classification models. Weights $\alpha_{k,i}$ of the employed Gaussian mixture model are based on the cluster-specific discriminative models f_k and the associated sample-dependent prior weights, see Eqs. (3) and (8). The training of f_k involves pedestrians and non-pedestrian samples which are readily available in great quantities at no additional cost and help to gain robustness by implicitly mapping out the feature space and the decision boundary. Using this scheme, the problems

Figure 2. Examples of training and test data for pedestrians in four view-related clusters and non-pedestrian samples.

of the *one vs. one* or *one vs. rest* strategies (see above) can be overcome.

Another aspect is computational efficiency. Our framework does not require to train an additional classifier for orientation estimation. Due to the integrated treatment, orientation estimation requires only little additional resources, since the main computational costs are introduced by the texture-based classifiers f_k, which are re-used.

4. Experiments

4.1. Experimental Setup

The proposed integrated framework was tested in large-scale experiments on pedestrian classification and orientation estimation. To illustrate the generality with respect to the discriminative models used, we chose two instances for experimental evaluation which exhibit a diverse set of features. First, we consider histograms of oriented gradients (9 orientation bins, 8×8 pixel cells) combined with a linear support vector machine classifier (HOG) [7]. Second, we evaluate adaptive local receptive field features (5×5 pixels) in a multi-layer neural network architecture (NN/LRF) [36]. Results are expected to generalize to other pedestrian classifiers that are sufficiently complex to represent the large training sets, e.g. [9, 18, 21, 27, 33, 35, 37, 39].

Training and test sets contain manually labeled pedestrian bounding boxes. We consider $K = 4$ view-related clusters Ψ_k, roughly corresponding to similarity in appearance to front, left, back and right views of pedestrians. For the non-pedestrian samples, we use the approximated cluster prior probability, see Sec. 3.2, as cluster membership weights for training:

$$z_k^i = w_k(\mathbf{x}_i) \approx P(\Psi_k|\mathbf{x}_i)\ ,\ \omega_i = \omega_1 \tag{11}$$

To compute $w_k(\mathbf{x}_i)$, a set of 10946 shape templates corresponding to clusters Ψ_k is used. Rather than Eq. (11), we use a manual assignment to clusters Ψ_k for pedestrian training samples, which we found to perform best in preliminary

	Pedestrians (labeled)	Pedestrians (jittered)	Non-Pedestrians
Training Set	42645	383805	342271
Test Set	7613	68517	73405

Table 1. Training and test set statistics.

experiments. A possible reason is that shape cannot provide a clear distinction between front and back views. Note that the approaches we compare against, i.e. [12, 26, 31], have similar requirements in terms of data labeling.

See Table 1 and Figure 2 for the dataset used. All training samples are scaled to 48×96 pixels (HOG) or 18×36 pixels (NN/LRF) with an eight-pixel border (HOG) or two-pixel border (NN/LRF), to retain contour information. Nine training (test) samples were created from each label by geometric jittering. Pedestrian samples depict non-occluded pedestrians in front of a changing background.

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. Training and test set were strictly separated: no instance of the same real-world pedestrian appears in both training and test set, similarly for the non-target samples.

4.2. Pedestrian Classification Performance

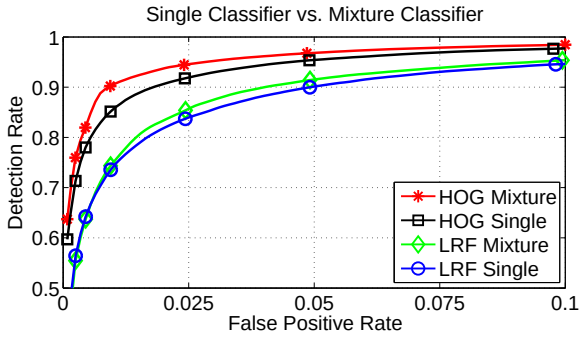
In our first experiment, we evaluate the classification performance of the proposed view-related mixture architecture in comparison to a single classifier trained on the whole dataset irrespective of view, i.e. the approach of [7, 36]. Cluster priors, see Sec. 3.2, are considered uniform. Results in terms of ROC performance are shown in Figure 3(a).

