A Mixed Generative-Discriminative Framework for Pedestrian Classification

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

[1] Image & Pattern Analysis Group, Dept. of Math. and Comp. Sc., Univ. of Heidelberg, Germany
[2] Environment Perception, Group Research, Daimler AG, Ulm, Germany
[3] Intelligent Systems Lab, Faculty of Science, Univ. of Amsterdam, The Netherlands
{uni-heidelberg.enzweiler,dariu.gavrila}@daimler.com

Abstract

This paper 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 captures prior knowledge about the pedestrian class 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.

In large-scale experiments on real-world datasets of tens of thousands of samples, we demonstrate a significant improvement in classification performance of the combined generative-discriminative approach over the discriminative-only approach (the latter exemplified by a neural network with local receptive fields and a support vector machine using Haar wavelet features).

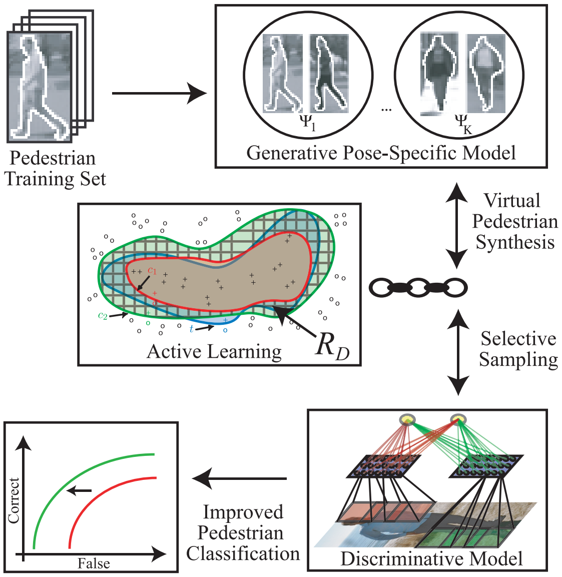

Figure 1. Framework overview. Utilizing the synthesized samples of a learned generative model to enhance the classification performance of a discriminative model.

1. Introduction

The ability to automatically detect pedestrians in images is key for a number of application domains such as surveillance and intelligent vehicles. Large variations in pedestrian appearance (e.g. clothing, pose) and environmental conditions (e.g. lighting, background) make this problem particularly challenging. A typical approach starts by identifying regions of interest in the image using a computationally efficient method (e.g. background subtraction, motion detection, obstacle detection) and thereafter moves on to a more expensive pattern classification step [8, 23, 31]. In this paper, we focus on the classification step, see also [5, 18].

Recently, an experimental study on pedestrian classification investigated the combination of several state-of-the-art features and classifiers [21]. Some combinations performed better than others, but interestingly, the benefit obtained by selecting the best combination was less pronounced than the gain obtained by increasing the training set (even though the latter was already quite large, involving many thousands of training samples). Methods to acquire additional training samples of the non-target class are commonly used [21, 23, 27], although it was observed that the performance tends to saturate fairly quickly, after a few bootstrapping iterations. The enlargement of the training set with respect to the target class yielded more gain [21], but this usually requires time consuming (and thus costly) manual labeling.

This paper proposes a novel combined generative-discriminative approach 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

The authors acknowledge the support of the *Studienstiftung des deutschen Volkes (German National Academic Foundation).*

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

to synthesize virtual samples of the target class, thus enlarging 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. The paper proposes the use of selective sampling, by means of probabilistic active learning, to guide the training process towards the most informative samples. See Figure 1.

The general idea is independent of the particular generative and discriminative model used, and can in principle extend to other object classes than pedestrians. In this paper, 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. 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.

2. Previous Work

A sizeable body of literature exists on the acquisition of object models from training data. Roughly, these models can be categorized into *generative* and *discriminative* models, see [29]. Generative models capture the (unknown) data generation process by representing the appearance of an object class in terms of its class-conditional density function. Using associated prior distributions, posterior probabilities for object classification can be inferred using Bayes rule. In contrast, discriminative models directly approximate posterior probabilities by learning the parameters of a discriminant function (decision boundary) between object classes. Regarding object classification, discriminative models are typically faster and more robust with regard to the prediction of class labels. On the other hand, generative models can handle partially labeled data and allow to synthesize new examples [29].

