Inaugural-Dissertation

zur
Erlangung der Doktorwürde
der
Naturwissenschaftlich–Mathematischen Gesamtfakultät
der
Ruprecht–Karls–Universität
Heidelberg

vorgelegt von
M. Comp. Sc. Markus Enzweiler
aus Augsburg

Tag der mündlichen Prüfung: 10.05.2011

Compound Models
for
Vision-Based Pedestrian Recognition

Gutachter:	**Prof. Dr. Christoph Schnörr** Universität Heidelberg
Zweitgutachter:	**Prof. Dr. Dariu M. Gavrila** Universität von Amsterdam

Zusammenfassung

Diese Arbeit beschäftigt sich mit bildgestützter Fußgängererkennung in realen, dynamischen Umgebungen mittels einer bewegten Kamera. Der Arbeitsschwerpunkt liegt nicht auf der Entwicklung neuer Merkmalstypen zur Klassifikation, sondern auf merkmals- und klassifikatorunabhängigen zusammengesetzten Ansätzen. Diese kombinieren komplementäre Informationen aus mehreren bildbasierten Informationsquellen mit dem Ziel einer verbesserten Fußgängererkennungsleistung.

Im Anschluss an die Etablierung einer Basiserkennungsleistung mit Hilfe einer ausführlichen Experimentalstudie im Bereich der monokularen Fußgängererkennung wird der Nutzen mehrerer Merkmale auf Modulebene untersucht. Hierbei wird ein bewegungsbasiertes Konzept zur Aufmerksamkeitssteuerung vorgestellt, welches auf einem wahrscheinlichkeitsbasierten, gelernten Fußgängerbewegungsmodell aufbaut. Dieses Modell dient zur Adaption der Suchbereiche nachgeschalteter form- und texturbasierter Klassifikationsmodule.

Im weiteren Verlauf dieser Arbeit liegt der Schwerpunkt auf der Integration komplementärer Informationen in den eigentlichen Mustererkennungsschritt. In diesem Sinne werden ansichtsspezifische generative Form- und Texturmodelle vorgestellt. Die Kombination dieser generativen Modelle mit diskriminativen Klassifikatoren erfolgt durch die Nutzung generativ erzeugter virtueller Trainingsbeispiele, um die Erkennungleistung der diskriminativen Modelle zu verbessern. Beide Modellarten sind durch Aktives Lernen verbunden, um den Trainingsprozess auf die wichtigsten und informativsten Trainingsbeispiele zu fokussieren.

Des Weiteren wird ein Mixture-of-Experts-System zur Klassifikation vorgeschlagen, welches auf lokalen ansichtsspezifischen Klassifikationsexperten basiert. Diese Experten nutzen mehrere Bildmodalitäten und -merkmale. Als Modalitäten werden Grauwertintensität, Tiefeninformation aus dichtem Stereosehen und Bewegungsinformation aus dichtem optischen Fluss betrachtet. Als Merkmale dienen sowohl formbasierte, gradientenbasierte als auch texturbasierte Merkmale. Gegenüber Methoden, die auf einem gemeinsamen Merkmalsraum beruhen, zeichnet sich das Mixture-of-Experts-Modell durch bessere Erkennungsleistung und bessere praktische Umsetzbarkeit aus.

Zu guter Letzt behandelt diese Arbeit die Erweiterung des Mixture-of-Experts-Modells im Hinblick auf die Behandlung von Teilverdeckungen und

die Schätzung der Körperorientierung der Fußgänger. Das entwickelte Verdeckungsmodell beruht auf der Untersuchung von Diskontinuitäten im Tiefen- und Bewegungsraum, welche durch Teilverdeckungen hervorgerufen werden. Abhängig von den Verdeckungen werden Gewichtungsfaktoren für einzelne Körperteile bestimmt, um die Gesamtentscheidung hauptsächlich auf sichtbare Körperteile zu stützen. Das ansichtsspezifische Mixture-of-Experts-Modell wird ebenfalls zur Schätzung der Dichtefunktion der Körperorientierung eines Fußgängers benutzt, auch hier unter Berücksichtigung von Form- und Texturinformation.

Im Rahmen dieser Arbeit wird besonderer Nachdruck auf ausführliche Systemevaluation gelegt, sowohl im Hinblick auf Evaluationsmethodik als auch unter Zuhilfenahme umfangreicher und anwendungsnaher Datensätze. Mehrere Datensätze werden öffentlich zu Vergleichszwecken zur Verfügung gestellt. Es konnten signifikante Verbesserungen in allen Teilbereichen dieser Arbeit, d.h. Fußgängererkennung, Behandlung von Teilverdeckungen und Schätzung der Körperorientierung, verglichen mit dem heutigen Stand der Technik erreicht werden. Dies gilt insbesondere für die Fußgängererkennungs-leistung; Falscherkennungen wurden bei gleicher Erkennungsrate um deutlich mehr als eine Größenordnung reduziert.

Abstract

This thesis addresses the problem of recognizing pedestrians in video images acquired from a moving camera in real-world cluttered environments. Instead of focusing on the development of novel feature primitives or pattern classifiers, we follow an orthogonal direction and develop feature- and classifier-independent compound techniques which integrate complementary information from multiple image-based sources with the objective of improved pedestrian classification performance.

After establishing a performance baseline in terms of a thorough experimental study on monocular pedestrian recognition, we investigate the use of multiple cues on module-level. A motion-based focus of attention stage is proposed based on a learned probabilistic pedestrian-specific model of motion features. The model is used to generate pedestrian localization hypotheses for subsequent shape- and texture-based classification modules.

In the remainder of this work, we focus on the integration of complementary information directly into the pattern classification step. We present a combination of shape and texture information by means of pose-specific generative shape and texture models. The generative models are integrated with discriminative classification models by utilizing synthesized virtual pedestrian training samples from the former to enhance the classification performance of the latter. Both models are linked using Active Learning to guide the training process towards informative samples.

A multi-level mixture-of-experts classification framework is proposed which involves local pose-specific expert classifiers operating on multiple image modalities and features. In terms of image modalities, we consider gray-level intensity, depth cues derived from dense stereo vision and motion cues arising from dense optical flow. We furthermore employ shape-based, gradient-based and texture-based features. The mixture-of-experts formulation compares favorably to joint space approaches, in view of performance and practical feasibility.

Finally, we extend this mixture-of-experts framework in terms of multi-cue partial occlusion handling and the estimation of pedestrian body orientation. Our occlusion model involves examining occlusion boundaries which manifest in discontinuities in depth and motion space. Occlusion-dependent weights which relate to the visibility of certain body parts focus the decision on unoccluded body components. We further apply the pose-specific

nature of our mixture-of-experts framework towards estimating the density of pedestrian body orientation from single images, again integrating shape and texture information.

Throughout this work, particular emphasis is laid on thorough performance evaluation both regarding methodology and competitive real-world datasets. Several datasets used in this thesis are made publicly available for benchmarking purposes. Our results indicate significant performance boosts over state-of-the-art for all aspects considered in this thesis, i.e. pedestrian recognition, partial occlusion handling and body orientation estimation. The pedestrian recognition performance in particular is considerably advanced; false detections at constant detection rates are reduced by significantly more than an order of magnitude.

Acknowledgements

This PhD thesis would not have been possible without the help and support of many people throughout the last years. First and foremost, I would like to sincerely thank Prof. Dr. Christoph Schnörr and Prof. Dr. Dariu M. Gavrila for guiding this work, providing inspiring ideas and their continuous support on so many different levels. I have benefited a lot from their knowledge, enthusiasm, motivation and dedication to teach me how to conduct scientific research. I am particularly thankful for the freedom to explore things on my own and the opportunity to work in a very inspiring and demanding research environment.

I would like to thank my friends and colleagues at the Image & Pattern Analysis Group for facilitating many fruitful discussions, insightful comments and advice. In particular, I want to thank Christoph Keller for our close collaboration and for making work fun.

Special thanks goes to Daimler R&D for providing some algorithms and unique data for experimental evaluation, as well as agreeing to publish some datasets for benchmarking. Furthermore, I would like to personally thank Dr. Stefan Munder, Prof. Dr. Christian Wöhler, Markus Gressmann, Alexander Barth, Dr. Tilo Schwarz, Dr. Fridtjof Stein, Dr. Martin Fritzsche, Dr. Ulrich Kressel, Dr. Uwe Franke, Stefan Hahn and Prof. Dr. Bernt Schiele for their collaboration and tremendous support.

Some parts of this work have been supported by students under my supervision. Thanks to Angela Eigenstetter, Zuzana Sulcova, Ina Bayer, Mia and Pascal Kanter, Marcus Rohrbach, Manuel Kugelmann, Wolfgang Schulz, Winn Voravuthikunchai and Mohamed Omran for their contributions and friendship.

I sincerely acknowledge the generous support of the "Studienstiftung des deutschen Volkes (German National Academic Foundation)" in terms of graduate and PhD scholarships.

Last but not least, I am profoundly thankful to my family for the continuous support, encouragement and the chance to pursue my own goals. My deepest thanks go to Simona for always being there for me, no matter when or where, at times across continents and oceans. Thanks for understanding, your encouragement and patience, for putting up with me and making me feel comfortable about doing all the things I have done.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Perception is one of the key elements for humans to survive in a dynamic environment. It allows us to extract information from our surroundings, interpret and understand the situation in order to act and interact appropriately. Vision has developed to be the primary and most important sensory cue in the human system. The perceptual process is taken to be organized in a cycle, the perception-action cycle, cf. [60]. A simplified version of the perception-action cycle is shown in Figure 1.1.

At the beginning of the perceptual process lies the stimulus from the environment which is received by the sensory system, e.g. the retina. The stimulus is processed by the neural system and generates perception. In case of visual stimuli, perception relates to the conscious experience of seeing *something*

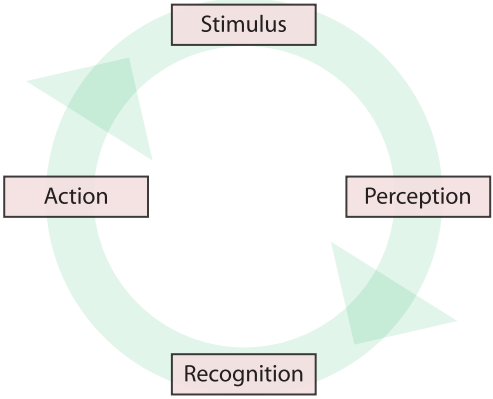


Figure 1.1: The human perceptual process is organized in a perception-action cycle. Adapted from [60].

without identifying what the object actually is. Recognition is the process of classifying the perceived object into a category, e.g. a cup or a human being. At this point, humans heavily rely on their knowledge of different object classes which results from a (possibly never-ending) learning process. The final step in the perceptual process is (re-)action derived from the recognized objects and the environment. Action can in turn effect new stimuli which makes the perception-action cycle continuous.

The ultimate design goal for autonomous systems is to mimic human behavior in terms of understanding and effortlessly acting within a dynamic human-inhabited environment. Although artificial sensors emulating the human sensory systems, e.g. cameras or microphones, are nowadays widely available, current autonomous systems are still far behind humans in terms of understanding and acting in real-world environments. The chief reason is the (theoretical and practical) unavailability of methods to reliably perform perception and recognition on a broad scale, i.e. not limited to isolated recognition problems. In addition, the human perception system is still not fully understood which makes it exceedingly difficult to duplicate artificially.

1.1 Motivation and Challenges

Most autonomous systems rely on visual cues derived from camera sensors to interpret and understand their environment. At the core, the visual perception and recognition problem can be formulated as an image-based multi-class object categorization problem. For most object classes, explicit models representing the structure and characteristics of that particular object category are not readily available. This has spawned the use of computer vision and pattern recognition techniques to learn an implicit representation of object classes from example images.

For autonomous systems situated in a human-inhabited environment, a key ability is to recognize and discriminate people from other object classes. While easy for humans, finding people in images is one of the most challenging problems from a machine vision perspective, notwithstanding years of methodical and technical progress, see Chapter 2.

Foremost, there is the wide range of possible pedestrian appearance. See Figure 1.2 for examples. The main cause for highly varying pedestrian appearance is clothing which can be of arbitrary color, style or may or may not exhibit (regular) patterns. Clothing can be tight-fitting and hence accentuat-

ing the body contours, see Figures 1.2a and 1.2c, or relatively loose, resulting in irregular pedestrian shape contours, as shown in Figures 1.2b and 1.2d.

Another challenge arises from the highly articulated body pose of pedestrians, see Figures 1.2a and 1.2b. Human pose changes considerably through a human gait cycle. People can bend down, tilt over, squat down and move their extremities independently from the torso. This makes it particularly hard to derive an explicit model for pedestrian recognition. The complexity of body articulations of pedestrians in 2D images increases significantly with the camera resolution and proximity to the camera which requires more complex models for close-range and high-resolution applications, e.g. component-based approaches, see Chapter 2. People can appear at multiple sizes (scales)

Figure 1.2: The appearance of pedestrians can change considerably due to clothing, scale, pose and contrast between the pedestrian and the background.

(a) (b)

Figure 1.3: In real-world settings, pedestrians are often partially occluded. Only parts of the body are visible.

in the image and are often partially occluded by other (static or moving) objects in the scene. See Figures 1.2a and 1.3.

In real-world applications, environmental conditions can also pose problems for pedestrian recognition systems. Pedestrians may appear at low contrast to the background, e.g. see Figure 1.2c, depending on the illumination conditions and the quality of camera sensors. Under low illumination, images tend to get noisy and motion blur can be induced due to high camera exposure times.

(a) (b)

Figure 1.4: (a) Difficult environmental conditions. (b) Highly textured and cluttered backgrounds.

Weather conditions can heavily influence recognition performance as well, see Figure 1.4a.

In case of a moving camera in a dynamic environment, problems arise from heavily cluttered and ever-changing backgrounds. Since pedestrians have mainly vertical contours, backgrounds with highly textured vertical structures are particularly challenging, as depicted in Figure 1.4b. Such areas can easily be mistaken for a pedestrian, especially in case of approaches that rely on (vertical) edge structure or shape.

At the same time, machine vision systems for people recognition are often subject to high performance demands. An ideal system should recognize any person and be avoid of any errors, i.e. mistaking an arbitrary object as a person. The system should be able to work robustly under various environmental conditions and be computationally efficient. Taking human recognition performance as a benchmark, real artificial systems available today paint an inferior performance picture. Most current systems are plagued by the large intra-class variability of persons which is hard to capture as a whole. Missing detections (false negatives) are the result. Real-world environments exhibit many structures that look similar to people at a lower-level, e.g. in terms of shape, size or structure, resulting in false detections (false positives). Deficiencies in the available sensory systems heavily decrease the robustness to environmental conditions, e.g. illumination or weather. All in all, the recognition performance is still orders of magnitudes away from human performance, when viewed in isolation from issues such as human reaction time or vigilance.

This thesis aims to raise the performance bar of state-of-the-art machine vision systems in terms of people recognition in dynamic real-world environments. A detailed overview of this thesis and its contributions is given in Chapter 3.

1.2 Applications

There are many potential application areas for the people recognition methods developed in this thesis. We are specifically concerned with those cases, where the human body to be detected covers a smaller portion of the image, i.e. is visible at lower resolution. Hence our use of the term “pedestrian” in the remainder of this thesis, rather than the more general “people” or “person.” We do not consider more detailed perception tasks, such as human

(a) (b)

Figure 1.5: An intelligent vehicle watches the road with cameras (marked in red) for possible collisions with pedestrians. © Daimler AG.

pose recovery or activity recognition [53, 107, 126].

Possible applications include outdoor settings such as visual surveillance, where a camera is watching down onto a street. Another application area is the field of intelligent vehicles, where an on-board camera watches the road ahead for possible collisions with pedestrians (and other traffic participants), see Figure 1.5. It also applies to indoor settings such as a robot recognizing a human walking down the hall or can be used as a proxy for human-computer interaction. It is further relevant to the fields of content-based image analysis and retrieval.

Although the methods presented in this work are fairly general with respect to the object class to be recognized, we choose the recognition of pedestrians in an urban environment from a moving vehicle as an experimental testbed. In this intelligent vehicles application, most of the previously mentioned challenges and difficulties are combined. In addition, it is arguably the most important application, given that worldwide fatality figures of pedestrians in traffic are estimated at 760.000 per year [51]. On average, this figure represents 65 % of all traffic-related deaths, including vehicle occupants. This percentage is particularly high in low-income countries.

Despite the inferiority of artificial pedestrian detectors relative to human vision and severe methodical challenges and performance demands, one of the central questions concerns the performance that is deemed necessary to deploy a pedestrian recognition system for real-world use. Besides the afore-

Figure 1.6: Automotive night vision with pedestrian recognition. © Daimler AG.

mentioned limitations in terms of recognition performance, artificial systems have an advantage over humans in that they do not fatigue, are always vigilant and can possible react in a small fraction of a second. Such benefits can outweigh the limitations to that extent, that real-world deployment becomes reasonable, depending on the exact application requirements. Recognizing pedestrians for content-based image analysis certainly has more relaxed performance requirements than automatic braking in the case of intelligent vehicles.

Taking the field of face recognition as an example, systems are widely used in real-world, although the recognition performance is not on par with human performance: Intelligent management software for digital photos can automatically sort and group images both in terms of the presence of faces and by the faces of individual persons [5]. Digital cameras (and even mobile phones) have face recognition software on-board to automatically control focus. Here, possible mistakes of the systems are not critical and can easily be corrected by the human user without severe consequences.

Regarding intelligent vehicles, the first night vision systems that detect and highlight pedestrians have reached the market (e.g. Mercedes-Benz E-Class 2009, BMW 7 series 2008 and Audi A8 2010), see Figure 1.6. Those systems use near-infrared vision combined with active illumination or far-

infrared vision (heat images) as the only sensory input and do not effect any vehicle actuation. The chief reason is the lacking recognition performance using vision only. As a way out, sensor fusion approaches have been pursued in the intelligent vehicles domain to boost the recognition performance. In the second half of 2010, Volvo introduced a collision mitigation system for pedestrians based on a fusion of vision and radar in their S60 limousine.

All of the previously mentioned applications would significantly benefit from more robust vision-based methods for pedestrian recognition to improve performance and to address a wider range of scenarios.

Chapter 2

Related Work

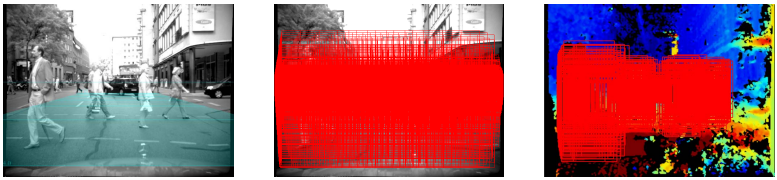
Pedestrian recognition has attracted a significant amount of interest from the computer vision and pattern recognition community over the past years. See [32, 37, 51, 58, 72] for recent surveys and performance studies. In this chapter, we focus on 2D approaches which are suitable for medium resolution pedestrian data (i.e. pedestrian height between 30 and 80 pixels). We do not cover higher-level recognition tasks such as human pose recovery or activity recognition [53, 107, 126]. A pedestrian classifier is typically part of an integrated system involving a pre-processing step to select initial object hypotheses and a post-processing step to integrate classification results over time (tracking). 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. This subdivision of the recognition problem is not specific to pedestrian recognition, but is a common concept for the recognition of arbitrary objects, see Figure 2.1.

Figure 2.1: Architecture of most object recognition systems involving four steps: Image acquisition, hypotheses selection, object classification and tracking.

2.1 Hypotheses Selection

The simplest technique to obtain initial object location hypotheses is the sliding window technique, where detector windows at various scales and locations are shifted over the image. The computational costs are often too high to allow for real-time processing. Significant speed-ups can be obtained by either coupling the sliding window approach with a classifier cascade of increasing complexity [47, 105, 116, 134, 135, 140, 159, 163, 164, 174, 175, 179, 182] or by restricting the search-space based on known camera geometry and prior information about the target object class. These include application-specific constraints such as the flat-world assumption, ground-plane based objects and common geometry of pedestrians, e.g. object height or aspect ratio [37, 40, 56, 92, 94, 110, 135, 139, 181]. In case of a moving camera in a real-world environment, varying pitch can be handled by relaxing the scene constraints [56] or by (re-)estimating the scene geometry on-line, using depth cues [41, 85, 92, 94].

Besides geometric constraints on possible pedestrian locations, cues derived directly from the image data are useful as early cueing mechanisms. Background subtraction, which is commonly used in static surveillance scenarios [107, 111, 147, 181], does not robustly generalize to a moving camera in a real-world setting. Some approaches have used stereo vision in combination with low-level segmentation or depth-filtering to further constrain the location of pedestrian candidates [2, 17, 41, 56, 85, 92, 94, 110, 112, 180], see Figure 2.2.



(a) **(b)** **(c)**

Figure 2.2: Example of hypotheses selection by scene constraints and depth-filtering, following [56]. (a) Input image with estimated scene geometry (ground-plane). (b) Dense grid of pedestrian hypotheses at various locations and scales, constrained to lie on the ground. (c) Remaining hypotheses after depth-filtering.

Motion cues resulting from the deviation of the observed optical flow from the expected ego-motion flow field [40, 124] can also provide hypotheses filtering. Another attention-focusing strategy involves interest point detectors to recover regions with high information content based on local discontinuities of the image brightness function which often occur at object boundaries [1, 92, 94, 98, 138].

2.2 Pedestrian Classification

After a set of initial object hypotheses has been acquired, further verification (classification) involves pedestrian appearance models, using various spatial and temporal cues. Following a rough categorization of such models into generative and discriminative models [160], we further introduce a delineation in terms of visual features and classification techniques. In both the generative and discriminative approach to pedestrian classification, a given image (or a sub-region thereof) is to be assigned to either the pedestrian or non-pedestrian class, depending on the corresponding class posterior probabilities. The main difference between generative and discriminative models is how posterior probabilities are estimated for each class.

2.2.1 Generative Models

Generative approaches to pedestrian classification model the appearance of the pedestrian class in terms of its class-conditional density function. In combination with the class priors, the posterior probability for the pedestrian class can be inferred using a Bayesian approach.

Shape Models

Shape cues are particularly attractive because of their property to reduce variations in pedestrian appearance due to lighting or clothing. At this point, we omit discussion of complex 3D human shape models [53, 107, 126] and focus on 2D pedestrian shape models which are commonly learned from shape contour examples. In this regard, both discrete and continuous representations have been introduced to model the shape space.

Discrete approaches represent the shape manifold by a set of exemplar shapes [54, 56, 69, 149, 155]. On the one hand, exemplar-based models imply a high specificity, since only plausible shape examples are included and

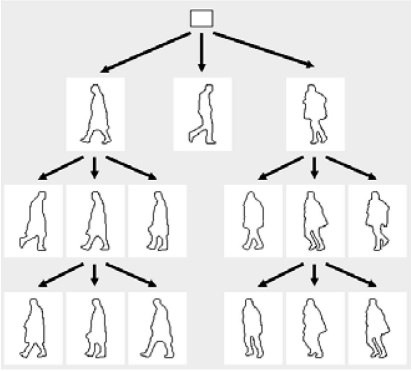


Figure 2.3: A hierarchical tree-like structure for discrete pedestrian shape exemplars. Adapted from [54].

changes of topology need not be explicitly modeled. On the other hand, such models require a large amount of example shapes (up to many thousands) to sufficiently cover the shape space due to transformations and intra-class variance. Particularly in close-range (high-resolution) settings, human body articulations become extremely diverse and often irregular. Exemplar-based models have difficulties to scale-up to such scenarios. The large amount of shape variation cannot be adequately represented by distinct shape examples.

From a practical point of view, exemplar-based models have to strike a balance between specificity and compactness to be used in real-world applications, particularly with regard to storage constraints and feasible on-line matching. Efficient matching techniques based on distance-transforms have been combined with pre-computed hierarchical structures, to allow for real-time on-line matching of many thousands of exemplars [54, 56, 69, 149], see Figure 2.3.

Continuous shape models involve a compact parametric representation of the class-conditional density, learned from a set of training shapes, given the existence of an appropriate manual [25, 63, 64] or automatic [9, 12, 36, 80, 110] shape registration method. Linear shape space representations which model the class-conditional density as a single Gaussian have been employed by [9, 25]. Forcing topologically diverse shapes (e.g. pedestrian with feet apart and with feet closed) into a single linear model may result in many intermediate

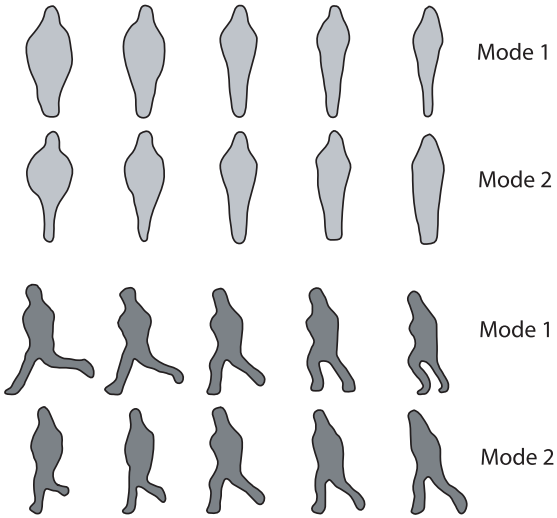


Figure 2.4: Continuous PCA-based pedestrian shape model depicting shape variation along two eigenvector modes for two orientation-specific models. Adapted from [34].

model instantiations that are physically implausible. To recover physically plausible regions in the linear model space, conditional-density models have been proposed [25, 34, 36], see Figure 2.4. Further, non-linear extensions have been introduced at the cost of requiring a larger number of training shapes to cope with the higher model complexity [25, 36, 63, 64, 110]. Rather than modeling the non-linearity explicitly, most approaches break up the non-linear shape space into piecewise linear patches. Techniques to determine these local sub-regions include fitting a mixture of Gaussians via the EM-algorithm [25] and K-means clustering in shape space [36, 63, 64, 110].

Compared to discrete shape models, continuous generative models can fill gaps in the shape representation using interpolation [36, 55]. However, online matching proves to be more complex, since recovering an estimate of the maximum-a-posteriori model parameters involves iterative parameter estimation techniques, i.e. Active Contours [25, 110].

A two-layer statistical field model has been proposed to increase the robust-

ness of shape representations to partial occlusions and background clutter by representing shapes as a distributed connected model [176]. Here, a hidden Markov field layer to capture the shape prior is combined with an observation layer which associates shape with the likelihood of image observations.

Combined Shape and Texture Models

A way to enrich the representation is to combine shape and texture information within a compound parametric appearance model [23, 25, 36, 43, 79, 80]. These approaches involve separate statistical models for shape and intensity variations. A linear intensity model is built from shape-normalized examples guided by sparse [25, 36, 43, 79] or dense correspondences [23, 80]. Model-fitting requires joint estimation of shape and texture parameters using iterative error minimization schemes [43, 79, 80]. To reduce the complexity of parameter estimation, the relation of the fitting error and associated model parameters can be learned from examples [25].

2.2.2 Discriminative Models

In contrast to generative models, discriminative models approximate the Bayesian maximum-a-posteriori decision by learning the parameters of a discriminant function (decision boundary) between the pedestrian and non-pedestrian class from training examples. We discuss the merits and drawbacks of several feature representations and continue with a review of classifier architectures and techniques to break down the complexity of the pedestrian class.

Features

Local filters operating on pixel intensities are a frequently used feature set [129]. Recently, local binary pattern (LBP) features [115] have been employed in the context of pedestrian classification [39, 167]. LBPs encode (thresholded) local gray-level differences into a binary number, followed by local histogramming. Their key advantage is the invariance against monotonic gray-level changes and noisy backgrounds which are common in cluttered environments.

Non-adaptive Haar wavelet features have been popularized by [120] and adapted by many others [108, 135, 142, 164, 165, 173], see Figure 2.5a. This

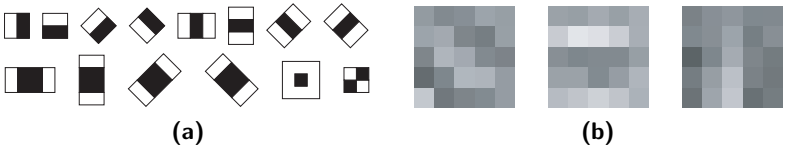


Figure 2.5: Examples of texture-based features. (a) Non-adaptive Haar wavelet features. (b) Adaptive 5×5 pixel local receptive field (LRF) features. Adapted from [37].

over-complete feature dictionary represents local intensity differences at various locations, scales and orientations. Their simplicity and fast evaluation using integral images [97, 164] contributed to the popularity of Haar wavelet features. However, the many-times redundant representation, due to overlapping spatial shifts, requires mechanisms to select the most appropriate subset of features out of the vast amount of possible features. Initially, this selection has been manually designed for the pedestrian class, by incorporating prior knowledge about the geometric configuration of the human body [108, 120, 142]. Later, automatic feature selection procedures, i.e. variants of AdaBoost [49], have been employed to select the most discriminative feature subset [164, 165, 166, 173].

The automatic extraction of a subset of non-adaptive features can be regarded as optimizing the features for the classification task. Likewise, the particular configuration of spatial features has been included in the actual optimization itself, yielding feature sets which adapt to the underlying dataset during training. Such features are referred to as local receptive fields (LRF), see Figure 2.5b, or convolutional networks [38, 50, 56, 109, 118, 135, 151, 170], in reference to neural structures in the human visual cortex [60]. Recent studies have empirically demonstrated the superiority of adaptive local receptive field features over non-adaptive Haar wavelet features with regard to pedestrian classification [109, 151].

Another class of local intensity-based features are codebook patches, extracted around interesting points in the image [1, 92, 93, 94, 138]. A codebook of distinctive object feature patches along with geometrical relations is learned from training data followed by clustering in the space of feature patches to obtain a compact representation of the underlying object class.

Based on this representation, feature vectors have been extracted including information about the presence and geometric relation of codebook patches [1, 92, 93, 94, 138].

Others have focused on discontinuities in the image brightness function in terms of models of local edge structure. Well-normalized image gradient orientation histograms, computed over local image blocks, have become popular in both dense [27, 35, 38, 39, 52, 102, 118, 137, 139, 140, 165, 166, 167, 173, 179, 182] (HOG, histograms of oriented gradients), and sparse representations [98] (SIFT, Scale-Invariant Feature Transform), where sparseness arises from pre-processing with an interest-point detector. Initially, dense gradient orientation histograms were computed using local image blocks at a single fixed scale [27] to limit the dimensionality of the feature vector and computational costs. Extensions to variable-sized blocks have been presented by [140, 173, 179, 182]. Results indicate a performance improvement over the original HOG approach. Recently, local spatial variation and correlation of gradient-based features have been encoded using covariance matrix descriptors which increase robustness towards illumination changes [137, 159, 173].

Yet others have designed local shape filters that explicitly incorporate the spatial configuration of salient edge-like structures. Multi-scale features based on horizontal and vertical co-occurrence groups of dominant gradient orientation have been introduced by [105]. Manually designed sets of Edgelets, representing local line or curve segments, have been proposed to capture edge structure [174]. An extension to these pre-defined edgelet features has been introduced with regard to adapting the local edgelet features to the underlying image data [134]. So called Shapelet features are assembled from low-level oriented gradient responses using AdaBoost, to yield more discriminative local features. Again, variants of AdaBoost are frequently used to select the most discriminative subset of features.

As an extension to spatial features, spatio-temporal features have been proposed to capture human motion [28, 35, 39, 143, 164, 165, 166, 173], especially gait [66, 91, 124, 170]. For example, Haar wavelets and local shape filters have been extended to the temporal domain by incorporating intensity differences over time [143, 164]. Local receptive field features have been generalized to spatio-temporal receptive fields [66, 170]. Histograms of oriented gradients (HOG) have been extended to histograms of differential optical flow [28, 35, 39, 165, 166, 173]. Several authors compared the performance of otherwise identical spatial and spatio-temporal features and reported superior

performance of the latter at the drawback of requiring temporally aligned training samples [28, 35, 39, 164, 165, 166, 173].

Classifier Architectures

Discriminative classification techniques aim at determining an optimal decision boundary between pattern classes in a feature space. Feed-forward multi-layer neural networks [78] (MLP, multi-layer perceptron) implement linear discriminant functions (top layer) in a feature space in which input patterns have been mapped non-linearly (hidden layer). Optimality of the decision boundary is assessed by minimizing an error criterion with respect to the network parameters, i.e. mean squared error [78]. In the context of pedestrian recognition, multi-layer neural networks have been applied primarily in conjunction with adaptive local receptive field features as non-linearities in the hidden network layer [50, 56, 109, 151, 170]. This architecture unifies feature extraction and classification within a single model. Other than that, MLPs can be combined with arbitrary feature sets and provide non-linear decision boundaries [39].

Support vector machines (SVM) [161] have evolved as a powerful tool to solve pattern classification problems. In contrast to neural networks, SVMs do not minimize some artificial error metric but maximize the margin of a linear decision boundary (hyperplane) to achieve maximum separation between the object classes. Regarding pedestrian classification, linear SVM classifiers have been used in combination with various (non-linear) feature sets [27, 28, 35, 38, 39, 109, 111, 137, 140, 142, 165, 166, 173, 179, 182].

Non-linear SVM classification, e.g. using polynomial or radial basis function kernels as implicit mapping of the samples into a higher-dimensional (and probably infinite) space, usually yields further performance boosts. These are however paid for with a significant increase in computational costs and memory requirements [2, 102, 108, 109, 111, 120, 137, 151]. Recent work presents efficient versions of non-linear SVMs for a specific class of kernels [102].

AdaBoost [49], which has been applied as automatic feature selection procedure (see above), has also been used to construct strong classifiers as weighted linear combinations of selected weak-learners, each involving a threshold on a single feature. Such boosting approaches require to map a multi-dimensional feature set to a single dimension, either by applying projections [159, 175]

or by treating each dimension as an individual feature [105, 116, 134, 135, 139, 165, 166, 173, 174]. An alternative is the use of more complex weak-learners that operate in a multi-dimensional space, e.g. support vector machines, [140, 179, 182].

To incorporate non-linearities and speed-up the classification process, boosted detector cascades have been introduced by [164] and adopted by many others [47, 105, 116, 134, 135, 140, 159, 163, 174, 175, 179, 182]. Motivated by the fact that the majority of detection windows in an image are non-pedestrians, the cascade structure is tuned to detect almost all pedestrians while rejecting non-pedestrians as early as possible. AdaBoost is used in each layer to iteratively construct a strong classifier guided by user-specified performance criteria. During training, each layer is focused on the errors the previous layers make. As a result, the whole cascade consists of increasingly more complex detectors. This contributes to the high processing speed of the cascade approach, since usually only a few feature evaluations in the early cascade layers suffice to quickly reject non-pedestrian examples.

2.2.3 Multi-Level Representations

Besides introducing new feature sets and classification techniques, many pedestrian recognition approaches attempt to break-down the complex appearance of the pedestrian class into better manageable sub-parts.

First, a mixture-of-experts strategy establishes local pose-specific pedestrian clusters, followed by the training of a specialized expert classifier for each subspace [38, 39, 56, 111, 139, 142, 174, 179]. Appropriate pose-based clustering involves both manually [111, 139, 142, 174] and automatically established [179] mutually exclusive clusters, as well as soft clustering approaches using a probabilistic assignment of pedestrian examples to pose clusters, obtained by a pre-processing step, e.g. shape matching [38, 39, 56]. An additional issue in mixture-of-experts architectures is how to integrate the individual expert responses to a final decision. Usually, all experts are run in parallel, where the final decision is obtained as a combination of local expert responses using techniques such as maximum selection [111, 174], majority voting [142], AdaBoost [139], trajectory-based data-association [179], and probabilistic shape-based weighting [38, 39, 56].

Second, component-based approaches decompose pedestrian appearance into parts. These parts are either semantically motivated (body parts such

as head, torso and legs) [2, 31, 35, 105, 108, 139, 143, 167, 174] or concern representations based on a collection of local parts in a deformable configuration [1, 46, 47, 48, 84, 92, 94, 95, 138]. A general trade-off is involved at the choice of the number and selection of the individual parts. On one hand, components should have as small spatial extent as possible, to succinctly capture articulated motion. On the other hand, components should have sufficiently large spatial extent to contain discriminative visual structure to allow reliable detection. Part-based approaches require assembly techniques to integrate the local part responses to a final detection, constrained by spatial relations among the parts.

Approaches using partitions into semantic sub-regions train a discriminative feature-based classifier (see above), specific to a single part, along with a model for geometric relations between parts. Techniques to assemble part-based detection responses to a final classification result include the training of a combination classifier [2, 108, 139], probabilistic inference to determine the most likely object configuration given the observed image features [105, 143, 174], voting schemes [31, 35] or heuristics [167].

Deformable part approaches, i.e. [1, 46, 47, 48, 84, 92, 94, 95, 138], represent pedestrians in a bottom-up fashion as assemblies of locally linked features, often augmented with a top-down verification step [92, 94, 95, 138]. Recently, it has been shown that context information can help to detect parts which cannot be reliably detected using their own appearance, e.g. because of low resolution or occlusions [84].

