A Multilevel Mixture-of-Experts Framework for Pedestrian Classification

Markus Enzweiler and Dariu M. Gavrila

***Abstract*—Notwithstanding many years of progress, pedestrian recognition is still a difficult but important problem. We present a novel multilevel Mixture-of-Experts approach to combine information from multiple features and cues with the objective of improved pedestrian classification. On pose-level, shape cues based on Chamfer shape matching provide sample-dependent priors for a certain pedestrian view. On modality-level, we represent each data sample in terms of image intensity, (dense) depth, and (dense) flow. On feature-level, we consider histograms of oriented gradients (HOG) and local binary patterns (LBP). Multilayer perceptrons (MLP) and linear support vector machines (linSVM) are used as expert classifiers.**

Experiments are performed on a unique real-world multimodality dataset captured from a moving vehicle in urban traffic. This dataset has been made public for research purposes. Our results show a significant performance boost of up to a factor of 42 in reduction of false positives at constant detection rates of our approach compared to a baseline intensity-only HOG/linSVM approach.

***Index Terms*—Mixture-of-experts, object detection, pedestrian classification.**

I. Introduction

PEDESTRIAN recognition is a key problem for a number of application domains, such as intelligent vehicles, surveillance, and robotics. Notwithstanding years of methodical and technical progress, e.g., see [10], [16], and [20], it is still a difficult task from a machine-vision point of view. There is a wide range of pedestrian appearance arising from changing articulated pose, clothing, lighting, and—in the case of a moving camera in a dynamic environment—ever-changing backgrounds. Explicit models to solve the problem are not readily available, so most research has focused on implicit learning-based representations [25].

Many interesting pedestrian classification approaches have been proposed; an overview is given in Section II. Most approaches follow a two-step approach involving feature extraction and pattern classification. In recent years, a multitude of (more or less) different feature sets has been used to discriminate pedestrians from nonpedestrian images. Most of those features operate on intensity contrasts in spatially restricted local parts of an image. As such, they resemble neural structures which exist in lower level processing stages of the human visual cortex [21]. In human perception, however, depth and motion are important additional cues to support object recognition. In particular, the motion flowfield and surface depth maps seem to be tightly integrated with spatial cues, such as shape, contrasts, or color [27].

With a few exceptions (see Section II), most spatial features used in machine vision for object classification are based on intensity cues only. If used at all, depth and motion cues merely provide information about scene geometry or serve as a selection mechanism for regions of interest in a segmentation rather than a classification context [12], [13], [19], [37].

In this paper, we propose to enrich intensity-based feature spaces for pedestrian classification with features operating on dense stereo (depth) and dense optical flow (motion). We show how to combine multifeature/multicue classifiers in a principled manner, using a classifier-independent Mixture-of-Experts framework which does neither suffer from the curse of dimensionality nor impractical training times, given our large high-dimensional dataset.

II. Previous Work

Pedestrian classification has attracted a significant amount of interest from the research community over the past years. See [10], [16], [20], and [23] for recent surveys and performance studies. In this work, we focus on 2-D approaches which are suitable for medium resolution pedestrian data (i.e., pedestrian height between 30 and 80 pixels). We do not consider more detailed perception tasks such as human pose recovery or activity recognition, e.g., [17], [34].

A pedestrian classifier is typically part of an integrated system involving a preprocessing step to select initial object hypotheses and a postprocessing step to integrate classification results over time (tracking); see [10] and [20]. The classifier itself is the most important module. Its performance accounts for the better part of the overall system performance and the majority of computational resources is spent here.

Most approaches for pedestrian classification follow a discriminative scheme by learning discriminative functions (decision boundaries) to separate object classes within a feature space. Prominent features can be roughly categorized into texture-based and gradient-based.

Nonadaptive texture-based Haar wavelet features have been popularized by [41] and used by many others [35], [50], [56]. Recently, local binary pattern (LBP) features [39] have also been

Manuscript received October 12, 2010; revised February 03, 2011; accepted March 23, 2011. Date of publication April 11, 2011; date of current version September 16, 2011. This work was supported by the Studienstiftung des deutschen Volkes (German National Academic Foundation). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Erhardt Barth.

M. Enzweiler is with the Environment Perception Department, Daimler AG Group Research & MCG Development, 89081 Ulm, Germany (e-mail: markus.enzweiler@daimler.com).

D. M. Gavrila is with the Environment Perception Department, Daimler AG Group Research & MCG Development, 89081 Ulm, Germany, and also with the Intelligent Autonomous Systems Group, University of Amsterdam, 1098 XH, Amsterdam, The Netherlands (e-mail: dariu.gavrila@daimler.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2011.2142006

employed in pedestrian classification [53]. The particular structure of local texture features has been optimized in terms of local receptive field (LRF) features [11], [19], [40], [55], which adapt to the underlying data during training. Other texture-based features are codebook patches, extracted around interest points in the image [1], [28], [45] 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) [30] and dense representations [histograms of oriented gradients (HOG)] [4], [9], [11], [32], [40], [51]–[53], [56], [59], [60]. Spatial variation and correlation of gradients have been encoded using covariance descriptors enhancing robustness towards brightness variations [48]. However, others have proposed local shape filters exploiting characteristic patterns in the spatial configuration of salient edges [33], [57].

Some of the presented spatial filters have been extended to the spatio-temporal domain by means of intensity differences over time [50], [55] or optical flow [5].

Regarding pattern classifiers, support vector machines (SVMs) have become very popular in the domain of pedestrian classification, in both linear [4], [5], [9], [11], [36], [51], [52], [56], [59], [60] and nonlinear variants [32], [35], [41]. However, performance boosts resulting from the nonlinear model are paid for with a significant increase in computational costs and memory. Recent work presented efficient versions of nonlinear SVMs for a specific class of kernels [32]. Other popular classifiers include neural networks [11], [19], [25], [36], [55] and boosted classifiers [33], [48], [50]–[52], [56], [57], [59], [60].

In the past years, many novel feature and classifier combinations were proposed to improve classification performance, along with corresponding experimental studies and benchmarks, e.g., [7], [10], [23], [36]. Orthogonal to such lower level performance boosts are improvements coming from higher level methods based on the fusion of multiple classifiers.

Several approaches have attempted to break down the complexity of the problem into subparts. One way is to represent each sample as an ensemble of components which are usually related to body parts. After detecting the individual body parts, detection results are fused using statistical models [15], [33], [57], learning or voting schemes [6], [9], [29], [35], [45], or heuristics [53].

Beside component-based approaches, multi-orientation models are relevant to current work. Here, local pose-specific clusters are established, followed by the training of specialized classifiers for each subspace. The final decision of the classifier ensemble involves maximum selection [57], trajectory-based data association [59], shape-based combination [11], [19], or a fusion classifier [46].

A recent trend in the community involves the combination of multiple features or modalities, e.g., intensity, depth. and motion. While some approaches utilize combinations on the module level [2], [12], [13], [19], [37], [47], others integrate multiple information sources directly into the pattern classification step [5], [9], [40], [43], [44], [49], [51]–[53], [56], [58].

To the best of our knowledge, our work in [9] presented the first use of appearance, motion, and stereo features for pedestrian classification. A similar approach was recently presented in [52]. Some approaches combine features in the intensity domain using a boosted cascade classifier [58] or multiple kernel learning [49]. One approach combines HOG, covariance, and edgelet features in the intensity domain into a boosted heterogenous cascade classifier with an explicit optimization with regard to runtime [58]. Others integrate intensity and flow features by boosting [51], [56] or by concatenating all features into a single feature vector which is then passed to a single classifier [5], [51], [56]. The work in [51] was recently extended to additionally include depth features [52]. A joint feature space approach to combine HOG and LBP features was used in [53]. [44] presents the integration of HOG features, co-occurrence features and color frequency descriptors into a very high-dimensional ($\approx$ 170 000 dimensions) joint feature space in which classical machine learning approaches are intractable. Hence, partial least squares is applied to project the features into a subspace with lower dimensionality which facilitates robust classifier learning. Boosting approaches require mapping the multidimensional features to a single dimension, either by applying projections [58] or treating each dimension as an individual feature [56]. An alternative is the use of more complex weak learners that operate in a multidimensional space, e.g., support vector machines, [60].

In contrast, [5], [9], [40], and [43] utilize fusion on the classifier level by training a specialized classifier for each cue. The work in [5] and [9] use a single feature (HOG) in two (intensity and depth) and three different modalities (intensity, depth, and motion), respectively. The work in [40] involves a combination of two features (HOG and LRF) with a single modality (intensity). Finally, the work in [43] presents a classifier-level combination of two features, where each feature operates in a different modality (HOG/intensity and LRF/depth). Classifier fusion is done using fuzzy integration [40], simple classifier combination rules [43], or a Mixture-of-Experts framework [5], [9], [24]. Our work in [9], [11], and [43] provides the foundation for this paper.

