In Good Shape: Robust People Detection based on Appearance and Shape

Leonid Pishchulin
leonid@mpi-inf.mpg.de
Arjun Jain
ajain@mpi-inf.mpg.de
Christian Wojek
cwojek@mpi-inf.mpg.de
Thorsten Thormählen
thormae@mpi-inf.mpg.de
Bernt Schiele
schiele@mpi-inf.mpg.de

Computer Vision and
Multimodal Computing
MPI Informatics
Saarbrücken, Germany

Abstract

Robustly detecting people in real world scenes is a fundamental and challenging task in computer vision. State-of-the-art approaches use powerful learning methods and manually annotated image data. Importantly, these learning based approaches rely on the fact that the collected training data is representative of all relevant variations necessary to detect people. Rather than to collect and annotate ever more training data, this paper explores the possibility to use a 3D human shape and pose model from computer graphics to add relevant shape information to learn more powerful people detection models. By sampling from the space of 3D shapes we are able to control data variability while covering the major shape variations of humans which are often difficult to capture when collecting real-world training images. We evaluate our data generation method for a people detection model based on pictorial structures. As we show on a challenging multi-viewpoint dataset, the additional information contained in the 3D shape model helps to outperform models trained on image data alone (see e.g. Fig. 1).

(a) (b) (c) (d)

Figure 1: Performance of our detector without (b,d) and with (a,c) complementary shape information

BMVC 2011 http://dx.doi.org/10.5244/C.25.5

1 Introduction

People detection is a challenging task in computer vision with applications to surveillance, automotive safety and human-computer interaction. Prominent challenges for this task are the high intra-class variability, the high degree of articulation and the varying shape from different viewpoints. As state-of-the-art approaches heavily rely on supervised learning methods it is important to collect sufficient amounts of training data that capture the essential variability of the complex people class distribution. However, collecting large amounts of relevant training data is not only tedious but often an ill-defined process as it is unclear which part of the people class distributions is well-represented and which other parts of the distribution are still insufficiently sampled. A typical approach is to simply collect more training data without any guarantee that it will cover the people class distribution any better. Current datasets are often limited to few thousand samples taken from consumer photographs without any statement which parts of the real distribution of human shapes are sampled.

Recent work [20, 22, 23, 25] has proposed to use synthetic data to increase the available amount of training data. Similar to these works we use a graphics based 3D human model [17] to enrich an existing training dataset with additional training samples. Contrary to existing works however, we do not aim to use visually appealing and photo-realisticly rendered data but instead focus on complementary and particularly important information for people detection, namely 3D human shape. The main intuition is that it is important to enrich image-based training data with the data that contains *complementary shape information* and that this *data is sampled from the underlying human 3D shape distribution*. Specifically, we sample the 3D human shape space to produce several thousand synthetic instances that cover a wide range of human poses and shapes. Next, we render non-photorealistic 2D edge images from these samples seen from a large range of viewpoints and compute the low-level edge-based feature representation [4] from these. It is important to point out, that in no stage of this pipeline photo-realistic images are produced. Instead, we show that by careful design of the rendering procedure our feature representation can generalize from synthetic training data to unseen real test data. Our experiments on people detection show that the combination of real and large amounts of synthetic data sampled from a previously learned 3D human shape distribution allows to train a detector which outperforms models trained from real data only and thus outperforms the state-of-the-art (see e.g. Fig. 1).

Related work. People detection is an active research area and a complete survey is beyond the scope of the paper. Instead we refer to two recent surveys [8, 9]. While the results reported here are based on a publicly available detector code by [1] preliminary experiments showed that similar conclusions also hold for other gradient based methods such as the deformable part model [12].