A main representative of generative models are linear shape models [3], learned from a set of training shapes, given the existence of an appropriate registration method. Extensions to cover non-rigid shape deformations involved using a single global shape model in combination with non-linear PCA techniques [25, 26] or a probabilistic model [4], as well as the use of multiple local-linear shape models [7, 11]. Other work considered joint linear shape and texture models [3, 6, 15] to capture underlying dependencies in a principled way. For this, texture models are derived from shape-normalized examples. The downside of the joint shape-texture approach is that it requires data normalization procedures and that it results in a significantly increased feature space dimensionality. Recently, layered shape-texture models have been introduced to add additional robustness against missing features and substantial spatial rearrangement of object parts [14]. See Table 1.

Authors	**Shape Model**	**Texture Model**	**Sample Plausibility**
Cootes et al. [3] Fan et al. [6] Jones et al. [15]	global linear (PCA)	global linear (PCA)	limit on deviation from mean
Jones et al. [14]	multi-layer global linear (weighted PCA)	multi-layer global linear (weighted PCA)	limit on deviation from mean
Gavrila et al. [7] Heap et al. [11]	pose-specific linear (PCA)	-	limit on deviation from mean
Romdhani et al. [25]	global non-linear (Kernel PCA)	-	limit on deviation from mean
Sozou et al. [26]	global non-linear (polynomial regression)	-	limit on deviation from mean
Cootes et al. [4]	global linear (PCA)	-	probabilistic (GMM)
current paper	**pose-specific linear (PCA)**	**pose-specific linear (PCA), decomposed**	**probabilistic (KDE)**

Table 1. Overview of existing and proposed generative shape and texture models.

While generative models implicitly establish a feature-space tailored to the object class under consideration, discriminative approaches to object classification involve a combination of a feature extraction method, e.g. local receptive fields, Haar wavelets, and a pattern classification technique, e.g. neural networks, support vector machines, [13, 21].

Several techniques which combine generative and discriminative models have recently been proposed [17, 20, 28, 33]. Discriminative models have been employed to learn a generative model in an iterative fashion [28]. 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 [20] and probabilistic [17] weighting schemes. Further, likelihood ratios in generative models have been replaced by more powerful discriminative models [33].

Aside from the particular models used, incorporating prior knowledge about the target class has been suggested as a way out of the bias-variance dilemma to increase robustness [22]. Prior knowledge can be both incorporated directly into the error function of a discriminative model (vicinal risk minimization) [30] and during training in terms of enlarging the training set with additional samples [21, 23, 24, 27, 30, 31]. While virtual samples of the non-target class can be easily collected using bootstrapping [21, 23, 27, 31], 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 [1, 2, 9]. Others utilize explicit 3D models

[12]. 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 [21, 24, 27, 30].

Finally, relevant to current work are sampling techniques which assess the information content of training samples and select the most informative training examples by bootstrapping [21, 23, 27, 31] or active learning [10, 16, 19].

The contributions of this paper are twofold. We consider our main contribution to be the novel framework illustrated in Figure 1. A learned generative model is used to enhance the performance of a discriminative model in terms of synthesizing virtual training samples combined with active learning. This is quite unlike previous combination strategies for generative and discriminative models [17, 20, 28, 33] and unlike previous applications of active learning. We neither require controlled data acquisition [1, 2, 9], nor do we have 3D models [12] to our disposition. At the same time, we go beyond the synthesis of samples based on simple transformations [21, 24, 27, 30] and take into account sample probabilities.

A secondary contribution concerns the generative pedestrian model proposed, see Table 1. Similar to [7, 11], 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 [7, 11] 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. Generative Pedestrian Model

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 [7] 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 1. 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 $\mathbf{s}_{k,i}$.

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.