III. Overview and Contributions

Our Mixture-of-Experts framework [24] for pedestrian classification combines four modalities (shape, intensity, depth, and motion) and three features (Chamfer distance, HOG, and LBP). We follow a multilevel approach by utilizing expert classifiers on pose, modality, and feature levels; see Fig. 1(a). The local experts are integrated in terms of a probabilistic pose-specific model based on fuzzy view-related clustering and associated sample-dependent cluster priors. K view-related models, specific to fuzzy clusters Ψ_k, are trained in an off-line step to discriminate between pedestrians and nonpedestrians. These models consist of sample-dependent cluster priors and multilevel (multicue/multifeature) expert classifiers. In the online application phase, cluster priors are computed using shape matching and used to fuse the multilevel expert classifiers to a combined decision; see Fig. 1(b). Details are given in Section IV.

The main contribution of this paper is the aforementioned pose-specific multilevel Mixture-of-Experts framework for pedestrian classification, which breaks down the complex classification problem into better manageable subproblems. To our

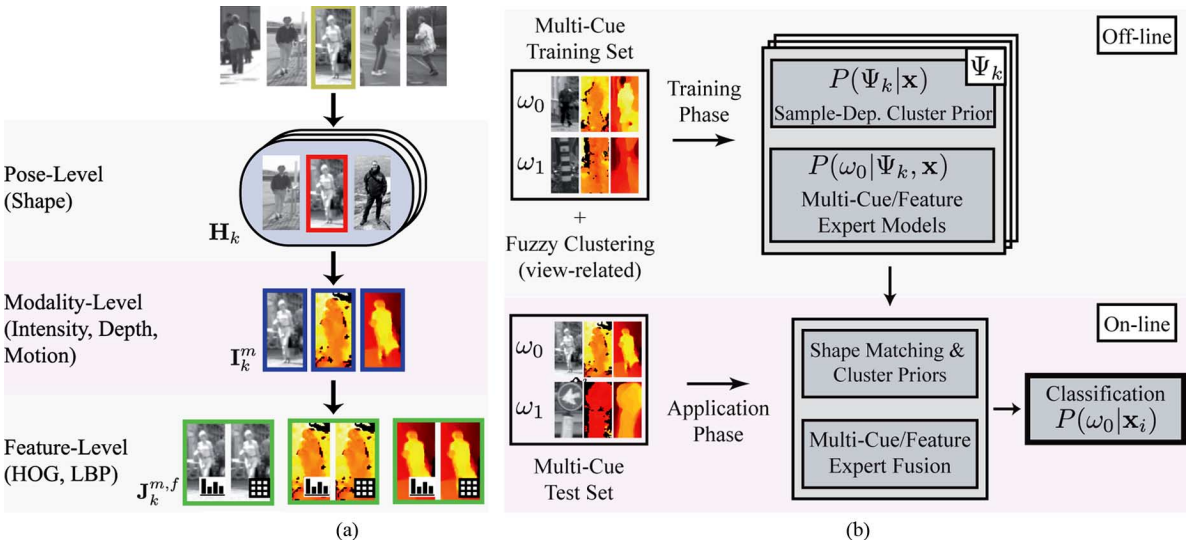

Fig. 1. Framework overview. (a) Multilevel object representation comprising Mixture-of-Experts on pose, (Ψ_k), modality, (Ψ_m), and feature levels (Ψ_f). (b) K view-related models specific to fuzzy clusters Ψ_k are used for pedestrian classification. The models consist of sample-dependent cluster priors and multicue/feature discriminative experts which are learned from pedestrian (class ω_0) and nonpedestrian (class ω_1) samples $\mathbf{x}$.

knowledge, this work represents the first integration of shape, intensity, depth and motion as features into a pattern classification framework. We observed that the same pedestrians appear with a different level of saliency in the gray-level intensity, depth, and motion images. This motivates our multimodality fusion approach, to benefit from the strengths of the individual cues. Our multicue dataset has been made public for evaluation purposes, see Section V-A.

In this paper, we are not concerned with establishing the best *absolute* performance given various state-of-the-art pedestrian classifiers. We refer the reader to recently proposed systems and benchmark studies, e.g., [4], [7], [10], [20], [23], [48], [50], [52], [57], [60]. Rather, our aim is to demonstrate the *relative* performance gain resulting from the proposed multilevel approach, exemplified using state-of-the-art feature sets and classifiers in our experiments. The proposed framework is independent of the actual feature sets and classifiers used. The experiments in this paper are designed to stimulate further research and provide an accessible baseline—we use publicly available data and software implementations wherever possible—to which the scientific community can benchmark additional feature-classifier combinations.

Our approach has a number of advantages compared to fusion approaches using a joint feature space, e.g., [44], [53], [56]. First, our individual expert classifiers operate on a local lower-dimensional feature subspace and are less prone to overfitting effects, given an adequate number of training samples. We do not need to apply dimensionality reduction techniques, e.g., [44], to robustly train our classifiers. Compared to multifeature boosting approaches, we also do not require techniques to map the multidimensional features to a single dimension, e.g., through projection [58] or selection of 1-D features [56].

Second, our Mixture-of-Experts framework alleviates practical problems arising from the use of large and high-dimensional datasets in pattern classification. Some authors reported that classical machine learning techniques do not scale up (on practical terms) to the use of many tens of thousands of high-dimensional training samples, due to excessive memory requirements, e.g., nonlinear SVMs [10] or even linear SVMs [4], [44]. In contrast, the local expert classifiers in our framework are trained on a lower dimensional subspace alleviating memory requirements. As a result, more complex classifiers and/or a larger amount of training samples can be used, which results in better performance.

A third issue is training time, which can be of the order of weeks on current hardware, particularly for boosting approaches, e.g., [10], [56], [58]. In our approach, training times are usually faster, given the lower dimensionality and inherent parallelism of training multiple local experts independently at the same time. Note, that the expert classifiers used in our experiments did not require more than one hour for each training run.

Finally, since our expert classifiers are independent from each other, they are not required to use exactly the same dataset for training. Given that most recently published datasets include samples from the intensity domain only [7], [10], [36], our approach could make maximum use of all available samples. For evaluation purposes, we utilize the same data samples for each cue/feature in our experiments to eliminate effects arising from imbalanced data.

This paper goes beyond our earlier work in [9], [11], and [43]. In [9], we focus on occlusion handling, whereas the main contribution of [11] is orientation estimation. In [43], we address intensity and depth based pedestrian classification, but take neither pose-specific Mixture-of-Experts nor motion-based features into account.

The remainder of this paper is structured as follows. In Section IV, our multilevel Mixture-of-Experts framework is introduced. Section V presents our dataset and experimental setup. In Section VI, we experimentally evaluate our approach, followed by a discussion in Section VII. We conclude in Section VIII.

(a) (b)

Fig. 2. (a) Average gradient magnitude of all pedestrian training samples for intensity, depth, and motion (left to right). (b) A difficult-to-recognize (low-contrast) pedestrian in the intensity domain can be very salient in other modalities.

IV. Multilevel Mixture-of-Experts

A. Object Representation

Input to our framework is a training set $\mathscr{D}$ of pedestrian (ω_0) and nonpedestrian (ω_1) samples $\mathbf{x}_i \in \mathscr{D}$. Each sample $\mathbf{x}_i = [\mathbf{x}_i^1; \mathbf{x}_i^2; \cdots; \mathbf{x}_i^M]$ consists of M different modalities Ψ_m. In each modality Ψ_m, a sample $\mathbf{x}_i^m \in \Psi_m$ is represented in terms of F features Ψ_f: $\mathbf{x}_i^m = [\mathbf{x}_i^{m,1}; \mathbf{x}_i^{m,2}; \cdots; \mathbf{x}_i^{m,F}]$.

In this work, we consider $M = 3$ different modalities, i.e., gray-level image intensity ($\mathbf{x}_i^1$), dense depth information via stereo vision ($\mathbf{x}_i^2$) [22] and dense optical flow ($\mathbf{x}_i^3$) [54]. We treat $\mathbf{x}_i^2$ and $\mathbf{x}_i^3$ similarly to gray-level intensity images $\mathbf{x}_i^1$, in that both depth and motion cues are represented as images, where pixel values encode distance from the camera and horizontal optical flow between two temporally aligned images.

Dense stereo provides information for most image areas, apart from regions which are visible only by one camera (stereo shadow). Spatial features can be based on either depth Z (in meters) or disparity d (in pixels). Both are inversely proportional, given the camera geometry with focal length f and the distance between the two cameras B:

$$Z(x,y) = \frac{fB}{d(x,y)} \text{at pixel } (x,y). \tag{1}$$

Objects in the scene have similar foreground/background gradients in depth space, irrespective of their location relative to the camera. In disparity space however, such gradients are larger, the closer the object is to the camera. To remove this variability, we derive spatial features from depth instead of disparity.

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. Besides, robust ego-motion compensation is a rather difficult task. As a positive side-effect, static pedestrians do not pose a problem in combination with a moving camera.