One of the first to employ artificially created training data were Grauman et al. [16], who used a commercially available tool to generate silhouettes for multi-view pose estimation with static cameras. Recently, Shotton et al. [24] use a commercial tool to re-target recorded motion capture (mocap) data and automatically synthesize a large number of depth images for human body pose estimation with a depth sensor. Similarly, Marin et al. [22] use a game engine to generate training samples from multiple viewpoints. Broggi et al. [5] aimed to synthetically model pedestrians' appearance for detection in infrared images, while [10] varied 2D pedestrian appearance by employing a generative model for shape and texture of pedestrians. Recently, Pishchulin et al. [23] use a 3D computer graphics model to re-render images in order to obtain a larger number of synthetic training samples with realistic appear-

ance. While all these works aim to directly model realistic appearance, depth or silhouettes, we on the contrary model the underlying pose and 3D shape space and thus can produce an appropriate distribution of low-level edge-based features by non-photorealistic rendering.

Other work includes [20] modeling cars and bicycles with 3D features obtained from CAD data, while the follow-up work [19] learns an appearance model from real world images and a geometric model from CAD data. Stark et al. [25] also use CAD data to learn appearance as well as part constellations from a small number of CAD models and apply their model to car detection. These works have not been shown to be applicable to articulated objects such as people yet. While our work uses similar features and rendering procedures, we additionally address the issue of intra-class and pose variation by automatically generating a large number of 3D models to detect people with large degrees of articulation.

2 Pictorial structures model

Here we briefly recapitulate the pictorial structures model [14] made popular by [1, 13].

Pictorial structures model. This model represents the human body as a flexible configuration $L = \{l_0, l_1, ..., l_N\}$ of N different parts. The state of each part i is provided by $l_i = (x_i, y_i, \theta_i, s_i)$, with (x_i, y_i) denoting its image position, θ_i the absolute part orientation, and s_i the part scale which is relative to the part size in the scale normalized training set.

Given image evidence E, the posterior probability of the part configuration L is provided by $p(L|E) \propto p(E|L)p(L)$, where $p(E|L)$ is the likelihood of image evidence E given a part configuration L and $p(L)$ is the kinematic tree prior describing dependencies between parts.

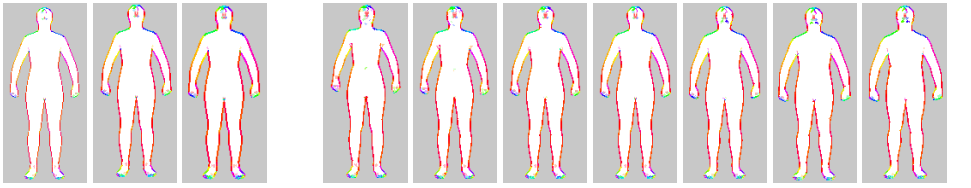
The prior can be factorized as $p(L) = p(l_0)\prod_{(i,j)\in G} p(l_i|l_j)$, with G denoting the set of edges representing kinematic dependencies between the parts, l_0 assigned to the root node (torso) and $p(l_i|l_j)$ are pairwise terms along the kinematic chains. $p(l_0)$ is assumed to be uniform, while pairwise terms are modeled to be Gaussians in the transformed space of joints between the adjacent parts [1, 13].

The likelihood term can be factorized into the product of individual part likelihoods $p(E|L) = p(l_0)\prod_{i=0}^{N} p(e_i(l_i))$, where $e_i(l_i)$ represents the evidence for part i at location l_i in the image. In the publicly available implementation provided by [1] part likelihoods are computed by boosted part detectors which use the output of AdaBoost classifiers [15] computed from dense shape context descriptors [4]. Marginal posteriors of individual body parts are found by sum-product belief propagation. Similar to the work of [1, 2], we use the marginal distribution of the torso location to predict the bounding box around pedestrian making a direct and fair comparison with this approaches feasible.

For our experiments we adjust the pictorial structures model of [1] to use 6 body parts which are relevant for pedestrian detection, namely right/left upper and lower legs, head and torso. We use a star prior on the part configuration, as it was shown to perform similar to the tree prior while making the inference more efficient. Individual part likelihoods along with the star prior are learned from real images, as it was done by [2], as well as from the rendered images produced by the 3D human shape model which we describe in Sec. 3.1.