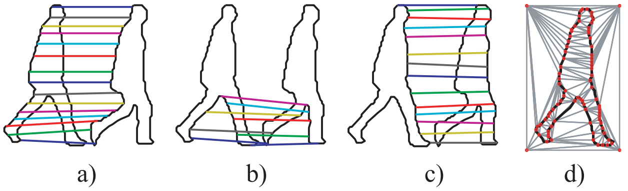

Figure 2. Shape registration. a) - c) automatically determined contour point correspondences, d) Delaunay triangulation

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

Shape Model-Space As a result of shape registration, [7], 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 2a-c. PCA is applied to the shape space to obtain a compact representation utilizing $N_{\mathbf{s}_k}$ dimensions (e.g. to model 95% of the total variance). The parametric representation $\mathbf{m}_{\mathbf{s}_{k,i}}$ of a pedestrian shape $\mathbf{s}_{k,i}$ in terms of shape model coordinates is given by

$$\mathbf{m}_{\mathbf{s}_{k,i}} = \mathbf{\Phi}_{\mathbf{s}_k}^T \left(\mathbf{s}_{k,i} - \bar{\mathbf{s}}_k \right). \tag{1}$$

Here, $\bar{\mathbf{s}}_k$ denotes the mean shape within Ψ_k, and $\mathbf{\Phi}_{\mathbf{s}_k}$ is a matrix containing $N_{\mathbf{s}_k}$ eigenvectors in its columns.

Foreground Texture Model-Space 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 3. 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 (Figure 2):

$$\hat{\mathbf{t}}_{k,i} = W_{\mathbf{s}_{k,i}}(\mathbf{t}_{k,i}) \tag{2}$$

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

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

$$\mathbf{m}_{\hat{\mathbf{t}}_{k,i}} = \mathbf{\Phi}_{\hat{\mathbf{t}}_k}^T \left(\hat{\mathbf{t}}_{k,i} - \bar{\hat{\mathbf{t}}}_k \right) \tag{3}$$

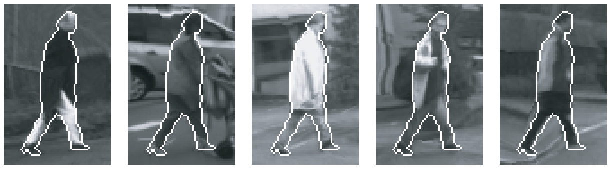

Figure 3. Shape-normalized examples for a pose-specific subspace.

Figure 4 depicts the mean texture along with the first four eigenvectors for a pose-specific texture model.

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 sizeable 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 $N_{\hat{\mathbf{t}}_k}$-dimensional texture model-space obtained by PCA into two subspaces. The first subspace represents coarse texture components (e.g. modeling overall appearance of clothing parts such as trousers and coat). Its dimensionality $N'_{\hat{\mathbf{t}}_k}$ 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 subspace captures fine texture components. Here no *pdf* estimation takes place, for synthesis (see Section 4) the associated entries are derived from particular training samples.

Hence, the parametric model-space representation $\mathbf{m}_{\hat{\mathbf{t}}_{k,i}}$ (cf. Eq. (3)) of a shape-normalized texture vector $\hat{\mathbf{t}}_{k,i}$ is decomposed into:

$$\mathbf{m}_{\hat{\mathbf{t}}_{k,i}} = \left(\mathbf{m}'_{\hat{\mathbf{t}}_{k,i}}, \mathbf{m}''_{\hat{\mathbf{t}}_{k,i}}\right) \tag{4}$$

with

$$\mathbf{m}'_{\hat{\mathbf{t}}_{k,i}} = \left(\mathbf{m}_{\hat{\mathbf{t}}_{k,i},1}, \ldots, \mathbf{m}_{\hat{\mathbf{t}}_{k,i},N'_{\hat{\mathbf{t}}_k}}\right) \tag{5}$$

$$\mathbf{m}''_{\hat{\mathbf{t}}_{k,i}} = \left(\mathbf{m}_{\hat{\mathbf{t}}_{k,i},N'_{\hat{\mathbf{t}}_k}+1}, \ldots, \mathbf{m}_{\hat{\mathbf{t}}_{k,i},N_{\hat{\mathbf{t}}_k}}\right) \tag{6}$$