A visual inspection of the intensity versus depth and flow images in Figs. 2 and 3 reveals that pedestrians have distinct contours and textures in each modality. Fig. 2(a) shows the average gradient magnitude of all pedestrian training samples for each modality. In intensity images, lower-body features (shape and appearance of legs) are the most significant features of a pedestrian (see results of part-based approaches, e.g., [35]). There is significant texture on the pedestrian due to different clothing.

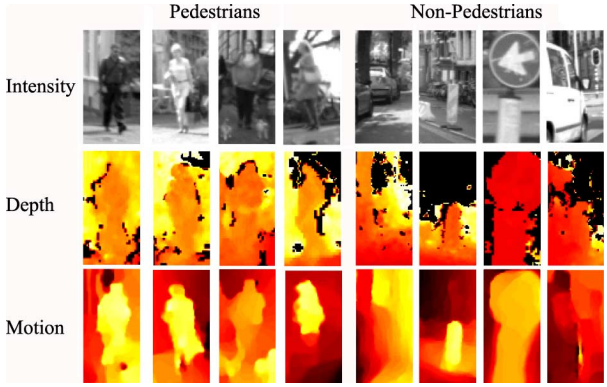

Fig. 3. Pedestrian and nonpedestrian samples in our dataset. In depth images, darker colors denote closer distances. Note that the background (large depth values) has been faded out for visibility. Optical flow images depict the horizontal component of flow vectors. Medium red colors denote close to zero flow, and darker and brighter colors indicate stronger motion (to the left and to the right, respectively).

In the depth image, the upper-body area has dominant foreground/background gradients and is particularly characteristic for a pedestrian. The depth texture on the pedestrian is fairly uniform, given that areas corresponding to the pedestrian are approximately in the same distance from the camera. Pedestrian gradients in flow images are particularly strong around the upper body and torso contours, resulting from motion discontinuities between the (uniformly moving) pedestrian and the background. Similar to the depth image, the pedestrian upper body area is fairly homogenous due to uniform pedestrian motion. Legs move nonrigidly and less uniform than the rest of the pedestrian body. As a result, the lower body area is more blurred and less significant in the average gradient image.

The various salient regions in intensity, depth, and flow images motivate our use of fusion approaches between those modalities to benefit from the individual strengths, see Section IV-C. A characteristic example is shown in Fig. 2(b). A pedestrian sample which is difficult to classify in the intensity domain due to low contrast may appear very salient in the depth and motion modalities. This highlights the complementary aspect of different modalities.

In our experiments, we consider $F = 2$ features per modality, that is, HOG features [4] and LBP features [39]. The motivation for this choice is twofold. First, recent studies have shown that HOG and LBP features are highly complementary regarding their sensitivity to noisy background edges which are common in cluttered backgrounds [53]. Second, despite the vast amount of features developed in recent years, HOG and LBP are still among the best features around [7], [10], [53]. Detailed parameterization of our feature set is given in Section V-B.

Associated with each sample $\mathbf{x}_i$ is a class label ω_i, (ω_0 for the pedestrian and ω_1 for the nonpedestrian 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 nonpedestrian training samples, where the image structure resembles a certain pedestrian view. Our definition of cluster membership $\mathbf{z}_i$ is given in Section V-A.

B. Pedestrian Classification

For pedestrian classification, our goal is to 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

$$\omega_i = \underset{\omega_j}{\operatorname{argmax}} P(\omega_j|\mathbf{x}_i). \tag{2}$$

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

$$P(\omega_0|\mathbf{x}_i) = \sum_k P(\Psi_k|\mathbf{x}_i) P(\omega_0|\Psi_k, \mathbf{x}_i) \tag{3}$$

$$\approx \sum_k w_k(\mathbf{x}_i) \mathbf{H}_k(\mathbf{x}_i). \tag{4}$$

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 a sample-dependent gating function $w_k(\mathbf{x}_i)$, with $0 \leq w_k(\mathbf{x}_i) \leq 1$ and $\sum_k w_k(\mathbf{x}_i) = 1$, as defined in (15), in Section IV-D.

$P(\omega_0|\Psi_k, \mathbf{x}_i)$ 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 $\mathbf{H}_k$. The classifier outputs $\mathbf{H}_k(\mathbf{x}_i)$ can be seen as approximation of the cluster-specific posterior probabilities $P(\omega_0|\Psi_k, \mathbf{x}_i)$.

C. Multimodality/Multifeature Expert Classifiers

Given our pose-specific Mixture-of-Experts formulation (4), we model the pose-specific expert classifiers $\mathbf{H}_k(\mathbf{x}_i)$ in terms of our multimodality dataset (intensity, depth, and flow). We extend the Mixture-of-Experts formulation by introducing individual classifiers for each modality m

$$\mathbf{H}_k(\mathbf{x}_i) = \sum_m v_k^m \mathbf{I}_k^m(\mathbf{x}_i^m). \tag{5}$$

In this formulation, $\mathbf{I}_k^m(\mathbf{x}_i^m)$ denotes a local expert classifier for the kth fuzzy pose cluster, which is represented in terms of the mth modality. v_k^m represents a pose- and modality-dependent weight.

Within each modality, we further introduce another level of expert classifiers, in that multiple feature sets f are considered. Following a similar Mixture-of-Experts principle, $\mathbf{I}_k^m(\mathbf{x}_i^m)$ is given by

$$\mathbf{I}_k^m(\mathbf{x}_i^m) = \sum_f u_k^{m,f} \mathbf{J}_k^{m,f}(\mathbf{x}_i^{m,f}) \tag{6}$$

$\mathbf{J}_k^{m,f}(\mathbf{x}_i^{m,f})$ represents a pose-, modality-, and feature-specific expert classifier with an associated weight $u_k^{m,f}$.

Plugging (5) and (6) into (4), we approximate $P(\omega_0|\mathbf{x}_i)$, the posterior probability that a given sample is a pedestrian, using our multilevel Mixture-of-Experts model as

$$P(\omega_0|\mathbf{x}_i) \tag{7}$$

$$\approx \sum_k w_k(\mathbf{x}_i) \left(\sum_m v_k^m \left(\sum_f u_k^{m,f} \mathbf{J}_k^{m,f}(\mathbf{x}_i^{m,f}) \right) \right) \tag{8}$$

$$= \sum_k w_k(\mathbf{x}_i) \left(\sum_m \sum_f v_k^m u_k^{m,f} \mathbf{J}_k^{m,f}(\mathbf{x}_i^{m,f}) \right) \tag{9}$$

$$= \sum_k w_k(\mathbf{x}_i) \left(\sum_{m,f} s_k^{m,f} \mathbf{J}_k^{m,f}(\mathbf{x}_i^{m,f}) \right),$$

$$\text{with } s_k^{m,f} = v_k^m \, u_k^{m,f} \text{ and } \sum_{m,f} s_k^{m,f} = 1. \tag{10}$$

As expert classifiers $\mathbf{J}_k^{m,f}$, we use pattern classifiers which are learned on the training set using data from the corresponding modality/feature only. Given K fuzzy pose clusters, M modalities, and F features, we train $K \times M \times F$ classifiers $\mathbf{J}_k^{m,f}$ on the full training set $\mathscr{D}$ to discriminate between the pedestrian and the nonpedestrian 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.

In principle, the proposed framework is independent from the actual discriminative models used [10]. We only require example-dependent weights during training and that the classifier outputs (decision value) relate to an estimate of posterior probability. For neural networks, example-dependent weights are incorporated using a weighted random sampling step to select the examples that are presented to the neural network during each learning iteration. In case of support vector machines, the approach of [3] can be used. In the limit of infinite data, the outputs of many state-of-the-art classifiers can be converted to an estimate of posterior probabilities [25], [42]. We use this in our experiments.

We compute $s_k^{m,f}$, the weights to the individual expert classifiers, by interpreting $\sum_{m,f} s_k^{m,f} \mathbf{J}_k^{m,f}(\mathbf{x}_i^{m,f}))$ [see (10)] as a dot-product in the $m \times f$-dimensional space of expert classifier posterior probabilities. To determine the weights $s_k^{m,f}$, we train a linear support vector machine (linSVM) $\mathbf{F}_k$ in the expert posterior space. With the linSVM bias term constrained to be zero [14], its decision function equals a dot-product

$$\mathbf{F}_k(\mathbf{x}_i) = \sum_{m,f} s_k^{m,f} \mathbf{J}_k^{m,f}(\mathbf{x}_i^{m,f})) \tag{11}$$

$$= \vec{\mathbf{s}} \cdot \vec{\mathbf{J}}(\mathbf{x}_i). \tag{12}$$

Inserting (11) into (10) then yields

$$P(\omega_0|\mathbf{x}_i) \approx \sum_k w_k(\mathbf{x}_i) \mathbf{F}_k(\mathbf{x}_i). \tag{13}$$

D. 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 (3). Note, that this prior is not a fixed prior, but depends on the sample $\mathbf{x}_i$ itself. As such, it represents the gating of the proposed Mixture-of-Experts architecture.