Component-based approaches have certain advantages compared to full-body classification. They do not suffer from the unfavorable complexity related to the number of training examples necessary to adequately cover the set of possible appearances, particularly in close-range and high-resolution scenarios. Furthermore, the expectation of missing parts due to scene occlusions or inter-object occlusions is easier addressed, particularly if explicit inter-object occlusion reasoning is incorporated into the model [35, 84, 92, 94, 138, 167, 174]. However, these advantages are paid for with higher complexity in both model generation (training) and application (testing). Their applicability to lower resolution images is limited since each component detector requires a certain spatial support for robustness.

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 module-level [40, 41, 56, 112], others inte-

Figure 2.6: Multi-modality pedestrian representation. Left to right: Intensity image, dense stereo (depth) image, dense flow image (motion). In depth images, darker colors denote closer distances. Optical flow images depict the horizontal component of flow vectors. Medium red colors denote close to zero flow, darker and brighter colors indicate stronger motion (to the left and to the right, respectively). Adapted from [35].

grate multiple information sources directly into the pattern classification step [28, 35, 39, 118, 131, 137, 163, 165, 166, 167, 173, 175].

Some approaches combine features in the intensity domain using boosting [175] or multiple kernel learning [163], e.g. by combining HOG, covariance and edgelet features into a boosted heterogenous cascade classifier with an explicit optimization with regard to runtime [175]. Others integrate intensity and flow features by boosting [165, 173] or concatenating all features into a single feature vector which is then passed to a single classifier [28, 165, 173]. The work in [165] was recently extended to additionally include depth features [166]. Note that the approach of [35] marks the first use of intensity, motion and depth features in the domain of pedestrian classification, see Figure 2.6. Boosting approaches require to map the multi-dimensional features to a single dimension, either by applying projections [175] or treating each dimension as an individual feature [165, 166, 173]. An alternative is the use of more complex weak-learners that operate in a multi-dimensional space, e.g. support vector machines [182].

A joint feature space approach to combine HOG and LBP features was used in [167]. [137] presents the integration of HOG features, co-occurrence features and color frequency descriptors into a very high-dimensional (≈ 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.

In contrast, [28, 35, 39, 118, 131] utilize fusion on classifier-level by training a specialized classifier for each cue. [28] and [35] use a single feature (HOG) in two (intensity and motion, [28]) and three different modalities (intensity, depth and motion, [35]), respectively. [118] involves a combination of two features (HOG and LRF) within a single modality (intensity). [131] presents a classifier-level combination of two features where each feature operates in a different modality (HOG / intensity and LRF / depth). Finally, a pose-specific mixture-of-experts framework using two features (HOG and LBP) in three different modalities (intensity, depth and motion) is proposed in [39]. Classifier fusion is done using fuzzy integration [118], simple classifier combination rules [131] or a mixture-of-experts framework [28, 35, 39, 77].

2.3 Tracking

There has been extensive work on the tracking of pedestrians to infer information on trajectory-level. Most approaches follow a tracking-by-detection paradigm which involves the association of detections made by an object detection system over time. This is a very challenging problem, given the uncertainty of estimated object positions, ever-changing backgrounds combined with a moving camera and occlusions.

One line of research has formulated tracking as frame-by-frame association of detections based on geometry and dynamics without particular pedestrian appearance models [2, 56]. Other approaches utilize pedestrian appearance models (Section 2.2) coupled with geometry and dynamics [9, 16, 64, 76, 92, 94, 99, 110, 122, 128, 143, 155, 174, 176, 179, 181]. Some approaches furthermore integrate detection and tracking in a Bayesian framework, combining appearance models with an observation density, dynamics and probabilistic inference of the posterior state density. For this, either single [9, 64, 122, 155, 174] or multiple cues [16, 76, 99, 110, 128, 143] are used.

The integration of multiple cues [147] involves combining separate models for each cue into a joint observation density. The inference of the posterior state density is usually formulated as a recursive filtering process [6]. Particle filters [74] are very popular due to their ability to closely approximate complex real-world multimodal posterior densities using sets of weighted random

samples. Extensions that are especially relevant for pedestrian tracking involve hybrid discrete / continuous state-spaces [64, 110] and efficient sampling strategies [29, 76, 86, 100].

An important issue in real-world pedestrian tracking problems is how to deal with multiple targets in the image. Two basic strategies with regard to the tracking of multiple objects have been proposed. First, the theoretically most sound approach is to construct a joint state-space involving the number of targets and their configurations which are inferred in parallel. Problems arise regarding the significantly increased and variable dimensionality of the state-space. Solutions to reduce the computational complexity have involved grid-based or pre-calculated likelihoods [76, 152] and sophisticated resampling techniques such as Metropolis-Hastings sampling [86], partitioned sampling [100], or annealed particle filters [29]. Second, some approaches have proposed to limit the number of objects to one per tracker and employ multiple tracker instances instead [16, 75, 82, 110, 116]. While this technique simplifies the state-space representation, a method for initializing a track along with rules to separate neighboring tracks is required. Typically, an independent detector process is employed to initialize a new track. To identify individual targets in the image sequence, appearance models have been learned on-line [3, 4, 16] which help to associate the object detections to the correct track. Other competition rules between multiple tracker instances have been formulated in terms of heuristics [82, 110].

A detection-by-tracking approach has been proposed, where temporal consistency of body articulations has been explicitly incorporated into the detection process using a hierarchical Gaussian process latent variable model [3]. The resulting system integrates detection and tracking into a single model and has shown to boost both detection and tracking performance at the same time. Further, the results show a certain robustness towards (partial and long-term) occlusions, due to the assumed regularity of the kinematic human gait model and the use of an on-line learned appearance model that is taken to remain constant during the occlusion. This work has recently been extended in terms of 3D pose estimation from multiple viewpoints [4].

Chapter 3

Outline and Contributions

The aim of this thesis is to develop novel vision-based methods for pedestrian recognition which can significantly improve the recognition performance. We do not focus on the development of new (and possibly improved) feature primitives to be used in the classification module. Instead, we follow an orthogonal direction and propose and evaluate feature- and classifier-independent compound techniques where the term "compound" refers to the combination of information sources on different levels. This involves multiple image modalities, e.g. gray-level intensity, dense stereo vision and dense optical flow, multiple features, e.g. shape-based, gradient-based and texture-based features and higher-level information, such as pedestrian pose and body orientation, body components and partial occlusions. Throughout this thesis, particular emphasis is laid on thorough performance evaluation, both from a methodical point-of-view and in terms of large and challenging datasets. Several datasets used in this thesis are made publicly available for benchmarking and to stimulate further research.

3.1 Monocular Pedestrian Recognition

Pedestrian recognition has attracted an extensive amount of interest from the computer vision community over the past few years. Many techniques have been proposed in terms of features, models and general architectures, see Chapter 2. The picture is increasingly blurred on the experimental side. Reported performances differ by up to several orders of magnitude (e.g. within the same study [164], or [92] vs. [164]). This stems from the different types of image data used (degree of background change), the limited size of the test datasets, and the different (often, not fully specified) evaluation criteria such as localization tolerance, coverage area, etc. Chapter 4 first covers an experimental study on monocular pedestrian recognition that provides a sound

performance baseline (Section 4.1). State-of-the-art detectors are evaluated using the same large real-world dataset with the same evaluation criteria. We present a thorough evaluation methodology for the evaluation of integrated multi-module pedestrian recognition systems and make our dataset publicly available for benchmarking purposes. Auxiliary effects, such as training sample resolution, the granularity of the detection grid, non-maximum suppression, tracking, as well as scene and processing time constraints are taken into account.

Chapter 4 then continues with work on the integration of multiple cues for pedestrian recognition (Section 4.2). Multiple cues in a classification setting will be a central topic later on in this thesis. At this point, multiple cues are used on module-level, in terms of a novel attentive strategy utilizing motion to recover meaningful pedestrian location hypotheses. Those hypotheses are processed by subsequent classification modules that combine shape and texture information.

3.2 A Mixed Generative-Discriminative Pedestrian Model

Starting with Chapter 5, we focus on the pedestrian classification component in isolation, i.e. the most important part of a full recognition system as described in Chapter 4.

Chapter 5 presents a novel approach to pedestrian classification which involves utilizing the synthesized virtual samples of a learned generative model to enhance the classification performance of a discriminative model. Our generative model combines shape and texture cues in terms of a number of probabilistic shape and texture models, each attuned to a particular pedestrian pose. Active learning provides the link between the generative and discriminative model, in the sense that the former is selectively sampled such that the training process is guided towards the most informative samples of the latter.

We consider the main contribution to be the novel mixed generative-discriminative framework for pedestrian classification where a generative model is used to enhance the performance of a discriminative model in terms of virtual training samples. This approach is quite unlike previous combination strategies for generative and discriminative models [90, 104, 156, 178] and unlike previous applications of active learning. We neither require controlled data acquisition [13, 20, 57], nor do we have 3D models [65, 103] to our disposition. At the

same time, we go beyond the synthesis of samples based on simple transformations [109, 125, 150, 162] and take into account sample probabilities.

A secondary contribution concerns the generative pedestrian model proposed. Similar to [55, 63], our approach uses separate feature-spaces to model topologically diverse shapes (e.g. pedestrian with feet apart and with feet closed), in order to increase model specificity. However, we extend the shape representation of [55, 63] with a texture component, distinguishing between texture variations at the coarse and the detail level. We establish a statistical shape-texture model along with the associated class-conditional density functions. This provides a sound basis for the synthesis of virtual pedestrian samples by means of three components: foreground shape, foreground texture and background texture.

3.3 Multi-Level Mixture-of-Experts for Pedestrian Classification

Most research in the field of pedestrian classification has focused on features operating on image intensity, as discussed in Chapter 2. In Chapter 6, we pursue an orthogonal direction and present a novel multi-level mixture-of-experts approach to combine information from multiple features and modalities 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 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). Multi-layer perceptrons (MLP) and linear support vector machines (linSVM) are used as expert classifiers.

The main contribution is the aforementioned multi-level mixture-of-experts framework for pedestrian classification, which breaks down the complex classification problem into better manageable sub-problems. To our knowledge, this work represents the first integration of shape, intensity, depth and motion as features into a pattern classification framework. We show how to combine multi-feature / multi-modality 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. Our multi-modality dataset is made public for evaluation purposes.

3.4 Multi-Modality Partial Occlusion Handling

In Chapter 7, we present an extension to the aforementioned mixture-of-experts framework in terms of a multi-modality model for partial occlusion handling. Our framework involves a set of component-based expert classifiers trained on features derived from intensity, depth and motion. To handle partial occlusion, we compute expert weights that are related to the degree of visibility of the associated component. This degree of visibility is determined by examining occlusion boundaries, i.e. discontinuities in depth and motion. Occlusion-dependent component weights focus the combined decision of the classifier on the unoccluded body parts.

We consider the mixture-of-experts extension with regard to partial occlusion handling as the main contribution of this chapter. In contrast to [174], we do neither require a particular camera set-up nor assume constant visibility of a certain body part. Our method is independent of the employed feature/classifier combination and the pedestrian component layout, unlike [167]. A secondary contribution involves the integration of intensity, depth and motion modalities throughout our approach. Off-line, we train multi-modality component-based expert classifiers involving feature spaces derived from gray-level images, depth maps (dense stereo vision) and motion (dense optical flow), cf. Chapter 6. On-line, we apply multi-modality (depth and motion) mean-shift segmentation to each test sample to recover occlusion-dependent component weights which are used to fuse the component-based expert classifiers to a joint decision with a focus on visible body parts.

3.5 Integrated Classification and Orientation Estimation

Chapter 8 extends the mixture-of-experts framework presented in Chapter 6 in terms of applying the pose-specific nature of the model towards single-frame estimation of pedestrian body orientation. We use the set of view-related expert models not only for classification as in Chapter 6, but also to approximate the probability density of pedestrian orientation. Sample-dependent priors are integrated in a Bayesian fashion and the approach scales-up to the use of multiple cameras.

We consider the main contribution of Chapter 8 to be the integrated framework for pedestrian classification and orientation. Previous approaches to orientation estimation, e.g. [52, 111, 142], 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 [52, 111, 142], 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 [52, 142], 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.

3.6 Evaluation Methodology

Performance evaluation of pedestrian classifiers is a major aspect of this work, both in terms of methodology and datasets used. Evaluation can be performed using a per-image measure (detection context) or a per-window measure (classification context). Per-image evaluation involves shifting a pedestrian classifier through location and scale across the whole test image. In per-window evaluation, the test data involves cut-out and scaled bounding boxes cropped from full test images.

We use both evaluation methods, depending on the application context of the systems to be evaluated. Per-image evaluation is used to evaluate (monocular) sliding-window classifiers, cf. Chapter 4 and [37, 72]. For the evaluation of integrated systems that include a hypotheses generation and tracking module, per-image (or even per-trajectory) evaluation is the only viable choice.

Most real-world systems however integrate several modules; i.e. they do not follow a brute-force sliding-window detection scheme, but use a pre-processing step to determine initial pedestrian location hypotheses for both enhanced performance and computational efficiency. This is done by background subtraction [107, 111, 147, 181], shape [40, 56], stereo [2, 17, 41, 56, 85, 92, 94, 110, 112, 180], motion [40] or non-vision sensors, such as radar or lidar [51]. As a result, the remaining object hypotheses are not random sub-windows, but contain meaningful structure that resembles pedestrians in some aspect.

Further, the number of hypotheses per image which are processed by the pedestrian classifier is greatly reduced (up to a factor of 10000) compared to dense sub-window scanning, resulting in a more even ratio between pedestrian and non-pedestrian samples. In this application context, we use the per-window measure to evaluate a classifier, since it more closely resembles the actual use of the classifier, cf. Chapters 5 - 8. Classification windows in the test set are not randomly selected sub-images, but result from a preprocessing step to focus on meaningful samples in the evaluation.

As opposed to Dollar et al. [32] who consider the per-window evaluation for classifiers flawed, since auxiliary effects (e.g. grid granularity or non-maximum suppression) are not taken into account, we regard both evaluation set-ups as viable. The choice has to be made depending on the actual application context of the pedestrian classifier.

3.7 Publications

This thesis has led to a number of publications that are listed in Appendix A. Note that the corresponding publications have been included in the discussion of related work in Chapter 2.

Chapter 4

Monocular Pedestrian Recognition

This chapter provides a performance baseline by experimentally evaluating state-of-the-art monocular pedestrian recognition systems using identical set-ups. The second part presents a novel multi-cue strategy to early focus the processing attention on image areas that likely contain pedestrians.

4.1 An Experimental Study

The experimental study presented in this section aims to increase visibility by providing a common point of reference of monocular pedestrian recognition performance from an experimental perspective. We evaluate a diverse set of state-of-the-art systems using identical test criteria and datasets:

- Haar wavelet-based AdaBoost cascade [164]
- Histograms of oriented gradient (HOG) features combined with a linear SVM [27]
- Neural network using local receptive fields (NN/LRF) [170]
- Combined hierarchical shape matching and texture-based NN/LRF classification [56]

In terms of evaluation, we consider both a generic and an application-specific test scenario. The generic test scenario is meant to evaluate the inherent potential of a pedestrian recognition method. It incorporates no prior scene knowledge as it uses a simple 2D bounding box overlap criterion for matching. Furthermore, it places no constraints on allowable processing times (apart from practical feasibility). The application-specific test scenario focuses on the case of pedestrian recognition from a moving vehicle, where

knowledge about camera calibration, location of the ground-plane, and sensible sensor coverage areas provide regions of interest. Evaluation takes place in 3D in a coordinate system relative to the vehicle. Furthermore, we place upper bounds on allowable processing times (250 ms vs. 2.5 s per frame). In both scenarios, we list recognition performance both at the frame and trajectory level.

4.1.1 Benchmark Dataset

The dataset is truly large-scale; it includes many tens of thousands of training samples as well as a test sequence consisting of 21790 monocular images at 640×480 resolution, captured from a vehicle in a 27 minute drive through urban traffic. See Table 4.1. Compared to previous pedestrian datasets, the availability of sequential images means that also hypothesis generation and tracking components of pedestrian systems can be evaluated, unlike with [73, 106, 109]. Furthermore, the dataset excels in complexity (dynamically changing background) and realism for the pedestrian protection application on-board vehicles. We release both training and test sets, so that other authors can independently evaluate their systems, in contrast to [32].

Figure 4.1 shows an excerpt from the Daimler pedestrian recognition benchmark dataset used in this work. Dataset statistics are shown in Table 4.1 (last row). Training images were recorded at various day times and locations with no constraints on illumination, pedestrian pose or clothing, except that pedestrians are fully visible in an upright position. 15660 pedestrian (positive) samples are provided as training examples. These samples were obtained by manually extracting 3915 rectangular position labels from video images. Four pedestrian samples were created from each label by means of mirroring and randomly shifting the bounding boxes by a few pixels in horizontal and vertical directions to account for localization errors in the application system. The addition of jittered samples was shown earlier to substantially improve performance [36]. Pedestrian labels have a minimum height of 72 pixels, so that there is no up-scaling involved in view of different training sample resolutions for the systems under consideration. Further, we provide 6744 full images not containing any pedestrians from which all approaches under consideration extract negative samples for training.

Our test dataset consists of an independent image sequence comprising 21790 images (640×480 pixels) with 56492 manual labels, including 259

Dataset	Training Set Pedestrian / Non-Pedestrian	Test Set Pedestrian / Non-Pedestrian	Comments
MIT CBCL Pedestrian Database [106]	924 / 0 (cut-outs)		single images, frontal and back views only
INRIA Person Dataset [73]	2416 (cut-outs) / 1218 (full images)	1132 (cut-outs) / 453 (full images)	single images (color)
Mobile Scene Analysis Dataset [41]	490 (full images), 1578 ped. labels	1803 (full images), 9380 ped. labels	camera at walking speed (stroller on urban sidewalks)
PETS Datasets (2001, 2003, 2004) [121]	-	2688, 2500, 13112 (full images)	16 image sequences from static cameras
DaimlerChrysler Pedestrian Classification Benchmark [109]	14400 / 15000 (cut-outs) + 1200 (full images)	9600 / 10000 (cut-outs)	single images
Caltech Pedestrian Dataset [32]	67.000 (full images), 192.000 ped. labels / 61.000 (full images)	not published	test sets not published
TU Darmstadt Pedestrian Dataset [173]	1092 (full images), 1776 ped. labels / 192 (full images)	508 (full-images), 1326 ped. labels	temporally aligned image pairs
Daimler Multi-Cue Pedestrian Classification Benchmark (this thesis, Chapters 6 and 7 and [35, 39])	**52112 / 32465 (cut-outs)**	**25608 fully visible + 11160 partially occluded ped. / 16235 (cut-outs)**	**intensity, dense stereo and dense flow images + partially occluded pedestrians**
Daimler Pedestrian Recognition Benchmark (this thesis, Chapter 4 and [37])	**15660 (cut-outs) / 6744 (full images)**	**21790 (full images), 56492 labels: 14132 fully visible ped. labels in 259 trajectories, 37236 partial ped. labels, 5124 other labels (bicyclists, motorcyclists, etc.)**	**test set corresponds to a 27 min drive through urban traffic**

Table 4.1: Overview of publicly available pedestrian datasets with ground-truth.

Figure 4.1: Overview of the Daimler pedestrian recognition benchmark dataset: Pedestrian training samples (top row), non-pedestrian training images (center row), test images with annotations (bottom row).

trajectories of fully visible pedestrians, captured from a moving vehicle in a 27 minute drive through urban traffic. In contrast to other established benchmark datasets (see Table 4.1), the size and complexity of the current data allows to draw meaningful conclusions without appreciable overfitting effects. The dataset has a total size of approximately 8.5 GB[1].

4.1.2 Selected Pedestrian Recognition Approaches

We select a diverse set of pedestrian recognition approaches in terms of features (adaptive, non-adaptive) and classifier architecture for evaluation: Haar wavelet-based cascade [164], neural network using LRF features [170], and histograms of oriented gradients combined with a linear SVM [27]. In addition to these approaches, used in sliding window fashion, we consider a system utilizing coarse-to-fine shape matching and texture-based classification, i.e. a monocular variant of [56]. Temporal integration is incorporated by coupling all approaches with a 2D bounding box tracker. We acknowledge that besides the selected approaches there exist many other interesting lines of research in the field of monocular pedestrian recognition (see Chapter 2). We encourage other authors to report performances using the proposed dataset and eval-

[1]The dataset is made available for research purposes at `http://www.science.uva.nl/research/isla/downloads/pedestrians/`

Figure 4.2: Overview of the employed set of Haar wavelets. Black and white areas denote negative and positive weights, respectively.

uation criteria for benchmarking. Here, we focus on the most widely-used approaches[2].

Our experimental set-up assigns the underlying system parameters (e.g. feature layout, training process, etc.) to the values reported to perform best in the original publications [27, 56, 109, 164, 170]. Two different resolutions of training samples are compared. We consider training samples with an actual pedestrian height of 32 pixels (small scale) and 72 pixels (medium scale). To this, a fixed fraction of border pixels (background) is added. Details are given below.

Haar Wavelet-Based Cascade

The Haar wavelet-based cascade framework [164] provides an efficient extension to the sliding window approach by introducing a degenerate decision tree of increasingly complex detector layers. Each layer employs a set of non-adaptive Haar wavelet features [108, 120]. We make use of Haar wavelet features at different scales and locations, comprising horizontal and vertical features, corresponding tilted features, as well as point detectors, see Figure 4.2. Sample resolution for the small scale training set is 18×36 pixels with a border of two pixels around the pedestrian. No constraints on scales or locations of wavelets are imposed, other than requiring the features to lie completely within our training samples. The total number of possible features is 154190. The medium scale training set consists of samples at 40×80 pixels with a border of four pixels around the pedestrian which leads to over 3.5 million possible features. Here, we have to constrain the features to allow for feasible training: we require a minimum area of 24 pixels with a two pixel scale step for each feature at a spatial overlap of 75 % which results in 134621 possible features. In each cascade layer, AdaBoost [49] is used to construct a classifier based on a weighted linear combination of selected features

[2]total processing time for training, testing and evaluation was several months of CPU time on a 2.66 GHz Intel processor, using implementations in C/C++.

which yield the lowest error on the training set consisting of pedestrian and non-pedestrian samples.

We investigated the performance after N_l layers and found that performance saturated after incorporating $N_l = 15$ layers for both training resolutions. Each cascade layer is trained on a new dataset consisting of the initial 15660 pedestrian training samples and a new set of 15660 non-pedestrian samples which is generated by collecting false positives of the cascade up to the previous layer on the given set of non-pedestrian images. Negative samples for the first layer are randomly sampled. Performance criteria for each layer are set to 50 % false positive rate at 99.5 % detection rate. Adding further cascade layers reduced the training error, but performance on the test set was observed to run in saturation. The total number of features selected by AdaBoost for the whole 15-layer cascade using small (medium) resolution samples is 4070 (3751), ranging from 15 (14) features in the first layer to 727 (674) features in the final layer. Experiments are conducted using the implementation found in the OpenCV library [119].

Neural Network using Local Receptive Fields (NN/LRF)

In contrast to multi-layer perceptrons (MLP), where the hidden layer is fully connected to the input layer, NN/LRFs introduce the concept of N_B branches B_i $(i = 1, \ldots, N_B)$, where every neuron in each branch only receives input from a limited local region of the input layer, its receptive field. See Figure 4.3. Since synaptical weights are shared among neurons in the same branch, every branch can be regarded as a spatial feature detector on the whole input pattern and the amount of parameters to be determined during training is reduced, alleviating susceptibility to overfitting.

Adaptive local receptive fields (LRF) [50] have shown to be powerful features in the domain of pedestrian recognition, in combination with a multi-layer feed-forward neural network architecture (NN/LRF) [170]. Although the combination of LRF features and non-linear support vector machine classification (SVM/LRF) yields slightly better performance [109], we opt for a NN/LRF in this study, since training a non-linear SVM/LRF classifier on our large dataset is infeasible due to excessive memory requirements.

We use a NN/LRF consisting of $N_B = 16$ branches B_i. For the small scale training samples at a resolution of 18×36 pixels with a two pixel border, 5×5 pixel receptive fields are utilized, shifted at a step size of two pixels

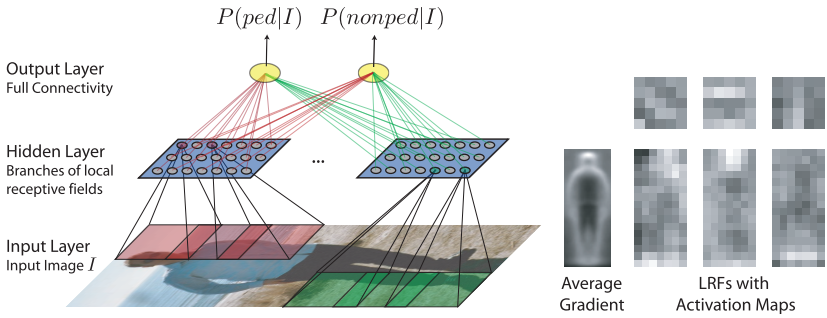


Figure 4.3: Overview of NN/LRF architecture (left). Average gradient image along with three exemplary 5×5 pixel local receptive field features (hidden layer weights) and their activation maps (output layer weights) for the "pedestrian" output neuron, highlighting regions where corresponding LRFs are most discriminative for the pedestrian class (right).

over the training images. 10×10 pixel receptive fields are shifted at a step size of five pixels over the medium scale training samples which are scaled to 40×80 pixels including a border of four pixels.

The output layer consists of two neurons, where the output of each neuron represents a (scaled) estimate of posterior probability for the pedestrian and non-pedestrian class, respectively. Initial training data consists of the given 15660 pedestrian samples, along with 15560 randomly selected samples from the set of negative images. We further apply a bootstrapping strategy by shifting the trained NN/LRF classifier over the images containing no pedestrians and augmenting the negative training set by collecting 15660 false positives in each iteration. Finally, the classifier is retrained using the extended negative training data. Bootstrapping is applied iteratively until test performance saturates. The higher complexity of the bootstrapped dataset is accounted for by incorporating additional eight branches in each iteration to increase classifier complexity.

Histograms of Oriented Gradients with Linear SVM (HOG/linSVM)

We follow the approach of Dalal and Triggs [27] to model local shape and appearance using well-normalized dense histograms of gradient orientation

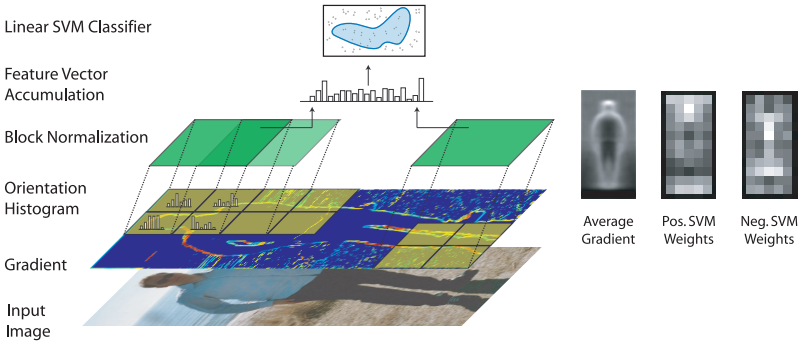


Figure 4.4: Overview of HOG/linSVM architecture. Cells on a spatial grid are shown in yellow, whereas overlapping normalization blocks are shown in green (left). Average gradient image along with visualization of positive and negative SVM weights, which highlight the most discriminative regions for both the pedestrian and non-pedestrian class (right).

(HOG), see Figure 4.4. Local gradients are binned according to their orientation, weighted by their magnitude, within a spatial grid of cells with overlapping block-wise contrast normalization. Within each overlapping block, a feature vector is extracted by sampling the histograms from the contributing spatial cells. The feature vectors for all blocks are concatenated to yield a final feature vector which is subject to classification using a linear support vector machine (linSVM).

Our choice of system parameters is based on the suggestions by [27]. Compared to the Haar wavelet-based cascade and the NN/LRF, we employ a larger border to ensure ample spatial support for robust gradient computation and binning at the pedestrian boundary. Hence, small scale training samples are utilized at a resolution of 22×44 pixels with a border of six pixels, whereas a resolution of 48×96 pixels with a border of twelve pixels is employed for medium scale training.

We utilize fine scale gradients ($[-1, 0, 1]$ masks without smoothing), fine orientation binning (9 bins), coarse spatial binning (2×2 blocks of either 4×4 pixel cells for small scale and 8×8 pixel cells for medium scale training) as well as overlapping block contrast normalization (L_2-norm). The descriptor stride

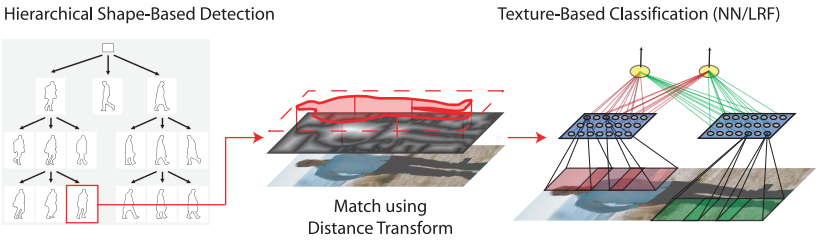


Figure 4.5: Overview of combined shape-based recognition and texture-based classification.

is set to half the block width, in order to have 50 % overlap. This amounts to four pixels for small scale and eight pixels for medium scale training.

Similar to the training of the NN/LRF, the initial 15560 negative samples are randomly sampled from the set of negative images. We apply bootstrapping by extending the training set by 15660 additional false positives in each iteration until test performance saturates. As opposed to the NN/LRF classifier, the complexity of the linear SVM is automatically adjusted during training by increasing the number of support vectors as the training set becomes more complex. Experiments are conducted using the implementation by [27].

Combined Shape-Texture-Based Pedestrian Recognition

We consider a monocular version of the real-time *PROTECTOR* system [56], by cascading shape-based pedestrian recognition with texture-based pedestrian classification. Shape-based recognition is achieved by coarse-to-fine matching of an exemplar-based shape hierarchy to the image data at hand. The shape hierarchy is constructed off-line in an automatic fashion from manually annotated shape labels, extracted from the 3915 pedestrian examples in the training set (see Chapter 2). On-line matching involves traversing the shape hierarchy with the Chamfer distance [14] between a shape template and an image sub-window as smooth and robust similarity measure. Image locations where the similarity between shape and image is above a user-specified threshold are considered recognitions. A single distance threshold applies for each level of the hierarchy. Additional parameters govern the edge density

on which the underlying distance map is based. All parameters have been optimized using a sequential ROC optimization technique [56].

Recognitions of the shape matching step are subject to verification by a texture-based pattern classifier. Here, we employ the multi-layer feed-forward neural network operating on local adaptive receptive field features (NN/LRF) on the small scale training set, with parameters as given above. See Figure 4.5. The initial negative training samples for the NN/LRF classifier were extracted by collecting false positives of the shape-based recognition module (with a relaxed threshold) on the given set of negative images. Finally, bootstrapping is applied to the NN/LRF, as described earlier.

Temporal Integration - Tracking

Temporal integration of recognition results allows to overcome gaps in recognition, suppress spurious false positives and provides higher-level temporally-fused trajectory information for detected objects. Recognitions on trajectory level are fundamental to many real-world attention-focusing or risk-assessment strategies, for instance in vehicle-based collision-mitigation systems or visual surveillance scenarios. In this study, we employ a rudimentary 2D bounding box tracker with an object state model involving bounding box position (x, y) and extent (w, h). Object state parameters are estimated using an $\alpha-\beta$ tracker, involving the classical Hungarian method for data assignment [89]. A new track is started whenever a new object appears in m successive frames and no active track fits to it. It ends, if the object corresponding to an active track has not been detected in n successive frames. We acknowledge the existence of more sophisticated trackers, see Section 2.3, whose performance evaluation remains for future work. The generality and simplicity of our tracker has the advantage to allow a straightforward integration into all detector approaches to be considered.

4.1.3 Experiments

Methodology

Performance evaluation of the pedestrian recognition systems is based on comparing system output (alarms) with manually labeled ground-truth (events) given by bounding box locations of pedestrians using the proposed benchmark test sequence consisting of 21790 monocular images (see Sec-

tion 4.1.1). We differentiate between the scenarios of generic pedestrian recognition and (near) real-time pedestrian recognition from a moving vehicle. There exists a wide range of possible applications of the first scenario, e.g. ranging from surveillance to advanced robotics. The second scenario is geared towards collision mitigation/avoidance in the context of intelligent vehicles [51, 56]. The two scenarios differ in the definition of the area of interest and match criteria. Additionally, the vehicle scenario involves restrictions on average processing time.

In both scenarios, we consider many-to-many data correspondences, that is, an event is matched if there is at least one alarm within localization tolerances, e.g. the systems are not required to detect each individual pedestrian in case of a pedestrian group. Multiple detector responses at near identical locations and scales are addressed in all approaches by applying confidence-based non-maximum suppression to the detected bounding boxes using pairwise box coverage: two system alarms a_i and a_j are subject to non-maximum suppression if their coverage $\Gamma(a_i, a_j)$ is above θ_n, with $\theta_n = 0.5$ in our evaluation. Coverage is defined as the ratio of intersection area and union area:

$$\Gamma(a_i, a_j) = \frac{A\left(a_i \cap a_j\right)}{A\left(a_i \cup a_j\right)} \tag{4.1}$$

The recognition with the lowest confidence is discarded, where confidence is assessed by the classifiers, i.e. cascade (final layer), NN/LRF and SVM decision values.

Performance is evaluated both at frame- and trajectory-level. Frame-level performance is measured in terms of sensitivity, precision and false positives per frame. Sensitivity relates to the percentage of true solutions that were detected, whereas precision corresponds to the percentage of system solutions that were correct. We visualize frame-level performance in terms of ROC curves, depicting the trade-off between sensitivity and false positives per frame based on the corresponding match criteria. ROC curves for the NN/LRF and HOG/linSVM technique are generated by varying the corresponding detector output thresholds along the curve. In case of the wavelet-based cascade and the cascaded shape-texture pedestrian recognition system, there are multiple thresholds (one for each cascade module) that can be varied simultaneously to determine ROC performance. Each multi-dimensional set of thresholds corresponds to a single point in ROC space, where the final ROC curve is computed as the Pareto-optimal frontier of this point cloud

[56].

After incorporating temporal integration (tracking), trajectory-level performance is evaluated in terms of the percentage of matched ground-truth trajectories (sensitivity), the percentage of correct system trajectories (precision) and the number of false trajectories per minute. We distinguish between two types of trajectories (see [56]): "class-B" and "class-A" trajectories that have at least one or at least 50 % of their events matched. All "class-A" trajectories are also "class-B" trajectories, but the former demand stronger application performance. Further, we quantify the reduction in frame-level false positives resulting from the incorporation of the tracking component.

Generic Pedestrian Recognition

In the evaluation of generic pedestrian recognition, no additional (3D) scene knowledge and constraints are employed. Instead, we consider pedestrian recognition solely as a 2D problem, where fully-visible ground-truth pedestrians (see Table 4.1) of at least 72 pixels height are marked as required, which corresponds to real-world pedestrians of 1.5 m height at a distance of 25 m in our camera set-up. Smaller or partially occluded pedestrians and bicyclists or motorcyclists are considered optional, in that the systems are not rewarded / penalized for correct / false / missing detections. In our experiments, we consider in isolation the resolution of the training data (see Section 4.1.2), the size of the detector grid, as well as the effect of adding additional negative training samples by bootstrapping or cascading.

Combined shape-texture-based recognition (Section 4.1.2) is disregarded here, since the shape-based recognition component, providing fast identification of possible pedestrian locations, is mainly employed because of processing speed, which is not considered in this evaluation scenario. We instead evaluate the NN/LRF classifier in isolation, which is the second (and more important) module of the combined shape-texture-based recognition system.

This leaves us with a total of three approaches: the Haar wavelet-based cascade, NN/LRF and HOG/linSVM (cf. Section 4.1.2) which are used in a multi-scale sliding window fashion. With s denoting the current scale, detector windows are both shifted through scale with a step factor of Δ_s and through location at fractions $s\Delta_x$ and $s\Delta_y$ of the base detector window size (see Section 4.1.2) in both x- and y-dimension. The smallest scale s_{min} corresponds to a detector window height of 72 pixels, whereas the largest scale

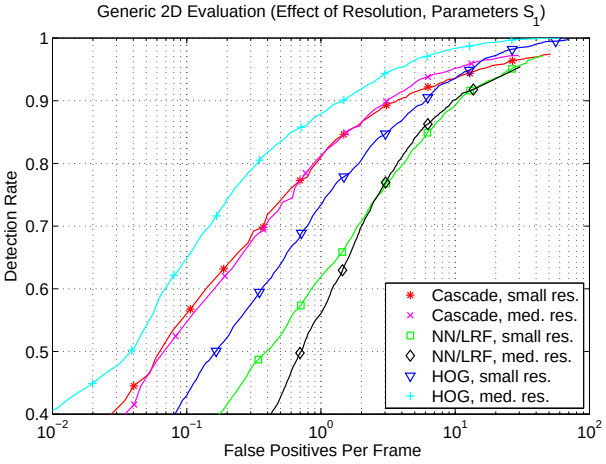
	S_1	S_2	S_3	S_4	S_5	S_6
Spatial Stride (Δ_x, Δ_y)	(0.1,0.025)	(0.15,0.05)	(0.3,0.075)	(0.1,0.025)	(0.15,0.05)	(0.3,0.075)
Scale Step Δ_s	1.1	1.1	1.1	1.25	1.25	1.25
# of detection windows	184392	61790	20890	90982	30608	10256

Table 4.2: Overview of sliding window parameter sets S_i for generic evaluation.