The mixture classifiers perform better than the corresponding single classifiers. The decomposition of the problem into view-related subparts simplifies the training of the expert classifiers, since a large part of the observable variation in the samples is already accounted for. Classification performance and robustness is increased by a combined decision of the experts. The performance benefit for the HOG classifier is approx. a factor of two in reduction of false positives at the same detection rate. Using LRF features, the benefit of the mixture classifier is less pronounced.

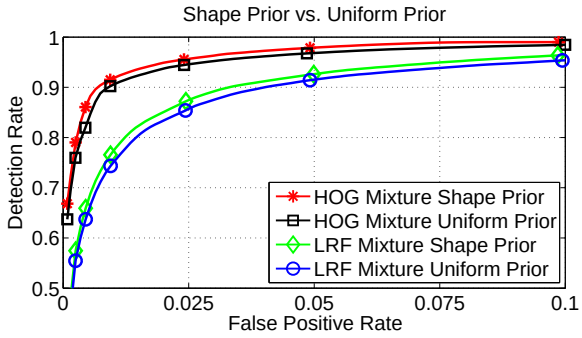
Figure 3(b) shows the effect of adding a sample-dependent cluster prior based on shape matching, see Sec. 3.2. For both HOG and LRF, only a small benefit is observed. This suggests, that the utilized classifiers are capable to adequately capture the structure of each cluster, based on the employed texture-based feature set alone.

4.3. Orientation Estimation Performance

Discrete Orientation Classes In our second experiment, we evaluate orientation estimation performance using the best performing system variant, as given in Figure 3: HOG



(a)



(b)

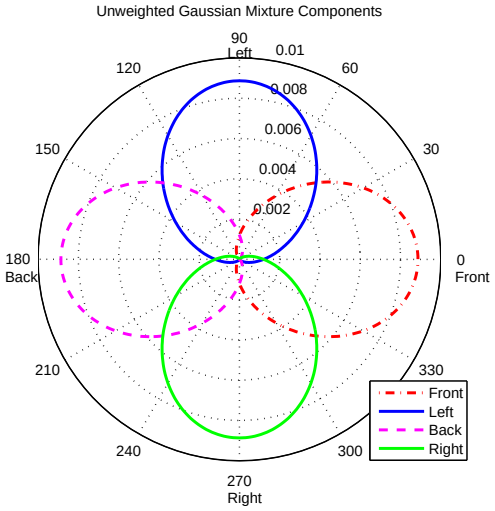
Figure 3. (a) Performance of single classifiers vs. the view-related mixture architecture. (b) Benefit of shape-based priors in comparison to non-informative priors.

mixture classifier with shape-based cluster priors. The Gaussian mixture components used to model the cluster-specific density of body orientation θ are empirically set as follows (cf. Sec. 4.1):

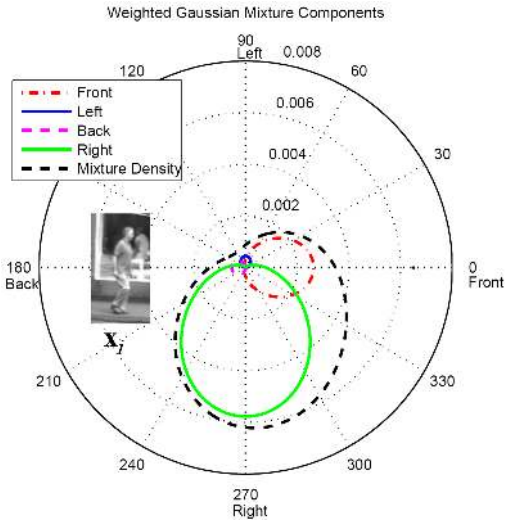
$$\Psi_i : \mu_i = i \cdot 90°, \sigma_i = 45°, \text{ for } i \in \{0, 1, 2, 3\} \qquad (12)$$

Figure 4 visualizes probability densities of body orientation θ using a polar coordinate system. The angular axis depicts orientation θ whereas the value of the densities is shown on the radial axis (i.e. distance from the center). In Figure 4(a), Gaussian mixture components $g_k(\theta|\mathbf{x}_i)$, see Eq. (8), are shown with parameters given in Eq. (12). Figure 4(b) depicts weighted mixture components and the resulting mixture density $p(\omega_0, \theta|\mathbf{x}_i)$. Weights $\alpha_{k,i}$ are derived from the given test sample $\mathbf{x}_i$ using Eq. (8). Note that the actual orientation of the pedestrian sample matches the mode of the recovered mixture density.