At this point, information from other cues besides texture (on which the discriminative models $\mathbf{H}_k$ are based) can be incorporated into our framework in a probabilistic manner. We propose to model cluster priors using a Bayesian approach as

$$P(\Psi_k|\mathbf{x}_i) = \frac{p(\mathbf{x}_i|\Psi_k)P(\Psi_k)}{\sum_l p(\mathbf{x}_i|\Psi_l)P(\Psi_l)}. \tag{14}$$

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 [5], or shape [19]. Likelihood estimation can be performed via histogramming on training data or fitting parametric models [19].

Here, we utilize shape cues to compute priors $P(\Psi_k|\mathbf{x}_i)$ for the membership of a sample $\mathbf{x}_i$ to a certain cluster Ψ_k: 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 (14), sample-dependent shape-based priors for cluster Ψ_k are approximated as

$$w_k(\mathbf{x}_i) = \frac{p(D_k(\mathbf{x}_i)|\Psi_k)P(\Psi_k)}{\sum_l p(D_l(\mathbf{x}_i)|\Psi_l)P(\Psi_l)} \approx P\left(\Psi_k|\mathbf{x}_i\right). \tag{15}$$

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

$$p(D_k(\mathbf{x}_i)|\Psi_k) \propto \lambda_k e^{-\lambda_k D_k(\mathbf{x}_i)}. \tag{16}$$

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

V. Experimental Setup

A. Dataset and Evaluation Methodology

The proposed multilevel Mixture-of-Experts framework is tested in experiments on pedestrian classification. We choose the application of pedestrian classification in complex urban traffic as an experimental testbed, since it is arguably one of the most challenging problems around. Because we require multicue (intensity, dense stereo, dense optical flow) training and test samples, we cannot use most established datasets for benchmarking, e.g., [4], [7], [10], [36]. The dataset introduced by [13] includes appearance and binocular image data, however actual depth maps and optical flow are not provided by the authors. While depth maps and flow images can be computed by other authors using this data, it is unclear, to what extent observed performance differences may result from different algorithms used to compute depth and motion data. The authors of [5], for example, demonstrated that the false positives at equal detection rate levels could be reduced by a factor of three, simply by exchanging the method of optical flow estimation. Moreover, the more sophisticated and visually better flow estimator resulted in worse classification performance [5]. Further, the dataset of [13] lacks realism given our experimental setup (urban traffic), since it has been captured at walking speeds on urban sidewalks.

Our experiments involve the recently introduced *Daimler Multi-Cue, Occluded Pedestrian Classification Benchmark* [9] (we do not use the partially occluded pedestrians additionally present in this dataset) which is publicly available to noncommercial entities for research purposes.[1] This dataset is captured from a moving vehicle in complex urban traffic. We provide gray-level intensity data, as well as precomputed dense depth maps and dense optical flow images, to eliminate any effects arising from differences in the computation of the latter.

Recently, an independently developed approach combining intensity, motion, and depth was presented in [52]. However, the dataset used in [52] is only partly publicly available (the training data is not public).

Performance evaluation of pedestrian classifiers can be performed using a per-image measure (detection context) or a per-window measure (classification context). Dollar *et al.* [7] consider the per-window evaluation for sliding-window detectors flawed, since auxiliary effects, such as grid granularity or nonmaximum suppression, are not taken into account. They mention as an additional pitfall the use of incorrectly cropped samples which skews performance due to boundary artifacts. We agree with Dollar *et al.* [7] that per-image evaluation should be the preferred methodology for the evaluation of (monocular) sliding-window or interest-point-based detectors [10], [23]. Images should be cropped in such a way to avoid boundary artifacts.

However, we do not consider the per-window evaluation measure as inherently flawed. Both evaluation setups have their justification, depending on the application context. Most real-world systems integrate several modules; they do not follow a brute-force sliding-window detection scheme, but use a preprocessing step to determine initial pedestrian location hypotheses for both enhanced performance and computational efficiency, e.g., using background subtraction [34], shape [12], [19], stereo [13], [19], [37], motion [12], or nonvision sensors, such as radar or lidar [16]. As a result, the remaining object hypotheses are not random subwindows, but contain a meaningful structure that resembles pedestrians in some aspect. Further, the number of hypotheses per image is greatly reduced (up to a factor of 10 000) compared with dense subwindow scanning, resulting in a more even ratio between pedestrian and nonpedestrian samples. In this application context, per-window evaluation should be the preferred method, since it more closely resembles the actual system setup.

[1][Online]. Available: http://www.science.uva.nl/research/isla/downloads/pedestrians/index.html

TABLE I
TRAINING AND TEST SET STATISTICS

	Pedestrians (labeled)	Pedestrians (jittered)	Non-Pedestrians
Training Set	6514	52112	32465
Test Set	3201	25608	16235

TABLE II
EXPERT WEIGHTS $s_k^{m,f}$ FOR FEATURES AND MODALITIES, ESTIMATED BY A LINEAR SVM ON THE TRAINING SET

	Intensity	Depth	Motion
HOG	0.27	0.14	0.08
LBP	0.24	0.11	0.16

Our training and test samples consist of manually labeled pedestrian and nonpedestrian bounding boxes in images captured from a vehicle-mounted calibrated stereo camera rig in an urban environment. For each manually labeled pedestrian, we create additional samples by geometric jittering. Nonpedestrian samples result from a pedestrian shape-detection preprocessing step [18] with a relaxed threshold setting (to not include largely uniform image patches, such as road surface or sky), as well as ground-plane constraints and prior knowledge about pedestrian geometry, i.e., containing a bias towards more "difficult" patterns, weakly resembling pedestrians in geometry and structure. Note that this selection strategy has already been performed for both the provided training and test data, i.e., it is not required to be implemented to reproduce and compare to the results presented in this paper.

Training and test samples have a resolution of 48×96 pixels with a 12-pixel border around the pedestrians; there is no artificial extension of the border (padding, mirroring) in our data. Dense stereo is computed using the semi-global matching algorithm [22]. To compute dense optical flow, we use the method of [54]. See Table I and Fig. 3 for an overview of the dataset.

We consider $K = 4$ view-related clusters Ψ_k, roughly corresponding to similarity in appearance to front, left, back and right views of pedestrians. We use the approximated cluster prior probability (see Section IV-D) as cluster membership weights for training

$$z_i^k = w_k(\mathbf{x}_i) \approx P(\Psi_k|\mathbf{x}_i). \tag{17}$$

To compute $w_k(\mathbf{x}_i)$, a set of 10 946 shape templates corresponding to clusters Ψ_k is used according to the methods outlined in Section IV-D.

B. Feature Extraction and Classification

Regarding features for our multicue classifiers, we choose histograms of oriented gradients (HOG) [4] and cell-structured local binary patterns (LBP) with uniformity constraints [39], [53] out of many possible feature sets [7], [10], [36]. The motivation for this choice is twofold. First, HOG and LBP are complementary in the sense that HOGs are gradient-based whereas LBPs are texture-based features. HOGs are sensitive to noisy background edges which often occur in cluttered backgrounds. LBPs can filter out background noise using uniformity constraints, see [53]. Second, HOG and LBP features are still among the best performing (and most popular) feature sets available [7], [10], [53].

We follow [4] and compute histograms of oriented gradients with nine orientation bins and 8×8 pixel cells, accumulated to overlapping 16×16 pixel blocks with a spatial shift of eight pixels. HOG features are computed using the implementation provided by [4]. To compute cell-structured LBPs, we adopt the terminology and method of [53] and compute L1-sqrt normalized $\mathrm{LBP}^2_{8,1}$ features, using 8×8 pixel cells and a maximum number of 0–1 transitions of 2. The resulting feature dimensionality is 1980 for HOG and 4248 for LBP. Note that the same HOG and LBP feature set is extracted from intensity, dense stereo and dense flow images.

For classification, we employ multilayer perceptrons (MLP) with one hidden layer consisting of eight neurons with sigmoidal transfer functions, trained stochastically using the online error back-propagation algorithm. We utilize the *FANN* library for MLP training [38]. Compared with the popular linSVMs, MLPs provide nonlinear decision boundaries which usually improve performance, see [36]. The training of nonlinear support vector machines was practically infeasible, given our large datasets.

Expert classifier weights $s_k^{m,f}$ [see (10) and (11)] are computed using the linear SVM approach given in Section IV-C, applied to the training set. We utilize the *LIBLINEAR* library for linear SVM training [14]. The actual weights for individual features and modalities are listed in Table II.

We reiterate that the proposed framework is independent from the actual feature set and discriminative models used. We encourage the scientific community to present results of other feature-classifier combinations on our multicue data.