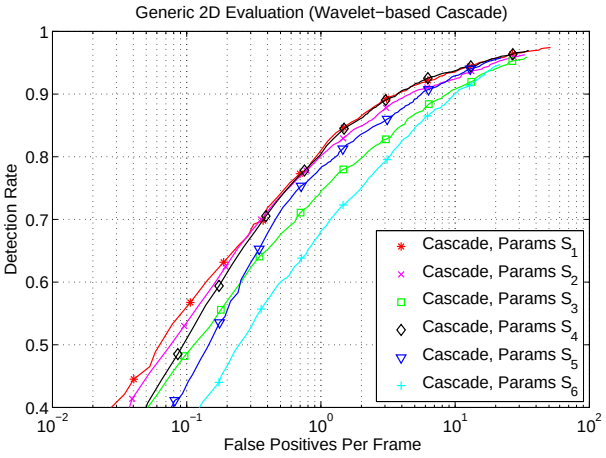
s_{max} has been chosen so that the detector windows still fit in the image. As a result, detector grids for all systems are identical. Several detector parameter settings $S_i = (\Delta_x^i, \Delta_y^i, \Delta_s^i)$, defining spatial stride (detector grid resolution) and scale, have been considered for all approaches, see Table 4.2. The 2D match criterion is based on bounding box coverage between a system alarm a_i and a ground-truth event e_j, where a correct recognition is given by $\Gamma(a_i, e_j) > \theta_m$, with $\theta_m = 0.25$. Results are given in Figures 4.6 - 4.9.

Figure 4.6a shows the effect of different training sample resolutions using detector parameters S_1. While the performance difference between small and medium resolution for the wavelet-based cascade and the NN/LRF detectors is minor, the HOG/linSVM approach performs significantly worse at a small resolution. The reason for that may lie in the reduced spatial support for histogramming. Further experiments involve only the best performing resolution for each system: small resolution for the wavelet-based cascade and the NN/LRF detector and medium resolution for the HOG/linSVM approach.

Figures 4.6b and 4.7 show the localization tolerance of each detector, that is the sensitivity to the granularity of the detection grid. Two observations can be made: First, all detectors perform best using the detection grid at the finest granularity (parameters S_1). Second, the localization tolerances of the approaches vary considerably. The NN/LRF detector performs almost identical for all parameter sets under consideration, with false positives per frame at constant detection rates being reduced by approx. a factor of 1.5, comparing the the best (S_1) and worst (S_6) setting. The wavelet-based cascade and HOG/linSVM approaches show a stronger sensitivity to the detection grid resolution, with a difference in false positives by approx. a factor of 3 and 5.5, respectively.

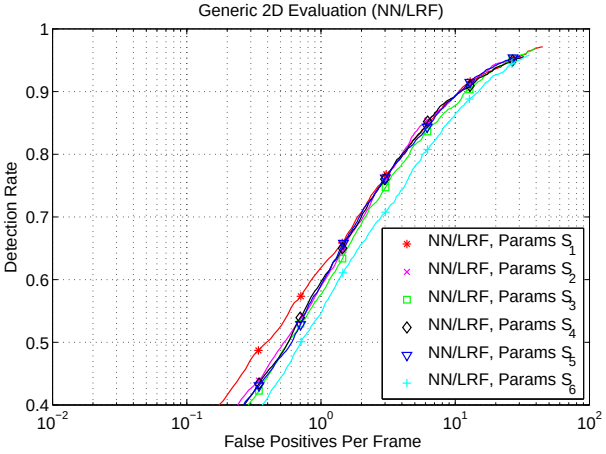


(a)

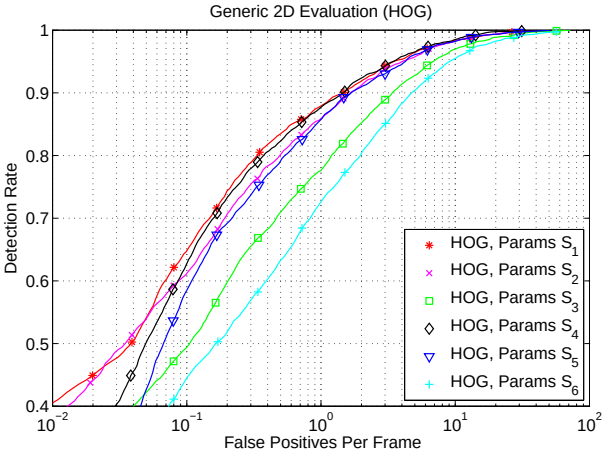


(b)

Figure 4.6: Evaluation of generic pedestrian recognition. (a) Effect of different training resolutions. (b) Effect of varying detector grid for wavelet-based cascade.

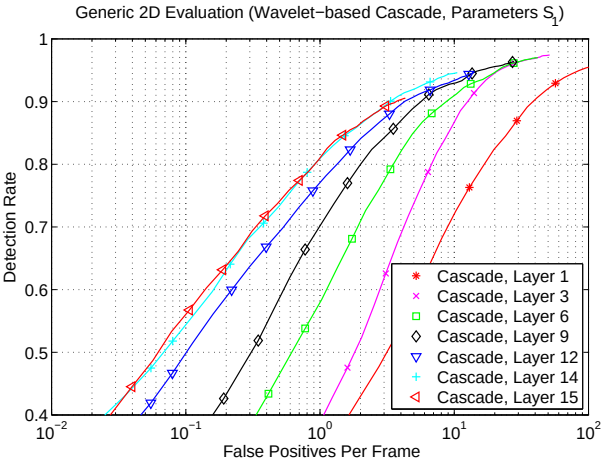


(a)

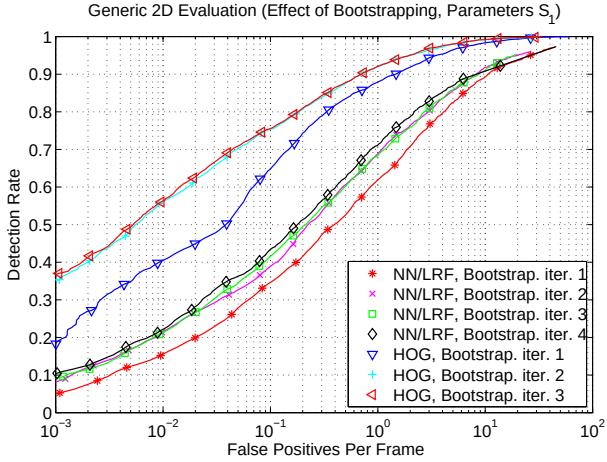


(b)

Figure 4.7: Evaluation of generic pedestrian recognition. Effect of varying detector grid for (a) NN/LRF (1 bootstrapping iteration) and (b) HOG/linSVM (1 bootstrapping iteration).



(a)



(b)

Figure 4.8: Evaluation of generic pedestrian recognition. (a) Performance of individual cascade layers. (b) Effect of bootstrapping on NN/LRF and HOG/linSVM.

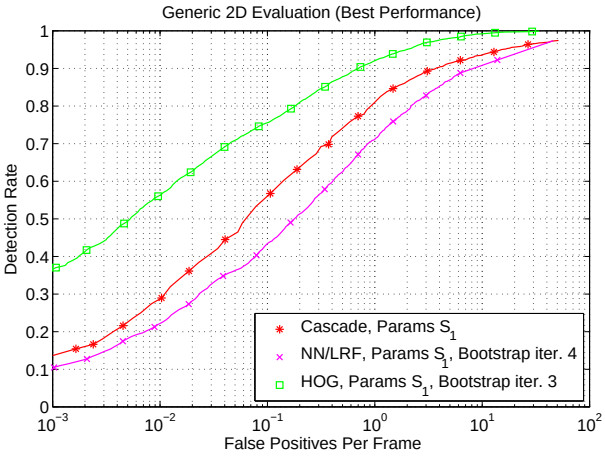


Figure 4.9: Evaluation of generic pedestrian recognition: best performance of each approach.

We attribute this to the fact that the NN/LRF uses comparatively the largest features (5×5 pixel receptive fields at a sample size of 18×36 pixels, see Section 4.1.2), whereas 8×8 pixel cells are used in the HOG/linSVM approach with a sample size of 48×96 pixels (see Section 4.1.2). The wavelet-based cascade employs features at different scales, as shown in Section 4.1.2.

In the following experiments, we restrict ourselves to the detector parameter set S_1 which was identified as the best setting for all techniques. We now evaluate the effect of adding negative samples to the training set, in terms of additional bootstrapping iterations for NN/LRF and HOG/linSVM and show the performance of individual layers of the wavelet-based cascade, each of which is trained on a different and increasingly more difficult set of negative samples. See Figure 4.8. All detectors show an initial performance improvement, but then saturate after 15 layers (wavelet-based cascade) or three (HOG/linSVM) and four (NN/LRF) bootstrapping iterations, respectively. The obtained performance improvements of the wavelet-based cascade and the NN/LRF detectors are paid for with an increase of computational costs, since the classifiers become more complex in case of more difficult training sets (recall that NN/LRF complexity was increased by design during

	Cascade			NN/LRF			HOG/linSVM		
	F	A	B	F	A	B	F	A	B
Sensitivity	65.4 %	61.9 %	73.0 %	65.3 %	69.8 %	81.7 %	64.1 %	61.6 %	76.2 %
Precision	56.1 %	47.3 %	53.8 %	33.5 %	27.5 %	33.3 %	90.2 %	84.9 %	87.2 %
FP 10^3 fr., min	156	19.0	16.7	307	35.7	35.1	16	2.0	1.7
Reduction False Positives	34.3 %	-	-	50.9 %	-	-	22.3 %	-	-
Avg. Proc. Time / 10^3 win.	20 ms			660 ms			430 ms		

Table 4.3: System performance after tracking. F/A/B denote frame- and trajectory-level performance. False positives "FP" are given per 10^3 frames and per minute for frame-level and trajectory performance.

bootstrapping, see Section 4.1.2). However, in the case of the HOG/linSVM detector, the processing time for the evaluation of a single detection window is constant. For a linear SVM, the processing time is independent from the actual number of support vectors [177], which becomes larger as more bootstrapping iterations are conducted. Figure 4.9 shows the best performance of each system on our test dataset. The HOG/linSVM approach clearly outperforms both the wavelet-based cascade and NN/LRF. At a detection rate of 70 %, false positives per frame for the HOG/linSVM detector amount to 0.045, compared to 0.38 and 0.86 for the wavelet-based cascade and NN/LRF. This is a reduction by a factor of 8 and 19, respectively.

Next, temporal integration is incorporated into all approaches using the 2D bounding box tracker (see Section 4.1.2) with parameters $m = 2$ and $n = 2$. Input to the tracker are system recognitions, with system parameterization selected from the corresponding ROC curves, as depicted in Figure 4.9, at a common reference point of 60 % sensitivity. Results are given in Table 4.3. One observes that the relative performance differences as shown in Figure 4.9 still apply after tracking. The HOG/linSVM approach achieves a significantly higher precision at the same sensitivity levels compared to the wavelet-based cascade and the NN/LRF detector.

On-Board Vehicle Application

In case of (near) real-time pedestrian recognition from a moving vehicle, application-specific requirements are specified in 3D. In particular, the sensor coverage area is defined in relation to the vehicle as 10 m - 25 m in longitudinal and ± 4 m in lateral direction. Given a system alarm a_i and

ground-truth event e_j, we enforce a maximum positional deviation in 3D to count the alarm as match, where both 2D ground-truth and 2D recognitions are back-projected into 3D using known camera geometry and the assumption that pedestrians are standing on the ground-plane (ground-plane constraint). Since this ground-plane assumption is only valid for fully-visible pedestrians, partially visible pedestrians are not back-projected into 3D, but matched in 2D with a box coverage of $\theta_m = 0.25$, as described earlier. Only fully-visible ground-truth pedestrians (see Table 4.1) within the sensor coverage area are considered required. Partially visible pedestrians and pedestrians outside the sensor coverage area are regarded as optional (i.e. recognitions are neither credited nor penalized).

Localization tolerances are defined as percentage of distance for lateral (X) and longitudinal (Z) direction with respect to the vehicle. Here, we consider tolerances of $X = 10$ % and $Z = 30$ % with a larger tolerance in longitudinal direction to account for non-flat road surface and vehicle pitch in case of back-projection of (monocular) ground-truth and recognitions into 3D, e.g. at 20 m distance, we tolerate a localization error of ± 2 m and ± 6 m in lateral and longitudinal direction.

All systems are evaluated by incorporating 3D scene knowledge into the recognition process: we assume pedestrians of heights 1.5 m - 2.0 m to be standing on the ground. Initial object hypotheses violating these assumptions are discarded. Non-flat road surface and vehicle pitch are modeled by relaxing the ground-plane constraint using a pitch angle tolerance of $\psi = \pm 2^\circ$.

We consider constraints on average processing time of 2.5 s and 250 ms (± 10 % tolerance) per image. To enforce these constraints, we choose to maintain the fundamental system parameters, e.g. sample resolution or feature layout, as reported by the original authors, see Section 4.1.2. Instead, we use the granularity of the detection grid as a proxy for processing speed.

Sliding window parameters T_i subject to processing time constraints are given in Table 4.4. The detector grids are finer grained in y-direction than in x-direction. This results in higher localization accuracy in y-direction which adds robustness to depth estimation by back-projecting recognitions into 3D. Instead of a sliding window approach, the combined shape-texture detector uses a coarse-to-fine hierarchical shape matching scheme yielding a variable number of hypotheses per image which are processed by the subsequent NN/LRF classifier. Hence, the hierarchy level thresholds of the shape matching module have the largest influence on processing time. We have in-

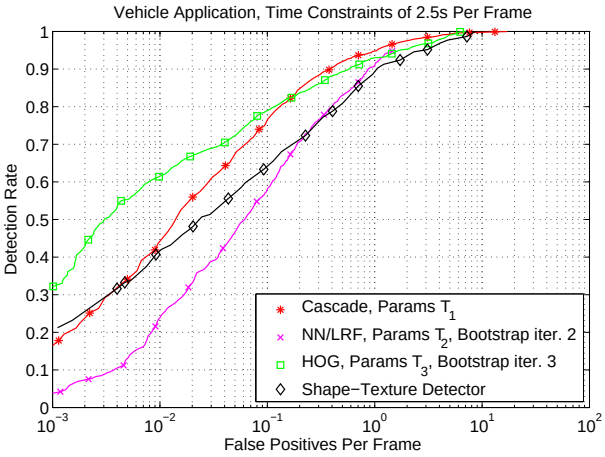
	Cascade		NN/LRF		HOG/linSVM	
	T_1 (2.5s)	T_4 (250ms)	T_2 (2.5s)	T_5 (250ms)	T_3 (2.5s)	T_6 (250ms)
Spatial Stride (Δ_x, Δ_y)	(0.05,0.025)	(0.05,0.025)	(0.1,0.025)	(0.3,0.08)	(0.1,0.025)	(0.3,0.08)
Scale Step Δ_s	1.05	1.05	1.1	1.25	1.1	1.25
# of detection windows	11312	11312	5920	617	5920	617

Table 4.4: Overview of sliding window parameter sets T_i for on-board vehicle evaluation.

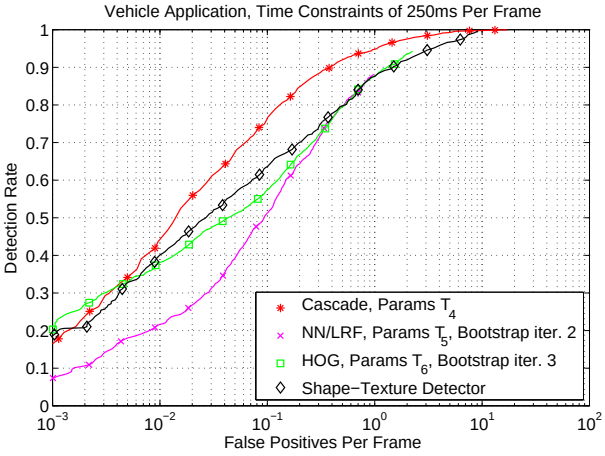
corporated time constraints into the parameter optimization [56], to optimize these thresholds for the given processing time requirements.

Performance is evaluated for the full 15-layer cascade, the shape-texture detector, as well as the HOG/linSVM and NN/LRF approaches after every bootstrapping iteration to find the best compromise between performance and processing speed under the given time constraints. In contrast to the results of the generic evaluation, the best performance of the NN/LRF classifier is reached after the second bootstrapping iteration, since the higher computational costs of more complex NN/LRF detectors require a too large reduction in detection grid resolution to meet the time constraints. In case of the wavelet-based cascade, identical parameter settings T_1 and T_4 are used for both time constraints settings. This is due to a very dense detection grid resolution even at time constraints of 250 ms per frame, since each detection window can be evaluated very rapidly. A further increase of grid resolution does not yield any performance improvements. We attribute this effect to the pre-processing of the training data, where robustness to localization errors is explicitly modeled in terms of shifting the training labels by a few pixels, as described in Section 4.1.1. Results are given in Figure 4.10.

With processing time constraints of 2.5 s per frame, the relative performance of all detector variants is similar to the case of generic evaluation, see Figures 4.9 and 4.10a. Compared to the application of the NN/LRF in isolation, the combined shape-texture detector further improves performance, particularly at low false positive rates. Further restricting processing time constraints to 250 ms per frame effects a massive drop in the performance



(a)



(b)

Figure 4.10: Results of on-board vehicle application using time constraints of (a) 2.5 s / frame and (b) 250 ms / frame.

of the HOG/linSVM detector, whereas the performance of the NN/LRF decreases only slightly. Again, this is an effect of the different localization tolerances. The performance of the combined shape-texture detector remains approximately constant. This indicates the powerful pruning capability of the shape recognition module which allows to quickly focus the subsequent costly texture classification on promising image regions, which reduces computational costs. At tight processing time constraints, the wavelet-based cascade significantly outperforms every other detector considered, benefiting from its high processing speed. The combined shape-texture detector delivers the second best performance, admittedly at a proper gap.

As in the case of generic pedestrian recognition, the bounding box tracker is incorporated. As a common reference point we again use 60 % sensitivity, obtained from the ROC curves depicted in Figure 4.10. Results are given in Table 4.5. For both time constraint settings, the relative performance order of various systems does not change in comparison to Figure 4.10. How-

		Cascade			NN/LRF		
		F	A	B	F	A	B
Sensitivity	(TC 2.5 s)	64.9 %	58.2 %	79.1 %	65.5 %	67.1 %	82.1 %
Precision	(TC 2.5 s)	77.2 %	71.5 %	75.5 %	53.4 %	58.3 %	63.1 %
FP 10^3 fr., min	(TC 2.5 s)	32	5.5	5.1	102	8.8	7.8
Reduction FP	(TC 2.5 s)	23.6 %	-	-	30.6 %	-	-
Sensitivity	(TC 250 ms)	64.9 %	58.2 %	79.1 %	67.0 %	71.6 %	80.6 %
Precision	(TC 250 ms)	77.2 %	71.5 %	75.5 %	43.4 %	45.6 %	52.2 %
FP 10^3 fr., min	(TC 250 ms)	32	5.5	5.1	171	17.2	15.0
Reduction FP	(TC 250 ms)	23.6 %	-	-	31.3 %	-	-
Avg. Proc. Time / 10^3 windows		20 ms			440 ms		

		HOG/linSVM			Shape-Texture Rec.		
		F	A	B	F	A	B
Sensitivity	(TC 2.5 s)	64.3 %	58.2 %	68.7 %	64.6 %	65.6 %	85.0 %
Precision	(TC 2.5 s)	88.7 %	81.2 %	84.8 %	59.3 %	52.7 %	62.1 %
FP 10^3 fr., min	(TC 2.5 s)	11.7	1.7	1.4	78	9.5	9.1
Reduction FP	(TC 2.5 s)	12.5 %	-	-	28.9 %	-	-
Sensitivity	(TC 250 ms)	67.4 %	65.7 %	79.1 %	63.1 %	65.2 %	80.1 %
Precision	(TC 250 ms)	47.6 %	50.8 %	55.8 %	59.2 %	51.3 %	61.9 %
FP 10^3 fr., min	(TC 250 ms)	143	14.5	13.0	81	9.1	8.7
Reduction FP	(TC 250 ms)	37.3 %	-	-	26.1 %	-	-
Avg. Proc. Time / 10^3 windows		430 ms			approx. 620 ms		

Table 4.5: System performance after tracking. F/A/B denote frame- and trajectory-level performance under processing time constraints "TC" of 2.5 s and 250 ms per image. False positives "FP" are given per 10^3 frames and per minute for frame-level and trajectory performance.

ever, differences in the beneficial effect of the tracker can be observed. For all systems except for HOG/linSVM, the benefit of the tracker is similar for the two time constraint settings, approx. 25 % - 35 %, see Table 4.5. For the HOG/linSVM detector at time constraints of 2.5 s per image, most false detections turn out to exhibit strong temporal coherence and cannot be eliminated by the tracker. The reduction in false positives only amounts to 12.5 %. The stronger benefit of the tracker for the HOG/linSVM detector at 250 ms per image can be explained by the fact that fewer detection windows can be evaluated per image. To reach a sensitivity of 60 %, a more relaxed threshold setting is required. As a result, additional spurious false positives are introduced which are observed to be less temporally coherent; these can be successfully suppressed by the tracker.

The average processing time per 10^3 detection windows is given in Table 4.5 using implementations in C/C++ on a 2.66 GHz Intel processor. In comparison to the other approaches, the wavelet-based cascade architecture has a massive advantage in processing time, i.e. it is approx. 20 times faster. Note that the combined shape-texture detector has the highest processing time per detection window. However, due the efficient pruning of the search space by the coarse-to-fine shape matching module, the number of detection windows per image is greatly reduced in comparison to the sliding window approaches, while maintaining similar performance levels.

4.1.4 Discussion

We obtained a nuanced picture regarding the relative performance of methods tested, where the latter depends on the pedestrian image resolution and the spatial grid size used for probing (used as proxy for processing speed). At low resolution pedestrian images (e.g. 18×36 pixels), dense Haar wavelet features represent the most viable option. HOG features, on the other hand, perform best at intermediate resolutions (e.g. 48×96 pixels). Their need for a larger spatial support limit their use in some application scenarios; for example in our camera set-up of Section 4.1.3, pedestrians further away than 25 m to the vehicle appear in the image with a height of less than 72 pixels. We would expect component-based, e.g. [2, 31, 35, 105, 108, 139, 143, 167, 174], or deformable part approaches, e.g. [1, 46, 47, 48, 84, 92, 94, 95, 138], to be the natural choice for those applications involving yet higher resolution pedestrian images.

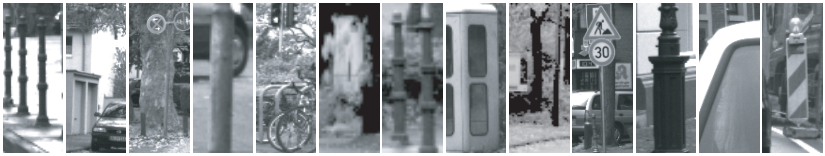


Figure 4.11: Typical false positives of all systems. Most errors occur in local regions with strong vertical structure.

In terms of overall systems, results indicate a clear advantage of the HOG-based linear SVM approach at intermediate pedestrian image resolutions and lower processing speeds, and a superiority of the wavelet-based AdaBoost cascade approach at lower pedestrian image resolutions and (near) real-time processing speeds. Not surprisingly, tracking improves the performances of all considered systems, it also decreases the absolute performance differences amongst the systems. We observe that the tested systems in this study tend to make rather similar mistakes, although they are based on different features. For all systems, typical false detections occur in local regions which are dominated by strong vertical structure, as shown in Figure 4.11.

It is instructive to place the best performance obtained in context, by comparing what would be necessary in a realistic application. Let us consider for this the intelligent vehicle application. If we assume a driver assistance system using monocular vision, that acoustically warns the driver for possible collisions with pedestrians, a correct detection rate upwards of 80 % on trajectory-level would be sensible, say, at a rate of less than one false alarms per 10 hours driving in urban traffic. Looking at the results currently obtained within 250 ms per frame (assuming that optimization would result in a real-time implementation), see Table 4.5, we see the best performance of approx. six false trajectories per minute at a detection rate of 60 % for the wavelet-based cascade. One might be tempted to conclude that a performance gap of three orders of magnitude exists. This would be overly pessimistic, though, since Table 4.5 reflects the average performance over all pedestrian trajectories within the defined coverage area (10 - 25 m in distance, up to ± 4 m laterally). In practice, trajectories that are collision-relevant tend to be longer and individual recognitions are easier, as they come closer to the vehicle. Our preliminary investigations show that recognition performance on such trajectory subsets can be up to one order of magnitude higher, leaving

a performance gap of two orders of magnitude.

How could one close the remaining performance gap? The most effective solution is to incorporate a pre-processing stage to constrain the image search space, based on alternate cues such as motion [40, 124], see Section 4.2, and depth [2, 17, 41, 56, 85, 92, 94, 110, 112, 180]. For example, [56] reports a performance gain of an order of magnitude by the inclusion of stereo-based obstacle detection (a similar boost can be expected in a surveillance setting by the incorporation of background subtraction).

Any remaining performance gain (i.e. one order of magnitude for the intelligent vehicle application listed above) would likely need to be derived from improving the actual classification methods. For example, in the shape-texture approach of Section 4.1.2, hierarchical shape matching can be performed probabilistically, with improved performance [54]. The particular shape template matched could furthermore index into a set of classifiers (experts), each attuned to a particular body pose. [56] reports a performance improvement of about 30 % from such a mixture-of-experts architecture. The cascade approach could be paired up with more powerful features, e.g. local receptive fields or gradient histograms (cf. Section 4.1.2). [182] presented initial work on cascade detectors using HOG features and reported real-time processing speeds at performance levels similar to the original HOG/linSVM approach [27]. Irrespective of the utilized feature set, the classification techniques could use multiple cues, features or modalities to improve performance [28, 35, 39, 118, 131, 137, 163, 165, 166, 167, 173, 175], see Chapter 6.

Or perhaps, it is the data that matters most, after all. A study on pedestrian classification [109] showed that the benefit of selecting the best combination of features and pattern classifiers was less pronounced than the gain obtained by increasing the training set, even though the base training set already involved many thousands of samples [109]. Adaptive feature sets in particular, e.g. LRF features, are expected to benefit more from an enlarged training set than non-adaptive features, e.g. HOG features, since the training data directly influences the development of the features.

4.1.5 Conclusion

This section presented an experimental study on monocular pedestrian recognition. In order to strike a suitable balance between generality and specificity, we considered two evaluation settings: a generic setting, where evaluation is

done without scene and processing constraints, and one specific to an application on-board a moving vehicle in traffic.

Results show a nuanced picture regarding the relative performance of methods tested, where the latter depends on the pedestrian image resolution and the spatial grid size used for probing (used as proxy for processing speed). The HOG-based linear SVM approach significantly outperformed all other approaches considered at little or no processing constraints (factors of 10 - 18 and 3 - 6 less false class-A trajectories at no time constraints and at 2.5 s per frame, respectively). This suggests that feature representations based on local edge orientation are well-suited to capture the complex appearance of the pedestrian object class. As tighter processing constraints are imposed, the Haar wavelet-based cascade approach outperforms all other detectors considered (factor of 2 - 3 less false class-A trajectories at 250 ms per frame).

For all systems, performance is enhanced by incorporating temporal integration and/or restrictions of the search space based on scene knowledge. The tracking component tends to decrease the absolute performance differences of the systems. From a real-world application perspective, the amount of false trajectories is too high by at least two orders of magnitude. Hence, this thesis will further present methods and techniques to boost performance. After evaluating a motion-based attention focusing strategy in Section 4.2, we will focus on the classification component in isolation in the remainder of this work.

4.2 Monocular Pedestrian Recognition Using Motion Parallax

4.2.1 Overview

This section aims at improving monocular pedestrian recognition performance by introducing an early attention stage to narrow down the hypotheses search space for subsequent complex pedestrian detectors, cf. Section 2.1. To that extent, we propose an attentive concept involving a probabilistic model of ego-motion corrected optical flow features, particularly attuned to the pedestrian class.

The general idea of early focus of attention is independent of the actual pedestrian recognition system used, cf. Section 2.2. We integrate the proposed hypotheses generation technique with real-time (monocular) shape-texture based pedestrian recognition and tracking [56], as presented in Sec-

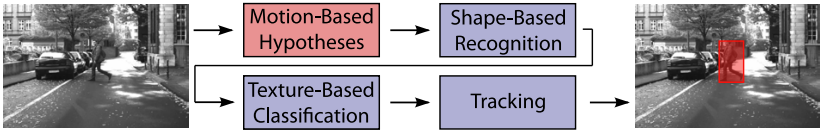


Figure 4.12: Overview of the integrated pedestrian recognition system comprising motion-based hypotheses generation, shape-based recognition, texture-based classification and tracking. We focus on the motion-based hypotheses generation module (red), but evaluate the whole integrated system.

tion 4.1.2. Shape-based recognition is achieved by efficient matching of an exemplar-based shape hierarchy to the generated hypotheses. Shape matches are verified by a texture-based pedestrian classifier, a neural network operating on local adaptive receptive fields. Temporal integration is provided by an $\alpha - \beta$ tracker. The integrated system combines three cues, i.e. motion, shape and texture, on module-level within a single system. See Figure 4.12.

Our attentive strategy utilizes a learned probabilistic model of motion-based features, which are particularly attuned to pedestrians. The features involve mean horizontal velocity and density of local parallax flow. The application of parallax flow, see Section 4.2.2, allows to focus on static non-planar or moving objects, while at the same time disregarding camera ego-motion. Further, this representation seamlessly extends to the recognition of static pedestrians, unlike previous approaches which require target motion [26, 45, 124, 139]. We employ a cascade structure with complementary cues for each module, as depicted in Figure 4.12, to successively narrow down the search space, see [56]. The proposed motion-model is utilized as hypotheses generation module for subsequent shape- and texture-based pedestrian classification, based on a sound Bayesian assessment of posterior probabilities for each hypothesis. Parameters of the integrated multi-cue system are optimized with regard to robustness and efficiency for maximum real-time performance, by employing sequential ROC optimization [56]. Details are given in the next section.

4.2.2 Motion-Based Pedestrian Model

The proposed probabilistic motion-based pedestrian model is based on sparse optical flow features, e.g. [10], induced by our moving camera. Rather than directly utilizing the observed intrinsic flow field, camera ego-motion is canceled out (estimated from inertial sensors) by computing the parallax flow field [7]. Parallax flow is the difference between the intrinsic optical flow field and estimated ground-plane flow. Residual parallax flow vectors are then induced by both static non-planar and moving objects. See Figure 4.13.

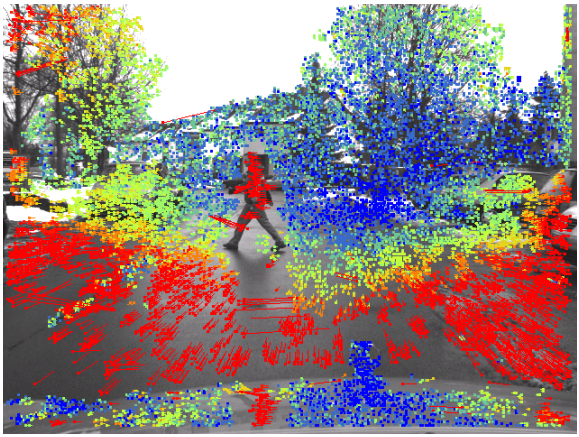
Given a sparse parallax flow field $F(\mathbf{p}) = \mathbf{f_p}$, let $\mathbf{f_p} = (f^u_\mathbf{p}, f^v_\mathbf{p})$ denote horizontal and vertical components of parallax flow at pixel location $\mathbf{p}$. Further, we introduce a function $S_F(\mathbf{p})$, assessing sparseness of a given parallax flow field F at location $\mathbf{p}$, with $S_F(\mathbf{p}) = 1$, if $\mathbf{f_p}$ exists, and $S_F(\mathbf{p}) = 0$ otherwise. For an arbitrary region of interest $R \subset F$, our aim is to estimate its posterior probability, $P(\omega_0|R)$, with respect to the pedestrian class ω_0. To that extent, we represent R in terms of a feature set φ_R based on parallax flow and follow a Bayesian approach:

$$P(\omega_0|R) = P(\omega_0|\varphi_R) = \frac{p(\varphi_R|\omega_0)P(\omega_0)}{\sum_{i=0}^{1} p(\varphi_R|\omega_i)P(\omega_i)} \tag{4.2}$$

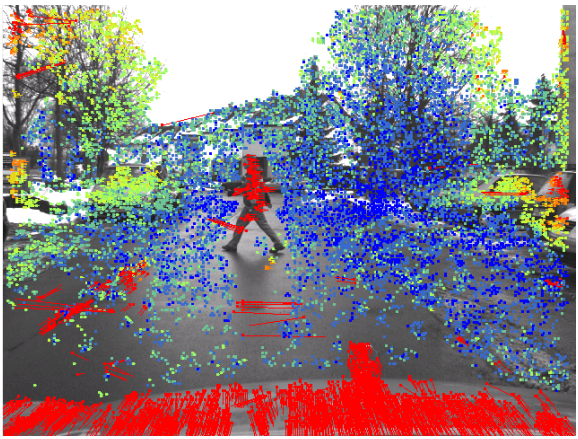
Priors for the pedestrian ω_0 and non-pedestrian class ω_1, $P(\omega_0)$ and $P(\omega_1)$, are assumed uniform. In the following, details on feature selection and the estimation of the likelihoods $p(\varphi_R|\omega_i)$ are given.

Selecting appropriate motion-based features φ_R involves a trade-off between generality and specificity. On the one hand, features should be general with respect to arbitrary pedestrian appearance and motion, while on the other hand, powerful distinction between pedestrians and background is desired. In view of utilizing the proposed model as a focus-of-attention strategy for subsequent classification modules, our main concern at this point is generality. Hence, the proposed motion-based features purposely involve rather generic measures to not reject potential pedestrian candidate regions too early in the processing cascade. We consider mean horizontal velocity V_R and density D_R within a local hypothesis R. To enhance specificity, features are particularly attuned to pedestrians in terms of both a probabilistic weighting scheme and statistical combination.

Pedestrian motion typically involves a characteristic velocity range, as opposed to other moving objects in urban traffic. Given that pedestrian motion



(a)



(b)

Figure 4.13: Visualization of the considered optical flow fields. Warmer colors encode longer flow vectors. (a) Observed intrinsic flow. (b) Ego-motion corrected parallax flow focusing on static non-planar and moving objects. Note that the resulting flow on the ground is close to zero (except for false correspondences).

is predominantly horizontal, we restrict ourselves to the horizontal velocity component $f^u_{\mathbf{p}}$ of parallax flow vectors. At the same time, errors induced by excessive pitch-movement of the camera are alleviated. Let $P_w(\mathbf{p}_r|\omega_0)$ denote a location-specific probabilistic weighting scheme for pixels $\mathbf{p}_r \in R$ which is employed to adapt the proposed features to the pedestrian class (the definition of $P_w(\mathbf{p}_r|\omega_0)$ is given below). Then, V_R involves the weighted mean of horizontal flow components $f^u_{\mathbf{p}_r}$ at locations $\mathbf{p}_r$:

$$V_R = \frac{\sum_{\mathbf{p}_r \in R} |f^u_{\mathbf{p}_r}| S_F(\mathbf{p}_r) P_w(\mathbf{p}_r|\omega_0)}{\sum_{\mathbf{p}_r \in R} S_F(\mathbf{p}_r) P_w(\mathbf{p}_r|\omega_0)} \tag{4.3}$$

As a second feature, we propose weighted mean flow-density D_R within R, again utilizing $P_w(\mathbf{p}_r|\omega_0)$ as weighting scheme:

$$D_R = \frac{\sum_{\mathbf{p}_r \in R} S_F(\mathbf{p}_r) P_w(\mathbf{p}_r|\omega_0)}{\sum_{\mathbf{p}_r \in R} P_w(\mathbf{p}_r|\omega_0)} \tag{4.4}$$

Local flow density in regions corresponding to pedestrians is expected to be rather sparse, in particular within the lower body area. The highly articulated and non-rigid pedestrian motion, combined with continuously appearing and disappearing background, as well as self-occlusions, negatively affects the computation of correspondences. Hence, the local density measure aims to distinct pedestrians from largely rigid objects, where recovered flow estimates are taken to be more dense. See Figure 4.13.