3 Synthetic data

In this section we introduce our novel approach to generate large amounts of synthetic data with a realistic shape distribution from a 3D human shape model. First we briefly describe



(a) Varying edge thickness (b) Varying waist

Figure 2: Examples of rendered images with (a) 1, 2 and 3 pixel thick edges and (b) deviations from the mean waist girth (middle). Shown are gradual shape changes in the range ± 3 standard deviations from the mean shape.

the statistical model of human shape [17]. Then we present our data generation method and evaluate its parameters on a challenging multi-viewpoint dataset of real pedestrians [2].

3.1 3D model of human shape

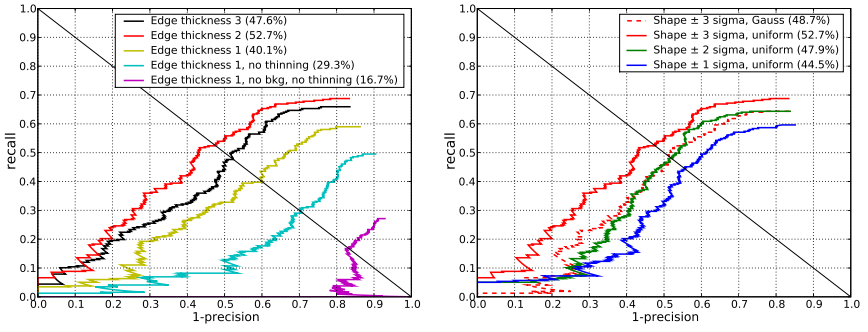
In order to generate non-photorealistic 2D edge images of pedestrians we employ a statistical model of 3D human shape [17] which is a variant of SCAPE model [3]. The model is learned from a publicly available database of 3D laser scans of humans and describes plausible shape and pose variations of human bodies. A total of 550 full body scans (with roughly 50% male and 50% female subjects, aged 17 to 61) in varied poses were used to generate the model.

The shape variation across individuals is represented via principal component analysis (PCA). We use the first 20 PCA components capturing 97% of the body shape variation. Unlike the SCAPE model, no per triangle deformation is learned to represent pose-specific surface deformations as for the rendering of edge images such level of detail is not necessary. Instead, the linear blend skinning is used to model shape motion. To this end, a kinematic skeleton was rigged into the average human shape model by a professional animation artist.

The motion of the model is represented by a kinematic skeleton comprising 15 joints, with a total of 22 degrees of freedom (DoF) plus 6 DoF for the whole body. The surface of the model consists of a triangle mesh with 6450 3D vertices and 12894 faces. In total the model has 28 pose parameters and 20 parameters representing the body shape variations. Following the approach of [18] we remap the 20 PCA shape parameters onto six meaningful semantic attributes, such as persons' height, weight, legs length, as well as waist, hips and breast girth, and directly change their values in our experiments.

3.2 Data generation

As argued before we propose to obtain synthetic training data by non-photorealistic rendering of the 3D human shape model described in Sec. 3.1. First, we uniformly sample shape parameters of six semantic attributes. Then we sample 3D joint angles from a set of 181 different walking poses. After shape and pose changes are applied, the 3D model is rendered from an arbitrary viewpoint into a 2D edge image by using an adapted rendering software kindly provided by the authors of [25]. Specifically, we render an edge created by two facets of the 3D model with respect to the normals of both facets. The closer the angle between camera view and normals to the right angle, the stronger the edge. Samples of rendered images are shown in Fig. 4(a). Prior to training of individual part detectors we combine the rendered edges with the background edges obtained by applying Canny edge detector [5] to an image without pedestrians. Finally, we perform edge thinning in the composed edge image, exactly as it is done in the edge detector, and use the obtained results as a direct input to



(a) Edge thickness and background composition (b) Sigma shape

Figure 3: Effects of (a) rendered edge thickness and composition with the background, and (b) allowed deviations from the mean shape. Composition of rendered and background edges following by edge thinning is essential for reasonable performance. Maximum range ±3 sigma of realistic shape variations leads to the best detection results.

shape context descriptors. All part annotations needed for training of the pictorial structures model are produced automatically from known 3D joint positions of the human shape model.