We compare our approach to our own implementations of two state-of-the-art approaches to recover discrete orientation classes (front, back, left and right), using the same data and evaluation criteria, in terms of confusion matrices. First, we consider the approach of Shimizu & Poggio [31] which involves Haar wavelet features with a set of support vector machines in a *one vs. rest* scheme. Second,



(a)



(b)

Figure 4. Orientation densities in polar coordinates. (a) Gaussian mixture components $g_k(\theta|\mathbf{x}_i)$, (b) Mixture density $p(\omega_0, \theta|\mathbf{x}_i)$ and components, weighted using $\alpha_{k,i}$ for sample $\mathbf{x}_i$ (as shown).

True Orientation

Estimated Orient.	Back	Front	Left	Right
Back	**0.67**	0.28	0.05	0.04
Front	0.30	**0.64**	0.06	0.06
Left	0.02	0.04	**0.85**	0.03
Right	0.01	0.04	0.04	**0.87**

correct (false) decision rate = 0.74 (0.26) per test sample

this paper (GMM)

(a)

True Orientation

Estimated Orient.	Back	Front	Left	Right
Back	**0.64**	0.28	0.06	0.06
Front	0.30	**0.61**	0.07	0.07
Left	0.03	0.05	**0.82**	0.04
Right	0.03	0.06	0.05	**0.83**

correct (false) decision rate = 0.70 (0.30) per test sample

this paper (maximum selection)

(b)

True Orientation

Estimated Orient.	Back	Front	Left	Right
Back	**0.65**	0.28	0.10	0.08
Front	0.19	**0.57**	0.05	0.04
Left	0.08	0.08	**0.77**	0.06
Right	0.08	0.07	0.08	**0.82**

correct (false) decision rate = 0.67 (0.33) per test sample

Gandhi & Trivedi

(c)

True Orientation

Estimated Orient.	Back	Front	Left	Right
Back	**0.56**	0.21	0.06	0.06
Front	0.18	**0.56**	0.08	0.09
Left	0.10	0.10	**0.65**	0.21
Right	0.16	0.13	0.21	**0.64**

correct (false) decision rate = 0.59 (0.41) per test sample

Shimizu & Poggio

(d)

Figure 5. Confusion matrices and correct / false decision rate per test sample for: (a) this paper (GMM), (b) this paper (max. selection), (c) Gandhi & Trivedi [12], (d) Shimizu & Poggio [31].

we evaluate the single-frame method of Gandhi & Trivedi [12]. This technique uses HOG features (we use identical HOG parameters as for our approach) and support vector machines in a *one vs. one* fashion, together with the estimation of pairwise cluster probabilities. Both approaches were trained on pedestrian data only. To obtain discrete orientation classes in our approach, we utilize Eq. (10). We additionally consider a variant of our framework involving maximum-selection over the expert classifiers, instead of the Gaussian mixture-model (GMM) formulation, cf. Sec. 3.3.

Results are given in Figures 5 and 6. Our approach reaches up to 67% accuracy for front/back views and up to 87% accuracy for left/right views, clearly outperforming previous work. The overall correct (false) decision rate is 0.74 (0.26) per test sample. This represents a reduction in false decision rate of more than 20% compared to Gandhi & Trivedi [12] and more than 35% compared to Shimizu & Poggio [31]. Note, that we use the same feature set for both our approach and for Gandhi & Trivedi [12]. The observed performance differences result from the proposed integration of orientation estimation and classification. Using maximum-selection decreases the performance over GMM.

While the errors in orientation estimation for left and

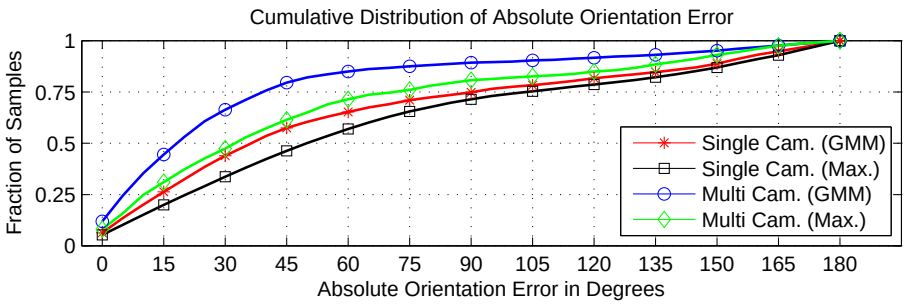


Figure 6. Cumulative distribution of absolute orientation error using different system variants, see text.

right views are evenly distributed among the other classes, front and back views are more often confused with each other. We attribute this to front and back views of pedestrians being highly similar both in shape and texture. The main distinguishing factor is the head/face area, which is very small compared to the torso/leg area, see Figure 2. In case of left and right views, characteristic leg posture and body tilt seem to be more discriminative cues.