To enhance specificity of the proposed motion-based features to the pedestrian class, $P_w(\mathbf{p}_r|\omega_0)$ has been introduced as a weighting paradigm, see Equations (4.3) and (4.4). $P_w(\mathbf{p}_r|\omega_0)$ denotes a two-dimensional probability mass function, representing the probability that a given location $\mathbf{p}_r \in R$ corresponds to a pedestrian. To estimate $P_w(\mathbf{p}_r|\omega_0)$, the superposition of a set of S aligned binary pedestrian foreground masks, $m_s(\mathbf{p}_r)$, as defined by manually labeled pedestrian contours, is utilized, see Figure 4.14:

$$P_w(\mathbf{p}_r|\omega_0) \sim \sum_{s=1}^{S} m_s(\mathbf{p}_r), \tag{4.5}$$

with P_w spatially scaled to the dimensions of the hypothesis R and normalized such that

$$0 \leq P_w(\mathbf{p}_r|\omega_0) \leq 1. \tag{4.6}$$

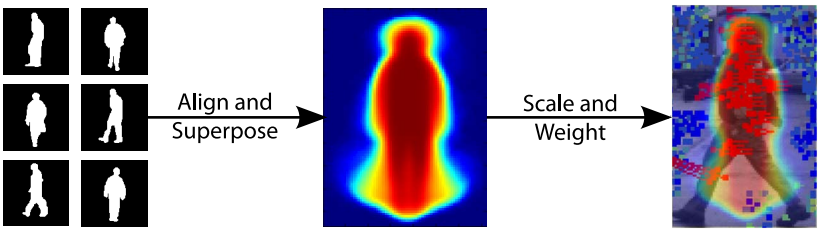


Figure 4.14: Overview of the probabilistic weighting scheme to adapt the motion-based flow-features to the pedestrian class. A two-dimensional probability mass function (center) is learned from binary pedestrian foreground masks (left) and utilized to weight feature computation (right). Warmer colors encode higher probability.

To further increase the discriminative power and robustness of the proposed features, we consider statistically combining V_R and D_R into a multidimensional feature $\varphi_R = V_R \wedge D_R$. Under the assumption of independence of V_R and D_R, the likelihood functions in Equation (4.2) can be decomposed into:

$$p(\varphi_R|\omega_i) = p(V_R \wedge D_R|\omega_i) = p(V_R|\omega_i)p(D_R|\omega_i) \tag{4.7}$$

Approximations of $p(V_R|\omega_i)$ and $p(D_R|\omega_i)$ are obtained via histogramming of training data with regard to the proposed features. In case of pedestrians, we utilize manually labeled bounding boxes, whereas non-pedestrian labels are randomly extracted from parallax flow fields of non-pedestrian images using ground-plane constraints.

4.2.3 System Integration

The probabilistic motion-based pedestrian model, as introduced in Section 4.2.2, is utilized as attentive method within a hypotheses generation module, see Figure 4.12. This module involves three components: optical (parallax) flow computation, generation of location hypotheses and filtering of hypotheses. The filtered hypotheses define initial search areas for the subsequent recognition module. A functional overview of these sub-components is given in Figure 4.15.

We consider the proposed flow-based features, see Section 4.2.2, as inde-

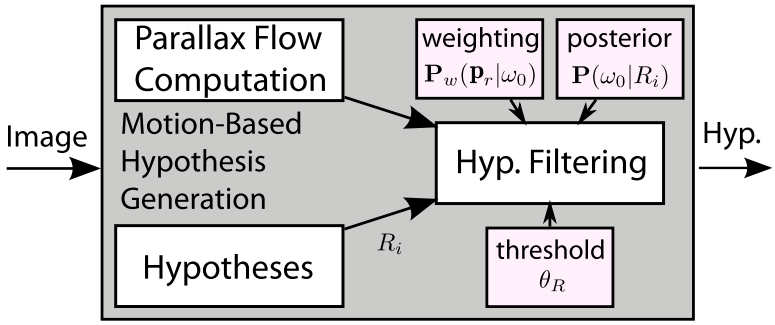


Figure 4.15: Overview of motion-based hypotheses generation, cf. Figure 4.12: parallax flow computation, generation of location hypotheses and filtering of hypotheses.

pendent from the actual algorithm to compute optical flow [10]. In our experiments, an efficient technique involving correspondences based on census transform signatures is utilized [148]. This allows for real-time flow computation (25 Hz). Parallax flow is obtained by canceling out camera ego-motion, as estimated from inertial sensors [7].

Initial object location hypotheses R_i are generated using the sliding window technique, where detector windows at various scales and locations are shifted over the image. Here, application-specific scene constraints, such as flat-world assumption, people standing on the ground or prior knowledge about the dimensions of target objects, are incorporated, see Figure 4.16a. Each pedestrian candidate region R_i is represented in terms of features V_{R_i} and D_{R_i}, followed by the estimation of posterior probability with respect to the pedestrian class, $P(\omega_0|R_i)$, see Equation (4.2). A threshold θ_R governs the amount of hypotheses which are committed to the subsequent module: Only hypotheses with $P(\omega_0|R_i) > \theta_R$, as shown in Figure 4.16b, trigger the evaluation of the next cascade module. Others are rejected immediately.

Pedestrian recognition proceeds with shape-based recognition, as shown in Figure 4.16c, involving coarse-to-fine matching of an exemplar-based shape hierarchy to the image data at hand [56]. Positional initialization is given by the output hypotheses of the motion-based attention stage. The shape hierarchy is constructed off-line in an automatic fashion from manually anno-

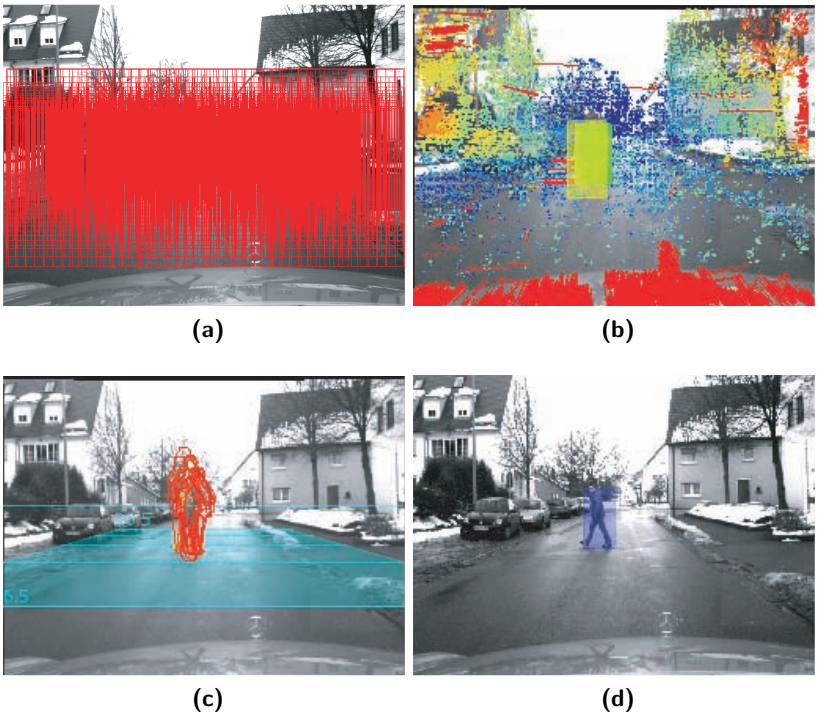


Figure 4.16: Results of the integrated pedestrian recognition system. (a) Initial location hypotheses. (b) Hypotheses filtered using the proposed motion-based focus of attention strategy. (c) Results of shape-based recognition. (d) System output after texture-based classification and tracking.

tated shape labels. On-line matching involves traversing the shape hierarchy with the Chamfer distance [14] between a shape template and an image sub-window as smooth and robust similarity measure. Image locations, where the similarity between shape and image is above a user-specified threshold, are considered recognitions. A single distance threshold applies for each level of the hierarchy. Additional parameters govern the edge density on which the

	dataset 1	dataset 2	dataset 3
# of images	2970	2000	2702
# of pedestrian trajectories	45	52	38
pedestrian labels	1 606	1 302	1 625
non-pedestrian labels	910 369	856 256	832 639

Table 4.6: Dataset statistics.

underlying distance map is based.

Recognitions of the shape matching step are subject to verification by a texture-based pattern classifier. Here, we employ a multi-layer feed-forward neural network operating on local adaptive receptive field features [56, 170]. Finally temporal integration of recognition results is employed to overcome gaps in recognition and suppress spurious false positives. A 2D bounding box tracker is utilized, with an object state model involving bounding box position and extent [56]. State parameters are estimated using an $\alpha - \beta$ tracker, see Figure 4.16d and Section 4.1.2.

4.2.4 Experiments

The proposed motion-based attention strategy is tested in experiments on pedestrian recognition from a moving vehicle. Datasets were acquired in daylight conditions in urban traffic and depict non-occluded pedestrians in front of a changing background. Pedestrian labels were manually extracted, whereas non-pedestrian labels were obtained randomly from non-pedestrian images using the sliding window technique in conjunction with ground-plane constraints. See Table 4.6 for the datasets used. In all experiments, we perform threefold cross-validation: Two datasets are utilized at a time to learn the probabilistic model of motion-features and to optimize parameters, respectively. Performance is evaluated on the remaining dataset.

In a first experiment, the proposed motion-based features, see Section 4.2.2, are evaluated. In particular, we consider both mean horizontal velocity and density as single features, $\varphi_R = V_R$ and $\varphi_R = D_R$, as well as the statistically combined multi-dimensional feature $\varphi_R = V_R \wedge D_R$, see Equation (4.7). To evaluate the inherent quality of flow features, the manually labeled pedestrians and corresponding non-pedestrians, see Table 4.6, are directly employed as training and test sets. That is, we consider pedestrian *classification* utiliz-

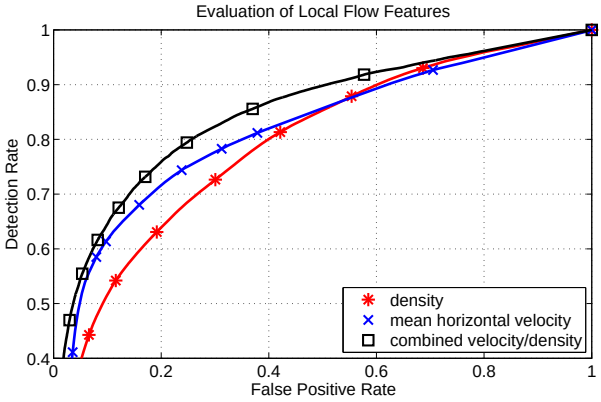
ing bounding box labels, instead of evaluating pedestrian *recognition* performance on an image sequence. A given test sample R is classified as pedestrian ω_0, if the associated posterior probability $P(\omega_0|\varphi_R) > \theta_R$, see Equation (4.2). Figure 4.17a shows the performances of different features sets φ_R, in terms of mean ROC curves of all three cross-validation runs, with the threshold θ_R varied along the curves. It is observed, that the mean horizontal velocity feature V_R is superior to the density feature D_R. Further performance boost is achieved by statistically combining both features to a robust multi-dimensional feature $\varphi_R = V_R \wedge D_R$, see Equation (4.7).

We now turn our attention to the problem of pedestrian recognition using test sequences consisting of entire images at a size of 640×480 pixels, see Table 4.6. The proposed integrated system using motion-based hypotheses generation, see Figure 4.12, is compared to an otherwise identical monocular recognition system without any hypotheses generation. Further, we compare to a stereo-based pedestrian recognition system, using depth information for hypotheses generation, see [56].

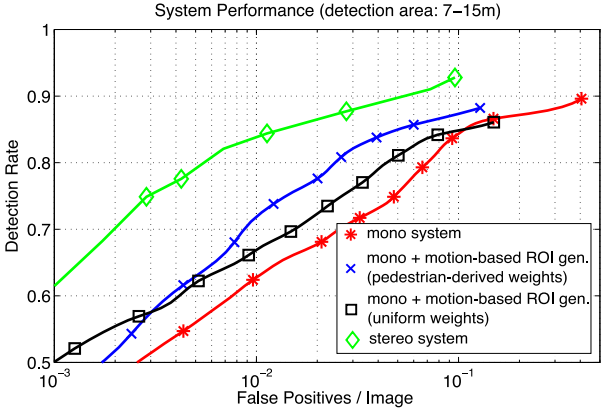
The motion-based hypotheses generation module utilizes the combined velocity/density feature (the best performing variant in Figure 4.17a) in conjunction with two different feature weighting strategies: First, we employ the proposed probabilistic weighting scheme derived from pedestrian foreground masks, see Figure 4.14, with the weight for each location $\mathbf{p}_r$ within an hypothesis R, $P_w(\mathbf{p}_r|\omega_0)$, determined using Equations (4.5) and (4.6). Further, we consider equal weights, with $P_w(\mathbf{p}_r|\omega_0)$ defined as a uniform distribution.

To balance efficiency and robustness for maximum performance of all considered systems, significant parameters of each module, see Figure 4.12 have been optimized using sequential ROC optimization [56]. Parameters subject to optimization include the posterior threshold θ_R for motion-based hypotheses generation, edge and distance thresholds for shape-based recognition, output threshold for texture-based classification, as well as track start and termination criteria for tracking. This technique avoids ad-hoc parameter tuning and provides tight module integration.

Evaluation criteria and application-specific requirements for pedestrian recognition are specified in 3D. In particular, the sensor coverage area is defined in relation to the vehicle as 7 m - 15 m in longitudinal and ± 3 m in lateral direction. Only fully-visible ground-truth pedestrians within the sensor coverage area are considered required, others are regarded as optional, in the sense that systems are not rewarded/penalized for correct/false/missing



(a)



(b)

Figure 4.17: Mean ROC performance of three cross-validation runs for (a) evaluation of local flow features and (b) different variants of pedestrian recognition systems.

detections.

Given a system alarm and a ground-truth event, we enforce a maximum positional deviation in 3D to count the alarm as match, where both 2D ground-truth and 2D recognitions are back-projected into 3D using known camera geometry and ground-plane constraints. Localization tolerances are defined as percentage of distance for lateral (X) and longitudinal (Z) direction with respect to the vehicle. Here, we consider tolerances of $X = 10$ % and $Z = 30$ % with a larger tolerance in longitudinal direction to account for non-flat road surface and vehicle pitch in case of back-projection of (monocular) ground-truth and recognitions into 3D.

Performance is given in terms of mean ROC curves over three cross-validation runs, depicting system performance (detection rate vs. false positives per frame) after the final module (tracking) for each system under consideration. From Figure 4.17b it is observed, that the presented attentive strategy involving motion-based hypotheses generation improves performance of an otherwise identical monocular pedestrian recognition system, even if uniform feature-weights are used (black curve vs. red curve). Additional performance gain is achieved by increasing feature-specificity in terms of the proposed probabilistic weighting scheme which is derived from pedestrian foreground masks (blue curve). Compared to the monocular system without any attention mechanism, false positives are significantly reduced by a factor of two, at equal detection rates (blue curve vs. red curve). The system variant utilizing stereo vision to obtain initial hypotheses, outperforms all other monocular approaches by an order of magnitude (green curve).

Processing time has been evaluated using implementations in C/C++ on an Intel 2.4 GHz processor, see Table 4.7. Compared to the regular monocular pedestrian recognition system, the proposed motion-based attention strategy yields a significant boost in recognition performance, paid for with only a minor increase in processing time (7.20 Hz vs. 8.10 Hz). Using a uniform feature-weighting strategy results in a significant reduction of computational resources (14.9 Hz vs. 7.20 Hz) at the cost of a decrease in recognition performance, since the motion-based features V_R and D_R are less specific to the pedestrian class. This cut of computational costs is due to the fact that uniform weighting allows to exploit integral images, as proposed by [164], to compute the motion-based features. The approach employing stereo vision exhibits both the best recognition performance and the lowest processing costs per image (15.5 Hz).

	mono	mono + motion-based hyp. gen. (pedestrian-derived weights)
images per second	8.10 Hz	7.20 Hz
processing time per image	123 ms	138 ms

(a)

	mono + motion-based hyp. gen. (unif. weights + integral img.)	stereo
images per second	14.9 Hz	15.5 Hz
processing time per image	67.1 ms	64.5 ms

(b)

Table 4.7: Processing speed of considered pedestrian recognition systems.

4.2.5 Conclusion

This section presented a novel attentive strategy for monocular pedestrian recognition involving a model of motion-based features learned from ego-motion corrected optical flow. Features are particularly attuned to the pedestrian class and modeled in a probabilistic fashion. In experiments on datasets captured from a moving vehicle in urban traffic, we obtained the result that pedestrian recognition performance is substantially enhanced by the proposed motion-based attention concept; false positives were reduced by a factor of two.

Chapter 5

A Mixed Generative-Discriminative Pedestrian Model

5.1 Overview

Several techniques which combine generative and discriminative models have been proposed [90, 104, 156, 178]. Discriminative models have been employed to learn a generative model in an iterative fashion [156]. One line of research has been concerned with designing objective functions which incorporate both generative and discriminative terms, where their balance is controlled by both heuristic [104, 178] and probabilistic [90] weighting schemes.

Aside from the particular models used, incorporating prior knowledge about the target class has been suggested to increase classification robustness [114]. Prior knowledge can be both incorporated directly into the error function of a discriminative model (vicinal risk minimization) [162] and during training in terms of enlarging the training set with additional samples [109, 120, 125, 150, 162, 164]. While samples of the non-target class can be easily collected using bootstrapping [109, 120, 150, 164], acquiring additional target class samples is typically burdensome. Besides the trivial approach of laborious manual labeling, a number of techniques to synthesize virtual patterns of the target class have been proposed. Some require controlled data acquisition (e.g. same individual with respect to changes in viewpoint, facial expression and lighting) to obtain prototypical images to be linearly combined [13, 20, 57]. Others utilize explicit 3D models [65, 103]. If such prerequisites cannot be satisfied, the synthesis of virtual examples has been limited to simple geometric and photometric jittering in terms of adding mirrored, rotated, shifted or intensity-manipulated versions of the original training patterns [109, 125, 150, 162].

This chapter proposes a novel combined generative-discriminative approach

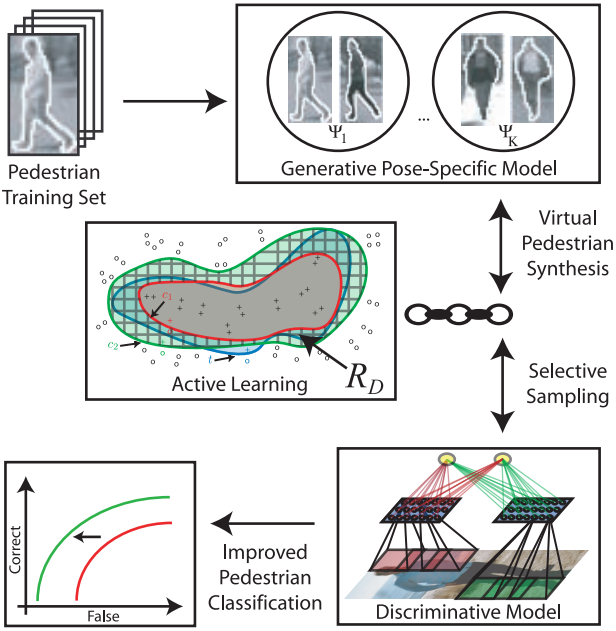


Figure 5.1: Framework overview. Utilizing the synthesized samples of a learned generative model to enhance the classification performance of a discriminative model.

to pedestrian classification, aimed at addressing the bottleneck caused by the scarcity of samples of the target class. A generative model is learned from a pedestrian dataset captured in real urban traffic and used to synthesize virtual samples of the target class that go way beyond simple transformations in terms of jittering, mirroring or rotating. The virtual samples enlarge the training set of a discriminative pattern classifier at little cost. This set of virtual samples can be considered as a regularization term to the real data to be fitted, which incorporates prior knowledge about the target object class. We propose the use of selective sampling, by means of probabilistic active learning, to guide the training process towards the most informative samples. See Figure 5.1.

The general idea is independent of the particular generative and discrim-

Authors	Shape Model	Texture Model	Sample Plausibility
Cootes et al. [22] Fan et al. [43] Jones et al. [80]	global linear (PCA)	global linear (PCA)	limit on deviation from mean
Jones et al. [79]	multi-layer global linear (weighted PCA)	multi-layer global linear (weighted PCA)	limit on deviation from mean
Gavrila et al. [55] Heap et al. [63]	pose-specific linear (PCA)	-	limit on deviation from mean
Romdhani et al. [132]	global non-linear (Kernel PCA)	-	limit on deviation from mean
Sozou et al. [146]	global non-linear (polynomial regression)	-	limit on deviation from mean
Cootes et al. [24]	global linear (PCA)	-	probabilistic (GMM)
this thesis	**pose-specific linear (PCA)**	**pose-specific linear (PCA), decomposed**	**probabilistic (KDE)**

Table 5.1: Overview of existing and proposed generative shape and texture models.

inative model used and can in principle extend to other object classes than pedestrians. We propose a generative model which consists of a number of probabilistic shape and texture models, each attuned to a generic object pose. For this, we require the existence of a registration method amongst samples associated with the same generic pose. See Table 5.1 for an overview of existing generative shape and texture models. Our use of active learning furthermore requires a confidence measure associated with the output of the discriminative model, but this assumption is easily met in practice.

5.2 Generative Pedestrian Model

5.2.1 Pedestrian Representation

Input to our pedestrian model is a set $\mathcal{D}$ of pedestrians $(\mathbf{x}_i, \omega_0) \in \mathcal{D}$ with class label ω_0. We apply an integrated shape registration and clustering approach with manual correction [55] to obtain a set of K view-specific clusters, Ψ_k, from the shapes underlying $\mathcal{D}$, with prototype shapes $\mathbf{p}_k$ (we use $K = 12$ in the experiments). See Figure 5.2. As a result of shape registration, [55], it is possible to embed the shapes within a cluster Ψ_k into a common feature-space. The features involve the pixel coordinates of corresponding points sampled at a given (arc-length normalized) distance along the contour. See

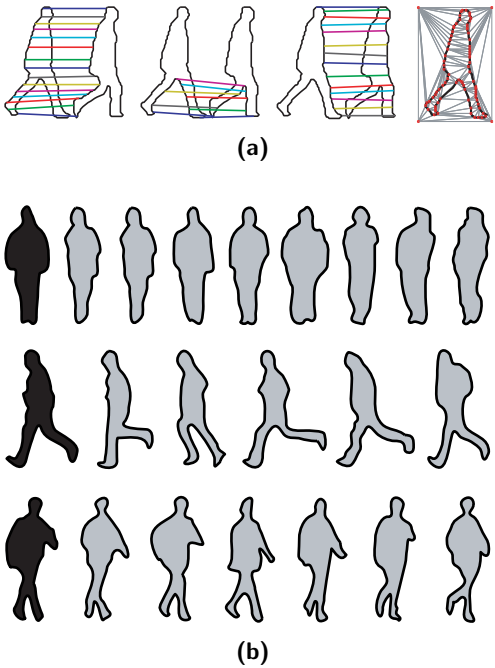


Figure 5.2: Integrated shape registration and clustering. (a) Shape registration. Automatically determined contour point correspondences (first three columns), Delaunay triangulation (last column). (b) Each row contains a set of randomly selected shapes from a pose-specific cluster (gray), along with the automatically determined prototype (black).

Figure 5.2a.

Let $\mathbf{x}_{k,i}$ denote the i-th example in the k-th pose-specific cluster Ψ_k, with $i = 1, \ldots, N_k$. A pedestrian sample $\mathbf{x}_{k,i} = (\mathbf{s}_{k,i}, \mathbf{t}_{k,i}) \oplus \mathbf{b}_{k,i}$ is represented as the composition $\oplus$ of a foreground texture $\mathbf{t}_{k,i}$ over a background $\mathbf{b}_{k,i}$, partitioned by a discrete shape contour $\mathbf{s}_{k,i}$.

After applying shape registration, each of the N_k discrete shape contours $\mathbf{s}_{k,i}$ in Ψ_k consists of l_k^s two-dimensional contour points (u, v) and is repre-

sented as a $2l_k^s$-dimensional shape vector:

$$\mathbf{s}_{k,i} = \left(\left(u_{i,1}, v_{i,1}\right), \left(u_{i,2}, v_{i,2}\right), \ldots, \left(u_{i,l_k^s}, v_{i,l_k^s}\right)\right)^T \quad (5.1)$$

$$= \left(u_{i,1}, v_{i,1}, u_{i,2}, v_{i,2}, \ldots, u_{i,l_k^s}, v_{i,l_k^s}\right)^T \quad (5.2)$$

The foreground texture $\mathbf{t}_{k,i}$ represents the area inside the shape contour in terms of l_k^t-dimensional vectors containing intensity values $\pi(u, v)$ at pixels $\mathbf{p} = (u, v)$:

$$\mathbf{t}_{k,i} = \left(\pi\left(u_{i,1}, v_{i,1}\right), \pi\left(u_{i,2}, v_{i,2}\right), \ldots, \pi\left(u_{i,l_k^t}, v_{i,l_k^t}\right)\right)^T, \ \forall (u_{i,j}, v_{i,j}) \text{ inside } \mathbf{s}_{k,i} \quad (5.3)$$

Note that the dimensionality of all $\mathbf{t}_{k,i} \in \Psi_k$ is the same, as a result of shape-normalization. Details are given in Section 5.2.3.

The introduction of pose-specific feature-spaces Ψ_k effectively reduces correlations between pedestrian texture and their pose or heading. Within each pose-specific space, a generative model is instantiated describing the pedestrian class-conditional density function for the shape and foreground texture component separately. Foreground and background are assumed uncorrelated, thus the background texture component $\mathbf{b}_k$ is not included into the generative model.

We now outline the learning procedure for the proposed pose-specific generative pedestrian shape-texture model involving the set-up of separate shape and texture model-spaces, as well as the estimation of the class-conditional densities therein.

5.2.2 Locally Linear Shape Model

Principal Component Analysis (PCA) is applied to each local shape space in Ψ_k to obtain a compact representation utilizing d_k^s dimensions (e.g. to model 95 % of the total variance).

Given a set of registered $2l_k^s$-dimensional shape vectors $\mathbf{s}_{k,i}$, with $i = 1, \ldots, N_k$, the mean shape $\bar{\mathbf{s}}_k$ is derived as:

$$\bar{\mathbf{s}}_k = \frac{1}{N_k} \sum_{i=1}^{N_k} \mathbf{s}_{k,i} \quad (5.4)$$

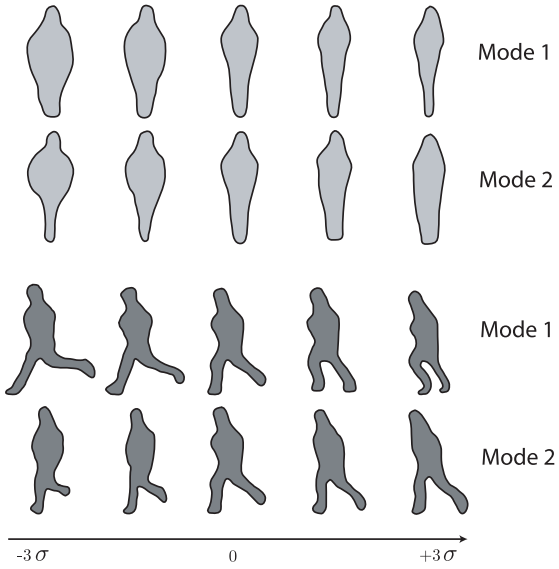


Figure 5.3: Shape variation along the first two PCA-dimensions within a $\pm 3\sigma$ range for two local shape models (light-gray and dark-gray.)

The covariance matrix $\mathbf{V}_k^s$ of the shape space is given by:

$$\mathbf{V}_k^s = \frac{1}{N_k} \sum_{i=1}^{N_k} (\mathbf{s}_{k,i} - \bar{\mathbf{s}}_k)(\mathbf{s}_{k,i} - \bar{\mathbf{s}}_k)^T \tag{5.5}$$

The principal axes are recovered by solving the eigensystem

$$\mathbf{V}_k^s \ \phi_{k,j}^s = \lambda_{k,j}^s \ \phi_{k,j}^s, \ j \in \{1, 2, \ldots, 2l_k^s\}, \tag{5.6}$$

where $\lambda_{k,j}^s$ denotes the j-th eigenvalue and $\phi_{k,j}^s$ denotes the j-th eigenvector of $\mathbf{V}_k^s$. Each eigenvector $\phi_{k,j}^s$ represents a set of displacement vectors along which the mean shape $\bar{\mathbf{s}}_k$ can be deformed. $\lambda_{k,j}^s$ defines the amount of variance along each principal axis $\phi_{k,j}^s$.

Any shape $\mathbf{s}_{k,i}$ can be reconstructed in terms of the mean shape $\bar{\mathbf{s}}_k$, a matrix of eigenvectors in each column $\mathbf{\Phi}_{\mathbf{s}_k}$ and a set of model parameters

$\mathbf{m}_{\mathbf{s}_{k,i}}$:

$$\mathbf{s}_{k,i} = \bar{\mathbf{s}}_k + \mathbf{\Phi}_{\mathbf{s}_k} \mathbf{m}_{\mathbf{s}_{k,i}} \tag{5.7}$$

This reconstruction is only exact, if all $2l_k^s$ dimensions are incorporated. We select a subset of $d_k^s \leq 2l_k^s$ dimensions to obtain a compact representation. In our experiments, we choose d_k^s so that 95 % of the total shape variance is explained.

The parametric representation $\mathbf{m}_{\mathbf{s}_{k,i}}$ of a pedestrian shape $\mathbf{s}_{k,i}$ in terms of shape model coordinates is then given by solving Equation (5.7) for $\mathbf{m}_{\mathbf{s}_{k,i}}$. Note that $\mathbf{\Phi}_{\mathbf{s}_k}$ is orthogonal, so $\mathbf{\Phi}_{\mathbf{s}_k}^{-1} = \mathbf{\Phi}_{\mathbf{s}_k}^T$:

$$\mathbf{m}_{\mathbf{s}_{k,i}} = \mathbf{\Phi}_{\mathbf{s}_k}^T (\mathbf{s}_{k,i} - \bar{\mathbf{s}}_k) \tag{5.8}$$

Figure 5.3 depicts the variation of the first two shape model parameters, i.e. the first two components of $\mathbf{m}_{\mathbf{s}_{k,i}}$, along the eigenvectors $\phi_{k,j}^s$ within a $\pm 3\sigma$ range, as defined by the square-root of the eigenvalue, $\sigma = \sqrt{\lambda_{k,j}^s}$, of the corresponding dimension. The more significant modes represent global variation due to pose changes, whereas the less significant modes are responsible for smaller local changes in pose.

The locally linear representation in terms of separate pose-specific spaces Ψ_k improves the specificity of the shape models involved. Forcing a topologically diverse set of shapes into a single global linear model, may result in physically implausible intermediate model instantiations, cf. Figure 5.4.

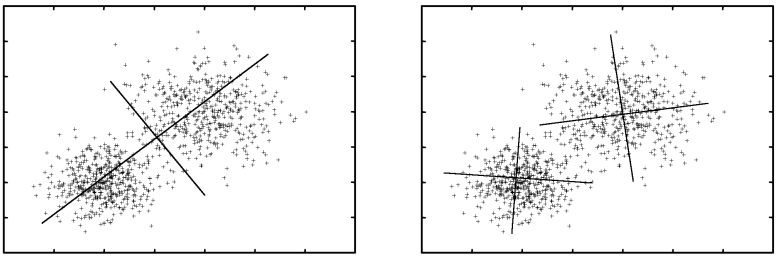


Figure 5.4: Single global linear (a) vs. two locally linear models (b) fitted to the same data. Adapted from [59].

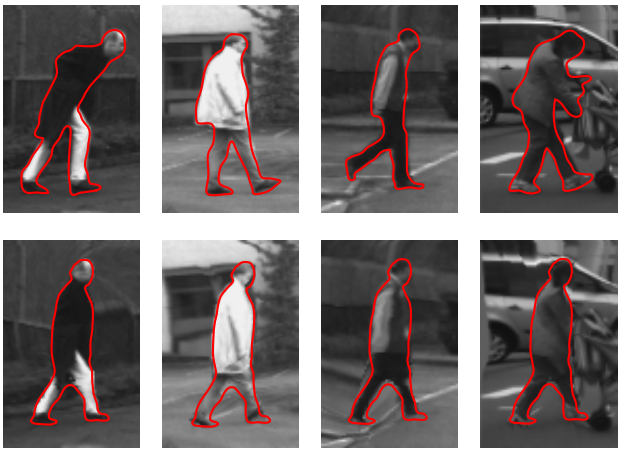


Figure 5.5: Shape-normalization. (top row) Original pedestrian examples. (bottom row) Shape-normalized examples for one pose-specific subspace.

5.2.3 Locally Linear Foreground Texture Model

To establish a foreground texture feature-space within each cluster Ψ_k, all texture vectors $\mathbf{t}_{k,i}$ are first shape-normalized to $\hat{\mathbf{t}}_{k,i}$ by warping them with respect to the cluster prototype $\mathbf{p}_k$, see Figure 5.5. A Delaunay triangulation-based piecewise-affine warping function $W_{\mathbf{s}_{k,i}}$ is employed, utilizing shape correspondences between shape $\mathbf{s}_{k,i}$ and prototype $\mathbf{p}_k$ to map triangles, see Figure 5.2a:

$$\hat{\mathbf{t}}_{k,i} = W_{\mathbf{s}_{k,i}}(\mathbf{t}_{k,i}) \tag{5.9}$$

Shape-normalization can be seen as a partial linearization of non-linear inter-dependencies within each pose-specific texture feature-space resulting from (slightly) different body poses and headings.

As before, PCA is applied to establish a parametric texture model-space representation of $\hat{\mathbf{t}}_{k,i}$ in terms of the mean texture $\bar{\hat{\mathbf{t}}}_k$ and eigenvectors $\mathbf{\Phi}_{\hat{\mathbf{t}}_k}$:

$$\mathbf{m}_{\hat{\mathbf{t}}_{k,i}} = \mathbf{\Phi}_{\hat{\mathbf{t}}_k}^T \left(\hat{\mathbf{t}}_{k,i} - \bar{\hat{\mathbf{t}}}_k \right) \tag{5.10}$$

Figure 5.6 depicts the mean texture along with the first four eigenvectors for a pose-specific texture model. Note the existence of pose-specific texture

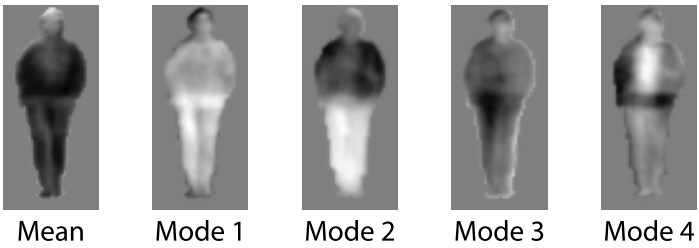


Figure 5.6: Mean texture and eigenvectors for a pose-specific texture model (background masked out). Note the pose-specific texture characteristics, e.g. different types of clothing in mode 2 and the coat-shirt pattern in mode 4.

characteristics, e.g. different types of clothing in mode 2 and the coat-shirt pattern in mode 4.

Given the scarcity of available texture samples (meanwhile subdivided by pose) and the high dimensionality of the shape-normalized texture model-space, we cannot reliably establish a generative texture model to capture a sizable amount of variance (e.g. 95 %), as done before for shape. Using solely a subspace spanned by fewer principal components is however not a viable option, as projection leads to subtle texture details being washed-out, which in large part determine pedestrian appearance. As a way out, we propose to decompose the full d_k^t-dimensional texture model-space obtained by PCA into two subspaces with dimensionality $d_k^{t\prime}$ and $d_k^{t\prime\prime}$:

$$d_k^t = d_k^{t\prime} + d_k^{t\prime\prime} \tag{5.11}$$

The first subspace represents coarse texture components (e.g. modeling overall appearance of clothing parts such as trousers and coat). Its dimensionality $d_k^{t\prime} < d_k^t$ is selected such that a reliable estimation of the relevant *pdf* from training data is possible (e.g. we model 65 % of the total variance). The second and complementary $d_k^{t\prime\prime}$-dimensional subspace captures fine texture components. Here no *pdf* estimation takes place, for synthesis (see Section 5.3) the associated entries are derived from particular training samples.

Hence, the parametric model-space representation $\mathbf{m}_{\tilde{\mathbf{t}}_{k,i}}$, as given in Equa-

tion (5.10), of a shape-normalized texture vector $\hat{\mathbf{t}}_{k,i}$ is decomposed into:

$$\mathbf{m}_{\hat{\mathbf{t}}_{k,i}} = \left(\mathbf{m}'_{\hat{\mathbf{t}}_{k,i}}, \mathbf{m}''_{\hat{\mathbf{t}}_{k,i}}\right) \tag{5.12}$$

with

$$\mathbf{m}'_{\hat{\mathbf{t}}_{k,i}} = \left(\mathbf{m}_{\hat{\mathbf{t}}_{k,i},1}, \ldots, \mathbf{m}_{\hat{\mathbf{t}}_{k,i},d_k^{t\prime}}\right) \tag{5.13}$$

$$\mathbf{m}''_{\hat{\mathbf{t}}_{k,i}} = \left(\mathbf{m}_{\hat{\mathbf{t}}_{k,i},d_k^{t\prime}+1}, \ldots, \mathbf{m}_{\hat{\mathbf{t}}_{k,i},d_k^{t}}\right) \tag{5.14}$$

5.2.4 Class-Conditional Density Estimation

After establishing K pose-specific shape and shape-normalized texture model-spaces, we estimate the class-conditional densities $p_{\mathbf{s}_k}\left(\mathbf{m}_{\mathbf{s}_k}|\omega_0\right)$ and $p_{\hat{\mathbf{t}}_k}\left(\mathbf{m}'_{\hat{\mathbf{t}}_k}|\omega_0\right)$ with respect to the pedestrian class ω_0 within each subspace. In preliminary experiments, we found Gaussian Kernel Density Estimation (KDE) to outperform Gaussian Mixture Models (GMM), based on the likelihood of model-fit.