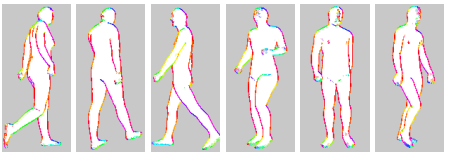
The thickness of rendered edges, composition with the background and the magnitude of shape variability may impact the detection performance. Therefore, we experimentally study the choice of these parameters. For each set of parameters we generate 6,000 (6K) synthetic images and train the pictorial structures model which we evaluate on the whole test set of the multi-view dataset [2] (see Sec. 4 for the dataset description and experimental setup).

Composition with background. First, we show the necessity of composing the rendered edge images with the background edges to achieve reasonable detection performance. The results are shown in Fig. 3(a). As it can be seen from the picture, combining rendered and background edges is absolutely essential as it helps to noticeably improve the results. Initial performance of the detector trained on the rendered edge images alone (purple curve, 16.6% EER – equal error rate) is significantly improved in case of composed images (cyan curve, 29.3% EER). This is due to the fact that dense part description covers a part of background in addition to the foreground structure, and thus the classifier learns that some edges come from the background. If these edges are completely missing in the training data, the boosted part detectors do not expect them during testing, which leads to frequent misclassifications.

Although the rendered edges are one pixel thick, their composition with the background edges can result in cluttered regions and thus edge thinning is needed. Fig. 3(a) shows that edge thinning in a Canny-detector-like manner helps to increase the performance from 29.3% EER to 40.1% EER. Thus, in the following experiments, we always perform edge thinning after composing rendered and background edge images.

Thickness of rendered edges. We observe that during rescaling and rotation of individual parts to the canonical orientation prior to learning of individual part detectors [1] some fragments of edges can be lost. We thus increase the thickness of rendered edges to two and three pixels and compare the results. As it can be seen from Fig. 3(a), rendering of two pixel edges helps to increase the performance from 40.1% to 52.7% EER. This result is also better than for three pixel edges. Therefore we render images with two pixel edges in the rest of our experiments. Examples of rendered images with various edges are shown in Fig. 2(a).

Shape variations. Another important parameter is the range of shape variations in the 3D human shape model. We study the optimal variations by uniformly sampling shape parameters from the range of ±1, 2 and 3 sigma (= standard deviation) from the mean shape.



(a) Rendered edge images

(b) Real pedestrians

Figure 4: Sample images of (a) rendered humans with shape, pose and viewpoint variations and (b) real pedestrians from the test set. Edge orientation is encoded as hue and edge strength as saturation in the HSV color system. Notice significant variability in clothing, shape, poses and viewpoints from which real pedestrians are seen.

Fig. 3(b) shows detection performances for those cases. It can be seen that the best result is achieved for the range ± 3 sigma (red curve, 52.7% EER). This is due to the fact that the data covers more variability then in case of ± 2 and ± 1 sigma (47.9% and 44.5% EER, respectively). We also compare this results to Gauss-sampling from the interval ± 3 sigma. As it can be seen, uniform sampling outperforms Gauss-sampling for the same interval (52.7% EER vs. 48.7% EER). Here the intuition is that the shape distribution learned from the database of human scans is narrower than the distribution in the test set of real pedestrians, and thus the tails of the shape distribution are essential and can be better represented by uniform sampling. This underlines the argument that it is important to vary the sample distribution systematically rather than blindly adding more training samples in order to cover essential parts of the distribution that improve detection performance. Fig. 2(b) shows sample images with various deviations from the mean waist girth.

4 Experimental Results

In this section we experimentally evaluate our approach of synthetic data generation described in Sec. 3.2 in various settings and compare it to the state-of-the-art performance obtained when using real training images only. First, we briefly introduce the datasets used for training and evaluation. Then we show the results when the detector is trained on synthetic data alone. Importantly and as argued before, joint training on real data with complementary shape information coming from the rendered samples allows to increase the performance over real data alone. Finally, we combine detectors trained on different datasets and outperform the state-of-the-art people detection results on the challenging multi-viewpoint dataset.