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-

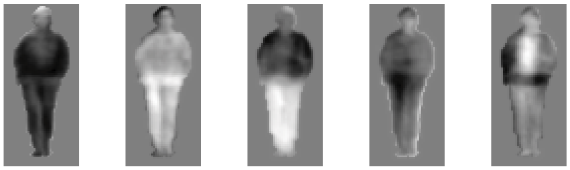

Figure 4. Mean texture and eigenvectors for a pose-specific texture model (background masked out).

conditional densities is given by:

$$p_k\left(\mathbf{m}|\omega_0\right) = \frac{1}{N_k}\sum_{n=1}^{N_k}\frac{1}{det(H)}\mathcal{K}\left\{H^{-1}(\mathbf{m} - \mathbf{m}_n)\right\} \tag{7}$$

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 [13], 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 [3, 6, 15], 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.

4. 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 variation of three components: shape, foreground texture and background texture. See Figure 5 for an overview.

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 $p_{\mathbf{s}_k}\left(\mathbf{m}_{\mathbf{s}_k}|\omega_0\right)$:

$$\mathbf{m}^*_{\mathbf{s}_{k,j}} \sim p_{\mathbf{s}_k}\left(\mathbf{m}_{\mathbf{s}_k}|\omega_0\right) \tag{8}$$

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:

$$\mathbf{s}^*_{k,j} = \bar{\mathbf{s}}_k + \mathbf{\Phi}_{\mathbf{s}_k}\mathbf{m}^*_{\mathbf{s}_{k,j}} \tag{9}$$

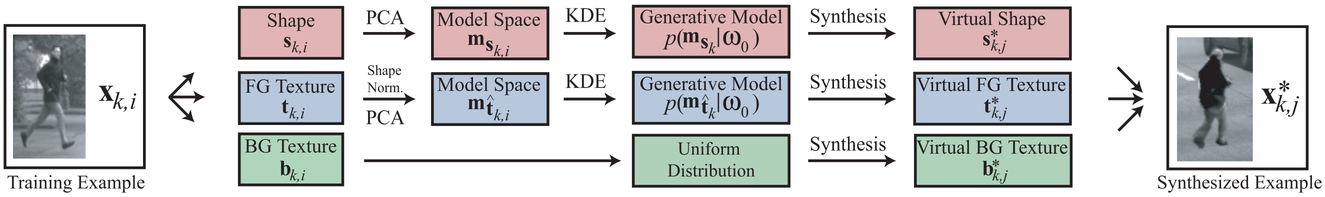
Training Example
$\mathbf{x}_{k,i}$
Shape $\mathbf{s}_{k,i}$
FG Texture $\mathbf{t}_{k,i}$
BG Texture $\mathbf{b}_{k,i}$
PCA
Shape Norm.
PCA
Model Space $\mathbf{m}_{\mathbf{s}_{k,i}}$
Model Space $\mathbf{m}_{\hat{\mathbf{t}}_{k,i}}$
KDE
KDE
Generative Model $p(\mathbf{m}_{\mathbf{s}_k}|\omega_0)$
Generative Model $p(\mathbf{m}_{\hat{\mathbf{t}}_k}|\omega_0)$
Uniform Distribution
Synthesis
Synthesis
Synthesis
Virtual Shape $\mathbf{s}^*_{k,j}$
Virtual FG Texture $\mathbf{t}^*_{k,j}$
Virtual BG Texture $\mathbf{b}^*_{k,j}$
$\mathbf{x}^*_{k,j}$
Synthesized Example

Figure 5. 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 re-sampled to create virtual pedestrian samples.

The virtual shape $\mathbf{s}^*_{k,j}$ is utilized to warp an existing pedestrian example into a new shape, as shown in Figure 6b.