Temporarily dropping the distinction between shape $\mathbf{s}_k$ and texture $\hat{\mathbf{t}}_k$, the Kernel Density estimate of the class-conditional densities is given by:

$$p_k\left(\mathbf{m}|\omega_0\right) = \frac{1}{N_k}\sum_{n=1}^{N_k}\frac{1}{det(H)}\mathcal{K}\left\{H^{-1}(\mathbf{m}-\mathbf{m}_n)\right\} \tag{5.15}$$

where $\mathcal{K}$ denotes the kernel function and H represents a diagonal matrix containing kernel bandwidths. We use anisotropic multivariate Gaussian kernels $\mathcal{K}$, with bandwidths optimized via maximum likelihood on the training set [78], for both the shape and shape-normalized texture space, respectively.

The class-conditional density functions $p_{\mathbf{s}_k}\left(\mathbf{m}_{\mathbf{s}_k}|\omega_0\right)$ and $p_{\hat{\mathbf{t}}_k}\left(\mathbf{m}'_{\hat{\mathbf{t}}_k}|\omega_0\right)$ provide the basis for the proposed synthesis of virtual pedestrians. As opposed to [22, 43, 80], where plausibility has been enforced by limiting the deviation of the model coordinates from the mean (which does not extend to a multimodal distribution), the probabilistic formulation allows for a direct assessment of plausibility for a given shape or texture vector.

5.3 Model-Based Virtual Pedestrian Synthesis

The model-based synthesis of virtual pedestrian samples utilizing the proposed pose-specific generative shape and texture models involves the varia-

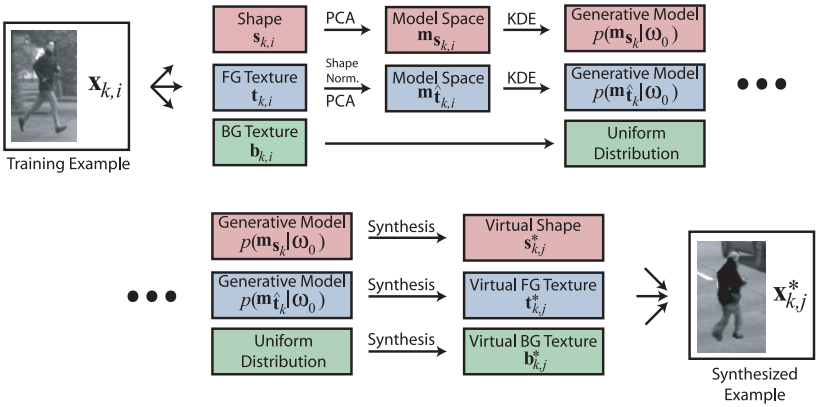


Figure 5.7: Overview of the proposed model-based pedestrian synthesis procedure within a pose-specific cluster Ψ_k. Existing pedestrian examples are projected onto a generative shape-texture model which is resampled to create virtual pedestrian samples.

tion of three components: shape, foreground texture and background texture. See Figure 5.7 for an overview.

5.3.1 Shape Variation

Model coordinates $\mathbf{m}^*_{\mathbf{s}_{k,j}}$ representing a new virtual shape $\mathbf{s}^*_{k,j}$ can be sampled directly from the generative shape model $p_{\mathbf{s}_k}\left(\mathbf{m}_{\mathbf{s}_k}|\omega_0\right)$:

$$\mathbf{m}^*_{\mathbf{s}_{k,j}} \sim p_{\mathbf{s}_k}\left(\mathbf{m}_{\mathbf{s}_k}|\omega_0\right) \tag{5.16}$$

Sampling the KDE estimate of $p_{\mathbf{s}_k}\left(\mathbf{m}_{\mathbf{s}_k}|\omega_0\right)$ involves uniformly selecting the j-th example $\mathbf{m}_{\mathbf{s}_{k,j}}$ in model-space and sampling from the local kernel $\mathcal{K}$, centered at $\mathbf{m}_{\mathbf{s}_{k,j}}$. Plausibility of the virtual shape model coordinates is enforced by requiring $p_{\mathbf{s}_k}\left(\mathbf{m}^*_{\mathbf{s}_{k,j}}|\omega_0\right) > c_{\mathbf{s}_k}$, with $c_{\mathbf{s}_k}$ a threshold parameter learned from the distribution of the training set so that the large majority of training samples (e.g. 99 %) are covered.

Transforming $\mathbf{m}^*_{\mathbf{s}_{k,j}}$ from shape model-space back to the shape feature-

space yields a new virtual shape contour:

$$\mathbf{s}^*_{k,j} = \bar{\mathbf{s}}_k + \mathbf{\Phi}_{\mathbf{s}_k} \mathbf{m}^*_{\mathbf{s}_{k,j}} \tag{5.17}$$

The virtual shape $\mathbf{s}^*_{k,j}$ is utilized to warp an existing pedestrian example into a new shape, as shown in Figure 5.8b.

5.3.2 Foreground Texture Variation

Regarding the synthesis of virtual texture samples for the pedestrian class, we utilize the proposed decomposed representation of the shape-normalized texture space in terms of coarse and detailed components, as outlined in Section 5.2. The main idea is, to employ the main modes of variation to control coarse appearance variations (e.g. individual clothing parts or global illumination) and induce pose-specific effects of different types of wear (e.g. closed coat vs. coat-shirt pattern, see Figure 5.6 mode 2 vs. mode 4, respectively), while at the same time retaining fine-scales details (e.g. internal body or face contours), which are crucial for pedestrian appearance.

Hence, to obtain virtual shape-normalized texture parameters $\mathbf{m}^*_{\hat{\mathbf{t}}_{k,j}}$, we first sample model parameters pertaining to coarse texture components $\mathbf{m}'^*_{\hat{\mathbf{t}}_{k,j}}$ from the generative texture model $p_{\hat{\mathbf{t}}_k}\big(\mathbf{m}'_{\hat{\mathbf{t}}_k}|\omega_0\big)$, by uniformly selecting the j-th example in model-space and sampling from the local kernel:

$$\mathbf{m}'^*_{\hat{\mathbf{t}}_{k,j}} \;\sim\; p_{\hat{\mathbf{t}}_k}\big(\mathbf{m}'_{\hat{\mathbf{t}}_k}|\omega_0\big) \tag{5.18}$$

Similar to the way the shape component is addressed, plausibility is enforced by applying a coverage threshold $c_{\hat{\mathbf{t}}_k}$ (e.g. 99 % coverage), with $p_{\hat{\mathbf{t}}_k}\big(\mathbf{m}'^*_{\hat{\mathbf{t}}_{k,j}}|\omega_0\big) > c_{\hat{\mathbf{t}}_k}$. Model parameters $\mathbf{m}''_{\hat{\mathbf{t}}_{k,j}}$ representing the original shape-normalized texture details of the j-th example $\mathbf{m}_{\hat{\mathbf{t}}_{k,j}}$ are retained and combined with the synthesized coarse model coordinates $\mathbf{m}'^*_{\hat{\mathbf{t}}_{k,j}}$ to yield (cf. Equation (5.12)):

$$\mathbf{m}^*_{\hat{\mathbf{t}}_{k,j}} = \big(\mathbf{m}'^*_{\hat{\mathbf{t}}_{k,j}}, \mathbf{m}''_{\hat{\mathbf{t}}_{k,j}}\big) \tag{5.19}$$

Thereafter, $\mathbf{m}^*_{\hat{\mathbf{t}}_{k,j}}$ is projected from the model-space back to the feature-space of shape-normalized texture:

$$\hat{\mathbf{t}}^*_{k,j} = \bar{\hat{\mathbf{t}}}_k + \mathbf{\Phi}_{\hat{\mathbf{t}}_k} \mathbf{m}^*_{\hat{\mathbf{t}}_{k,j}} \tag{5.20}$$

Finally, the inverse of the shape-normalization operator, $W^{-1}_{\mathbf{s}^*_{k,j}}$, is applied to warp the virtual shape-normalized texture $\hat{\mathbf{t}}^*_{k,j}$ to a shape $\mathbf{s}^*_{k,j}$ (which can be a new virtual shape or an existing shape) within the same pose-specific model, see Equation (5.9):

$$\mathbf{t}^*_{k,j} = W^{-1}_{\mathbf{s}^*_{k,j}}(\hat{\mathbf{t}}^*_{k,j}) \tag{5.21}$$

An example of this technique is depicted in Figures 5.8c - 5.8e. Note how fine-scale details, e.g. the internal contour of the right arm (Figures 5.8c - 5.8e,

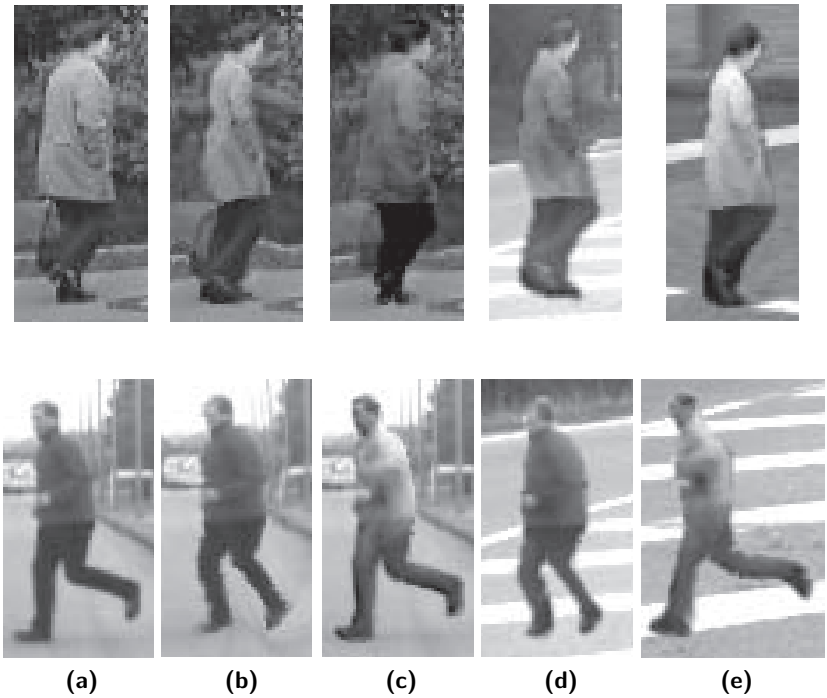


Figure 5.8: Example of virtual pedestrian synthesis. (a) Original pedestrian examples. (b) Shape variation. (c) Foreground texture variation. (d) - (e) Joint variation of shape, foreground and background texture.

first row) are preserved, while the overall texture exhibits sensible variations.

5.3.3 Background Texture Variation

The background texture component is assumed independent from pedestrian appearance and is represented by a non-parametric exemplar-based model. Virtual background texture vectors $\mathbf{b}^*_{k,j}$ are uniformly sampled ($\mathcal{U}$) from a set of non-pedestrian images $\mathcal{B}$ that can be obtained at low cost:

$$\mathbf{b}^*_{k,j} \sim \mathcal{U}(\mathcal{B}) \tag{5.22}$$

Application-specific constraints regarding likely target locations (e.g. flat-world assumption, people standing on the ground) can be incorporated at this point.

5.3.4 Joint Variation and Compositing

Joint variation of shape, foreground and background texture involves sampling virtual examples for each component. Virtual texture $\mathbf{t}^*_{k,j}$ is sampled from the generative texture model $p_{\mathbf{t}_k}\left(\mathbf{m}'_{\mathbf{t}_k}|\omega_0\right)$, see Equations (5.18) - (5.21), and warped to a virtual shape $\mathbf{s}^*_{k,j}$, sampled from the generative shape model $p_{\mathbf{s}_k}\left(\mathbf{m}_{\mathbf{s}_k}|\omega_0\right)$ (cf. Equations (5.16) - (5.17)). Finally, background $\mathbf{b}^*_{k,j}$ is sampled from the the non-parametric background model (cf. Equation (5.22)) and a virtual pedestrian example $\mathbf{x}^*_{k,j}$ is obtained by compositing the textured pedestrian shape over the background, see Figure 5.8:

$$\mathbf{x}^*_{k,j} = (\mathbf{s}^*_{k,j}, \mathbf{t}^*_{k,j}) \oplus \mathbf{b}^*_{k,j} \tag{5.23}$$

5.4 Probabilistic Selective Sampling

A probabilistic least-certain querying scheme, an instance of an active learning algorithm [62, 83, 96], is utilized to directly link the discriminative with the generative model in terms of assessing the information content of virtual pedestrian samples. Resampling a generative model allows to create a virtually infinite number of training samples for a discriminative model. Here, selective sampling becomes a necessity to remove redundancy from the training set and focus the resources of the discriminative learning procedure on the examples with the highest information content. In classification tasks,

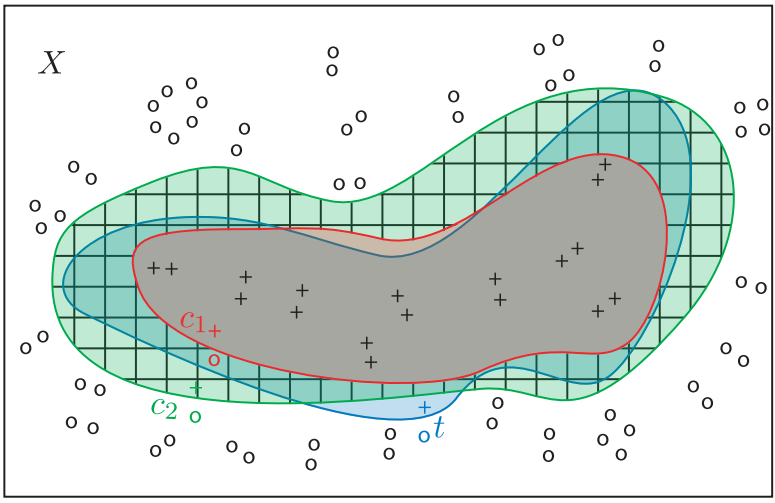


Figure 5.9: Region of uncertainty in an exemplary two-dimensional classification problem (+ vs. ○). Two decision boundaries c_1 (red) and c_2 (green) are shown which are consistent with the training set. The unknown true decision boundary t is depicted in blue. The region of uncertainty R_D is marked with a hatched pattern, see text.

there exists a region of uncertainty R_D, where the classification result is not unambiguously defined, see the hatched area in Figure 5.9. That is, the discriminative model can learn a multitude of decision boundaries which are consistent with the given training patterns, but yet disagree in some regions of the decision space. If a sample is drawn from R_D, the size of R_D and thus the global uncertainty can be reduced.

In our probabilistic least-certain querying scheme, we approximate R_D using the probability of error for each sample $\mathbf{x}_i$. Given a two-class problem with classes ω_0 (target class) and ω_1 (non-target class), we assume the discriminative model to approximate posterior probabilities and to make a Bayesian decision, i.e. $\mathbf{x}_i$ is classified as ω_0, if $P(\omega_0|\mathbf{x}_i) > P(\omega_1|\mathbf{x}_i)$. Then, the probability of error $P(error|\mathbf{x}_i)$ is given by

$$P(error|\mathbf{x}_i) = \min\left\{P(\omega_0|\mathbf{x}_i), P(\omega_1|\mathbf{x}_i)\right\}. \tag{5.24}$$

Obviously, $P(error|\mathbf{x}_i)$ has a peak at $P(\omega_0|\mathbf{x}_i) = P(\omega_1|\mathbf{x}_i) = 0.5$, which represents the decision boundary. To base uncertainty on $P(error|\mathbf{x}_i)$, we introduce a threshold $\Theta \in [0, 0.5]$ on $P(error|\mathbf{x}_i)$ and consider only those samples $\mathbf{x}_i$ as informative examples, where $P(error|\mathbf{x}_i) > \Theta$. This is equivalent to putting a threshold on the absolute difference of the posterior probabilities:

$$0 \leq |P(\omega_0|\mathbf{x}_i) - P(\omega_1|\mathbf{x}_i)| \leq 1 - 2\Theta \tag{5.25}$$

Hence, the approximation of the region of uncertainty R_D is defined as a symmetric region centered at $P(\omega_0|\mathbf{x}) = P(\omega_1|\mathbf{x}) = 0.5$, the decision boundary of the discriminative model. This technique requires an estimate of the underlying (unknown) probabilities. 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 [78, 83, 96]. We use this in our experiments.

The aforementioned selective sampling strategy is used in an iterative scheme to link the training of the discriminative model with the generative pedestrian synthesis. In each iteration l, the set of virtual examples $\mathcal{D}_l^*$ is resampled to $\widehat{D}_l^*$ by retaining only the informative samples $\mathbf{x}_j^* \in \mathcal{D}_l^*$, as evaluated by the discriminative model trained on D_l, using Equation (5.25). Finally, the discriminative model is retrained on the joint dataset $D_{l+1} = D_l \cup \widehat{D}_l^*$.

5.5 Experiments

The proposed generative-discriminative framework is tested in large-scale experiments on pedestrian classification. Our purpose is not to establish the best *absolute* classification performance amongst the various state-of-the-art methods, see Chapter 2. Rather, our aim is to examine the *relative* performance gain that can be obtained by using the proposed mixed generative-discriminative framework over a particular discriminative-only approach. To illustrate the generality with respect to the discriminative model used, we consider two diverse instances: a neural network with local receptive fields of size 5×5 pixels (NN/LRF) [170] and a linear[1] support vector machine using Haar wavelet features at scales of 4×4 and 8×8 pixels (Haar/linSVM)

[1] training a non-linear SVM on our large datasets was not feasible due to excessive memory requirements

	Pedestrians (labeled)	Pedestrians (jittered)	Non-Pedestrians
Initial Training Set	10946	43784	82698
Test Set	13971	251478	133813

Table 5.2: Training and test set statistics.

[120]. Results are expected to generalize to other pedestrian classifiers that are sufficiently complex to represent the large training datasets, cf. Chapter 2.

See Table 5.2 for the datasets used. Training and test sets contain manually labeled pedestrian bounding boxes with additional contour labels for the training set. All training samples are scaled to 18×36 pixels with a two pixel border in order not to lose contour information. The samples were acquired in daylight conditions from a moving vehicle and depict non-occluded pedestrians in front of a changing background. The non-pedestrian samples were the result of a pedestrian shape recognition pre-processing step with relaxed threshold setting, i.e. containing a bias towards more "difficult" patterns, similar to [109]. Training and test set are strictly separated: no instance of the same real-world pedestrian appears in both training and test set, similarly for the non-target samples. See Figure 5.10 for an overview of the dataset. Discriminative models trained on this dataset are referred to as *base classifiers*.

We examine the effect of introducing jittering to pedestrian training samples; this represents the applicable state-of-the-art, see Section 5.1. Geometric jittering is introduced in terms of creating four patterns from each pedestrian sample in the training set by applying a random shift (± 2 pixels) and mirroring. Since we employ contrast normalization during training of the classifiers, photometric jittering is not considered. Discriminative models utilizing this dataset are referred to as *jittered classifiers*.

In all experiments with our mixed generative-discriminative framework (Figure 5.1), we perform several iterations of virtual sampling and discriminative model retraining, up to performance saturation. In each such iteration, the training set is extended by 10946 synthesized pedestrians (plus additional four jittered versions of each virtual pedestrian), guided by selective sampling (Equation (5.25)), with $\Theta = 0.35$. For the case of non-targets, we perform a similar iterative dataset extension approach (4×10946 samples, now obtained by selective sampling on images not containing targets, without

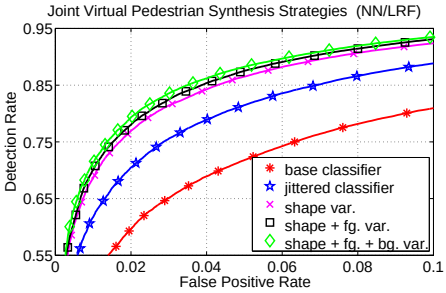
Figure 5.10: Dataset overview. (a) Training set examples. (b) Test set examples. Top and bottom rows show target and non-target samples, respectively.

jittering).

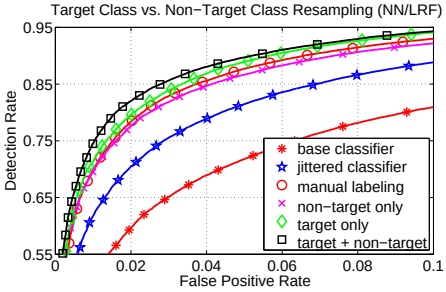
In a first experiment with an NN/LRF classifier (Figure 5.11a), the number of non-target training samples is kept constant and the benefit of jittering and virtual pedestrian synthesis is studied. From Figure 5.11a one observes that jittering leads to a significant performance improvement over the base classifier (more jittered samples did not yield further improvement). Yet we obtained additional performance gain using the proposed framework, by incrementally incorporating shape, foreground and background texture variation.

Furthermore, we compare target class resampling involving joint shape, foreground and background variation (the best performing synthesis variant in Figure 5.11a) to non-target class resampling, see Figures 5.11b and 5.11c. The total performance gain by adding non-target training samples only is significant, yet less than in the case of augmenting the pedestrian set only (Figures 5.11b and 5.11c, magenta vs. green curve). Best performance is reached by joint augmentation of the pedestrian and non-pedestrian class. This variant saturated after three iterations, compared to two iterations for all others.

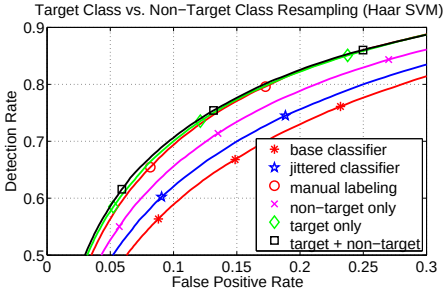
For comparison, we added 10946 real pedestrian samples plus four jittered versions, manually labeled from an auxiliary data pool, to the base dataset



(a)



(b)



(c)

Figure 5.11: ROC performance for classification experiments. (a) Virtual pedestrian synthesis (NN/LRF), (b) - (c) Target class vs. non-target class resampling for NN/LRF and Haar/linSVM.

(without synthetic samples and active learning). Remarkably, the proposed generative-discriminative framework even outperforms the manual approach (see Figures 5.11b and 5.11c, green vs. red circled curve). This is not an aberration caused by overfitting; the datasets used are truly large. Rather, it is the consequence of the fact that, although the manually labeled samples are more realistic, they are not necessarily more informative (we tediously label samples that the classifier already knows). Of course, the aim of our proposed generative-discriminative framework is to avoid this additional manual labeling in the first place.

We finally note that, although absolute performances for the two considered discriminative models are different, the relative order in which the various resampling techniques perform is identical, see Figures 5.11b and 5.11c.

5.6 Conclusion

This chapter presented a novel framework for pedestrian classification which involves utilizing the synthesized samples of a learned generative model to enhance the classification performance of a discriminative model. In extensive experiments, we obtained the non-trivial result that classification performance is substantially enhanced by the augmented training set; the false positive rate of the mixed generative-discriminative approach was reduced by up to a factor of two compared to discriminative-only approach, at the same detection rate. Our approach also outperformed classifiers bootstrapped by non-target data or by jittered samples of the target class. Remarkably, high-informative virtual samples proved to have a similar value than additional (random) real pedestrian samples. We take this as evidence of the strength of our generative pedestrian model and selective sampling method. Future work could involve feedback mechanisms to allow the selective sampling procedure to guide the generative model, i.e. to create new virtual samples in areas, where original samples are sparse. Further, the extension to other object classes is desired.

Chapter 6

Multi-Level Mixture-of-Experts for Pedestrian Classification

6.1 Overview

In recent years, a multitude of (more of less) different feature sets has been used to discriminate pedestrians from non-pedestrian images, as discussed in Chapter 2. Most of these 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 [60]. 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 [88].

The mixture-of-experts framework, cf. [77], for pedestrian classification presented in this chapter combines four modalities (shape, intensity, depth and motion) and three features (Chamfer distance [14], histograms of oriented gradients (HOG) [27] and local binary patterns (LBP) [115, 167], cf. Section 2.2). We follow a multi-level approach by utilizing expert classifiers on pose-, modality- and feature-levels, see Figure 6.1a. The local experts are integrated in terms of a probabilistic model based on fuzzy view-related clustering and associated sample-dependent cluster priors. K view-related models are trained in an off-line step to discriminate between pedestrians and non-pedestrians. These models consist of sample-dependent cluster priors and multi-level (multi-modality / multi-feature) expert classifiers. In the on-line application phase, cluster priors are computed using shape matching and used to fuse the multi-level expert classifiers to a combined decision, see Figure 6.1b. Details are given in Section 6.2.

Our approach has a number of advantages compared to fusion approaches

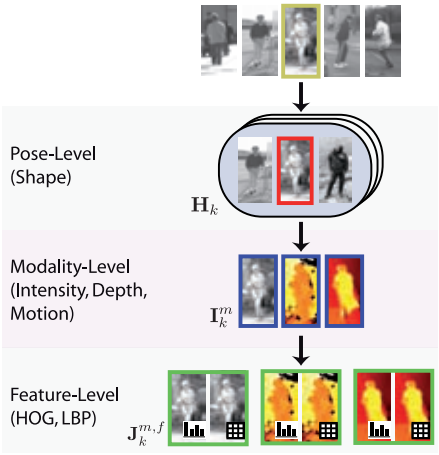


Figure 6.1: Framework overview. (a) Multi-level object representation comprising mixture-of-experts on pose-level, modality-level and feature-level. (b) K view-related models specific to fuzzy clusters Ψ_k are used for classification. The models consist of sample-dependent cluster priors and multi-modality/feature discriminative experts which are learned from pedestrian (class ω_0) and non-pedestrian (class ω_1) samples $\mathbf{x}$.

using a joint feature space, e.g. [137, 167, 173]. 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. [137], to robustly train our classifiers. Compared to multi-feature boosting approaches, we also do not require techniques to map the multi-dimensional features to a single dimension, e.g. through projection [175] or selection of 1D features [165, 166, 173].

Second, our mixture-of-experts framework alleviates practical problems arising from the use of large and high-dimensional datasets. 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. non-linear SVMs [37] or even linear SVMs [27, 137]. 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 on the order of weeks on current hardware, particularly for boosting approaches, e.g. [37, 165, 166, 173, 175]. 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, cf. [32, 37, 109], our approach could make maximum use of all available samples. For evaluation purposes, we utilize the same data samples for each modality/feature in our experiments to eliminate effects arising from imbalanced data.

6.2 Multi-Level Mixture-of-Experts

6.2.1 Object Representation

Input to our framework is a training set $\mathcal{D}$ of pedestrian (ω_0) and non-pedestrian (ω_1) samples $\mathbf{x}_i \in \mathcal{D}$. Each sample $\mathbf{x}_i = [\mathbf{x}_i^1; \mathbf{x}_i^2; \ldots; \mathbf{x}_i^M]$ 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: $\mathbf{x}_i^m = [\mathbf{x}_i^{m,1}; \mathbf{x}_i^{m,2}; \ldots; \mathbf{x}_i^{m,F}]$.

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 ($\mathbf{x}_i^2$) [68] and dense optical flow ($\mathbf{x}_i^3$) [168]. Other alternatives include near/far infrared (NIR/FIR) [101] or time-of-flight imagery [130]. 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, see below.

Dense stereo provides information for most image areas, apart from regions which are visible only by one camera (stereo shadow), see Figure 6.3. 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:

$$Z(u,v) = \frac{fB}{d(u,v)} \text{ at pixel } \mathbf{p} = (u,v) \tag{6.1}$$

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. In contrast to the approach described in Section 4.2, we do not compensate for the ego-motion of the camera at this point, since we are only interested in local differences in flow between a pedestrian and the environment. As a positive side-effect, static pedestrians do not pose a problem in combination with a moving camera.

A visual inspection of the intensity vs. depth and flow images in Figures 6.2 and 6.3 reveals that pedestrians have distinct contours and textures in each modality. Figure 6.2a 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. [108]). There is significant

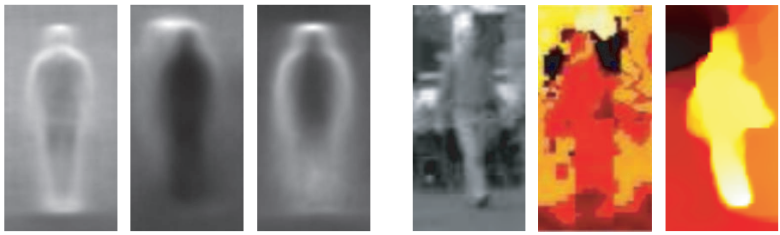


Figure 6.2: Multi-modality sample visualization. (a) Average gradient magnitude of all pedestrian training samples for intensity, depth and motion (left to right). (b) A difficult to recognize (low-contrast) pedestrian in the intensity domain can be very salient in other modalities.

texture on the pedestrian due to different clothing. 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 homogeneous due to uniform pedestrian motion. Legs move non-rigidly 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 6.2.3. A characteristic example is shown in Figure 6.2b. 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 histograms of oriented gradients (HOG) features [27] and local binary pattern (LBP) features [115, 167]. The motivation for this choice is two-fold. First, recent studies have shown that HOG and LBP features are highly com-

plementary regarding their sensitivity to noisy background edges which are common in cluttered backgrounds, cf. [167]. Second, despite the vast amount of features developed in recent years, HOG and LBP are still among the best features around [32, 37, 167]. Detailed parameterization of our feature set is given in Section 6.3.2.

Associated with each sample $\mathbf{x}_i$ 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 probabilistic membership to a set of K clusters Ψ_k, which relate to the similarity in appearance to a certain view and pose of a pedestrian. Note that the same also applies to non-pedestrian training samples, where the image structure resembles a certain pedestrian view. Our definition of cluster membership $\mathbf{z}_i$ is given in Section 6.3.1.

6.2.2 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:

$$\omega_i = \underset{\omega_j}{\operatorname{argmax}}\, P(\omega_j|\mathbf{x}_i) \tag{6.2}$$

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:

$$P(\omega_0|\mathbf{x}_i) = \sum_k P(\Psi_k|\mathbf{x}_i) P(\omega_0|\Psi_k, \mathbf{x}_i) \tag{6.3}$$

$$\approx \sum_k w_k(\mathbf{x}_i) \mathbf{H}_k(\mathbf{x}_i) \tag{6.4}$$

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 Equation (6.14) in Section 6.2.4.

$P(\omega_0|\Psi_k, \mathbf{x}_i)$ 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 $P(\omega_0|\Psi_k, \mathbf{x}_i)$.

6.2.3 Multi-Modality / Multi-Feature Expert Classifiers

Given our pose-specific mixture-of-experts formulation, cf. Equation (6.4), we model the pose-specific expert classifiers $\mathbf{H}_k(\mathbf{x}_i)$ in terms of our multi-modality dataset (intensity, depth, flow). We extend the mixture-of-experts formulation by introducing individual classifiers for each modality m:

$$\mathbf{H}_k(\mathbf{x}_i) = \sum_m v_k^m \ \mathbf{I}_k^m(\mathbf{x}_i^m) \tag{6.5}$$

In this formulation, $\mathbf{I}_k^m(\mathbf{x}_i^m)$ denotes a local expert classifier for the k-th fuzzy pose cluster, which is represented in terms of the m-th 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:

$$\mathbf{I}_k^m(\mathbf{x}_i^m) = \sum_f u_k^{m,f} \ \mathbf{J}_k^{m,f}(\mathbf{x}_i^{m,f}) \tag{6.6}$$

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

Plugging Equations (6.5) and (6.6) into Equation (6.4), we approximate $P(\omega_0|\mathbf{x}_i)$, the posterior probability that a given sample is a pedestrian, using our multi-level mixture-of-experts model as:

$$\begin{aligned} P(\omega_0|\mathbf{x}_i) &\approx \sum_k w_k(\mathbf{x}_i)\Big(\sum_m v_k^m \ \big(\sum_f u_k^{m,f} \ \mathbf{J}_k^{m,f}(\mathbf{x}_i^{m,f})\big)\Big) && (6.7)\\ &= \sum_k w_k(\mathbf{x}_i)\Big(\sum_m \sum_f v_k^m u_k^{m,f} \ \mathbf{J}_k^{m,f}(\mathbf{x}_i^{m,f})\Big) && (6.8)\\ &= \sum_k w_k(\mathbf{x}_i)\Big(\sum_{m,f} s_k^{m,f} \ \mathbf{J}_k^{m,f}(\mathbf{x}_i^{m,f})\Big), && (6.9)\\ &\quad \text{with } s_k^{m,f} = v_k^m u_k^{m,f} \text{ and } \sum_{m,f} s_k^{m,f} = 1. \end{aligned}$$

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 $\mathcal{D}$ 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.

In principle, the proposed framework is independent from the actual discriminative models used, cf. [37]. We only require example-dependent weights during training, and that the classifier outputs (decision values) 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 [15] 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 [78, 123]. 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 Equation (6.9), 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, cf. [44], its decision function equals a dot-product:

$$\mathbf{F}_k(\mathbf{x}_i) = \sum_{m,f} s_k^{m,f} \mathbf{J}_k^{m,f}(\mathbf{x}_i^{m,f})) \tag{6.10}$$

$$= \vec{\mathbf{s}} \cdot \vec{\mathbf{J}}(\mathbf{x}_i) \tag{6.11}$$

Plugging Equation (6.10) into Equation (6.9) then yields:

$$P(\omega_0|\mathbf{x}_i) \approx \sum_k w_k(\mathbf{x}_i)\mathbf{F}_k(\mathbf{x}_i) \tag{6.12}$$

6.2.4 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 Equation (6.3). 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 choose to model cluster priors using a Bayesian approach as:

$$P(\Psi_k|\mathbf{x}_i) = \frac{p(\mathbf{x}_i|\Psi_k)P(\Psi_k)}{\sum_l p(\mathbf{x}_i|\Psi_l)P(\Psi_l)} \tag{6.13}$$

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 [28], or shape [56]. Likelihood estimation can be performed via histogramming on training data or fitting parametric models [56].