Rendered dataset. We generate 15,000 non-photorealisticly rendered edge images, as described in section 3.2. We uniformly sample shape parameters of six semantic attributes within ± 3 standard deviation from the mean shape and set the edge width to two pixels as these parameters led to the best detection performance (c.f. Sec. 3.2). Each training sample is rendered with 200 pixels height. As background dataset we use a set of 350 images without pedestrians. See for samples without background Fig. 4(a) where edge orientation is color-coded.

Multi-viewpoint dataset. The second dataset used in our experiments is the multi-view data proposed by [2]. The dataset contains images of real pedestrians from arbitrary viewpoints. It includes 1486 images of part-annotated pedestrians for training, 248 for testing and 248 for validation. In order to increase the amount of training data the images from the training set were mirrored. Samples from the multi-viewpoint dataset are shown in Fig. 4(b).

Experimental setup. For evaluation of trained detection models we use the test set of the multi-view dataset and thus are able to directly compare the obtained results to the state-of-

the-art performance of [2] which we denote as *Andriluka* in the following and to the method of [23]. In our experiments with rendered data we use the negative set of the multi-viewpoint dataset. We provide all results as precision-recall curves and use equal error rates (EER) to compare the performance. To match ground truth annotations to object detections we use the PASCAL criterion [11] which implies at least 50% overlap between ground truth and detection. We consider only detections which are at least 100 pixels high and additionally annotate all smaller pedestrians which were not annotated in the test set and thus when detected counted as false positives. For fairer comparison, we re-evaluate detectors of both [2] and [23] for this setting.

4.1 Results using rendered data alone and combined with image data

First we train the pictorial structures model on non-photorealisticly rendered edge images alone. We also compare the results to the performance obtained while training on real and joint real/rendered data. Fig. 5 shows the results when using 3K, 6K and 12K synthetic images. It can be seen that training on the rendered data alone achieves reasonable performance of 53.3% EER (dotted green and red curves), especially when taking into account that internal edges are missing in the synthetic data. However, the difference in performance due to various amounts of synthetic data is insignificant. In contrast, by adding different numbers of synthetic samples to the real ones and thus bringing various portions of complementary shape information into the training data we are able to noticeably improve detection performance over real data alone. The best result is achieved when jointly trained on 3K real and 12K synthetic samples, improving the detection rate from 80.8% (real data alone) to a remarkable 86.1% EER (combined synthetic with real data). This improvement clearly shows the power of using our synthetic data in addition to real images as it helps to increase the variability of training data and thus allows for better generalization of people detection models. The obtained result corresponds to the best ratio (12:3) between synthetic and real data. Increasing the amount of synthetic data leads to worse performance (84.9% EER) due to overfitting of the detection model to synthetic samples while decreasing reduces the shape variability gained through additional samples (c.f. Fig. 5 blue and green curves).

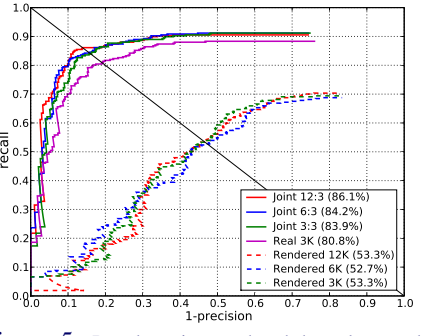
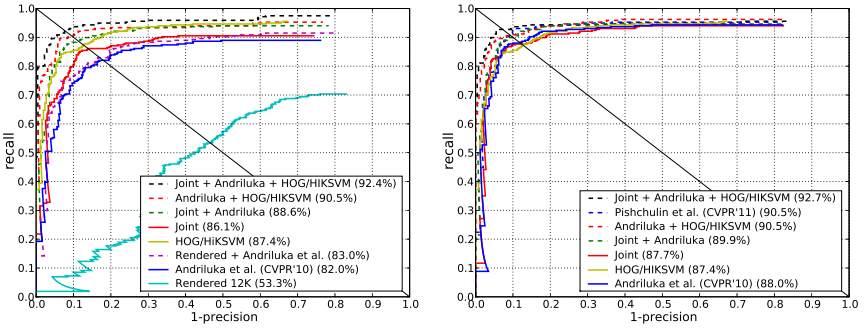


Figure 5: Results using rendered data alone and jointly with real data. Our detector trained on joint data outperforms the one trained on real data alone.