VI. EXPERIMENTS

Our experiments are designed to evaluate the different levels of our proposed Mixture-of-Experts framework [see Fig. 1(a)], both in isolation and in combination, to quantify the contribution of the individual cues to the overall performance. After presenting the experimental results for pedestrian classification in terms of ROC performance, we analyze the correlation of classifier outputs in different modalities/features to gain further insight into the observed performance.

A. Pose-Level Mixture-of-Experts

In our first experiment, we evaluate the benefit of our Mixture-of-Experts architecture on pose-level only. For that, we compare the proposed pose-specific mixture architecture to single "monolithic" classifiers trained on the whole dataset irrespective of view. We do not consider multimodality or multifeature classifiers yet. For this experiment, we utilize HOG and LBP features separately, operating in the intensity domain only. Regarding classifiers, we compare linear support vector machines (linSVM) to multilayer perceptrons (MLPs). Note that the monolithic HOG/linSVM approach corresponds to the method proposed by Dalal and Triggs [4]. Results are shown in Fig. 4(a) for HOG and in Fig. 4(b) for LBP features.

Irrespective of the employed feature set, the pose-level mixture classifiers perform better than the corresponding monolithic

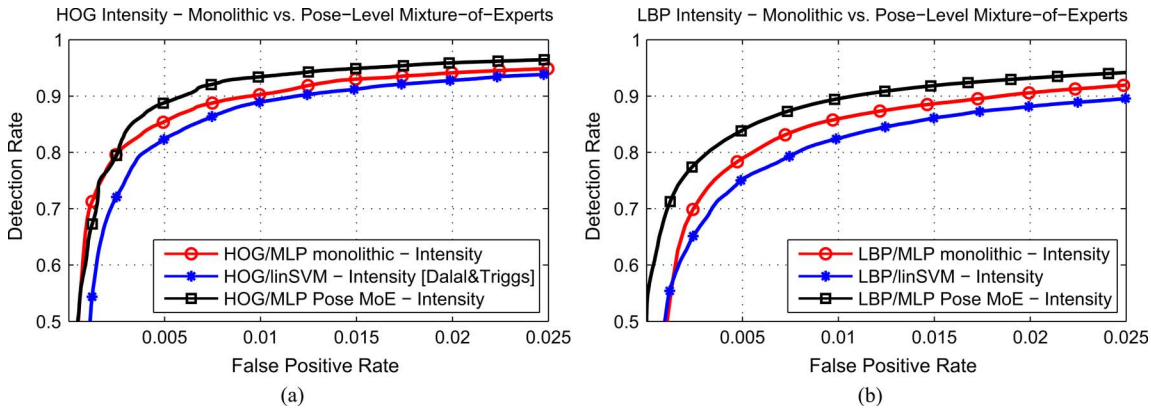

Fig. 4. Pose-level Mixture-of-Experts versus monolithic classifier. (a) HOG features in intensity modality. (b) LBP features in intensity modality.

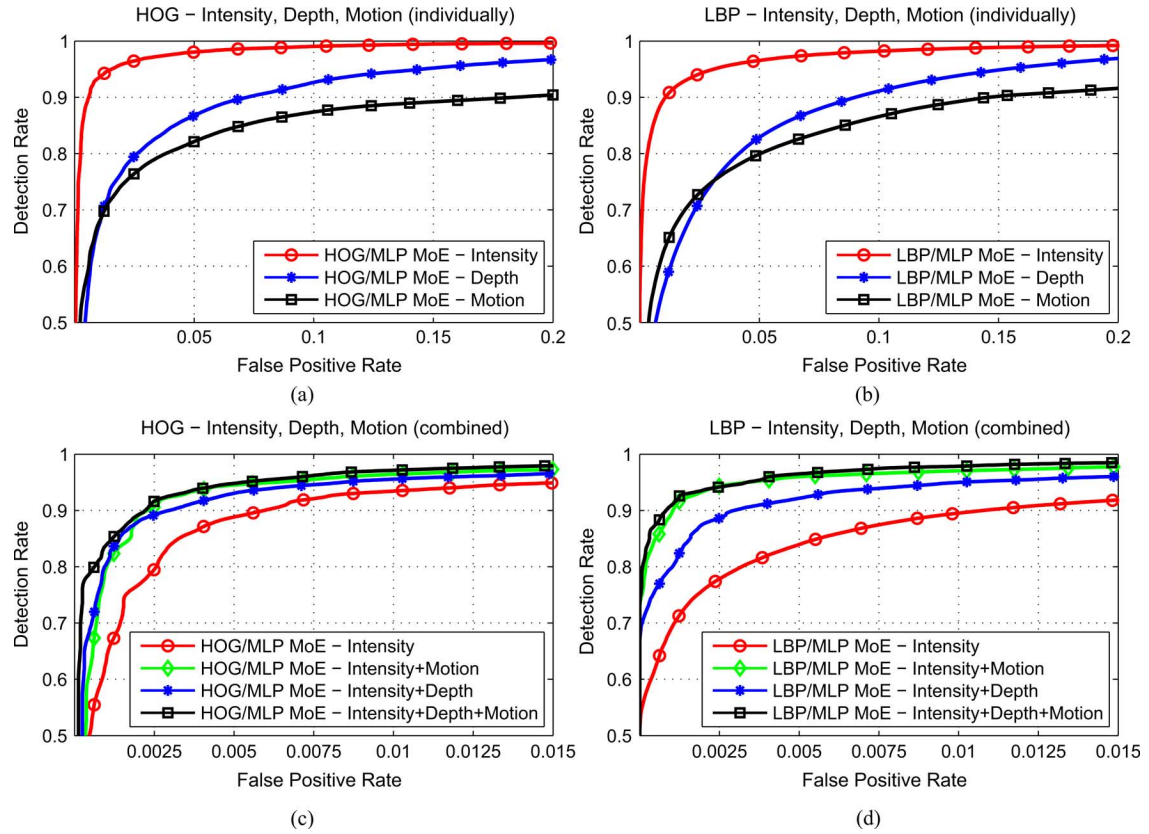

Fig. 5. Modality-level Mixture-of-Experts. Individual classification performance of (a) HOG and (b) LBP features in intensity, depth, and motion modality. Combined classification performance of (c) HOG and (d) LBP features in intensity, depth, and motion modality. Note the different scaling on the x-axis.

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 pose-level mixture classifier is up to a factor of two in reduction of false positives at the same detection rate. Further, multilayer perceptrons outperform linear support vector machines, because of their nonlinearities in decision space. Except for some experiments in Section VI-E, we utilize pose-level Mixture-of-Experts classification throughout the following experiments.

B. Modality-Level Mixture-of-Experts

In our second experiment, we evaluate the performance of modality-level classifiers, as presented in Section IV-C, compared with intensity-only classifiers. Pose-level mixtures are also used, that is, the first two levels of our framework [see Fig. 1(a)] are in place in this experiment. Performance is evaluated for both HOG and LBP features individually. In each feature-space, we first evaluate all modalities separately and incrementally add depth and motion to the baseline intensity cue. Results are shown in Fig. 5(a) and (c) for HOG and in Fig. 5(b) and (d) for LBP features.

The relative performance of classifiers trained on intensity, depth and motion features only is consistent across the two different feature spaces, cf. Fig. 5(a) (HOG) versus Fig. 5(b) (LBP). Classifiers in the intensity modality have the best performance, by a large margin. In depth and motion modalities, performance is similar for both feature sets with depth features performing better then motion features at higher false positive rates and worse at lower false positive rates. Note, that these performance relations are also apparent in the individual expert classifier weights; see Table II.

Fig. 5(c) and (d) show the effect of incrementally adding depth and motion to the intensity modality. Here, the best performance is reached when all modalities are taken into account. However, the observable performance boosts are different for HOG compared with LBP features. The HOG classifier using intensity, depth, and motion has approximately a factor of four less false positives than a comparable HOG classifier using intensity only [see Fig. 5(c)]. From Fig. 5(d) we observe, that in the case of LBP features, the performance boost resulting from utilizing all modalities versus intensity-only is approximately a factor of 12 in reduction of false positives at equal detection rates.

C. Feature-Level Mixture-of-Experts

Similar to analyzing the effect of modality-level Mixture-of-Experts , we now evaluate the effect of feature-level Mixture-of-Experts. To that extent, we combine pose-level Mixture-of-Experts with feature-level Mixture-of-Experts and evaluate the performance of the multifeature approach in all three modalities, i.e., intensity, depth, motion, individually. Recalling our framework architecture [see Fig. 1(a)], this corresponds to having levels 1 (pose) and 3 (features) in place. Results are given in Fig. 6(a) (intensity), Fig. 6(b) (depth), and Fig. 6(c) (motion).

In all modalities, one can observe that combining HOG and LBP improves performance over using both features individually. The largest performance boost coming from the feature-level Mixture-of-Experts exists in the intensity modality. Here, the combined HOG+LBP classifier has up to a factor of four less false positives than the HOG classifier, which in turn outperforms the LBP classifier at higher detection rates. In depth and motion modalities, the corresponding performance boosts amount to factors of 2 (motion) and 1.5 (depth) at equal detection rate levels. Compared with the performance improvement obtained by combining different modalities, as shown in Section VI-B, the effect of feature-level Mixture-of-Experts is less pronounced, but still significant.