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 $D_k(\mathbf{x}_i) \geq 0$ denote the residual shape distance, e.g. the Chamfer distance [54], 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 Equation (6.13), sample-dependent shape-based priors for cluster Ψ_k are approximated as:

$$P\left(\Psi_k|\mathbf{x}_i\right) \approx \frac{p(D_k(\mathbf{x}_i)|\Psi_k)P(\Psi_k)}{\sum_l p(D_l(\mathbf{x}_i)|\Psi_l)P(\Psi_l)} = w_k(\mathbf{x}_i) \tag{6.14}$$

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

$$p(D_k(\mathbf{x}_i)|\Psi_k) \approx \widehat{\alpha}_k e^{-\widehat{\alpha}_k D_k(\mathbf{x}_i)} \quad , \quad \widehat{\alpha}_k > 0 \tag{6.15}$$

Parameters $\widehat{\alpha}_k$ of the exponential distributions are learned via maximum-likelihood on the training set, as follows. With Equation (6.15), the likelihood function is given by:

$$\mathcal{L}(\alpha_k|D_k(\mathbf{x}_i)) = \prod_{\mathbf{x}_i \in \Psi_k} p(D_k(\mathbf{x}_i)|\Psi_k) \tag{6.16}$$

$$\overset{(6.15)}{\approx} \prod_{\mathbf{x}_i \in \Psi_k} \alpha_k e^{-\alpha_k D_k(\mathbf{x}_i)} \tag{6.17}$$

Let $N_k > 0$ be the number of samples in cluster Ψ_k. Instead of maximizing $\mathcal{L}(\alpha_k|D_k(\mathbf{x}_i))$, we apply a logarithmic transformation and maximize the log-

likelihood $\Lambda(\alpha_k|D_k(\mathbf{x}_i))$. Since this transform is monotonically increasing, the maximum remains at the same value of α_k:

$$\begin{aligned}
\Lambda(\alpha_k|D_k(\mathbf{x}_i)) &= \ln\Big(\mathcal{L}\left(\alpha_k|D_k(\mathbf{x}_i)\right)\Big) && (6.18)\\
&\overset{(6.17)}{\approx} \ln\left(\prod_{\mathbf{x}_i\in\Psi_k} \alpha_k e^{-\alpha_k D_k(\mathbf{x}_i)}\right) && (6.19)\\
&= \ln\left(\prod_{\mathbf{x}_i\in\Psi_k} \alpha_k\right) + \ln\left(\prod_{\mathbf{x}_i\in\Psi_k} e^{-\alpha_k D_k(\mathbf{x}_i)}\right) && (6.20)\\
&= N_k \ln\left(\alpha_k\right) - \alpha_k \sum_{\mathbf{x}_i\in\Psi_k} D_k(\mathbf{x}_i) && (6.21)\\
&\rightarrow \max
\end{aligned}$$

To determine the maximum of the log-likelihood function, we compute its first derivative:

$$\begin{aligned}
\frac{\partial}{\partial\alpha_k}\Lambda(\alpha_k|D_k(\mathbf{x}_i)) &\overset{(6.18)}{=} \frac{\partial}{\partial\alpha_k}\ln\Big(\mathcal{L}\left(\alpha_k|D_k(\mathbf{x}_i)\right)\Big) && (6.22)\\
&\overset{(6.21)}{\approx} \frac{\partial}{\partial\alpha_k}\left(N_k\ln\left(\alpha_k\right) - \alpha_k\sum_{\mathbf{x}_i\in\Psi_k} D_k(\mathbf{x}_i)\right) && (6.23)\\
&= \frac{N_k}{\alpha_k} - \sum_{\mathbf{x}_i\in\Psi_k} D_k(\mathbf{x}_i) && (6.24)
\end{aligned}$$

Setting the derivative to zero and solving for α_k yields the maximum-likelihood parameter estimate $\widehat{\alpha}_k$:

$$\frac{\partial}{\partial\alpha_k}\Lambda(\alpha_k|D_k(\mathbf{x}_i)) \overset{(6.24)}{\approx} \frac{N_k}{\alpha_k} - \sum_{\mathbf{x}_i\in\Psi_k} D_k(\mathbf{x}_i) = 0 \qquad (6.25)$$

$$\Leftrightarrow \quad \widehat{\alpha}_k = \frac{N_k}{\sum_{\mathbf{x}_i\in\Psi_k} D_k(\mathbf{x}_i)} \qquad (6.26)$$

The second derivative of the log-likelihood function $\Lambda(\alpha_k|D_k(\mathbf{x}_i))$ is always less than zero, given that N_k, the number of samples in cluster Ψ_k, is a

positive number. Thus, $\widehat{\alpha}_k$ is indeed a maximum point:

$$\frac{\partial^2}{\partial {\alpha_k}^2}\Lambda(\alpha_k|D_k(\mathbf{x}_i)) \overset{(6.18)}{=} \frac{\partial^2}{\partial {\alpha_k}^2}\ln\Big(\mathcal{L}\left(\alpha_k|D_k(\mathbf{x}_i)\right)\Big) \tag{6.27}$$

$$\overset{(6.24)}{\approx} \frac{\partial}{\partial \alpha_k}\left(\frac{N_k}{\alpha_k} - \sum_{\mathbf{x}_i \in \Psi_k} D_k(\mathbf{x}_i)\right) \tag{6.28}$$

$$= \frac{-N_k}{{\alpha_k}^2} \tag{6.29}$$

$$< 0, \qquad \forall N_k > 0 \tag{6.30}$$

6.3 Experimental Set-Up

6.3.1 Dataset and Evaluation Methodology

The proposed multi-level mixture-of-experts framework is tested in experiments on pedestrian classification. Since we require multi-cue (intensity, dense stereo, dense optical flow) training and test samples, we cannot use established datasets for benchmarking, e.g. [27, 32, 37, 109]. Recently, an independently developed approach combining intensity, motion and depth was presented in [166]. However, the dataset used in [166] is only partly publicly available (the training data is not public). We make our full multi-cue training and test dataset publicly available to non-commercial entities for research purposes.[1]

Our training and test samples consist of manually labeled pedestrian and non-pedestrian bounding boxes in images captured from a vehicle-mounted calibrated stereo camera rig in an urban environment. For each manually labeled pedestrian, we create additional samples by geometric jittering. Non-pedestrian samples result from a pedestrian shape recognition pre-processing step with a relaxed threshold setting, as well as ground-plane constraints and prior knowledge about pedestrian geometry, i.e. containing a bias towards more "difficult" patterns, resembling pedestrians in geometry and structure. 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 [68]. To compute dense optical flow, we use the

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

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

Table 6.1: Training and test set statistics.

method of [168]. See Figure 6.3 and Table 6.1 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 6.2.4, as cluster membership weights for training:

$$z_i^k = w_k(\mathbf{x}_i) \approx P(\Psi_k|\mathbf{x}_i) \tag{6.31}$$

To compute $w_k(\mathbf{x}_i)$, a set of 10946 shape templates corresponding to clusters Ψ_k is used according to the methods outlined in Section 6.2.4.

6.3.2 Feature Extraction and Classification

Regarding features for our multi-modality classifiers, we choose histograms of oriented gradients (HOG) [27] and cell-structured local binary patterns (LBP) with uniformity constraints [115, 167] out of many possible feature sets, cf. [32, 37, 109]. The motivation for this choice is two-fold: 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 [167]. Second, HOG and LBP features are still among the best performing (and most popular) feature sets available, cf. [32, 37, 167].

We follow [27] and compute histograms of oriented gradients with 9 orientation bins and 8×8 pixel cells, accumulated to overlapping 16×16 pixel blocks with a spatial shift of 8 pixels. See Section 4.1.2 for more details on the HOG feature extraction algorithm.

A single LBP feature is based on a local comparison of n pixels $\mathbf{p}_i$ within a given region to the center pixel $\mathbf{c}$ of the region. Each region is described as an n-bit string, where each bit denotes the relation of $\mathbf{p}_i$ and $\mathbf{c}$. If the pixel intensity $\pi(\mathbf{p}_i)$ is larger than $\pi(\mathbf{c})$, 1 is added to the bit string and

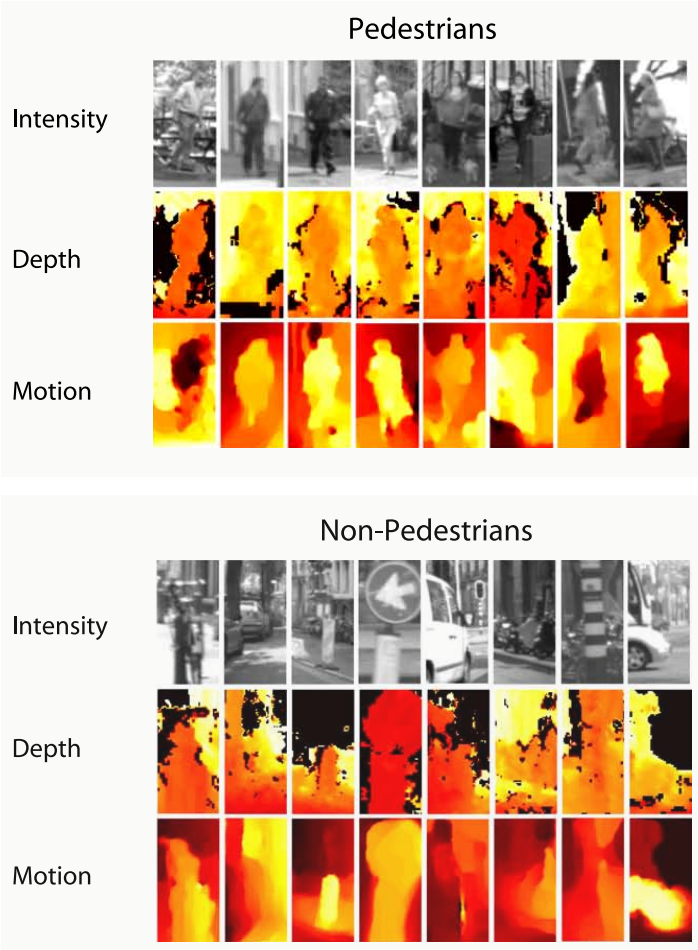


Figure 6.3: Multi-modality pedestrian and non-pedestrian samples in our dataset. In depth images, darker colors denote closer distances. Note that the background (large depth values) has been faded out for visibility. Optical flow images depict the horizontal component of flow vectors. Medium red colors denote close to zero flow, darker and brighter colors indicate stronger motion (to the left and to the right, respectively).

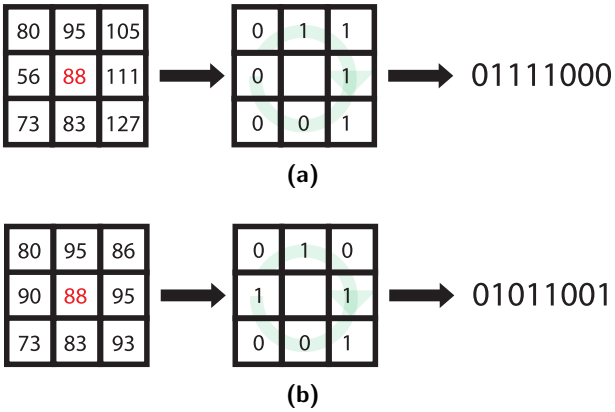


Figure 6.4: Extraction of $\text{LBP}^2_{8,1}$ features in an 8 pixel neighborhood. (a) A uniform $\text{LBP}^2_{8,1}$ feature with two 0-1 transitions. (b) A non-uniform $\text{LBP}^2_{8,1}$ feature with five 0-1 transitions.

0 otherwise. Generally, $\text{LBP}^u_{n,r}$ denotes LBP features that use n sample points with a radius r and the number of 0-1 transitions in the bit string no more than u. Patterns that satisfy this constraint are referred to as *uniform* patterns. Here, we use $\text{LBP}^2_{8,1}$ features, as shown in Figure 6.4.

To compute cell-structured LBPs, we first divide the input sample (48×96 pixels) into 8×8 pixel cells, similar to HOG features. Within each 8×8 pixel cell, 64 single LBP features can be extracted (every pixel in the cell can be regarded as a center pixel **c** for a single LBP feature). The feature vector for a cell is then given by building a histogram which counts the occurrence of each LBP bit string. To filter out noise in uniform image areas, only uniform LBPs are voted into different bins. All non-uniform patterns are voted into a single bin. The individual cell feature vectors are then concatenated into a single feature vector for the whole 48×96 pixel input sample, followed by L_1-sqrt normalization (other normalization variants did not improve performance).

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 multi-layer perceptrons (MLP) with one hidden layer consisting of eight neurons with sigmoidal transfer functions, trained

	Intensity	Depth	Motion
HOG	0.27	0.14	0.08
LBP	0.24	0.11	0.16

Table 6.2: Mean weights $\overline{s}^{\,m,f}$ for features and modalities, estimated by a linear SVM on the training set and averaged over view-clusters Ψ_k.

stochastically using the on-line error back-propagation algorithm. We utilize the *FANN* library for MLP training [113]. Compared to the popular linear support vector machines (linSVM), MLPs provide non-linear decision boundaries which usually improve performance, see [109]. The training of non-linear support vector machines was practically infeasible, given our large datasets.

Expert classifier weights $s_k^{m,f}$, see Equations (6.9) and (6.10), are computed using the linear SVM approach given in Section 6.2.3, applied to the training set. We utilize the *LIBLINEAR* library for linear SVM training [44]. Table 6.2 lists $\overline{s}^{\,m,f}$, the actual weights for individual features and modalities averaged over view-clusters Ψ_k:

$$\overline{s}^{\,m,f} = \frac{1}{K} \sum_{k=1}^{K} s_k^{m,f} \tag{6.32}$$

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 multi-modality data.

6.4 Experiments

Our experiments are designed to evaluate the different levels of the proposed mixture-of-experts framework, see Figure 6.1a, 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.

6.4.1 Pose-Level Mixture-of-Experts

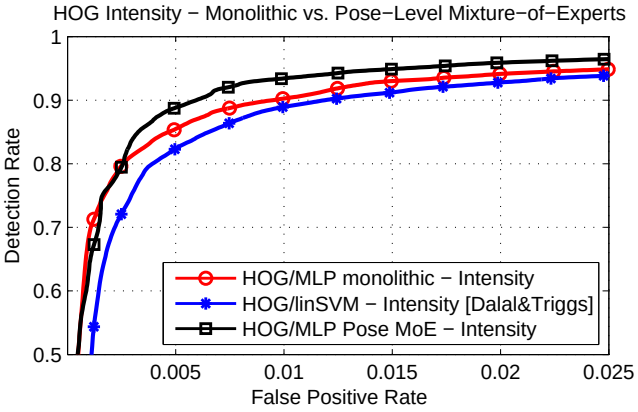
In our first experiment, we evaluate the benefit of our mixture-of-experts architecture on pose-level only. For that, we compare the proposed pose-specific mixture architecture to single "monolithic" classifiers trained on the whole dataset irrespective of view. We do not consider multi-modality or multi-feature 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 multi-layer perceptrons (MLP). Note that the monolithic HOG/linSVM approach corresponds to the method proposed by Dalal & Triggs [27]. Results are shown in Figure 6.5a for HOG and Figure 6.5b for LBP features.

Irrespective of the employed feature set, the pose-level mixture classifiers perform better than the corresponding monolithic classifiers. The decomposition of the problem into view-related sub-parts 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, multi-layer perceptrons outperform linear support vector machines, because of their non-linearities in decision space. Except for some experiments in Section 6.4.5, we utilize pose-level mixture-of-experts classification throughout the following experiments.

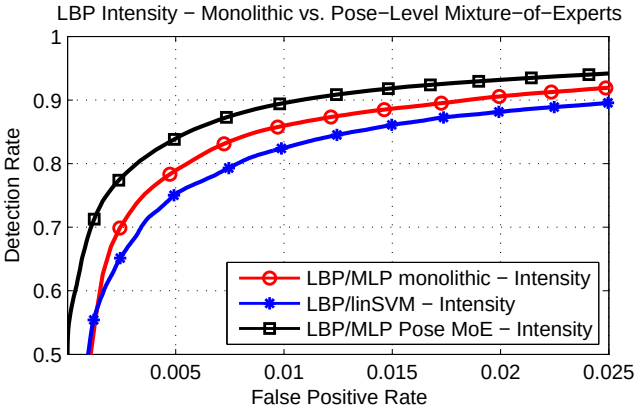
6.4.2 Modality-Level Mixture-of-Experts

In our second experiment, we evaluate the performance of modality-level classifiers, as presented in Section 6.2.3, compared to intensity-only classifiers. Pose-level mixtures are also used, that is, the first two levels of our framework, see Figure 6.1a, 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 Figures 6.6a and 6.7a for HOG and Figures 6.6b and 6.7b 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. Figure 6.6a (HOG) vs. Figure 6.6b (LBP). Classifiers in the intensity modality have the best performance, by a large margin. In depth and motion modal-

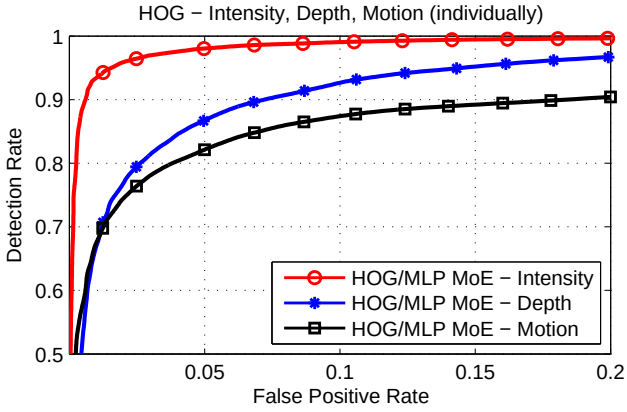


(a)

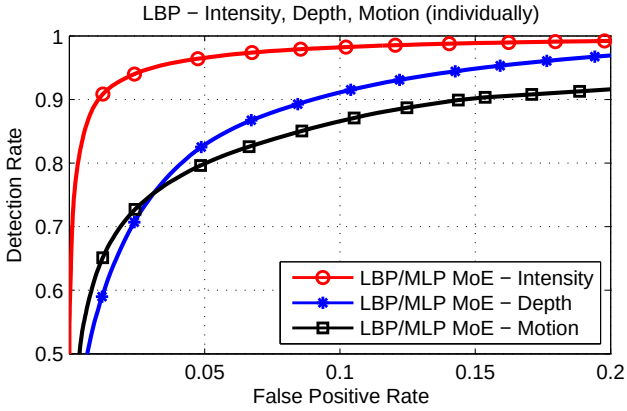


(b)

Figure 6.5: Pose-level mixture-of-experts vs. monolithic classifier. (a) HOG features in intensity modality. (b) LBP features in intensity modality.

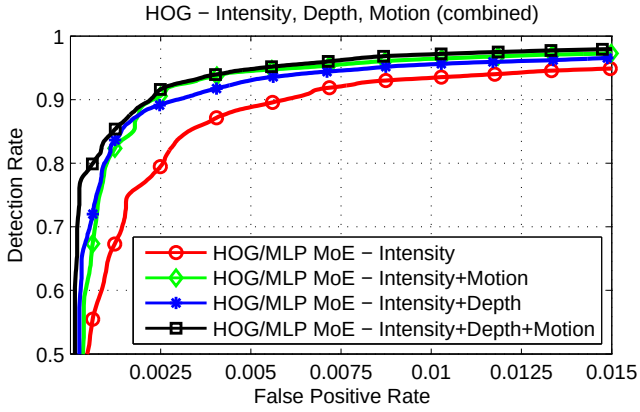


(a)

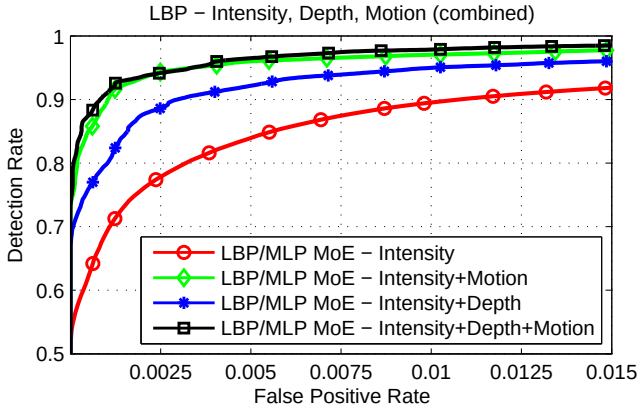


(b)

Figure 6.6: Modality-level mixture-of-experts. (a)-(b) Individual classification performance of HOG (a) and LBP (b) features in intensity, depth and motion modality.



(a)



(b)

Figure 6.7: Modality-level mixture-of-experts. (a)-(b) Combined classification performance of HOG (a) and LBP (b) features in intensity, depth and motion modality.

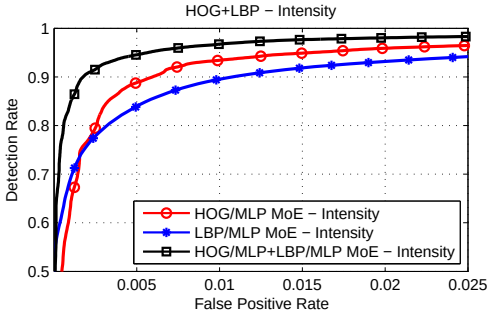
ities, performance is similar for both feature sets with depth features performing better then motion features at higher false positive rates and worse at lower false positive rates. Note that these performance relations are also apparent in the individual expert classifier weights, see Table 6.2.

Figure 6.7 shows 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 to LBP features. The HOG classifier using intensity, depth and motion has approx. a factor of four less false positives than a comparable HOG classifier using intensity only, cf. Figure 6.7a. From Figure 6.7b we observe, that in case of LBP features, the performance boost resulting from utilizing all modalities vs. intensity-only is approx. a factor of twelve in reduction of false positives at equal detection rates.

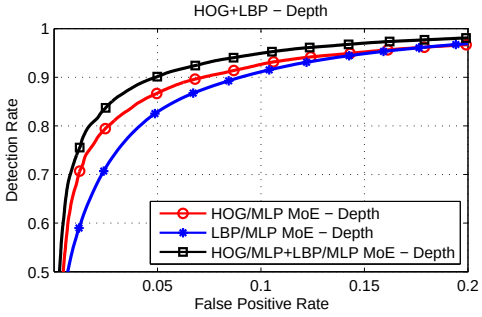
6.4.3 Feature-Level Mixture-of-Experts

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 multi-feature approach in all three modalities, i.e. intensity, depth, motion, individually. Recalling our framework architecture, see Figure 6.1a, this corresponds to having levels 1 (pose) and 3 (features) in place. Results are given in Figures 6.8a (intensity), 6.8b (depth) and 6.8c (motion).

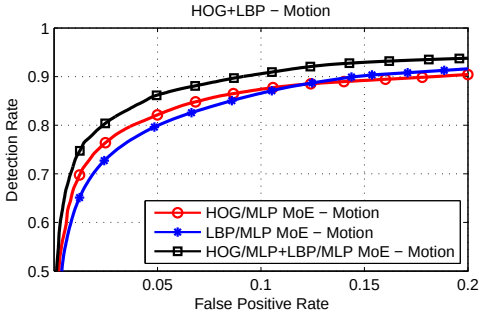
In all modalities, one can observe that combining HOG and LBP improves performance over using both features individually. The largest performance boost coming from the feature-level mixture-of-experts exists in the intensity modality. Here, the combined HOG+LBP classifier has up to a factor of four less false positives than the HOG classifier, which in turn outperforms the LBP classifier at higher detection rates. In the depth and motion modalities, the corresponding performance boosts amount to factors of 2 (motion) and 1.5 (depth) at equal detection rate levels. Compared to the performance improvement obtained by combining different modalities, as shown in Section 6.4.2, the effect of feature-level mixture-of-experts is less pronounced, but still significant.



(a)

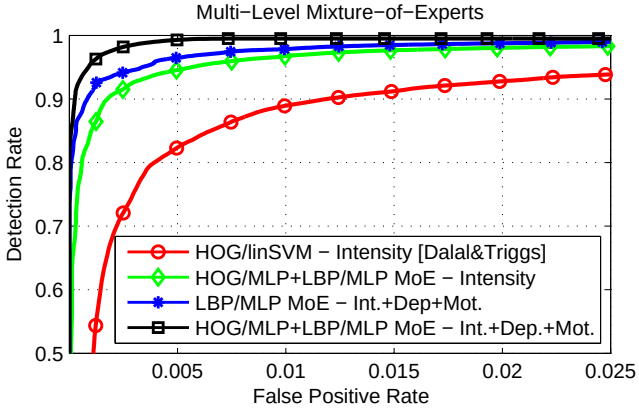


(b)

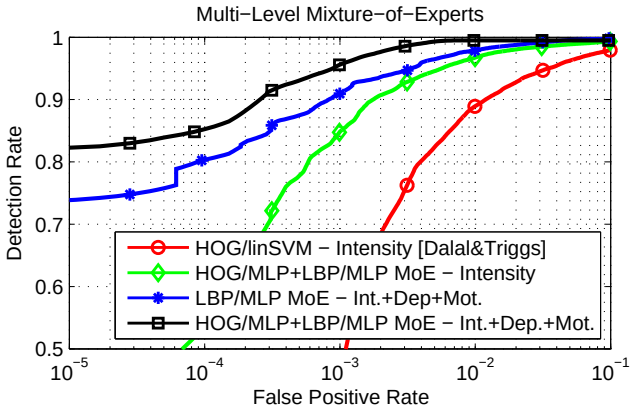


(c)

Figure 6.8: Feature-level mixture-of-experts. Individual classification performance of HOG, LBP and HOG+LBP in intensity (a), depth (b) and motion (c) modality. Note the different scaling on the x-axis.



(a)



(b)

Figure 6.9: Multi-level mixture-of-experts evaluation and performance overview. (a) Monolithic HOG classifier in intensity domain, best feature-level MoE (HOG+LBP, intensity), best modality-level MoE (LBP, intensity+depth+motion), multi-level MoE (HOG+LBP, intensity+depth+motion). (b) Logarithmic plot of (a), focusing on low false-positive rates.

6.4.4 Multi-Level Mixture-of-Experts

We now evaluate the performance of our full multi-level 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 [27], 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 modalities, cf. Section 6.4.2, as well as feature-level mixture-of-experts using HOG+LBP mixture-of-experts in the intensity domain only, cf. Section 6.4.3.

ROC performance is given in Figure 6.9. We observe that our combined multi-level 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 [27]. To quantify performance, Table 6.3 lists the false positive rates of all approaches shown in Figure 6.9 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 positives of more than a factor of 6 over the Dalal & Triggs HOG/linSVM approach. The use of pose-level and modality-level experts (LBP/MLP, intensity + depth + motion modalities) reduces false positives by more than a factor of 13 compared to the HOG/linSVM baseline. Our full multi-level mixture-of-experts approach (HOG/MLP + LBP/MLP, intensity + depth + motion modalities) further boosts performance up to a reduction in false positives by a factor of 42.

The results clearly show the benefit of our integrated multi-level architec-

	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 6.3: Performance of approaches in Figure 6.9 using 90 % detection rate as a common reference point, see text.

	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)

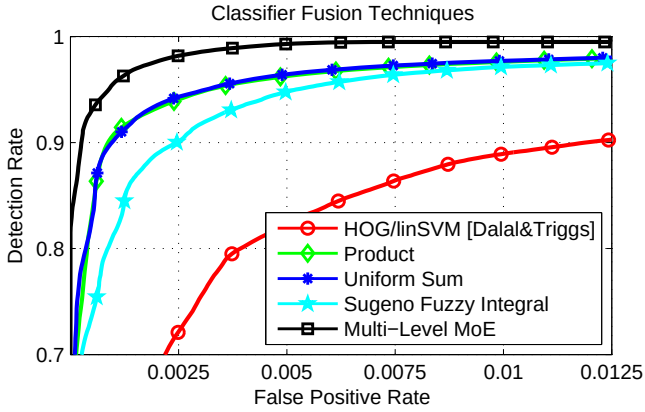
Table 6.4: Correlation of classifier outputs in (a) different modalities and (b) different features.

ture. 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, multi-modality classification approaches seem to be a promising direction for future research in the domain of object classification to boost overall performance.

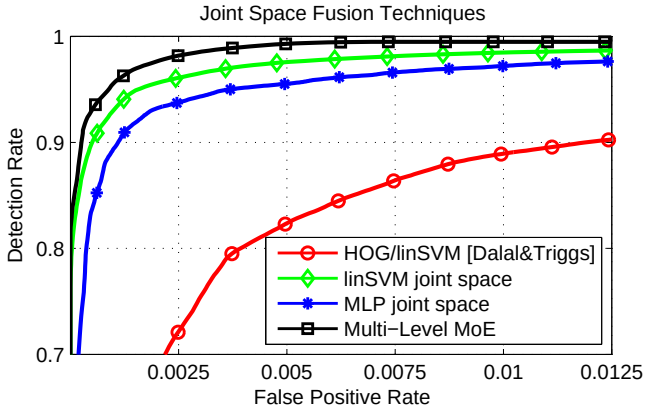
To gain further insight, we compute the correlation of classifier outputs (decision values) for the individual modality/feature expert classifiers, computed for pedestrian and non-pedestrian samples individually and then averaged over the two classes, see Table 6.4. The correlation analysis shows, that classifier outputs are far less correlated across different modalities (Table 6.4a) than across different features (Table 6.4b). Here, the less correlated two modalities/features are, the larger the benefits obtained in classification performance, cf. Figures 6.6, 6.7 and 6.8.

6.4.5 Classifier Fusion

In our final experiments, we compare our multi-level 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 classifiers using a joint feature space which consists of all features in all modalities L_2-normalized and concatenated into a single feature vector, cf. [173]. Given our feature set-up as presented in Section 6.3.2, the total dimensionality of the joint feature space is 18684. For comparison, the performance of the Dalal & Triggs HOG/linSVM baseline [27] is also given. Results are shown in Figure 6.10a for the multi-classifier fusion and in Figure 6.10b for



(a)



(b)

Figure 6.10: Performance of different classifier fusion techniques. (a) multi-classifier fusion. (b) joint feature space with single classifiers.

the joint space fusion approaches.

The multi-classifier 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, cf. [87], as well as a fuzzy integration using Sugeno integrals, cf. [118]. 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 & Triggs HOG/linSVM classifier [27], our multi-level mixture-of-experts classifier has a much better performance. This clearly shows the benefit of gating on pose-level, see Equation (6.4), and the learned classifier combination weights in Equation (6.11).

In terms of joint space approaches, we train both a multi-layer perceptron (MLP) and a linear support vector machine (linSVM) in the enlarged 18684-dimensional joint feature space (training a non-linear SVM was not feasible given our large dataset). While one could expect the MLP to improve performance over the linSVM, due to the non-linear 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. [33], 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 training [33]. This rule is severely violated in case of the MLP in the 18684-dimensional joint feature space with 149489 free parameters and 84577 training samples. The linear support vector machine 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 multi-level mixture-of-experts framework using MLPs as expert classifiers outperforms the joint space linSVM. We can afford to use more complex sub-classifiers in our model, since each MLP is an expert in a lower-dimensional modality/feature subspace, weighted by the contribution of the shape cues.

6.5 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) [101], to further increase performance. Multi-modality classifiers particularly outperform multi-feature classifiers in a single modality. Yet, modalities and features are orthogonal, so that a combined multi-modality / multi-feature 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 from 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, cf. Section 6.2.1, 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, cf. [131], where the relative order of feature/classifier performance reverses with respect to intensity and depth.

Orthogonal to the improvements presented in this work are benefits resulting from an increased training set, cf. [37, 109]. 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, cf. [37, 109]. 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 [32]. 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 set-up.

Certainly, more research is necessary to fully explore the benefits of multi-modality / multi-feature classification. For that purpose, we provide our

multi-modality dataset not only as a means for benchmarking but also to stimulate further research on the issues mentioned above.

6.6 Conclusion

This chapter presented a probabilistic multi-level mixture-of-experts framework involving a view-related and sample-dependent combination of multi-modality / multi-feature 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 sub-problems, 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 from both the fuzzy sub-division 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 multi-modality dataset publicly available for benchmarking purposes and to stimulate further research to address open issues with regard to multi-modality / multi-feature classification.

Chapter 7

Multi-Modality Partial Occlusion Handling

7.1 Overview

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

Component-based approaches which represent a pedestrian as an ensemble of parts, see Section 2.2.3, can only alleviate this problem to some extent without prior knowledge. The key to successful recognition of partially occluded pedestrians is additional information about which body parts are occluded. Classification can then rely on the unoccluded pedestrian components to obtain a robust decision.

In this chapter, we present a multi-modality component-based mixture-of-experts framework for pedestrian classification with partial occlusion handling. The multi-level mixture-of-experts framework, as introduced in Chapter 6, is employed. We do not consider view-specific experts on "pose-level", cf. Figure 6.1a in Section 6.1, but replace this level with a component-based approach which represents a pedestrian as an ensemble of body parts. Pose-level experts could be additionally incorporated, given a method to extend the shape-based computation of view priors, as outlined in Section 6.2.4, to operate on body components instead of fully visible pedestrians, e.g. part-specific hierarchical shape matching, see the discussion in [54]. In this chapter, we focus on the method for partial occlusion handling. At the core of our framework is a set of component-based expert classifiers trained on intensity, depth and motion features. Occlusions of individual body parts manifest in local

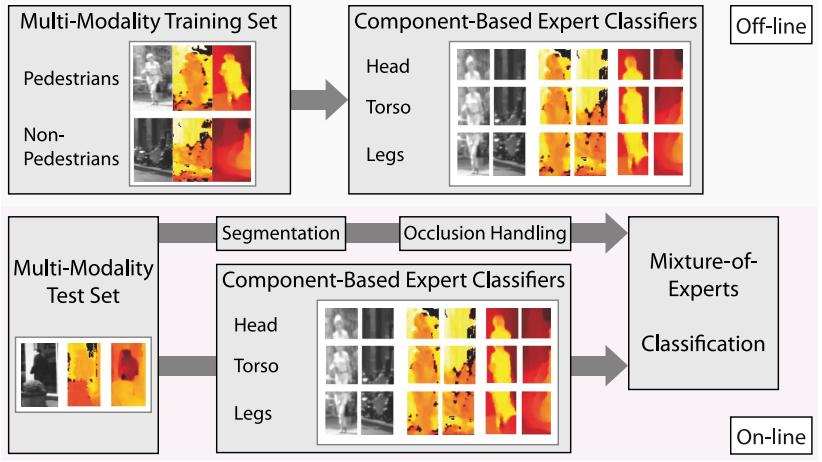


Figure 7.1: Framework overview. Multi-modality component-based expert classifiers are trained off-line on features derived from intensity, depth and motion. On-line, multi-modality segmentation is applied to determine occlusion-dependent component weights for expert fusion. Data samples are shown in terms of intensity images, dense depth maps and dense optical flow (left to right).

depth- and motion-discontinuities. In the application phase, a segmentation algorithm is applied to extract areas of coherent depth and motion. Based on the segmentation result, we determine occlusion-dependent weights for our component-based expert classifiers to focus the combined decision on the visible parts of the pedestrian. See Figure 7.1.

In view of recognizing partially occluded pedestrians, component-based classification seems an obvious choice. Yet, only a few approaches have used techniques to infer a measure of (partial) occlusion from the image data [145, 167, 174]. Sigal and Black proposed a technique for articulated 3D body pose estimation which is able to handle self-occlusion of body parts [145]. In our application however, we are not interested in (self-)occlusion handling of articulated 3D pose but focus on partial occlusions observed in 2D images of pedestrians. Particularly relevant to current work are the approaches of Wu and Nevatia [174] and Wang et al. [167]. They explicitly incorporate a

model of partial occlusion into their 2D classification framework. However, both methods make some restrictive assumptions, as follows.

The approach of Wu and Nevatia requires a particular camera set-up, where the camera looks down on the ground-plane [174]. Consequently, they assume that the head of a pedestrian in the scene is always visible. They further apply a binary threshold to ignore occluded components in their component-fusion algorithm.

Wang et al. use a monolithic (full-body) HOG/linSVM classifier to determine occlusion maps from the responses of the underlying block-wise feature set [167]. Based on the spatial configuration of the recovered occlusion maps, they either apply a full-body classifier or activate part-based classifiers in non-occluded regions or heuristically combine both full-body and part-based classifiers. Since their method depends on the block-wise responses of HOG features combined with linear SVMs, it is unclear how to extend their approach to other popular features or classifiers.

Unlike [174], our method does neither pose restrictions on the camera set-up nor assumes constant visibility of a certain body part. In contrast to [167], our approach does not depend on a particular feature/classifier combination or a certain pedestrian component layout.

7.2 Pedestrian Classification

Input to our framework is a training set $\mathcal{D}$ of pedestrian (ω_0) and non-pedestrian (ω_1) samples $\mathbf{x}_i \in \mathcal{D}$. Similar to Section 6.2.1, each sample $\mathbf{x}_i = [\mathbf{x}_i^1; \mathbf{x}_i^2; \mathbf{x}_i^3]$ consists of three different modalities, i.e. gray-level image intensity ($\mathbf{x}_i^1$), dense depth information via stereo vision ($\mathbf{x}_i^2$) [68] and dense optical flow ($\mathbf{x}_i^3$) [168]. See Figure 7.5 in Section 7.3.1.

7.2.1 Component-Based Classification

For classification, we approximate the posterior probability that an unseen sample $\mathbf{x}_i$ is a pedestrian, $P(\omega_0|\mathbf{x}_i)$, in terms of a component-based model. Each sample $\mathbf{x}_i$ is composed out of C components which are usually related to body parts. With $\mathbf{C}_c(\mathbf{x}_i)$ representing a local expert classifier for the c-th component of $\mathbf{x}_i$ and $w_c^o(\mathbf{x}_i)$ denoting its weight, we approximate $P(\omega_0|\mathbf{x}_i)$

using:

$$P(\omega_0|\mathbf{x}_i) \approx \sum_{c=1}^{C} w_c^o(\mathbf{x}_i)\mathbf{C}_c(\mathbf{x}_i) \tag{7.1}$$

Note that the weight $w_c^o(\mathbf{x}_i)$ for each component expert classifier is not a fixed component prior, but depends on the sample $\mathbf{x}_i$ itself. These component weights allow to incorporate a model of partial occlusion into our framework (hence the "o" superscript), as shown in Section 7.2.3.

7.2.2 Multi-Modality Component Expert Classifiers

Given our component-based mixture-of-experts model, cf. Equation (7.1), we model the component expert classifiers $\mathbf{C}_c(\mathbf{x}_i)$ in terms of our multi-modality (intensity, depth, flow) dataset. As in Section 6.2.3, we extend the mixture-of-experts formulation by introducing individual component-based classifiers for each modality:

$$\mathbf{C}_c(\mathbf{x}_i) = \sum_m v_c^m \ \mathbf{D}_c^m(\mathbf{x}_i^m) \tag{7.2}$$

In this formulation, $\mathbf{D}_c^m(\mathbf{x}_i^m)$ denotes a local expert classifier for the c-th component of $\mathbf{x}_i$, which is represented in terms of the m-th modality. As expert classifiers, we use feature-based pattern classifiers which are learned on the training set using data from the corresponding component and modality only. Each component/modality classifier is trained to discriminate between the pedestrian and non-pedestrian class in its local area of the feature space. Similar to Equation (6.11) in Section 6.2.3, we estimate weights v_c^m to each modality classifier on the training set using a linear support vector machine.

7.2.3 Occlusion-Dependent Component Weights

Weights $w_c^o(\mathbf{x}_i)$ for component classifiers were introduced in Section 7.2.1. We derive $w_c^o(\mathbf{x}_i)$ from each example $\mathbf{x}_i$ to incorporate a measure of occlusion of certain pedestrian components into our model. Expert classifier outputs, related to occluded components, should have a low weight in the combined decision of the expert classifiers, cf. Equation (7.1). We propose to extract visibility information from each sample $\mathbf{x}_i$ using the depth (stereo vision) and motion (optical flow) modalities. Partially occluded pedestrians, e.g. a walking pedestrian behind a static object, exhibit significant depth and motion discontinuities at the occlusion boundary, as shown in Figures 7.2 and 7.5.