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 3. 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 4 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 $p_{\hat{\mathbf{t}}_k}\left(\mathbf{m}'_{\hat{\mathbf{t}}_k}|\omega_0\right)$, by uniformly selecting the j-th example in model-space and sampling from the local kernel:

$$\mathbf{m}'^*_{\hat{\mathbf{t}}_{k,j}} \;\sim\; p_{\hat{\mathbf{t}}_k}\left(\mathbf{m}'_{\hat{\mathbf{t}}_k}|\omega_0\right) \tag{10}$$

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 $p_{\hat{\mathbf{t}}_k}\left(\mathbf{m}'^*_{\hat{\mathbf{t}}_{k,j}}|\omega_0\right) > c_{\hat{\mathbf{t}}_k}$. 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. Eq. (4)):

$$\mathbf{m}^*_{\hat{\mathbf{t}}_{k,j}} = \left(\mathbf{m}'^*_{\hat{\mathbf{t}}_{k,j}}, \mathbf{m}''_{\hat{\mathbf{t}}_{k,j}}\right) \tag{11}$$

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

$$\hat{\mathbf{t}}^*_{k,j} = \bar{\hat{\mathbf{t}}}_k + \mathbf{\Phi}_{\hat{\mathbf{t}}_k}\mathbf{m}^*_{\hat{\mathbf{t}}_{k,j}} \tag{12}$$

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 (cf. Eq. (2)):

$$\mathbf{t}^*_{k,j} = W^{-1}_{\mathbf{s}^*_{k,j}}\left(\hat{\mathbf{t}}^*_{k,j}\right) \tag{13}$$

An example of this technique is depicted in Figure 6c-e. Note how fine-scale details, e.g. the internal contour of the right arm (Figure 6c-e, first row) are preserved, while the overall texture exhibits sensible variations.

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:

$$\mathbf{b}^*_{k,j} \;\sim\; \mathcal{U}(\mathcal{B}) \tag{14}$$

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

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 $p_{\hat{\mathbf{t}}_k}\left(\mathbf{m}'_{\hat{\mathbf{t}}_k}|\omega_0\right)$ (cf. Eqs. (10)-(13)) 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. Eqs. (8)-(9)). Finally, background $\mathbf{b}^*_{k,j}$ is sampled from the the non-parametric background model (cf. Eq. (14)) and a virtual pedestrian example $\mathbf{x}^*_{k,j}$ is obtained by compositing the textured pedestrian shape over the background, see Figure 6:

$$\mathbf{x}^*_{k,j} = \left(\mathbf{s}^*_{k,j}, \mathbf{t}^*_{k,j}\right) \;\oplus\; \mathbf{b}^*_{k,j} \tag{15}$$

5. Probabilistic Selective Sampling

A probabilistic least-certain querying scheme, an instance of an active learning algorithm [10, 16, 19], 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

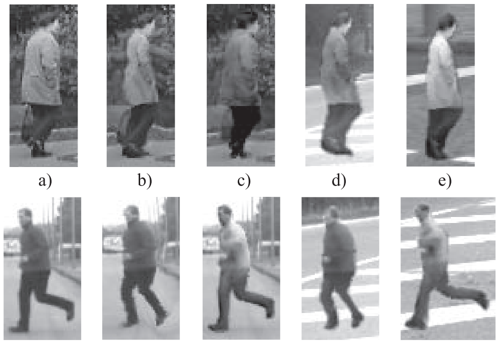

Figure 6. Example of virtual pedestrian synthesis. a) original pedestrian examples, b) shape variation, c) foreground texture variation, d) - e) joint variation of shape, foreground and background texture

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, there exists a region of uncertainty R_D, where the classification result is not unambiguously defined (see the hatched area in Figure 1, Active Learning). 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

$$P(error|\mathbf{x}_i) = \min\left\{P(\omega_0|\mathbf{x}_i), P(\omega_1|\mathbf{x}_i)\right\}. \tag{16}$$

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:

$$0 \leq |P(\omega_0|\mathbf{x}_i) - P(\omega_1|\mathbf{x}_i)| \leq 1 - 2\Theta \tag{17}$$

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 [13, 16, 19]. We use this in our experiments.