4.2 Combination of people detection models

Good results for joint training on synthetic and real data motivate experiments to combine different and potentially complementary people detectors trained on both datasets. In this section we explore various combinations of people detection models by following a detector stacking strategy. For that purpose we train the pictorial structures model on different datasets and then combine detectors by a linear SVM trained on vectors of mean-variance normalized detector outputs. For combination we use dense detection score maps which contain position and scale of detection hypotheses produced by individual detectors. For each detection score of each detector we build a feature vector by concatenating this score with



(a) Generic detectors (b) Multi-viewpoint detectors

Figure 6: Combination of (a) generic and (b) multi-viewpoint detectors. Combination of our detector jointly trained on the real and synthetic data with the one trained on the real data helps to improve detection performance.

the scores of all other detectors at current scale and position respecting the overlap between detections' bounding boxes. We use the validation set of multi-view data for SVM training.

To combine people detection models we consider two settings. In the first setting we train generic detectors – i.e. single detectors that are trained simultaneously on all viewpoints – and combine them with the best generic detector of [2] which we refer to as *Andriluka*. In the following discussion *Andriluka* corresponds to the detector trained on real data only, *Rendered* to the detector trained on synthetic data only and *Joint* trained on the combination of both data. The results are shown in Fig. 6(a). As it can be seen from the plot the combination *Andriluka* and *Rendered* improves the performance over *Andriluka* alone (83.0 vs. 82.0% EER) which means that the *Rendered* is complementary to the *Andriluka* despite missing internal edges and clothes. However, combination of *Joint* and *Andriluka* helps to increase the detection rate from 86.1% EER to 88.6% EER. This improvement means that indeed the detector trained on joint data is complementary to the one trained on real data alone. We compare the obtained result to the monolithic sliding window-based detector computed from histograms of oriented gradients (HOG) [7] and used for SVM training with the histogram intersection kernel (HIKSVM) [21]. The *HOG/HIKSVM* (yellow line, 87.1% EER) performs worse than our *Joint+Andriluka* detector (88.6% EER). The performance improves in case of *HOG/HIKSVM+Andriluka* (90.5% EER), while the combination of *HOG/HIKSVM* with *Joint* and *Andriluka* allows to achieve the best detection results (black dotted curve, 92.4% EER) outperforming *Joint+Andriluka* and *Andriluka* alone. This can be explained by more accurate scale estimation by *HOG/HIKSVM* detector which resolves the first false positives by the pictorial structures model such as pedestrian detections on the wrong scale.

In the second setting we first train individual viewpoint-specific detectors on appropriate subsets of the respective real and synthetic data and combine them again by means of an SVM classifier. In contrast to the previous case, where only one generic detector was trained on each subset, training the pictorial structures model separately for each single view allows for better adjustment of the individual part detectors and the tree prior for each viewpoint, making the detection model more discriminative. The results are shown in Fig. 6(b). First, the combination of 8 viewpoint-specific and one generic detector all trained on joint real and synthetic data outperforms the generic detector alone (87.7% EER vs. 86.1% EER, c.f. Fig. 6(a)), which supports the intuition that viewpoint-specific part detectors are more discriminative. The obtained results are on par with [2] (88.0% ERR) who not only combined 8 viewpoint-specific and one generic detector, but also used 2 side-view detectors additionally containing feet. By enriching this set of detectors by 8 viewpoint-specific detectors and