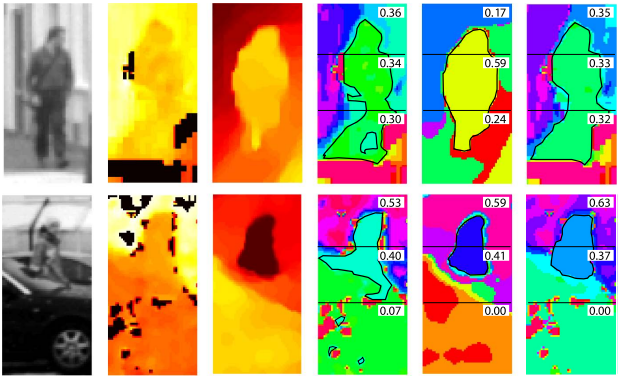


Figure 7.2: Segmentation results for a non-occluded (first row) and partially occluded pedestrian (second row). From left to right, the columns show: intensity image, stereo image, flow image, segmentation on stereo, segmentation on flow, combined segmentation on stereo and flow. Clusters are color-coded and the cluster chosen as pedestrian cluster $\vec{\phi}_{ped}$, cf. Equation (7.7), is outlined in black. The computed occlusion-dependent component weights $w_c^o(\mathbf{x}_i)$, cf. Equation (7.8), are also shown.

Visible parts of a pedestrian are assumed to be in approximately the same distance from the camera (pedestrian standing upright on the ground) and move uniformly.

We employ a three-step procedure to derive component weights $w_c^o(\mathbf{x}_i)$ from an unseen sample $\mathbf{x}_i$: First, we apply a segmentation algorithm, cf. [42], to the dense stereo and optical flow images of $\mathbf{x}_i$. Second, we select the segmented cluster which likely corresponds to the visible area of a pedestrian. For this, a measure of similarity of a cluster to a generic model of pedestrian geometry in terms of pedestrian shape, size and location is utilized. Third, we estimate the degree of visibility of each component given the selected cluster.

For segmentation, we choose the mean-shift algorithm, cf. [21], out of many possible choices. As shown in [42], mean-shift provides a good balance between segmentation accuracy and processing efficiency. The result of the mean-shift segmentation is a set of l clusters ϕ_l with $l = 1, \ldots, L$, as shown in Figure 7.2. The actual number of clusters L is optimized during mean-shift

itself [21]. We evaluate both single-modality segmentation using depth or motion and simultaneous multi-modality segmentation using both modalities in our experiments, as shown in Section 7.3.

Let $\vec{\phi}_l$ and $\vec{\gamma}_c$ denote binary vectors defining the membership of pixel-locations of the sample $\mathbf{x}_i$ to the l-th cluster ϕ_l and c-th component γ_c, respectively. Note that $\vec{\phi}_l$ results from the segmentation algorithm, whereas $\vec{\gamma}_c$ is given by the geometric component layout. Further, we utilize a two-dimensional probability mass function $\mu_v(\mathbf{p}|\omega_0)$ which represents the probability that a given pixel $\mathbf{p} \in \mathbf{x}_i$ corresponds to a pedestrian ω_0, solely based on its location within $\mathbf{x}_i$. $\mu_v(\mathbf{p}|\omega_0)$ is obtained from the normalized superposition of a set of S aligned binary pedestrian foreground masks $m_s(\mathbf{p})$, obtained from manually labeled pedestrian shapes, cf. Equation (4.5) in Section 4.2.2:

$$\mu_v(\mathbf{p}|\omega_0) \sim \sum_{s=1}^{S} m_s(\mathbf{p}), \quad 0 \leq \mu_v(\mathbf{p}|\omega_0) \leq 1 \tag{7.3}$$

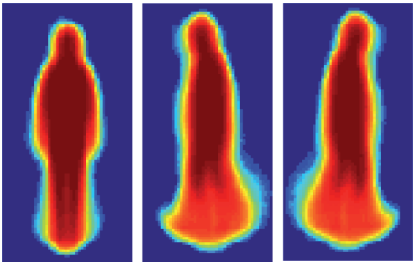
To increase specificity, we use view-dependent probability masks $\mu_v(\mathbf{p}|\omega_0)$ in terms of separate masks for front/back, left and right views. Those probability masks represent a view-dependent model of pedestrian geometry in terms of shape, size and location. See Figure 7.3a. Again, a vectorized representation of μ_v is denoted as $\vec{\mu}_v$.

To select the segmented cluster, which corresponds to the visible area of a pedestrian, we utilize a correlation-based similarity measure Γ, as defined in Equation (7.4). Our similarity measure employs the cluster information and the probability masks to assess the likelihood that a cluster ϕ_l corresponds to the visible parts of a pedestrian. We model Γ as the sum of two terms, Γ_{in} and Γ_{out}:

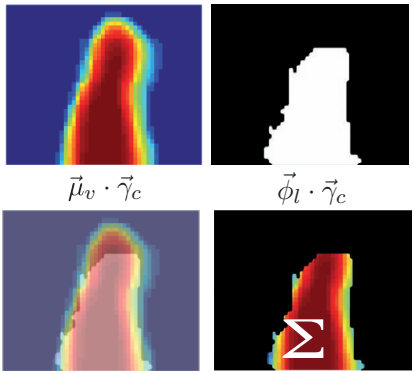
$$\Gamma(\vec{\phi}_l, \vec{\gamma}_c, \vec{\mu}_v) = \Gamma_{in}(\vec{\phi}_l, \vec{\gamma}_c, \vec{\mu}_v) + \Gamma_{out}(\vec{\phi}_l, \vec{\gamma}_c, \vec{\mu}_v) \tag{7.4}$$

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

$$\Gamma_{in}(\vec{\phi}_l, \vec{\gamma}_c, \vec{\mu}_v) = \frac{(\vec{\mu}_v \cdot \vec{\gamma}_c) \circ (\vec{\phi}_l \cdot \vec{\gamma}_c)}{\vec{\mu}_v \circ \vec{\gamma}_c} \tag{7.5}$$



(a)



(b)

Figure 7.3: (a) Probability masks for front/back, left and right view. The values of the probability masks are in the range of zero (dark blue) to one (dark red). The values specify the probability of a certain pixel to be part of a pedestrian with the corresponding view. (b) Visualization of the correlation-based similarity measure $\Gamma_{in}(\vec{\phi}_l, \vec{\gamma}_c, \vec{\mu}_v)$ for the head component, see text.

Here, $\cdot$ denotes point-wise multiplication of vectors, while $\circ$ denotes a dot product. Note that the main purpose of $\vec{\gamma}_c$ in this formulation is to restrict computation to a local body component γ_c. See Figure 7.3b.

The second measure $\Gamma_{out}(\vec{\phi}_l, \vec{\gamma}_c, \vec{\mu}_v)$ relates to the specificity of the cluster

ϕ_l. The idea is to penalize clusters which extend too far beyond a typical pedestrian shape. For that we perform similar correlation using an "inverse" probability mask $\vec{\nu}_v = 1 - \vec{\mu}_v$:

$$\Gamma_{out}(\vec{\phi}_l, \vec{\gamma}_c, \vec{\mu}_v) = 1 - \frac{(\vec{\nu}_v \cdot \vec{\gamma}_c) \circ (\vec{\phi}_l \cdot \vec{\gamma}_c)}{\vec{\nu}_v \circ \vec{\gamma}_c} \tag{7.6}$$

The cluster similarity measure $\Gamma(\vec{\phi}_l, \vec{\gamma}_c, \vec{\mu}_v)$, see Equation (7.4), is computed per cluster, component and view-dependent probability mask. To choose the cluster $\vec{\phi}_{ped}$ which most likely corresponds to visible parts of the pedestrian, we apply a maximum operation over components and views:

$$\vec{\phi}_{ped} = \underset{\vec{\phi}_l}{\operatorname{argmax}} \left(\max_{\vec{\gamma}_c \vec{\mu}_v} \left(\Gamma(\vec{\phi}_l, \vec{\gamma}_c, \vec{\mu}_v) \right) \right) \tag{7.7}$$

From our experiments we observed that the visible parts of a pedestrian do not significantly disintegrate in the mean-shift segmentation results, see Figure 7.2. Hence, we only consider single clusters ϕ_l and pairs of clusters merged together as possible candidates.

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

$$w_c^o(\mathbf{x}_i) \sim \frac{n(\vec{\phi}_{ped} \cdot \vec{\gamma}_c)}{\sum_{\vec{\phi}_j \in \Upsilon} \left(n(\vec{\phi}_j \cdot \vec{\gamma}_c) \right) + n(\vec{\phi}_{ped} \cdot \vec{\gamma}_c)} \tag{7.8}$$

See Figure 7.2 for a visualization of the cluster $\vec{\phi}_{ped}$, corresponding to visible parts of the pedestrian, and the recovered occlusion-dependent component weights $w_c^o(\mathbf{x}_i)$.

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

Table 7.1: Training and test set statistics.

7.3 Experiments

7.3.1 Experimental Set-Up

As fully visible pedestrian and non-occluded pedestrian samples, we use the dataset described in Section 6.3.1. Partially occluded pedestrians have been acquired in a similar fashion. Training and test samples have a resolution of 36×84 pixels with a 6 pixel border around the pedestrians. In our experiments, we use $C = 3$ components γ_c, corresponding to head/shoulder (36×24 pixels), torso (36×36 pixels) and leg (36×48 pixels) regions, see Figure 7.4. Note that our components vertically overlap by 12 pixels, i.e. each component has a 6 pixel border around the associated body part. In preliminary experiments, we determined this overlap to improve performance. To train the component classifiers, only non-occluded pedestrians (and non-pedestrian samples) are used. For testing, we evaluate performance using two different test sets: one involving non-occluded pedestrians and one consisting of par-

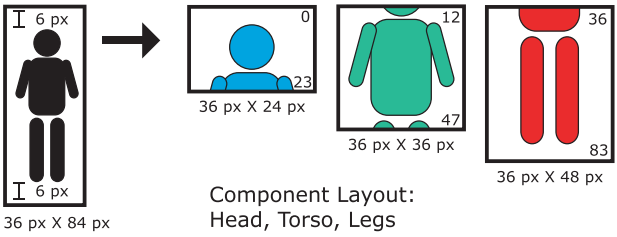


Figure 7.4: Component layout as used in our experiments. We employ three overlapping components, corresponding to head, torso and leg regions, see text.

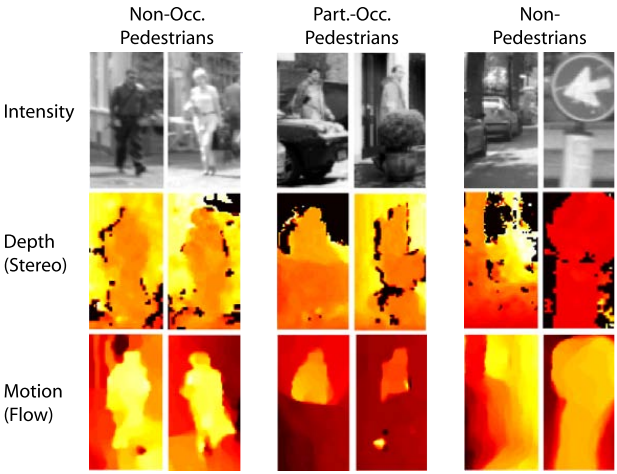


Figure 7.5: Non-occluded pedestrians, partially occluded pedestrian and non-pedestrian samples in our data. In depth (stereo) images, darker colors denote closer distances. Note that the background (large depth values) has been faded out for visibility. Optical flow images depict the magnitude of the horizontal component of flow vectors. Medium red colors denote close to zero flow, darker and brighter colors indicate stronger motion (to the left and to the right, respectively).

tially occluded pedestrians. The non-pedestrian samples are the same for both test sets. See Table 7.1 and Figure 7.5 for an overview of the dataset.

Regarding features for the component/modality expert classifiers $\mathbf{D}_c^m$, see Equation (7.2), we utilize histograms of oriented gradients (HOG). This allows us to compare our framework to the approach of Wang et al. [167] which explicitly requires and operates on the block-wise structure of HOG features. We compute histograms of oriented gradients with 12 orientation bins and 6×6 pixel cells, accumulated to overlapping 12×12 pixel blocks with a spatial shift of 6 pixels. For classification, we employ linear support vector machines (SVMs). Note that the same HOG feature set is extracted from intensity, dense stereo and dense flow images, cf. Chapter 6. In our implementation of [167], we use the occlusion handling of Wang et al. together with

	Intensity	Depth	Motion
Head	0.49	0.34	0.17
Torso	0.61	0.28	0.11
Legs	0.73	0.14	0.13

Table 7.2: Component-specific modality expert weights v_c^m estimated by a linear SVM on the training set.

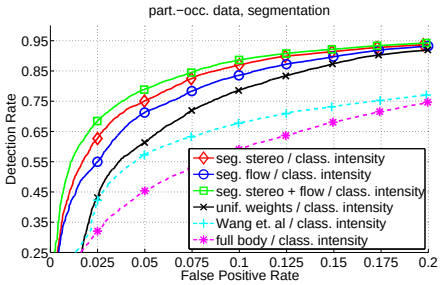
the same component layout (head, torso, legs), features (HOG) and classifiers (linear SVMs) as in our approach, but only for the intensity modality (as in the original publication). Table 7.2 lists the component-specific modality expert weights v_c^m as estimated on the training set, see Section 7.2.2.

7.3.2 Performance on Partially Occluded Test Data

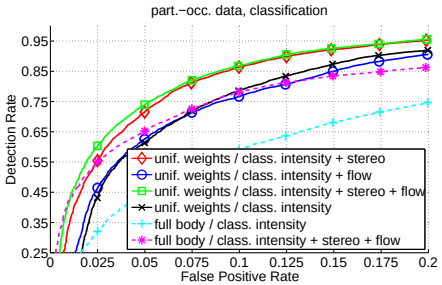
Partial Occlusion Handling

In our first experiment, we evaluate the effect of different models of partial occlusion handling. We do not consider multi-modality classifiers yet. All expert component classifiers are trained on intensity images only. As baseline classifiers, we evaluate the full-body HOG approach of [27] (we use the code provided by the original authors) and the approach of [167], which uses an occlusion model based on the block-wise response of a full-body HOG classifier to activate part-based classifiers in areas corresponding to non-occluded pedestrian parts. Our framework is evaluated using four different strategies to compute occlusion-dependent component weights $w_c^o(\mathbf{x}_i)$ for $\mathbf{x}_i$, as defined in Section 7.2.3: We consider weights resulting from mean-shift segmentation using depth only, flow only and a combination of both depth and flow. Additionally, we consider uniform weights $w_c^o(\mathbf{x}_i)$, i.e. no segmentation. Note that weights v_c^m, as given in Equation (7.2), are still in place. Results in terms of ROC performance are given in Figure 7.6a.

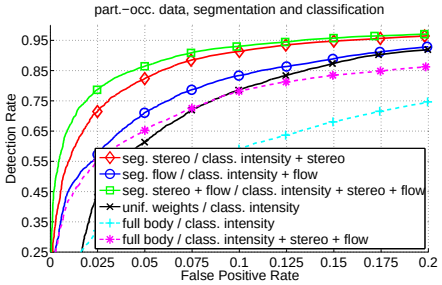
All component-based approaches outperform the full-body HOG classifier (magenta *). The approach of Wang et al. [167] (cyan +) significantly improves performance over the full-body HOG classifier by a factor of two (reduction in false positives at constant detection rates). All variants of our framework in turn outperform the method of Wang et al. [167], with segmentation on combined depth and flow (green □) performing best. Compared to the use of uniform weights $w_c^o(\mathbf{x}_i)$ (black ×), the addition of multi-modality



(a)



(b)



(c)

Figure 7.6: Classification performance on the partially occluded test set. (a) Evaluation of partial occlusion handling strategies. (b) Multi-modality classification in comparison to intensity-only classification. (c) Combined multi-modality partial occlusion handling and classification.

segmentation to compute component weights (green □) improves performance by approximately a factor of two.

Multi-Modality Classification

In our second experiment, we evaluate the performance of multi-modality component classifiers, as presented in Section 7.2.2, compared to intensity-only component classifiers. Uniform component weights $w_c^o(\mathbf{x}_i)$, i.e. no segmentation, were used throughout all approaches. Results are given in Figure 7.6b (solid lines). As baseline classifiers, we use a full-body intensity-only HOG classifier and a multi-modality full-body HOG classifier trained on intensity, stereo and flow data (dashed lines). Multi-modality classification significantly improves performance both for the full-body and for the component-based approach. The best performance (particularly at low false positive rates) is reached by the component-based approach involving intensity, stereo and flow (green □). The performance improvement over a corresponding component-based classifier using intensity-only (black ×) is up to a factor of two reduction in false positives.

Multi-Modality Classification with Partial Occlusion Handling

In the next experiment, we evaluate the proposed multi-modality framework involving occlusion-dependent component weights derived from mean-shift segmentation combined with multi-modality classification. Instead of presenting results for all possible combinations of modalities for segmentation and classification, we chose to use the same modalities for both segmentation and classification. We did evaluate all modality-combinations and found no better performing combination. Similar to the previous experiment, the baseline is given by full-body classifiers (cyan + and magenta *), as well as a component-based intensity-only classifier using uniform weights (black ×). See Figure 7.6c.

The best performing system variant is the proposed component-based mixture-of-experts architecture using stereo and optical flow concurrently to determine occlusion-dependent weights $w_c^o(\mathbf{x}_i)$ and for multi-modality classification (green □). Compared to a corresponding multi-modality full-body classifier (magenta *), the performance boost is approximately a factor of four. A similar performance difference exists between

our best approach (green □) and a component-based intensity-only classifier using uniform component weights (black ×).

7.3.3 Performance on Non-Occluded Test Data

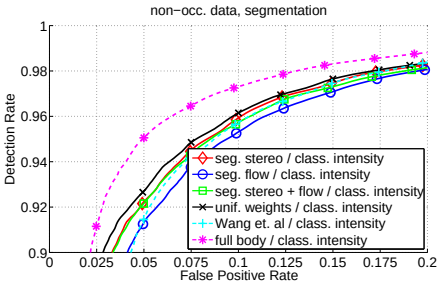
After demonstrating significant performance boosts on partially occluded test data, we evaluate the performance of the proposed approach using non-occluded pedestrians (and non-pedestrians) as test set. Similar to our previous experiments, we evaluate the effect of partial occlusion handling independently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

Figure 7.7a shows the effect of different models of partial occlusion handling combined with intensity-only component-based classifiers. The full-body HOG classifier (magenta *), as well as the approach of Wang et al. [167] (cyan +), serve as baselines. The best performance is reached by the full-body HOG classifier. All component-based approaches perform slightly worse. Of all component-based approaches, uniform component weights $w_c^o(\mathbf{x}_i)$, i.e. no occlusion handling, yields the best performance by a small margin. This is not surprising, since all components are visible to the same extent. On non-occluded test samples, our best approach with occlusion handling (green □) gives the same performance as Wang et al. [167] (cyan +).

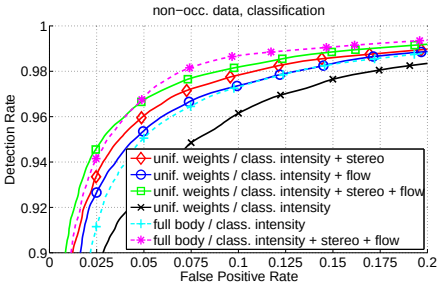
Multi-modality classification, as shown in Figure 7.7b, yields similar performance boosts compared to intensity-only classification as observed for the test on partially occluded data, cf. Section 7.3.2. Figure 7.7c depicts results of our integrated multi-modality mixture-of-experts framework with partial occlusion handling. Compared to a full-body classifier involving intensity, stereo and flow (magenta *), our best performing mixture-of-experts approach gives only slightly worse performance, particularly at low false positive rates. In relation to intensity-only full-body classification (cyan +), i.e. the approach of [27], our multi-modality framework improves performance by up to a factor of two.

7.4 Conclusion

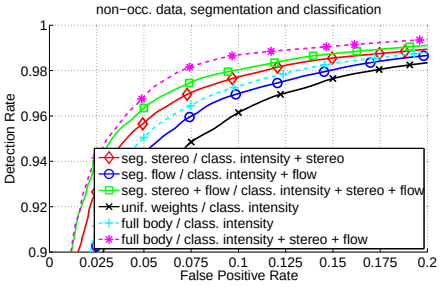
This chapter presented a multi-modality mixture-of-experts framework for component-based pedestrian classification with partial occlusion handling. For the partially occluded dataset, we obtained in the case of depth- and motion-based occlusion handling an improvement of more than a factor of



(a)



(b)



(c)

Figure 7.7: Classification performance on the non-occluded test set. (a) Evaluation of partial occlusion handling strategies. (b) Multi-modality classification in comparison to intensity-only classification. (c) Combined multi-modality partial occlusion handling and classification.

two versus the baseline (component-based, no occlusion handling) and state-of-the-art [167]. We obtained in the case of multi-modality (intensity, depth, motion) classification an additional improvement of a factor of two versus the baseline (intensity only). The full-body classifiers performed worse than the aforementioned baselines. For the non-occluded dataset, occlusion handling does not appreciably deteriorate results, while multi-modality classification improves performance by a factor of two.

Chapter 8

Integrated Classification and Orientation Estimation

8.1 Overview

Beyond recognizing a pedestrian in the scene, many application areas benefit from knowledge of body orientation of a pedestrian. In the domain of intelligent vehicles [51], known pedestrian orientation can enhance path prediction, to improve risk assessment. Other applications include perceptual interfaces [158], where body orientation can be used as a proxy for human-computer-interaction.

Orientation could be inferred by trajectory information (tracking) over time, assuming that pedestrians move forward. 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.

As a way out, methods to infer pedestrian orientation have been proposed. Besides work in the domain of 3D human pose estimation [107], few approaches have tried to recover an estimate of pedestrian orientation based on 2D lower-resolution images [52, 111, 142]. Existing approaches re-used popular features, i.e. Haar wavelets [142] or gradient histograms [52], and applied them in a different classification scheme. While pedestrian classification usually involves a two-class model (pedestrian vs. non-pedestrian), [52, 111, 142] did not use non-pedestrian training samples for orientation estimation. Instead, *one vs. one* [52] and *one vs. rest* [111, 142] multi-class schemes have been trained using pedestrian data only. Recovering the most likely discrete orientation class then involved maximum-selection over the

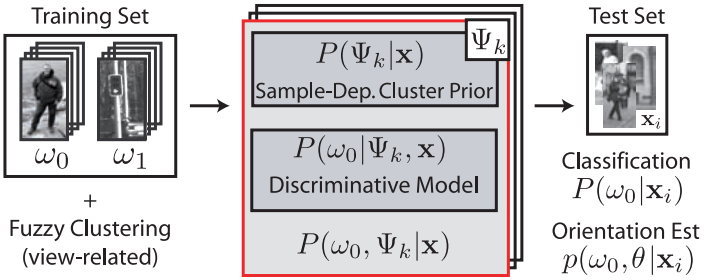


Figure 8.1: Framework overview. K view-related models specific to fuzzy clusters Ψ_k are used for pedestrian classification and orientation estimation. The models capture sample-dependent cluster priors and discriminative experts which are learned from pedestrian (class ω_0) and non-pedestrian (class ω_1) samples $\mathbf{x}$.

associated multi-class model.

In this chapter, 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. We re-use the view-specific set-up of the multi-level mixture-of-experts framework described in Chapter 6. However, to focus on orientation estimation, we dispense with the use of multiple features and multiple-cues in this chapter, i.e. only the "pose-level" of the mixture-of-experts framework is used, see Figure 6.1a in Section 6.1. The extension to multi-level orientation estimation is straightforward, similar to Section 6.2.3.

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 8.1 for an overview.

The general approach is independent from the actual type of discriminative models used and can be extended to other object classes. 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 Section 8.3).

8.2 Classification and Orientation Estimation

Similar to Section 6.2.1, the input data consists of 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 probabilistic 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 8.2. Our definition of cluster membership $\mathbf{z}_i$ is given in Section 8.3.1.

8.2.1 Pedestrian Classification

To classify a sample $\mathbf{x}_i$, we apply Equations (6.2), (6.3) and (6.4), see Section 6.2.2:

$$P(\omega_0|\mathbf{x}_i) = \sum_k P(\Psi_k|\mathbf{x}_i) P(\omega_0|\Psi_k, \mathbf{x}_i) \tag{8.1}$$

$$\approx \sum_k w_k(\mathbf{x}_i) \mathbf{H}_k(\mathbf{x}_i) \tag{8.2}$$

The sample-dependent priors for the membership to a certain cluster Ψ_k of an unseen sample $\mathbf{x}_i$, $P(\Psi_k|\mathbf{x}_i)$ are determined using the method outlined in Section 6.2.4. In our experiments, we additionally consider uniform priors.

8.2.2 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:

$$p(\omega_0, \theta|\mathbf{x}_i) \approx \sum_k \alpha_{k,i}\ g_k(\theta|\mathbf{x}_i) \tag{8.3}$$

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 $w_k(\mathbf{x}_i)\mathbf{H}_k(\mathbf{x}_i)$, the weighted classifier outputs, as defined in Equation (8.2):

$$g_k(\theta|\mathbf{x}_i) = \mathcal{N}(\theta|\mu_k, \sigma_k^2) \;\; ; \;\; \alpha_{k,i} = w_k(\mathbf{x}_i)\mathbf{H}_k(\mathbf{x}_i) \tag{8.4}$$

The most likely pedestrian orientation $\hat{\theta}_i$ can be recovered by finding the mode of the density in Equation (8.3), e.g. [19]:

$$\hat{\theta}_i = \underset{\theta}{\operatorname{argmax}} \left(p\left(\omega_0, \theta|\mathbf{x}_i\right)\right) \tag{8.5}$$

Besides estimating $p(\omega_0, \theta|\mathbf{x}_i)$, our framework allows to recover so-called orientation classes, similar to [52, 111, 142]: 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:

$$P(\omega_0, \theta \in [\tilde{\theta}_a, \tilde{\theta}_b] \mid \mathbf{x}_i) = \int_{\tilde{\theta}_a}^{\tilde{\theta}_b} p(\omega_0, \theta|\mathbf{x}_i) d\theta \tag{8.6}$$

We do not use *one vs. one*, e.g. [52, 111], or *one vs. rest*, e.g. [111, 142], 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 8.2. 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 $\mathbf{H}_k$ and the associated sample-dependent prior weights, see Equations (8.2) and (8.4). The training of $\mathbf{H}_k$ 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 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 $\mathbf{H}_k$, which are re-used.

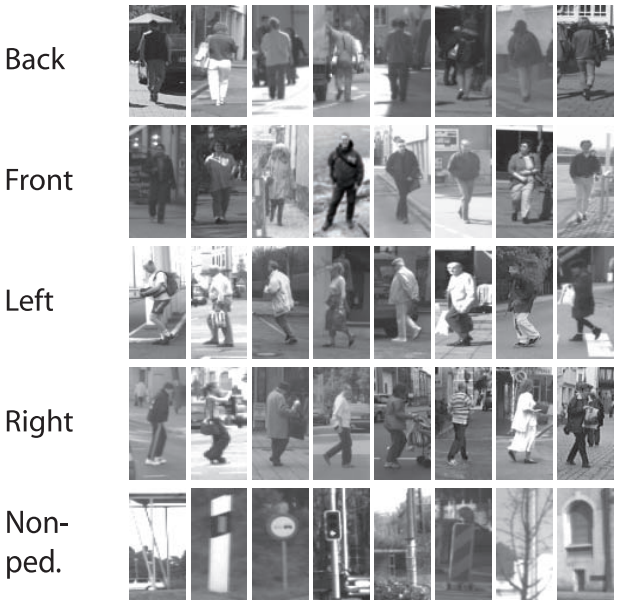


Figure 8.2: Examples of training and test data for pedestrians in four view-related clusters and non-pedestrian samples.

8.3 Experiments

8.3.1 Experimental Set-Up

The proposed integrated framework is 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) [27]. Second, we evaluate adaptive local receptive field features (5×5 pixels) in a multi-layer neural network architecture (NN/LRF) [171]. Results are expected to generalize to other pedestrian classifiers that are sufficiently complex to represent the large training sets, see Chapter 2.

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

Table 8.1: Training and test set statistics.

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 Section 6.2.4, as cluster membership weights for training:

$$z_i^k = w_k(\mathbf{x}_i) \approx P(\Psi_k|\mathbf{x}_i) \ , \ \omega_i = \omega_1 \tag{8.7}$$

To compute $w_k(\mathbf{x}_i)$, a set of 10946 shape templates corresponding to clusters Ψ_k is used. Rather than Equation (6.31), we use a manual assignment to clusters Ψ_k for pedestrian training samples, which we found to perform best in preliminary 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. [52, 111, 142], have similar requirements in terms of data labeling.

See Table 8.1 and Figure 8.2 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 recognition preprocessing step with relaxed threshold setting, i.e. containing a bias towards more "difficult" patterns. Training and test set are strictly separated: no instance of the same real-world pedestrian appears in both training and test set, similarly for the non-target samples.

8.3.2 Pedestrian Classification Performance

In our first experiment, we evaluate the classification performance of the proposed view-related mixture architecture in comparison to a single monolithic

classifier trained on the whole dataset irrespective of view, i.e. the approach of [27, 171]. Cluster priors, see Sections 6.2.4 and 8.2.1, are considered uniform. Results in terms of ROC performance are shown in Figure 8.3a. Note that this experiment is similar to the experiments in Section 6.4.1, but with different classifiers and datasets. In a qualitative manner, results are identical, cf. Figures 8.3a and 6.5.

The mixture classifiers perform better than the corresponding single classifiers. The decomposition of the problem into view-related sub-parts 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 8.3b shows the effect of adding a sample-dependent cluster prior for the test samples based on shape matching, see Sections 6.2.4 and 8.2.1. Note that the pose-based weighting for classifier training is still in place. For both HOG and LRF, only a small benefit is observed. This suggests, that the larger part of the observed benefit in Figure 8.3a comes from the use of multiple pose-specific classifiers for *training*. How exactly the fusion of those classifiers is done in *testing*, e.g. pose-based vs. uniform, seems less important.

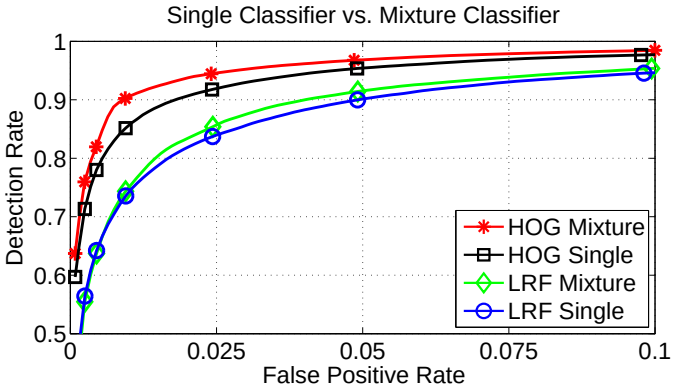
8.3.3 Pedestrian Orientation Estimation Performance

Discrete Orientation Classes

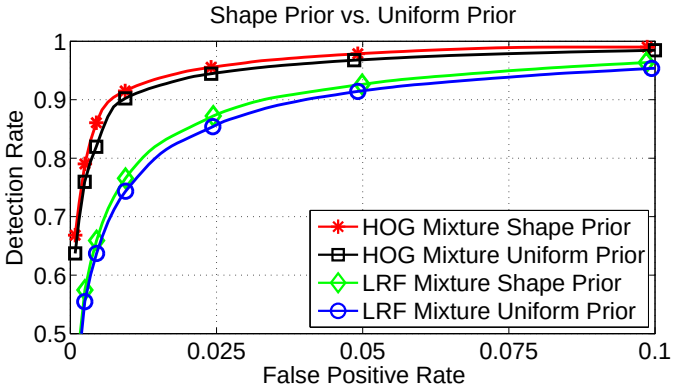
In our second experiment, we evaluate orientation estimation performance using the best performing system variant, as given in Figure 8.3: HOG mixture classifier with shape-based cluster priors. The Gaussian mixture components used to model the cluster-specific density of body orientation θ are empirically set as follows (cf. Section 8.3.1):

$$\Psi_i : \mu_i = i \cdot 90^\circ, \sigma_i = 45^\circ, \text{ for } i \in \{0, 1, 2, 3\} \tag{8.8}$$

Figure 8.4 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).

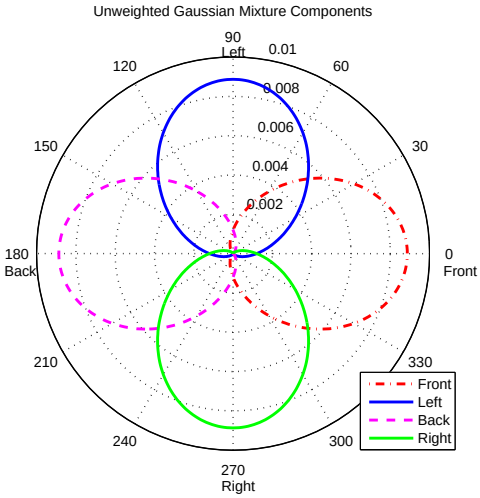


(a)

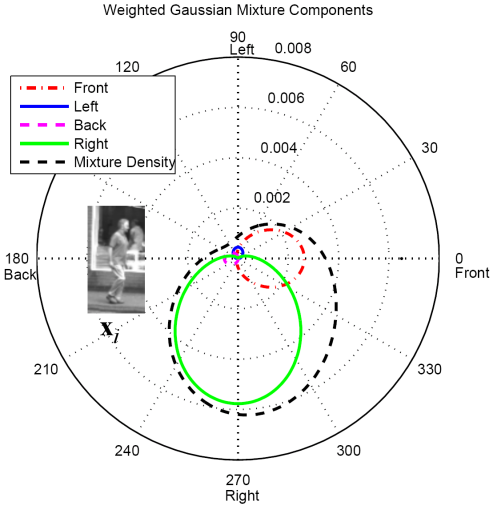


(b)

Figure 8.3: Classification performance. (a) Performance of monolithic classifiers vs. the view-related mixture architecture. (b) Benefit of shape-based priors in comparison to non-informative priors.



(a)



(b)

Figure 8.4: Visualization of 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).

In Figure 8.4a, Gaussian mixture components $g_k(\theta|\mathbf{x}_i)$, see Equation (8.4), are shown with parameters given in Equation (8.8). Figure 8.4b 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 Equation (8.4). 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 [142] which involves Haar wavelet features with a set of support vector machines in a *one vs. rest* scheme. Second, we evaluate the single-frame method of Gandhi & Trivedi [52]. 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 Equation (8.6). We additionally consider a variant of our framework involving maximum-selection over the expert classifiers, instead of the Gaussian mixture-model (GMM) formulation, cf. Section 8.2.2.

Results are given in Figure 8.5. 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 [52] and more than 35 % compared to Shimizu & Poggio [142]. Note that we use the same feature set for both our approach and for Gandhi & Trivedi [52]. 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 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 8.2. In case of left and right views, characteristic leg posture and body tilt seem to be more discriminative cues.

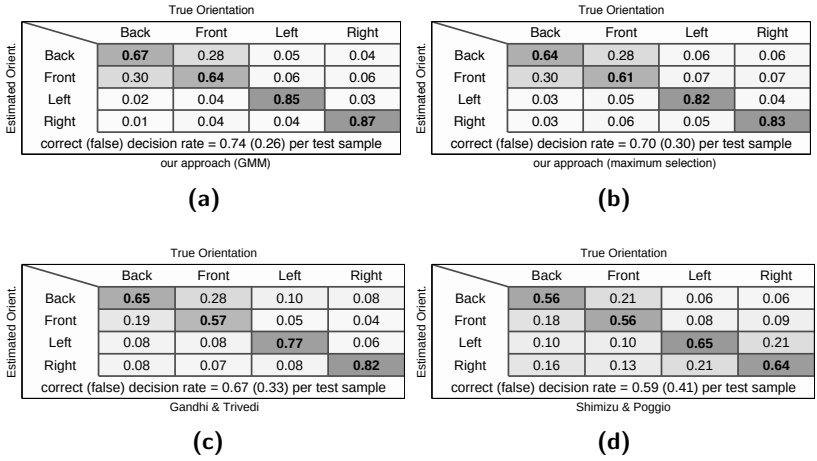


Figure 8.5: Confusion matrices and correct / false decision rate per test sample for: (a) Our approach (GMM). (b) Our approach (max. selection). (c) Gandhi & Trivedi [52]. (d) Shimizu & Poggio [142].

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 [69]. 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 Equation (8.5). 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

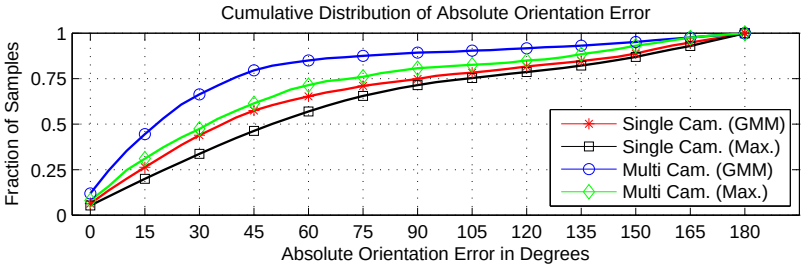


Figure 8.6: Cumulative distribution of absolute orientation error using different system variants, see text.

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 models into a single 12-component GMM model, see Section 8.2.2, 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.

Results are shown in Figure 8.6, 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 8.5. Further, no temporal filtering of the recovered orientation densities was applied, which would presumably further improve absolute performance.