	Pedestrians (labeled)	Pedestrians (jittered)	Non-Pedestrians
Init. Train Set	10946	43784	82698
Test Set	13971	251478	133813

Table 2. Training and test set statistics.

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 Eq. (17). Finally, the discriminative model is retrained on the joint dataset $D_{l+1} = D_l \cup \widehat{D}_l^*$.

6. Experiments

The proposed generative-discriminative framework was 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 (Section 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 considered two diverse instances: a neural network with local receptive fields of size 5×5 pixels (NN/LRF) [32] and a linear[1] support vector machine using Haar wavelet features at scales of 4×4 and 8×8 pixels (Haar SVM) [23]. Results are expected to generalize to other pedestrian classifiers that are sufficiently complex to represent the large training datasets e.g. [5, 13, 18].

See Table 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 detection pre-processing step with relaxed threshold setting, i.e. containing a bias towards more ”difficult” patterns, similar to [21]. Training and test set were strictly separated: no instance of the same real-world pedestrian appears in both training and test set, similarly for the non-target samples. See Figure 7 for some examples of the dataset. Discriminative models trained on this dataset are referred to as *base classifiers*.

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

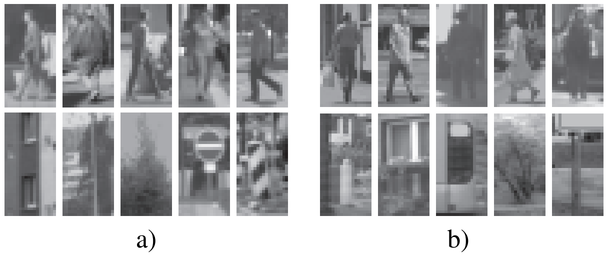

Figure 7. Dataset overview. a) training set examples, b) test set examples. Top and bottom rows show target and non-target samples, respectively.

We examine the effect of introducing jittering to pedestrian training samples; this represents the applicable state-of-the-art, see Section 2. 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 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 (Eq. (17)), 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 jittering).

In a first experiment with a NN/LRF classifier (Figure 8a), the number of non-target training samples is kept constant and the benefit of jittering and virtual pedestrian synthesis is studied. From Figure 8a 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 8a) to non-target class resampling, see Figure 8b and 8c. 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 (Figure 8b and 8c, 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.