D. Multilevel Mixture-of-Experts

We now evaluate the performance of our full multilevel Mixture-of-Experts framework combining pose-, modality-, and feature-level expert classifiers. As baseline performance, the monolithic (i.e., no delineation of classifiers at pose-level) HOG/linSVM approach of [4], as well the best performing variants from the previous two experiments are utilized: modality-level Mixture-of-Experts using LBP/MLP in intensity, depth and motion (see Section VI-B) as well as feature-level

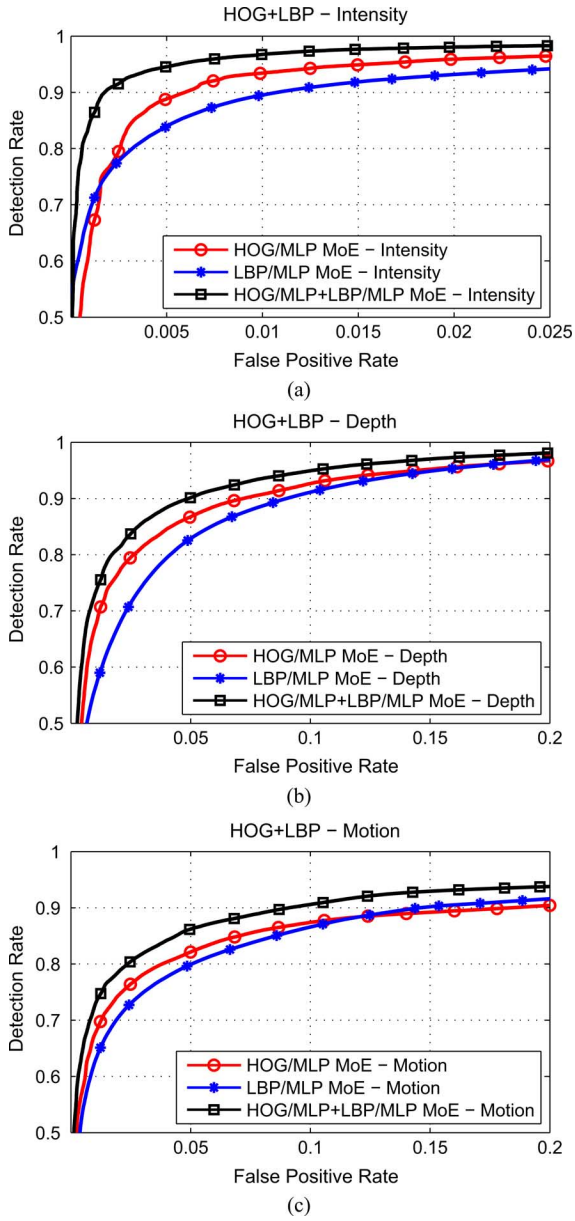

Fig. 6. Feature-level Mixture-of-Experts. Individual classification performance of HOG, LBP, and HOG+LBP in (a) intensity, (b) depth, and (c) motion modality. Note the different scaling on the x-axis.

Mixture-of-Experts using HOG+LBP Mixture-of-Experts in intensity domain only (see Section VI-C).

ROC performance is given in Fig. 7. We observe that our combined multilevel Mixture-of-Experts approach significantly outperforms both variants using either modality-level or feature-level fusion, as well as the state-of-the-art monolithic HOG/linSVM approach [4]. To quantify performance, Table III lists the false positive rates of all approaches shown in Fig. 7 using a detection rate of 90% as a common reference point. We further indicate the resulting reduction in false positives, in comparison to the monolithic HOG/linSVM classifier as baseline.

If we combine experts on pose-level with experts on feature-level (HOG/MLP + LBP/MLP, intensity modality), we achieve a reduction in false positive of more than a factor of 6 over the Dalal and Triggs HOG/linSVM approach. The use of pose-level

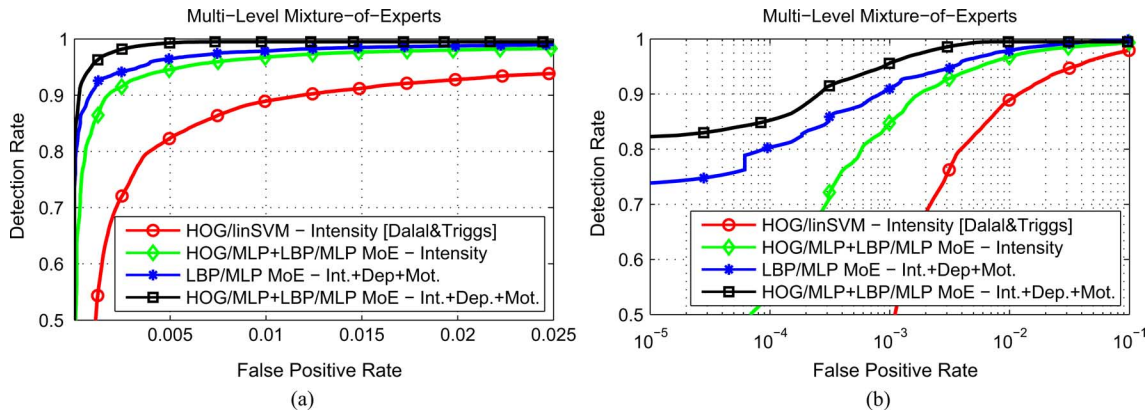

Fig. 7. Performance overview. (a) Monolithic HOG classifier in intensity domain, best feature-level MoE (HOG+LBP, intensity), best modality-level MoE (LBP, intensity+depth+motion), multilevel MoE (HOG+LBP, intensity+depth+motion). (b) Logarithmic plot of (a), focusing on low false-positive rates.

TABLE III
PERFORMANCE OF APPROACHES IN FIG. 7 USING 90% DETECTION RATE AS A COMMON REFERENCE POINT

	FP Rate	Factor
HOG/linSVM - Intensity [Dalal & Triggs]	1.1e-2	1
HOG+LBP/MLP MoE - Intensity	1.7e-3	6.4
LBP/MLP MoE - Int.+Dep.+Mot.	8.2e-4	13.4
HOG+LBP/MLP MoE - Int.+Dep.+Mot.	2.6e-4	42.0

TABLE IV
CORRELATION OF CLASSIFIER OUTPUTS IN (A) DIFFERENT MODALITIES AND (B) DIFFERENT FEATURES

	HOG	LBP
Intensity / Depth	0.21	0.21
Intensity / Motion	0.19	0.01
Depth / Motion	0.25	0.13

(a)

	HOG / LBP
Intensity	0.52
Depth	0.61
Motion	0.62

(b)

and modality-level experts (LBP/MLP, intensity+depth+motion modalities) reduces false positives by more than a factor of 13 compared with the HOG/linSVM baseline. Our full multilevel Mixture-of-Experts approach (HOG/MLP + LBP/MLP, intensity+depth+motion modalities) further boost performance up to a reduction in false positives by a factor of 42.

The results clearly show the benefit of our integrated multilevel architecture. Additionally, we observe that the combination of different modalities attributes more to the overall performance, than the use of multiple features within a single modality. Given that most recent research has focused on developing yet another feature to be used in the intensity domain, multicue classification approaches seem to be a promising direction for future research in the domain of object classification to boost overall performance.

To gain further insight, we compute the correlation of classifier outputs (decision values) for the individual modality/feature expert classifiers, computed for pedestrian and nonpedestrian samples individually and then averaged over the two classes, see Table IV. The correlation analysis shows, that classifier outputs are far less correlated across different modalities (Table IV-B) than across different features (Table IV-A). Here, the less correlated two modalities/features are, the larger the benefits obtained in classification performance (see Figs. 5 and 6).

E. Classifier Fusion

In our final experiments, we compare our multilevel Mixture-of-Experts fusion approach to other techniques for classifier fusion. First, we analyze fusion approaches involving a combination of different classifiers in other ways than our Mixture-of-Experts framework. Second, we compare our approach against a single classifier using a joint feature space which consists of all features in all modalities L^2-normalized and concatenated into a single feature vector [56]. Given our feature setup as presented in Section V-B, the total dimensionality of the joint feature space is 18 684. For comparison, the performance of the Dalal and Triggs HOG/linSVM baseline [4] is also given. Results are shown in Fig. 8(a) for the multiclassifier fusion and in Fig. 8(b) for the joint space fusion approaches.

The multiclassifier fusion approaches (entitled "Uniform Sum", "Product" and "Sugeno Fuzzy Integral") involve individual classifiers for each feature (HOG and LBP) and modality (intensity, depth and motion). Altogether, there are six classifiers to be combined, using the sum and product of the individual decision values [26], as well as a fuzzy integration using Sugeno integrals [40]. Fuzzy integration involves treating the individual classifier outputs as a fuzzy set and aggregating them into a single value using the Sugeno integral. While those approaches improve performance over the state-of-the-art Dalal and Triggs HOG/linSVM classifier [4], our multilevel Mixture-of-Experts classifier has a much better performance. This clearly shows the benefit of gating on pose-level [see (4)] and the learned classifier combination weights in (12).