Continuous Orientation To evaluate the quality of our continuous orientation estimate, we utilize 14118 2D images of fully visible pedestrians from a realistic multi-camera (3 cameras at different view-points, 4706 images per camera) 3D human pose estimation dataset, see [16][1]. Since ground-truth 3D pose is available, we can obtain exact ground-truth body orientation for all 2D images to compare against. We evaluate the two best performing systems from the previous experiment: our approach using GMM and maximum-selection. Our evaluation measure is absolute difference of estimated orientation and ground-truth orientation.

First, we treat all images independently, irrespective of which camera they come from (simulating a single camera) and perform orientation estimation using Eq. (9). Second, we take into account that each pedestrian is visible in three cameras at the same time from different view-points. One camera serves as a reference camera and the rotational offsets of the other cameras are known through camera calibration. For orientation estimation, we establish $K = 4$ view-related models (related to front, back, left and right) per camera and incorporate all clusters into a single 12-component GMM model, see Sec. 3.3, with orientations normalized to the reference camera. For maximum-selection using multiple cameras, we perform orientation estimation using maximum-selection over the expert classifiers independently for each camera and average (normalized) orientations over all three cameras. This technique performs better than maximum-selection over all 12 models.

[1]Thanks to the authors of [16] for making the dataset publicly available.

Results are shown in Figure 6, in terms of cumulative distributions of absolute orientation error which are obtained using histogramming. All GMM variants outperform the maximum-selection variants. Multi-camera fusion significantly improves performance. The benefit is more significant for the GMM approach (blue curve vs. red curve) than for the maximum-selection approach (green curve vs. black curve) which demonstrates the strength of the proposed GMM-based orientation estimation technique. Covering the same fraction of samples, orientation errors for the multi-camera GMM approach are up to 50% less than for the corresponding maximum-selection technique (blue curve vs. green curve).

Note that the presented results were obtained by considering orientation errors for all views. Results on a subset consisting of left and right views are significantly better, cf. Figure 5. Further, no temporal filtering of the recovered orientation densities was applied, which would presumably further improve absolute performance.

5. Conclusion

This paper presented a novel integrated approach for pedestrian classification and orientation estimation. Our probabilistic model does not restrict the estimated pedestrian orientation to a fixed set of orientation classes but directly approximates the probability density of body orientation. Cluster priors can be incorporated using a Bayesian model. In large-scale experiments, we showed that the proposed integrated approach reduces the error rate for classification and orientation estimation by up to 50%, compared to state-of-the-art. We take this as evidence for the strength of the proposed integrated approach. Future work deals with additional cues as priors (e.g. motion) and full integration into a pedestrian recognition system.

References

[1] S. Agarwal, A. Awan, and D. Roth. Learning to detect objects in images via a sparse, part-based representation. *IEEE PAMI*, 26(11):1475–1490, 2004. 2

[2] S. Beauchemin and J. Barron. The computation of optical flow. *ACM Computing Surveys*, 27(3):433–467, 1995. 3

[3] J. Berclaz, F. Fleuret, and P. Fua. Principled detection-by-classification from multiple views. *Proc. of the Third Int. Conf. on Computer Vision Theory and Applications*, pages 375–382, 2008. 2

[4] U. Brefeld, P. Geibel, and F. Wysotzki. Support vector machines with example dependent costs. *Proc. ECML*, pages 23–34, 2003. 3

[5] M. A. Carreira-Perpinan. Mode-finding for mixtures of Gaussian distributions. *IEEE PAMI*, 22(11):1318–1323, 2000. 4

[6] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE PAMI*, 23(6):681–685, 2001. 2

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. CVPR*, pages 886–893, 2005. 2, 4, 5

[8] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *Proc. ECCV*, pages 428–441, 2006. 2, 3