8.4 Conclusion

This chapter 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.

Chapter 9

Discussion and Perspectives

The central focus of this thesis are methods for vision-based pedestrian recognition. To that extent, Chapter 4 presented an experimental study involving state-of-the-art pedestrian detectors. The results obtained serve as a performance baseline. We showed that HOG features in combination with linear support vector machines (HOG/linSVM) outperform all other approaches considered. Similar results were obtained in another more recent benchmark study [32]. In Chapter 4, we addressed the issue of sample resolution in terms of training all systems at a low- and a medium-scale resolution. The test data was identical for both cases. Our results indicate that HOG features perform best when trained on higher resolutions. The authors of [32] evaluated the performance of several detectors on different scales in the test data which includes pedestrians between 10 pixels and 256 pixels height, corresponding to a distance to the camera between 7 m and 180 m. While HOG/linSVM is not the best performing system on their whole test set irrespective of pedestrian height, a significant performance advantage of HOG/linSVM over all other systems in close-range scenarios is reported. They defined "close-range" to consist of pedestrians larger than 80 pixels which corresponds to a distance to the camera of up to 22.5 m in their set-up. They further demonstrated that the performance of all systems is closely related to the available image resolution. Pedestrian recognition performance beyond distances of 60 m is reported to be orders of magnitudes worse than in close-range settings [32]. Hence, particularly for the intelligent vehicles application where the main focus is on pre-crash collision mitigation close to the vehicle, HOG is still one of the best features available and presents a solid and challenging performance baseline for this thesis.

One of the main contributions of this thesis is the integration of information from multiple sources into the actual pedestrian classification step. Chapter 6 presented a multi-level mixture-of-experts model which combines

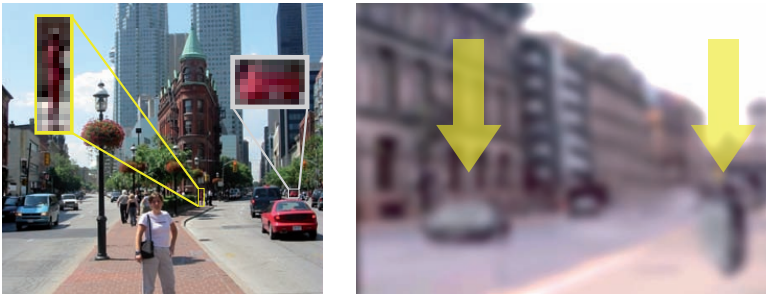
several features and image modalities. We showed that the benefit of using multiple (complementary) image modalities for classification, such as gray-level intensity, dense stereo and dense optical flow, is larger than the use of multiple feature sets in the same image modality, see Table 6.3. Our full pose-specific multi-level mixture-of-experts approach reduced the false positives of the state-of-the-art HOG/linSVM approach at the same detection rate levels by a factor of 42. We followed a mixture-of-experts strategy in terms of treating each feature/modality separately. An alternative is the construction of a joint feature/modality space at the expense of a very high-dimensional feature space. By design, the mixture-of-experts approach is less susceptive to overfitting effects in high-dimensional spaces resulting from the scarcity of training samples. However, it has the disadvantage that correlations between individual feature and modality dimensions cannot be learned. In practice, the choice between mixture-of-experts or joint space approaches has to be made in view of the actual dimensionality and number of training samples available. The more features and modalities are included, the less robust classical machine learning techniques become. The application of dimensionality reduction techniques to the joint feature space [137] can possibly alleviate the problems to some extent. Feature selection approaches, e.g. boosting [49], could also help in that regard, but they require mappings of multi-dimensional features to be used with one-dimensional weak-learners and are often plagued by practically infeasible training times on the order of weeks or months, see the discussion in Section 6.1.

Chapters 7 and 8 focused on extensions to the mixture-of-experts framework involving higher-level information such as partial occlusions or pedestrian body orientation. Although performance improvements over state-of-the-art in terms of occlusion handling and orientation recovery could be demonstrated, several open issues remain. In its current state, our model for occlusion handling, see Chapter 7, is tied to the actual layout of the decomposition of pedestrians into body parts. We chose to subdivide a pedestrian into a representation involving head, torso and leg components. This choice is motivated by observing partial occlusions in the real-world which are mostly horizontal and result from other static or moving ground-based objects in the scene, see Figure 7.5. However, in some application scenarios, such as a pedestrian stepping onto the road behind a large object, vertical occlusions can be present, e.g. parts of the pedestrian facing the road are non-occluded. Ultimately, a part-based approach that scales-up to arbitrary occlusions is de-

sired. Local deformable part approaches, i.e. [1, 46, 47, 48, 84, 92, 94, 95, 138], which build up their evidence in a bottom-up scheme, might be better suited to handle arbitrary occlusions. However, they are usually less discriminative than systems involving dense feature sets, e.g. HOG, and hence often require several verification stages. It is currently unclear, how to integrate the handling of arbitrary partial occlusions into approaches using dense feature sets for classification. Detection-by-tracking approaches, e.g. [3], which explicitly include temporal consistency of body articulations in terms of human gait and appearance, e.g. clothing, can help to recover from partial occlusions. However, such approaches require an initialization phase where the pedestrian is fully visible, to build-up the corresponding models.

Our approach to the estimation of body orientation, as outlined in Chapter 8, currently operates on single images only. A straightforward extension is the incorporation of motion information under the assumption that pedestrians usually move forward. The recovered body orientation densities could be tracked over time, assuming local orientation constancy. Initial research has been done on extracting yet higher-level models in view of object-specific (learned) motion models for enhanced path prediction, activity recognition and risk assessment, e.g. [61, 71, 144, 154, 157]. We consider this a worthwhile and promising future research direction which can make use of the proposed powerful single-frame methods for classification, partial occlusion handling and orientation estimation.

Besides the selection of features and classifiers, the quality and size of the training dataset are essential factors contributing to system performance. In Chapter 5, we proposed an automatic method to generate high informative virtual pedestrian samples that proved to be more valuable than real randomly selected pedestrian samples, most of which the classifier already "knew". We did not explicitly model the synthesis of a pedestrian but made use of statistical models describing the shape and appearance variance observed on a set of training samples. A drawback of this approach is that it can hardly extend beyond the shape and appearance variations present in the training set, e.g. as opposed to [103], where virtual pedestrians are generated by rendering an arbitrarily textured 3D pedestrian model. Further, we follow a "generate and test" approach, in that each virtually generated pedestrian is filtered by the discriminative classifier and is either included in the training set for the next round of training (if it is close to the decision boundary in classifier feature space) or discarded (if it is too far beyond the decision



(a) (b)

Figure 9.1: The influence of scene context on human object recognition. (a) The recognition of the highlighted person and car becomes much easier if scene context, i.e. the location and shape of the road, are incorporated. (b) Visually identical objects can be recognized as different objects (a car and a person in this case) when they appear in a different context.

boundary), see Section 5.4. This approach has the disadvantage, that the number of informative virtual samples which are not discarded decreases as the classifiers get better and better. Of interest would be an approach to generate informative samples more purposefully. A further extension could involve non-linear shape and texture models instead of the PCA-based models in our approach, e.g. kernel PCA [132] or manifold learning techniques, such as Isomap [153] or locally linear embedding (LLE) [133]. Most manifold learning techniques do not provide an easy generalization to discover the low-dimensional embedding for new data points, i.e. virtual samples in our case. As a result, extensions to alleviate or circumvent this so called "out-of-sample problem" are necessary, e.g. [11, 141].

In human visual processing, knowledge of scene context has a tremendous effect on object recognition [117]. Objects do typically not appear in isolation but interact actively or passively with the environment, e.g. in terms of location and scale relative to other objects. In Figure 9.1a, it is hard to recognize the person and the car in isolation. Once scene context is available - in this case, the relative position of the objects to the road - the interpretation becomes much easier. The visual system further uses context to distinguish

between similar objects. In the image shown in Figure 9.1b, most people recognize a person and a car on a road. However, both marked objects have in fact the same shape and appearance. They only differ in a 90° rotation. This highlights the significant influence of context on human object recognition and categorization. Most current artificial object recognition systems do not consider an image as a whole but operate locally on a constricted area of the image, e.g. in a sliding window approach. Hence, significant performance boosts could be expected from incorporating a model of contextual feedback on object recognition and hypotheses generation. While initial research on this topic has recently been presented, several open issues remain [30, 70].

In this work, we did not particularly focus on real-time processing time constraints, e.g. 25 Hz, 40 ms per image, and assumed that software optimization or hardware implementation would result in real-time applicability of the proposed algorithms. In case of HOG/linSVM several real-time implementations have recently been proposed [8, 18, 67, 81, 127, 169, 172].

Finally, in Section 4.1.4 we concluded that for HOG/linSVM a performance gap of about a factor of 10 exists, in case of an intelligent vehicle application with an acoustical driver warning for collisions with pedestrians. With other improvements and constraints already factored in, this improvement needed to be derived from the actual classification component. We obtained a performance boost of a factor of 42, stemming from the mixture-of-experts framework presented in Chapter 6, not taking benefits from virtual training samples (Chapter 5) into account. Collision mitigation systems with automatic emergency braking on the other hand have much higher performance constraints in terms of false activations. Here, systems using camera input as the only sensory input are still lacking the necessary performance. However, sensor fusion approaches, e.g. with radar or laser scanners, can provide that additional level of robustness and reliability which is required for viable commercial deployment.

Chapter 10

Conclusion

This thesis addressed the problem of vision-based pedestrian recognition in real-world environments using compound models involving the combination of several complementary cues, modalities and features. Multiple integration approaches, both on module-level and in terms of direct integration into the pattern classification step, were presented. Higher-level extensions involving partial occlusion handling and pedestrian body orientation estimation have been developed.

In extensive experiments on large real-world datasets, we obtained significant performance boosts over state-of-the-art for all aspects considered in this thesis, i.e. pedestrian recognition, partial occlusion handling and body orientation estimation. The pedestrian recognition performance in particular was considerably advanced; false detections at constant detection rates were reduced by significantly more than an order of magnitude compared to state-of-the-art, finally reaching performance levels that are viable for the commercial deployment of pedestrian recognition systems.

Appendix A

Publications

This thesis has led to the following publications:

Journal Publications

[Enzweiler2011a] M. Enzweiler and D. M. Gavrila. A Multi-Level Mixture-of-Experts Framework for Pedestrian Classification. *IEEE Transactions on Image Processing*, in press, 2011.

[Keller2011a] C. G. Keller, M. Enzweiler, M. Rohrbach, D. F. Llorca, C. Schnörr, and D. M. Gavrila. The Benefits of Dense Stereo for Pedestrian Recognition. *IEEE Transactions on Intelligent Transportation Systems*, in press, 2011.

[Enzweiler2009a] M. Enzweiler and D. M. Gavrila. Monocular Pedestrian Detection: Survey and Experiments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(12):2179 - 2195, 2009.

Conference Publications

[Enzweiler2010a] M. Enzweiler, A. Eigenstetter, B. Schiele, and D. M. Gavrila. Multi-Cue Pedestrian Classification With Partial Occlusion Handling. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[Enzweiler2010b] M. Enzweiler and D. M. Gavrila. Integrated Pedestrian Classification and Orientation Estimation. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[Rohrbach2009a] M. Rohrbach, M. Enzweiler, and D. M. Gavrila.

High-Level Fusion of Depth and Intensity for Pedestrian Classification. *Proc. of the DAGM Symposium on Pattern Recognition*, pages 101 - 110, 2009.

[Enzweiler2008a] M. Enzweiler and D. M. Gavrila. A Mixed Generative-Discriminative Framework for Pedestrian Classification. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[Enzweiler2008b] M. Enzweiler, P. Kanter, and D. M. Gavrila. Monocular Pedestrian Recognition Using Motion Parallax. *Proc. of the IEEE Intelligent Vehicles Symposium*, pages 792 - 797, 2008.

[Schulz2007a] W. Schulz, M. Enzweiler, and T. Ehlgen. Pedestrian Recognition from a Moving Catadioptric Camera. *Proc. of the DAGM Symposium on Pattern Recognition*, pages 456 - 465, 2007.

Bibliography

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

[2] I. P. Alonso, D. F. Llorca, M. A. Sotelo, L. M. Bergasa, P. Revenga de Toro, J. Nuevo, M. Ocana, and M. A. G. Garrido. Combination of feature extraction methods for SVM pedestrian detection. *IEEE Transactions on Intelligent Transportation Systems*, 8(2):292–307, 2007.

[3] M. Andriluka, S. Roth, and B. Schiele. People-tracking-by-detection and people-detection-by-tracking. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[4] M. Andriluka, S. Roth, and B. Schiele. Monocular 3D pose estimation and tracking by detection. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[5] Apple Inc. Aperture 3 - Image management software. `http://www.apple.com/aperture/`, 2010.

[6] S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for on-line non-linear/non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2):174–188, 2002.

[7] D. Baehring, S. Simon, W. Niehsen, and C. Stiller. Detection of close cut-in and overtaking vehicles for driver assistance based on planar parallax. *Proc. of the IEEE Intelligent Vehicles Symposium*, pages 290–295, 2005.

[8] S. Bauer, U. Brunsmann, and S. Schlotterbeck-Macht. FPGA implementation of a HOG-based pedestrian recognition system. *MPC-Workshop*, pages 49–58, 2009.

[9] A. Baumberg. Hierarchical shape fitting using an iterated linear filter. *Proc. of the British Machine Vision Conference (BMVC)*, pages 313–323, 1996.

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

[11] Y. Bengio, J.-F. Paiement, P. Vincent, O. Delalleau, N. Le Roux, and M. Ouimet. Out-of-sample extensions for LLE, Isomap, MDS, Eigenmaps, and Spectral Clustering. *Advances in Neural Information Processing Systems (NIPS)*, 2003.

[12] M. Bergtholdt, D. Cremers, and C. Schnörr. Variational segmentation with shape priors. In N. Paragios, Y. Chen, and O. Faugeras, editors, *Mathematical Models in Computer Vision: The Handbook*. Springer, 2005.

[13] D. Beymer and T. Poggio. Face recognition from one example view. *Proc. of the International Conference on Computer Vision (ICCV)*, pages 500–507, 1995.

[14] G. Borgefors. Distance transformations in digital images. *Computer Vision, Graphics and Image Processing*, 34(3):344–371, 1986.

[15] U. Brefeld, P. Geibel, and F. Wysotzki. Support vector machines with example dependent costs. *Proc. of the European Conference on Machine Learning (ECML)*, pages 23–34, 2003.

[16] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool. Robust tracking-by-detection using a detector confidence particle filter. *Proc. of the International Conference on Computer Vision (ICCV)*, 2009.

[17] A. Broggi, A. Fascioli, I. Fedriga, A. Tibaldi, and M. Del Rose. Stereo-based preprocessing for human shape localization in unstructured environments. *Proc. of the IEEE Intelligent Vehicles Symposium*, pages 410–415, 2003.

[18] T. P. Cao, G. Deng, and D. Mulligan. Implementation of real-time pedestrian detection on FPGA. *Image and Vision Computing New Zealand*, 2008.

[19] M. A. Carreira-Perpinan. Mode-finding for mixtures of Gaussian distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1318–1323, 2000.

[20] H.-P. Chiu, T. Lozano-Perez, and L. Pack Kaelbling. Virtual training for multi-view object class recognition. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[21] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002.

[22] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.

[23] T. F. Cootes, S. Marsland, C. J. Twining, K. Smith, and C. J. Taylor. Groupwise diffeomorphic non-rigid registration for automatic model building. *Proc. of the European Conference on Computer Vision (ECCV)*, pages 316–327, 2004.

[24] T. F. Cootes and C. J. Taylor. A mixture model for representing shape variation. *Image and Vision Computing*, 17(8):567–574, 1999.

[25] T. F. Cootes and C. J. Taylor. Statistical models of appearance for computer vision. Technical report, University of Manchester, 2004.

[26] C. Curio, J. Edelbrunner, T. Kalinke, C. Tzomakas, and W. von Seelen. Walking pedestrian recognition. *IEEE Transactions on Intelligent Transportation Systems*, 1(3):155–163, 2000.

[27] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 886–893, 2005.

[28] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. *Proc. of the European Conference on Computer Vision (ECCV)*, pages 428–441, 2006.

[29] J. Deutscher, A. Blake, and I. D. Reid. Articulated body motion capture by annealed particle filtering. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 126–133, 2000.

[30] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, and M. Hebert. An empirical study of context in object detection. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[31] P. Dollar, B. Babenko, S. Belongie, P. Perona, and Z. Tu. Multiple component learning for object detection. *Proc. of the European Conference on Computer Vision (ECCV)*, pages 211–224, 2008.

[32] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[33] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification (2nd ed.)*. Wiley Interscience, 2000.

[34] M. Enzweiler. Resampling techniques for pedestrian classification. Master's thesis, University of Ulm, Faculty of Computer Science, 2005.

[35] M. Enzweiler, A. Eigenstetter, B. Schiele, and D. M. Gavrila. Multi-Cue pedestrian classification with partial occlusion handling. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[36] M. Enzweiler and D. M. Gavrila. A mixed generative-discriminative framework for pedestrian classification. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[37] M. Enzweiler and D. M. Gavrila. Monocular pedestrian detection: Survey and experiments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(12):2179–2195, 2009.

[38] M. Enzweiler and D. M. Gavrila. Integrated pedestrian classification and orientation estimation. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[39] M. Enzweiler and D. M. Gavrila. A multi-level Mixture-of-Experts framework for pedestrian classification. *IEEE Transactions on Image Processing*, in press, 2011.

[40] M. Enzweiler, P. Kanter, and D. M. Gavrila. Monocular pedestrian recognition using motion parallax. *Proc. of the IEEE Intelligent Vehicles Symposium*, pages 792–797, 2008.

[41] A. Ess, B. Leibe, and L. van Gool. Depth and appearance for mobile scene analysis. *Proc. of the International Conference on Computer Vision (ICCV)*, 2007.

[42] F. J Estrada and A. D. Jepson. Benchmarking image segmentation algorithms. *International Journal of Computer Vision*, 85(2):167–181, 2009.

[43] L. Fan, K.-K. Sung, and T.-K. Ng. Pedestrian registration in static images with unconstrained background. *Pattern Recognition*, 36:1019–1029, 2003.

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

[45] B. Fardi, I. Seifert, G. Wanielik, and J. Gayko. Motion-based pedestrian recognition from a moving vehicle. *Proc. of the IEEE Intelligent Vehicles Symposium*, pages 219–224, 2006.

[46] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester. Cascade object detection with deformable part models. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[47] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:1627–1645, 2010.

[48] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55–79, 2005.

[49] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Proc. of the European Conference on Computational Learning Theory*, pages 23–37, 1995.

[50] K. Fukushima, S. Miyake, and T. Ito. Neocognitron: A neural network model for a mechanism of visual pattern recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, 13:826–834, 1983.

[51] T. Gandhi and M. M. Trivedi. Pedestrian protection systems: Issues, survey, and challenges. *IEEE Transactions on Intelligent Transportation Systems*, 8(3):413–430, 2007.

[52] T. Gandhi and M. M. Trivedi. Image based estimation of pedestrian orientation for improving path prediction. *Proc. of the IEEE Intelligent Vehicles Symposium*, pages 506–511, 2008.

[53] D. M. Gavrila. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, 73(1):82–98, 1999.

[54] D. M. Gavrila. A Bayesian, exemplar-based approach to hierarchical shape matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(8):1408–1421, 2007.

[55] D. M. Gavrila and J. Giebel. Virtual sample generation for template-based shape matching. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 676–681, 2001.

[56] D. M. Gavrila and S. Munder. Multi-Cue pedestrian detection and tracking from a moving vehicle. *International Journal of Computer Vision*, 73(1):41–59, 2007.

[57] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):643–660, 2001.

[58] D. Geronimo, A. M. Lopez, A. D. Sappa, and T. Graf. Survey on pedestrian detection for advanced driver assistance systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(7):1239–1258, 2010.

[59] J. Giebel, D. M. Gavrila, and C. Schnörr. A Bayesian framework for multi-Cue 3D object tracking. *Proc. of the European Conference on Computer Vision (ECCV)*, pages 241–252, 2004.

[60] B. E. Goldstein. *Sensation and Perception (6th ed.)*. Wadsworth, 2002.

[61] D. A. V. Govea and T. Fraichard. Motion prediction for moving objects: A statistical approach. *Proc. of the International Conference on Robotics and Automation (ICRA)*, pages 3931–3936, 2004.

[62] M. Hasenjaeger and H. Ritter. Active learning in neural networks. *New learning paradigms in soft computing*, pages 137–169, 2002.

[63] T. Heap and D. Hogg. Improving specificity in PDMs using a hierarchical approach. *Proc. of the British Machine Vision Conference (BMVC)*, pages 80–89, 1997.

[64] T. Heap and D. Hogg. Wormholes in shape space: Tracking through discontinuous changes in shape. *Proc. of the International Conference on Computer Vision (ICCV)*, pages 344–349, 1998.

[65] B. Heisele, T. Serre, M. Pontil, T. Vetter, and T. Poggio. Categorization by learning and combining object parts. *Advances in Neural Information Processing Systems (NIPS)*, pages 1239–1245, 2001.

[66] B. Heisele and C. Wöhler. Motion-based recognition of pedestrians. *Proc. of the International Conference on Pattern Recognition (ICPR)*, pages 1325–1330, 1998.

[67] M. Hiromoto and R. Miyamoto. Hardware architecture for high-accuracy real-time pedestrian detection with CoHOG features. *Proc. of The Fifth IEEE Workshop on Embedded Computer Vision*, pages 894–899, 2009.

[68] H. Hirschmüller. Stereo processing by semi-global matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):328–341, 2008.

[69] M. Hofmann and D. M. Gavrila. Multi-view 3D human upper body pose estimation combining single-frame recovery, temporal integration and model adaptation. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[70] D. Hoiem, A. A. Efros, and M. Hebert. Putting objects in perspective. *International Journal of Computer Vision*, 80(1):3–15, 2008.

[71] W. Hu, T. Tan, L. Wang, and S. Maybank. A survey on visual surveillance of object motion and behaviors. *IEEE Transactions on Systems, Man, and Cybernetics*, 34:334–352, 2004.

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

[73] INRIA Person Dataset. `http://pascal.inrialpes.fr/data/human/`, 2007.

[74] M. Isard and A. Blake. CONDENSATION - Conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1):5–28, 1998.

[75] M. Isard and A. Blake. ICONDENSATION: Unifying low-level and high-level tracking in a stochastic framework. *Proc. of the International Conference on Computer Vision (ICCV)*, pages 893–908, 1998.

[76] M. Isard and J. MacCormick. BraMBLE: A Bayesian multiple-blob tracker. *Proc. of the International Conference on Computer Vision (ICCV)*, pages 34–41, 2001.

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

[78] A. K. Jain, R. P. W. Duin, and J. Mao. Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37, 2000.

[79] E. Jones and S. Soatto. Layered active appearance models. *Proc. of the International Conference on Computer Vision (ICCV)*, pages 1097–1102, 2005.

[80] M. J. Jones and T. Poggio. Multidimensional morphable models. *Proc. of the International Conference on Computer Vision (ICCV)*, pages 683–688, 1998.

[81] R. Kadota, H. Sugano, M. Hiromoto, H. Ochi, R. Miyamoto, and Y. Nakamura. Hardware architecture for HOG feature extraction. *Proc. of the Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pages 1330–1333, 2009.

[82] H. Kang and D. Kim. Real-time multiple people tracking using competitive condensation. *Pattern Recognition*, 38(7):1045–1058, 2005.

[83] A. Kapoor, K. Grauman, R. Urtasun, and T. Darrell. Active learning with Gaussian processes for object categorization. *Proc. of the International Conference on Computer Vision (ICCV)*, 2007.

[84] L. Karlinsky, M. Dinerstein, D. Harari, and S. Ullman. The chains model for detecting parts by their context. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[85] C. Keller, M. Enzweiler, M. Rohrbach, D. F. Llorca, C. Schnörr, and D. M. Gavrila. The benefits of dense stereo for pedestrian recognition. *IEEE Transactions on Intelligent Transportation Systems*, in press, 2011.

[86] Z. Khan, T. Balch, and F. Dellaert. MCMC-based particle filtering for tracking a variable number of interacting targets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(11):1805–1819, 2005.

[87] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.

[88] 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. *The Journal of Neuroscience*, 23(4):1451–1463, 2003.

[89] H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97, 1955.

[90] J. A. Lasserre, C. M. Bishop, and T. P. Minka. Principled hybrids of generative and discriminative models. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[91] S. Lee, Y. Liu, and R. Collins. Shape variation-based frieze pattern for robust gait recognition. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[92] B. Leibe, N. Cornelis, K. Cornelis, and L. Van Gool. Dynamic 3D scene analysis from a moving vehicle. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[93] B. Leibe, A. Leonardis, and B. Schiele. Robust object detection with interleaved categorization and segmentation. *International Journal of Computer VisionSpecial Issue on Learning for Recognition and Recognition for Learning*, 77(1-3):259–289, 2008.

[94] B. Leibe, K. Schindler, N. Cornelis, and L. Van Gool. Coupled object detection and tracking from static cameras and moving vehicles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(10):1683–1698, 2008.

[95] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 878–885, 2005.

[96] M. Li and I. K. Sethi. Confidence-based active learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(8):1251–1261, 2006.

[97] R. Lienhart and J. Maydt. An extended set of Haar-like features for rapid object detection. *Proc. of the International Conference on Image Processing (ICIP)*, pages 900–903, 2002.

[98] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[99] J. MacCormick and A. Blake. Partitioned sampling, articulated objects and interface-quality hand tracking. *Proc. of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2000.

[100] J. MacCormick and A. Blake. A probabilistic exclusion principle for tracking multiple objects. *International Journal of Computer Vision*, 39(1):57–71, 2000.

[101] M. Mählisch, M. Oberländer, O. Löhlein, D. M. Gavrila, and W. Ritter. A multiple detector approach to low-resolution FIR pedestrian recognition. *Proc. of the IEEE Intelligent Vehicles Symposium*, 2005.

[102] S. Maji, A. C. Berg, and J. Malik. Classification using intersection kernel support vector machines is efficient. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[103] J. Marin, D. Vazquez, D. Geronimo, and A. M. Lopez. Learning appearance in virtual scenarios for pedestrian detection. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[104] A. McCallum, C. Pal, G. Druck, and X. Wang. Multi-conditional learning: Generative/discriminative training for clustering and classification. *Proc. of the AAAI Conference on Artificial Intelligence*, 2006.

[105] K. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust part detectors. *Proc. of the European Conference on Computer Vision (ECCV)*, pages 69–81, 2004.

[106] MIT CBCL Pedestrian Database. `http://cbcl.mit.edu/cbcl/software-datasets/PedestrianData.html`.

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

[108] A. Mohan, C. Papageorgiou, and T. Poggio. Example-based object detection in images by components. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(4):349–361, 2001.

[109] S. Munder and D. M. Gavrila. An experimental study on pedestrian classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(11):1863–1868, 2006.

[110] S. Munder, C. Schnörr, and D. M. Gavrila. Pedestrian detection and tracking using a mixture of view-based shape-texture models. *IEEE Transactions on Intelligent Transportation Systems*, 9(2):333–343, 2008.

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

[112] S. Nedevschi, S. Bota, and C. Tomiuc. Stereo-based pedestrian detection for collision-avoidance applications. *IEEE Transactions on Intelligent Transportation Systems*, 10(3):380–391, 2009.

[113] S. Nissen. Implementation of a fast artificial neural network library (FANN). Technical report, Department of Computer Science, University of Copenhagen, Denmark, 2003.

[114] P. Niyogi, F. Girosi, and T. Poggio. Incorporating prior information in machine learning by creating virtual examples. *IEEE Proceedings on Intelligent Signal Processing*, pages 2196–2209, 1998.

[115] T. Ojala, M. Pietikäinen, and D. Harwood. A comparative study of texture measures with classification based on feature distributions. *Pattern Recognition*, 29:51–59, 1996.

[116] K. Okuma, A. Taleghani, N. de Freitas, J. Little, and D. Lowe. A boosted particle filter: Multitarget detection and tracking. *Proc. of the European Conference on Computer Vision (ECCV)*, pages 28–39, 2004.

[117] A. Oliva and A. Torralba. The role of context in object recognition. *Trends in Cognitive Sciences*, 11(12):520–527, 2007.

[118] L. Oliveira, U. Nunes, and P. Peixoto. On exploration of classifier ensemble synergism in pedestrian detection. *IEEE Transactions on Intelligent Transportation Systems*, 11(1):16–27, 2010.

[119] OpenCV Library. `http://opencv.willowgarage.com/`, 2010.

[120] C. Papageorgiou and T. Poggio. A trainable system for object detection. *International Journal of Computer Vision*, 38:15–33, 2000.

[121] PETS Datasets. `http://www.cvg.rdg.ac.uk/slides/pets.html`, 2007.

[122] V. Philomin, R. Duraiswami, and L. S. Davis. Quasi-random sampling for condensation. *Proc. of the European Conference on Computer Vision (ECCV)*, pages 134–149, 2000.

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

[124] R. Polana and R. Nelson. Low-level recognition of human motion. *IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, pages 77–92, 1994.

[125] D. Pomerleau. Neural network vision for robot driving. In *The Handbook of Brain Theory and Neural Networks*. M. Arbib, ed., 1995.

[126] R. Poppe. Vision-based human motion analysis: An overview. *Computer Vision and Image Understanding*, 108:4–18, 2007.

[127] V. Prisacariu and I. Reid. FastHOG - a real-time GPU implementation of HOG. Technical Report 2310/09, Department of Engineering Science, Oxford University, 2009.

[128] D. Ramanan, A. D. Forsyth, and A. Zisserman. Strike a pose: Tracking people by finding stylized poses. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 271– 278, 2005.

[129] T. Randen and J. H. Husøy. Filtering for texture classification: A comparative study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(4):291–310, 1999.

[130] M. Rapus, S. Munder, G. Baratoff, and J. Denzler. Pedestrian recognition using combined low-resolution depth and intensity images. *Proc. of the IEEE Intelligent Vehicles Symposium*, 2008.

[131] M. Rohrbach, M. Enzweiler, and D. M. Gavrila. High-level fusion of depth and intensity for pedestrian classification. *Proc. of the DAGM Symposium on Pattern Recognition*, pages 101–110, 2009.

[132] S. Romdhani, S. Gong, and A. Psarrou. A multi-view nonlinear active shape model using kernel PCA. *Proc. of the British Machine Vision Conference (BMVC)*, pages 483–492, 1999.

[133] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[134] P. Sabzmeydani and G. Mori. Detecting pedestrians by learning shapelet features. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[135] W. Schulz, M. Enzweiler, and T. Ehlgen. Pedestrian recognition from a moving catadioptric camera. *Proc. of the DAGM Symposium on Pattern Recognition*, pages 456–465, 2007.

[136] J. Schürmann. *Pattern Classification: A Unified View of Statistical and Neural Approaches*. Wiley Interscience, 1996.

[137] W. R. Schwartz, A. Kembhavi, D. Harwood, and L. S. Davis. Human detection using partial least squares analysis. *Proc. of the International Conference on Computer Vision (ICCV)*, 2009.

[138] E. Seemann, M. Fritz, and B. Schiele. Towards robust pedestrian detection in crowded image sequences. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

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

[140] V. D. Shet, J. Neumann, V. Ramesh, and L. S. Davis. Bilattice-based logical reasoning for human detection. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[141] L.-K. Shi, P.-L. He, B. Liu, K. Fu, and Q. Wu. A robust generalization of isomap for new data. *Proc. of the International Conference on Machine Learning and Cybernetics*, 3:1707–1712, 2005.

[142] H. Shimizu and T. Poggio. Direction estimation of pedestrian from multiple still images. *Proc. of the IEEE Intelligent Vehicles Symposium*, pages 596–600, 2004.

[143] H. Sidenbladh and M. J. Black. Learning the statistics of people in images and video. *International Journal of Computer Vision*, 54(1/2/3):183–209, 2003.

[144] H. Sidenbladh, M. J. Black, and L. Sigal. Implicit probabilistic models of human motion for synthesis and tracking. *Proc. of the European Conference on Computer Vision (ECCV)*, pages 784–800, 2002.

[145] L. Sigal and M. J. Black. Measure locally, reason globally: Occlusion-sensitive articulated pose estimation. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2041–2048, 2006.

[146] P. D. Sozou, T. F. Cootes, C. J. Taylor, and E. C. Di-Mauro. A non-linear generalisation of PDMs using polynomial regression. *Proc. of the British Machine Vision Conference (BMVC)*, pages 397–406, 1994.

[147] M. Spengler and B. Schiele. Towards robust multi-cue integration for visual tracking. *Machine Vision and Applications*, 14(1):50–58, 2003.

[148] F. Stein. Efficient computation of optical flow using the census transform. *Proc. of the DAGM Symposium on Pattern Recognition*, pages 79–86, 2004.

[149] B. Stenger, A. Thayananthan, P. H. S. Torr, and R. Cipolla. Model-based hand tracking using a hierarchical Bayesian filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(9):1372–1385, 2006.

[150] K.-K. Sung and T. Poggio. Example-based learning for view-based human face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):39–51, 1995.

[151] M. Szarvas, A. Yoshizawa, M. Yamamoto, and J. Ogata. Pedestrian detection with convolutional neural networks. *Proc. of the IEEE Intelligent Vehicles Symposium*, pages 223–228, 2005.

[152] L. Taycher, G. Shakhnarovich, D. Demirdjian, and T. Darrell. Conditional random people: Tracking humans with CRFs and grid filters. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 222–229, 2006.

[153] J. B. Tenenbaum, V. de Silva, and H. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[154] N. Tiemann, W. Branz, and D. Schramm. Predictive pedestrian protection - Sensor requirements and risk assessment. *International Technical Conference on the Enhanced Safety of Vehicles*, 2009.

[155] K. Toyama and A. Blake. Probabilistic tracking with exemplars in a metric space. *International Journal of Computer Vision*, 48(1):9–19, 2002.

[156] Z. Tu. Learning generative models via discriminative approaches. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[157] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea. Machine recognition of human activities: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(11):1473–1488, 2008.

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

[159] O. Tuzel, F. Porikli, and P. Meer. Human detection via classification on Riemannian manifolds. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[160] I. Ulusoy and C. M. Bishop. Generative versus discriminative methods for object recognition. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 258–265, 2005.

[161] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

[162] A. Vedaldi, P. Favaro, and E. Grisan. Boosting invariance and efficiency in supervised learning. *Proc. of the International Conference on Computer Vision (ICCV)*, 2007.

[163] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman. Multiple kernels for object detection. *Proc. of the International Conference on Computer Vision (ICCV)*, 2009.

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

[165] S. Walk, N. Majer, K. Schindler, and B. Schiele. New features and insights for pedestrian detection. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[166] S. Walk, K. Schindler, and B. Schiele. Disparity statistics for pedestrian detection: Combining appearance, motion and stereo. *Proc. of the European Conference on Computer Vision (ECCV)*, 2010.

[167] X. Wang, T.X. Han, and S. Yan. A HOG-LBP human detector with partial occlusion handling. *Proc. of the International Conference on Computer Vision (ICCV)*, 2009.

[168] A. Wedel, D. Cremers, T. Pock, and H. Bischof. Structure- and motion-adaptive regularization for high accuracy optic flow. *Proc. of the International Conference on Computer Vision (ICCV)*, 2009.

[169] T. Wilson, M. Glatz, and M. Hödlmoser. Pedestrian detection implemented on a fixed-point parallel architecture. *Proc. of the 13th IEEE International Symposium on Consumer Electronics*, pages 47–51, 2009.

[170] C. Wöhler and J. K. Anlauf. An adaptable time-delay neural-network algorithm for image sequence analysis. *IEEE Transactions on Neural Networks*, 10(6):1531–1536, 1999.

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

[172] C. Wojek, G. Dorko, A. Schulz, and B. Schiele. Sliding-windows for rapid object class localization: A parallel technique. *Proc. of the DAGM Symposium on Pattern Recognition*, pages 71–81, 2008.

[173] C. Wojek, S. Walk, and B. Schiele. Multi-cue onboard pedestrian detection. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

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

[175] B. Wu and R. Nevatia. Optimizing discrimination-efficiency tradeoff in integrating heterogeneous local features for object detection. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[176] Y. Wu and T. Yu. A field model for human detection and tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(5):753–765, 2006.

[177] K. Zapien, J. Fehr, and H. Burkhardt. Fast support vector machine classification using linear SVMs. *Proc. of the International Conference on Pattern Recognition (ICPR)*, pages 366–369, 2006.

[178] D.-Q. Zhang and S.-F. Chang. A generative-discriminative hybrid method for multi-view object detection. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[179] L. Zhang, B. Wu, and R. Nevatia. Detection and tracking of multiple humans with extensive pose articulation. *Proc. of the International Conference on Computer Vision (ICCV)*, 2007.

[180] L. Zhao and C. Thorpe. Stereo and neural network-based pedestrian detection. *IEEE Transactions on Intelligent Transportation Systems*, 1(3):148–154, 2000.

[181] T. Zhao and R. Nevatia. Tracking multiple humans in complex situations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1208–1221, 2004.

[182] Q. Zhu, S. Avidan, M. Ye, and K.-T. Cheng. Fast human detection using a cascade of histograms of oriented gradients. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1491–1498, 2006.