In terms of joint space approaches, we train both a MLP and a linSVM in the enlarged 18 684-dimensional joint feature space (training a nonlinear SVM was not feasible given our large dataset). While one could expect the MLP to improve performance over the linSVM, due to the nonlinear decision boundary, our results paint a different performance picture. The MLP classifier is outperformed by the linSVM by a significant margin. We attribute this to the so-called "curse of dimensionality," e.g., [8], which relates the number of free parameters in a classifier (as given by feature space dimensionality) to the amount of available training samples. As a rule of thumb, the number of training samples should be a factor-of-10 larger than the number of free parameters to be estimated during

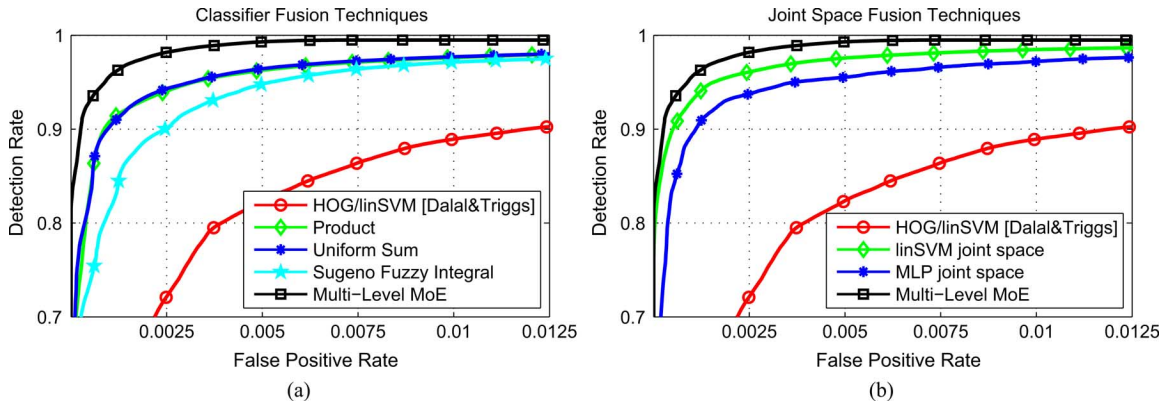

Fig. 8. Performance of different classifier fusion techniques. (a) Multiclassifier fusion. (b) Joint feature space with single classifiers.

training [8]. This rule is severely violated in case of the MLP in the 18 684-dimensional joint feature space with 149 489 free parameters and 84 577 training samples. The linSVM can better cope with the higher dimensionality given its maximum-margin constraint at the core which is less susceptible to overfitting effects in high-dimensional spaces. Still, our multilevel Mixture-of-Experts framework using MLPs as expert classifiers outperforms the joint space linSVM. We can afford to use more complex subclassifiers in our model, since each MLP is an expert in a lower dimensional modality/feature subspace, weighted by the contribution of the shape cues.

VII. Discussion

We obtained a significant boost in pedestrian classification performance from the use of multiple modalities and features in a Mixture-of-Experts setting. Our experiments show that the largest performance gain stems from the combination of intensity features with depth and motion features. We expect the use of additional modalities, e.g., far-infrared (FIR) [31], to further increase performance. Multimodality classifiers particularly outperform multifeature classifiers in a single modality. However, modalities and features are orthogonal, so that a combined multimodality/multifeature approach can further boost performance.

In this work, we did not heavily optimize the feature sets with regard to the different modalities. Instead, we transferred general knowledge and experience from the behavior of features and classifiers in the intensity domain to the depth and motion domains. At this point, it is not clear if (and how) additional modification and adaptation of the feature sets to the different characteristics found in depth and motion data (see Section IV-A) can further improve performance. While the HOG/MLP classifier outperforms the LBP/MLP classifier in all modalities in our experiments, this may not be generally true. See, for example, [43], where the relative order of feature/classifier performance reverses with respect to intensity and depth.

Orthogonal to the improvements presented in this paper are benefits resulting from an increased training set [10], [36]. In the intensity domain, feature-classifier combinations respond differently to an increased training set (in both size and dimensionality), e.g., in terms of classifier complexity, discriminative power, practical feasibility and saturation effects [10], [36]. It is currently unknown to what extent similar (or different) effects are present for features and classifiers in other modalities.

Recent work analyzed the dependence of classification performance and pedestrian image size (as a proxy for distance to the camera) in the intensity domain [7]. Results show significant relative performance differences of the evaluated classifiers across multiple scales. Similar effects may also be found in depth and motion features, particularly since depth and motion measurements tend to get noisy at larger distances to the camera. In case of stereo vision, the range of measurements is further limited by the camera setup.

Certainly, more research is necessary to fully explore the benefits of multimodality/multifeature classification. For that purpose, we provide our multicue dataset not only as a means for benchmarking but also to stimulate further research on the issues mentioned above.

VIII. Conclusion

This paper presented a probabilistic multilevel Mixture-of-Experts framework involving a view-related and sample-dependent combination of multicue/multifeature pedestrian classifiers. We use highly complementary Chamfer distance, HOG, and LBP features that are extracted from intensity, dense depth and dense flow data. The pose-specific Mixture-of-Experts formulation, which divides the complex pedestrian classification problem into better manageable subproblems, is feature- and classifier-independent, practically feasible and does not suffer from overfitting effects in high-dimensional spaces.

Results show a significant performance boost of up to a factor of 42 in reduction of false positives at constant detection rates over a state-of-the-art intensity-only classifier using HOG features and linear SVM classification. The observed performance improvements stem both from the fuzzy subdivision of our data in terms of pose and the combination of multiple features and modalities. In our experiments, we identified the use of multiple modalities as the most benefiting factor which is confirmed by a correlation analysis. We make our multicue dataset publicly available for benchmarking purposes and to stimulate further research to address open issues with regard to multicue/multifeature classification.

ACKNOWLEDGMENT

The authors would like to thank Prof. C. Schnörr (Image and Pattern Analysis Group, University of Heidelberg, Germany), who provided helpful comments and discussions.

REFERENCES

[1] S. Agarwal, A. Awan, and D. Roth, "Learning to detect objects in images via a sparse, part-based representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 11, pp. 1475–1490, Nov. 2004.

[2] M. Bajracharya, B. Moghaddam, A. Howard, S. Brennan, and L. Matthies, "A fast stereo-based system for detecting and tracking pedestrians from a moving vehicle," *Int. J. Robot. Res.*, vol. 28, pp. 1466–1485, 2009.

[3] U. Brefeld, P. Geibel, and F. Wysotzki, "Support vector machines with example dependent costs," in *Proc. Eur. Conf. Mach. Learning (ECML)*, 2003, pp. 23–34.

[4] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. Int. Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 886–893.

[5] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 428–441.

[6] P. Dollar, B. Babenko, S. Belongie, P. Perona, and Z. Tu, "Multiple component learning for object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 211–224.

[7] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *Proc. Int. Conf. Comput. Vis. Pattern Recognit.*, 2009.

[8] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley Interscience, 2000.

[9] M. Enzweiler, A. Eigenstetter, B. Schiele, and D. M. Gavrila, "Multi-cue pedestrian classification with partial occlusion handling," in *Proc. Int. Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 990–997.

[10] M. Enzweiler and D. M. Gavrila, "Monocular pedestrian detection: Survey and experiments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 12, pp. 2179–2195, Dec. 2009.

[11] M. Enzweiler and D. M. Gavrila, "Integrated pedestrian classification and orientation estimation," in *Proc. Int. Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 982–989.

[12] M. Enzweiler, P. Kanter, and D. M. Gavrila, "Monocular pedestrian recognition using motion parallax," in *Proc. IEEE Intell. Vehicles Symp.*, 2008, pp. 792–797.

[13] A. Ess, B. Leibe, and L. van Gool, "Depth and appearance for mobile scene analysis," in *Proc. Int. Conf. Comput. Vis.*, 2007, pp. 1–8.

[14] R. E. Fan, K. W. Chang, C. J. Hsieh, X. R. Wang, and C. J. Lin, "Liblinear: A library for large linear classification," *J. Mach. Learning Res.*, vol. 9, pp. 1871–1874, 2008.

[15] P. Felzenszwalb, R. Girshick, and D. M. Ramanan, "Object detection with discriminatively trained part based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[16] T. Gandhi and M. M. Trivedi, "Pedestrian protection systems: Issues, survey and challenges," *IEEE Trans. Intell. Transp. Syst.*, vol. 8, no. 3, pp. 413–430, Sep. 2007.