[9] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. *Proc. CVPR*, 2009. 2, 4

[10] M. Enzweiler and D. M. Gavrila. A mixed generative-discriminative framework for pedestrian classification. In *Proc. CVPR*, 2008. 2

[11] M. Enzweiler and D. M. Gavrila. Monocular pedestrian detection: Survey and experiments. *IEEE PAMI*, 31(12):2179–2195, 2009. 2

[12] T. Gandhi and M. Trivedi. Image based estimation of pedestrian orientation for improving path prediction. In *IEEE IV Symposium*, pages 506–511, 2008. 1, 2, 3, 4, 5, 6

[13] D. M. Gavrila and S. Munder. Multi-cue pedestrian detection and tracking from a moving vehicle. *IJCV*, 73(1):41–59, 2007. 1, 2, 3

[14] D. Geronimo, A. M. Lopez, A. D. Sappa, and T. Graf. Survey on pedestrian detection for advanced driver assistance systems. *IEEE PAMI, available online: http://doi.ieeecomputersociety.org/10.1109/TPAMI.2009.122*, 2009. 2

[15] T. Heap and D. Hogg. Improving specificity in PDMs using a hierarchical approach. In *Proc. BMVC*, pages 80–89. A. F. Clark (ed.), 1997. 2

[16] M. Hofmann and D. M. Gavrila. Multi-view 3D human upper body pose estimation combining single-frame recovery, temporal integration and model adaptation. *Proc. CVPR*, 2009. 2, 7

[17] 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. 2

[18] A. K. Jain, R. P. W. Duin, and J. Mao. Statistical pattern recognition: A review. *IEEE PAMI*, 22(1):4–37, 2000. 1, 2, 3, 4

[19] E. Jones and S. Soatto. Layered active appearance models. In *Proc. ICCV*, pages 1097– 1102, 2005. 2

[20] B. Leibe, N. Cornelis, K. Cornelis, and L. V. Gool. Dynamic 3d scene analysis from a moving vehicle. In *Proc. CVPR*, 2007. 2

[21] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In *Proc. CVPR*, pages 878–885, 2005. 4

[22] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 2

[23] 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. 2

[24] 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. 2

[25] A. Mohan, C. Papageorgiou, and T. Poggio. Example-based object detection in images by components. *IEEE PAMI*, 23(4):349–361, 2001. 2

[26] C. Nakajima, M. Pontil, B. Heisele, and T. Poggio. Full-body recognition system. *Pattern Recognition*, 36:1997–2006, 2003. 1, 2, 4, 5

[27] C. Papageorgiou and T. Poggio. A trainable system for object detection. *IJCV*, 38:15–33, 2000. 2, 4

[28] J. C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances In Large Margin Classifiers*, pages 61–74, 1999. 3

[29] E. Seemann, M. Fritz, and B. Schiele. Towards robust pedestrian detection in crowded image sequences. In *Proc. CVPR*, 2007. 2

[30] A. Shashua, Y. Gdalyahu, and G. Hayon. Pedestrian detection for driving assistance systems: Single-frame classification and system level performance. In *IEEE IV Symposium*, pages 1–6, 2004. 2

[31] H. Shimizu and T. Poggio. Direction estimation of pedestrian from multiple still images. In *IEEE IV Symposium*, pages 596–600, 2004. 1, 2, 3, 4, 5, 6

[32] M. Turk and M. Kölsch. Perceptual interfaces. In G. Medioni and S. Kang, editors, *Emerging Topics in Computer Vision*. Prentice Hall, 2004. 1

[33] O. Tuzel, F. Porikli, and P. Meer. Human detection via classification on Riemannian manifolds. In *Proc. CVPR*, 2007. 2, 4

[34] I. Ulusoy and C. M. Bishop. Generative versus discriminative methods for object recognition. In *Proc. CVPR*, pages 258–265, 2005. 2

[35] P. Viola, M. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. *IJCV*, 63(2):153–161, 2005. 2, 4

[36] C. Wöhler and J. K. Anlauf. A time delay neural network algorithm for estimating image-pattern shape and motion. *IVC*, 17:281–294, 1999. 2, 4, 5

[37] 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. 2, 4

[38] L. Zhang, B. Wu, and R. Nevatia. Detection and tracking of multiple humans with extensive pose articulation. In *Proc. ICCV*, 2007. 2

[39] 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. 2, 4