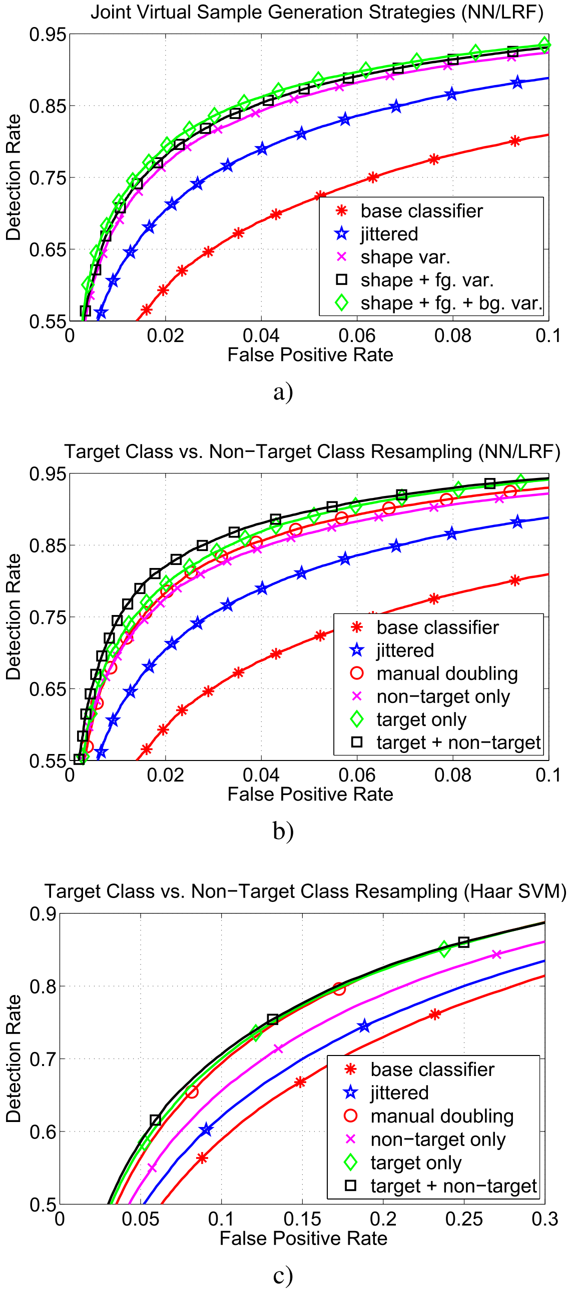

Figure 8. ROC performance for classification experiments. a) virtual pedestrian synthesis (NN/LRF), b)+c) target class vs. non-target class resampling for NN/LRF and Haar SVM

For comparison, we added 10946 real pedestrian samples plus four jittered versions, manually labeled from an auxiliary data pool, to the base dataset (without synthetic samples and active learning). Remarkably, the proposed generative-discriminative framework even outperforms the manual approach (see Figure 8b and 8c, 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 Figure 8b vs. Figure 8c.

7. Conclusion

This paper 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. We take this as evidence of the strength of our generative pedestrian model and selective sampling method. Future work involves applying the proposed framework to other object classes.

References

[1] D. Beymer and T. Poggio. Face recognition from one example view. In *Proc. ICCV*, pages 500–507, 1995.

[2] H.-P. Chiu et al. Virtual training for multi-view object class recognition. In *Proc. CVPR*, 2007.

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

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

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

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

[7] D. M. Gavrila and J. Giebel. Virtual sample generation for template-based shape matching. In *Proc. CVPR*, pages 676–681, 2001.

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

[9] A. S. Georghiades et al. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE PAMI*, 23(6):643–660, 2001.

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

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

[12] B. Heisele et al. Categorization by learning and combining object parts. In *NIPS*, pages 1239–1245, 2001.

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

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

[15] M. J. Jones and T. Poggio. Multidimensional morphable models. In *Proc. ICCV*, pages 683–688, 1998.

[16] A. Kapoor et al. Active learning with Gaussian processes for object categorization. In *Proc. ICCV*, 2007.

[17] J. A. Lasserre, C. M. Bishop, and T. P. Minka. Principled hybrids of generative and discriminative models. In *Proc. CVPR*, 2006.

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

[19] M. Li and I. K. Sethi. Confidence-based active learning. *IEEE PAMI*, 28(8):1251–1261, 2006.

[20] A. McCallum et al. Multi-conditional learning: Generative/discriminative training for clustering and classification. In *Proc. AAAI*, 2006.

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

[22] P. Niyogi, F. Girosi, and T. Poggio. Incorporating prior information in machine learning by creating virtual examples. In *IEEE Proc. Int. Sig. Proc.*, pages 2196–2209, 1998.

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

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

[25] S. Romdhani, S. Gong, and A. Psarrou. A multi-view nonlinear active shape model using kernel PCA. In *Proc. BMVC*, pages 483–492. A. F. Clark (ed.), 1999.

[26] P. D. Sozou et al. A non-linear generalisation of PDMs using polynomial regression. In *Proc. BMVC*, pages 397–406, 1994.

[27] K.-K. Sung and T. Poggio. Example-based learning for view-based human face detection. *IEEE PAMI*, 20(1):39–51, 1995.

[28] Z. Tu. Learning generative models via discriminative approaches. In *Proc. CVPR*, 2007.

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

[30] A. Vedaldi, P. Favaro, and E. Grisan. Boosting invariance and efficiency in supervised learning. In *Proc. ICCV*, 2007.

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

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

[33] D.-Q. Zhang and S.-F. Chang. A generative-discriminative hybrid method for multi-view object detection. In *Proc. CVPR*, 2006.