[17] D. M. Gavrila, "The visual analysis of human movement: A survey," *Comput. Vis. Image Understanding*, vol. 73, no. 1, pp. 82–98, 1999.

[18] D. M. Gavrila, "A Bayesian, exemplar-based approach to hierarchical shape matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 8, pp. 1408–1421, Aug. 2007.

[19] D. M. Gavrila and S. Munder, "Multi-cue pedestrian detection and tracking from a moving vehicle," *Int. J. Comput. Vis.*, vol. 73, no. 1, pp. 41–59, 2007.

[20] D. Geronimo, A. M. Lopez, A. D. Sappa, and T. Graf, "Survey on pedestrian detection for advanced driver assistance systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 7, pp. 1239–1258, Jul. 2010.

[21] B. E. Goldstein, *Sensation and Perception*, 6th ed. Belmont, CA: Wadsworth, 2002.

[22] H. Hirschmüller, "Stereo processing by semi-global matching and mutual information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 328–341, Feb. 2008.

[23] M. Hussein, F. Porikli, and L. Davis, "A comprehensive evaluation framework and a comparative study for human detectors," *IEEE Trans. Intell. Transp. Syst.*, vol. 10, no. 3, pp. 417–427, Sep. 2009.

[24] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural Computation*, vol. 3, no. 1, pp. 79–87, 1991.

[25] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, pp. 4–37, Jan. 2000.

[26] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 226–239, Mar. 1998.

[27] N. Kriegeskorte, B. Sorger, M. Naumer, J. Schwarzbach, E. van den Boogert, W. Hussy, and R. Goebel, "Human cortical object recognition from a visual motion flowfield," *J. Neurosci.*, vol. 23, no. 4, pp. 1451–1463, 2003.

[28] B. Leibe, N. Cornelis, K. Cornelis, and L. V. Gool, "Dynamic 3-D scene analysis from a moving vehicle," in *Proc. Int. Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

[29] B. Leibe, E. Seemann, and B. Schiele, "Pedestrian detection in crowded scenes," in *Proc. Int. Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 878–885.

[30] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[31] M. Mählisch, M. Oberländer, O. Löhlein, D. M. Gavrila, and W. Ritter, "A multiple detector approach to low-resolution FIR pedestrian recognition," in *Proc. IEEE Intell. Vehicles Symp.*, 2005, pp. 1–6.

[32] S. Maji, A. C. Berg, and J. Malik, "Classification using intersection kernel support vector machines is efficient," in *Proc. Int. Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[33] K. Mikolajczyk, C. Schmid, and A. Zisserman, "Human detection based on a probabilistic assembly of robust part detectors," in *Proc. Eur. Conf. Comput. Vis.*, 2004, pp. 69–81.

[34] T. B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Comput. Vis. Image Understanding*, vol. 104, no. 2–3, pp. 90–126, 2006.

[35] A. Mohan, C. Papageorgiou, and T. Poggio, "Example-based object detection in images by components," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 4, pp. 349–361, Apr. 2001.

[36] S. Munder and D. M. Gavrila, "An experimental study on pedestrian classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 11, pp. 1863–1868, Nov. 2006.

[37] S. Nedevschi, S. Bota, and C. Tomiuc, "Stereo-based pedestrian detection for collision-avoidance applications," *IEEE Trans. Intell. Transp. Syst.*, vol. 10, no. 3, pp. 380–391, Sep. 2009.

[38] S. Nissen, "Implementation of a fast artificial neural network library (FANN)," Dept. Comput. Sci., Univ. of Copenhagen, Denmark, 2003.

[39] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on feature distributions," *Pattern Recognit.*, vol. 29, pp. 51–59, 1996.

[40] L. Oliveira, U. Nunes, and P. Peixoto, "On exploration of classifier ensemble synergism in pedestrian detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 1, pp. 16–27, Mar. 2010.

[41] C. Papageorgiou and T. Poggio, "A trainable system for object detection," *Int. J. Comput. Vis.*, vol. 38, pp. 15–33, 2000.

[42] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Adv. Large Margin Classifiers*, pp. 61–74, 1999.

[43] M. Rohrbach, M. Enzweiler, and D. M. Gavrila, "High-level fusion of depth and intensity for pedestrian classification," in *Proc. DAGM Symp. Pattern Recognit.*, 2009, pp. 101–110.

[44] W. R. Schwartz, A. Kembhavi, D. Harwood, and L. S. Davis, "Human detection using partial least squares analysis," in *Proc. Int. Conf. Comput. Vis.*, 2009, pp. 24–31.

[45] E. Seemann, M. Fritz, and B. Schiele, "Towards robust pedestrian detection in crowded image sequences," in *Proc. Int. Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

[46] A. Shashua, Y. Gdalyahu, and G. Hayon, "Pedestrian detection for driving assistance systems: Single-frame classification and system level performance," in *Proc. IEEE Intell. Vehicles Symp.*, 2004, pp. 1–6.

[47] A. Talukder and L. Matthies, "Real-time detection of moving objects from moving vehicles using dense stereo and optical flow," in *Proc. IEEE Conf. Intell. Robots Syst. (IROS)*, 2004, pp. 3718–3725.

[48] O. Tuzel, F. Porikli, and P. Meer, "Human detection via classification on riemannian manifolds," in *Proc. Int. Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

[49] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman, "Multiple kernels for object detection," in *Proc. Int. Conf. Comput. Vis.*, 2009, pp. 606–613.

[50] P. Viola, M. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," *Int. J. Comput. Vis.*, vol. 63, no. 2, pp. 153–161, 2005.

[51] S. Walk, N. Majer, K. Schindler, and B. Schiele, "New features and insights for pedestrian detection," in *Proc. Int. Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 1030–1037.
[52] S. Walk, K. Schindler, and B. Schiele, "Disparity statistics for pedestrian detection: Combining appearance, motion and stereo," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 182–195.
[53] X. Wang, T. Han, and S. Yan, "A HOG-LBP human detector with partial occlusion handling," in *Proc. Int. Conf. Comput. Vis.*, 2009, pp. 32–39.
[54] A. Wedel, D. Cremers, T. Pock, and H. Bischof, "Structure- and motion-adaptive regularization for high accuracy optic flow," in *Proc. Int. Conf. Comput. Vis.*, 2009, pp. 1663–1668.
[55] C. Wöhler and J. K. Anlauf, "A time delay neural network algorithm for estimating image-pattern shape and motion," *Image Vis. Computing*, vol. 17, pp. 281–294, 1999.
[56] C. Wojek, S. Walk, and B. Schiele, "Multi-cue onboard pedestrian detection," in *Proc. Int. Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 794–801.
[57] B. Wu and R. Nevatia, "Detection and tracking of multiple, partially occluded humans by Bayesian combination of edgelet based part detectors," *Int. J. Comput. Vis.*, vol. 75, no. 2, pp. 247–266, 2007.
[58] B. Wu and R. Nevatia, "Optimizing discrimination-efficiency tradeoff in integrating heterogeneous local features for object detection," in *Proc. Int. Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.
[59] L. Zhang, B. Wu, and R. Nevatia, "Detection and tracking of multiple humans with extensive pose articulation," in *Proc. Int. Conf. Comput. Vis.*, 2007, pp. 1–8.
[60] Q. Zhu, S. Avidan, M. Ye, and K.-T. Cheng, "Fast human detection using a cascade of histograms of oriented gradients," in *Proc. Int. Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 1491–1498.

Markus Enzweiler received the M.Sc. degree from the University of Ulm, Ulm, Germany, in 2005, and the Ph.D. degree at the University of Heidelberg, Heidelberg, Germany, in 2011, both in computer science.

In 2002 and 2003, he was a Visiting Student Researcher with the Centre for Vision Research, York University, Toronto, ON, Canada. From 2006 to 2010, he was with the Image and Pattern Analysis Group, University of Heidelberg, Heidelberg, Germany. Since 2010, he has been a Research Scientist with Daimler Research, Ulm, Germany. His current research focuses on statistical models of human appearance with application to pedestrian recognition in the domain of intelligent vehicles.

Mr. Enzweiler was the recipient of a Ph.D. scholarship from the Studienstiftung des deutschen Volkes (German National Academic Foundation).

Dariu M. Gavrila received the M.Sc. degree from the Free University, Amsterdam, The Netherlands, in 1990, and the Ph.D. degree from the University of Maryland, College Park, in 1996, both in computer science.

He was a Visiting Researcher with the MIT Media Laboratory in 1996. Since 1997, he has been a Senior Research Scientist with Daimler Research, Ulm, Germany. In 2003, he became a Professor with the Faculty of Science, University of Amsterdam, Amsterdam, The Netherlands, chairing the area of Intelligent Perception Systems (part time). Over the last decade, he has focused on visual systems for detecting human presence and recognizing activity, with application to intelligent vehicles and surveillance. He has authored or coauthored more than 50 papers in this area.

Prof. Gavrila was the recipient of the I/O Award 2007 from the Netherlands Organisation for Scientific Research (NWO).