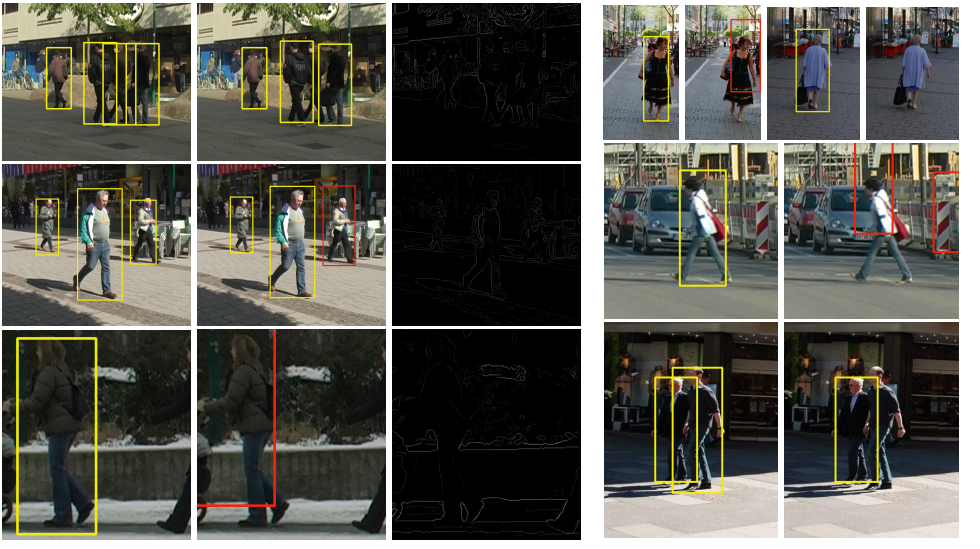
one generic detector trained on joint data, we outperform the results of [2] achieving 89.9% EER. Moreover, adding a *HOG/HIKSVM* to the bank of these detectors helps to further boost the performance up to 92.7% ERR outperforming *HOG/HIKSVM+Andriluka* (90.5% EER). We compare the results to the method of [23] (blue dotted curve) who also use morphable 3d body model for generating photorealsitic synthetic training samples and employ different detector combinations to boost the performance. It can be seen that we outperform the results of [23] (92.7% vs 90.5 %ERR) despite using non-photorealistically rendered images of pedestrians with missing clothes and internal edges. It can be explained by the fact that our approach is more flexible and allows to easily produce lots of training data with much variation in shape and pose whereas [23] is unable to represent pedestrians in unseen poses and appearances. Overall our method achieves the best known detection result on this dataset.

Discussion. In order to understand the reasons for better performance of joint training on synthetic and real data compared to the training on real data alone, we analyze typical failures of *Andriluka* and compare those to the detections by our *Joint* detector. Fig. 7(a) shows sample detections of both models at the EER, as well as edge images obtained by Canny edge detector [6]. It can be seen that *Andriluka* (middle column) fails when edge evidence for several body parts is missing or only partial e.g. due to occlusion (top row) or poor contrast (middle and bottom row). However, our detector (left column) is able to cope with such hard cases by focusing on shape evidence taken from external (human shape) edges. This underlines the argument that the synthetic data does indeed contain complementary information, namely the additional shape information, w.r.t. the real data alone. This clearly shows the advantage of using our approach for rapid generation of synthetic data with relevant shape distribution to increase the variability covered by the limited number of real samples.

We also analyze the advantages of using the combination of generic detectors *Joint* + *Andriluka* + *HOG/HIKSVM* over *Andriluka* alone. Sample detections at the EER in both cases are shown in Fig. 7(b). It can clearly be seen that our stacked detector is able to more accurately detect pedestrians when *Andriluka* fails. This suggests that our stacked detector generalizes better and can even detect pedestrians having unusual shapes (top row). This again means that both *HOG/HIKSVM* and our *Joint* detector are complementary to *Andriluka* and provide additional information helping to improve detection performance.

5 Conclusion

In this work we explored the potential of synthetically generated data from a 3D human shape model in order to enrich image-based data with complementary shape information. As we use a 3D human shape and pose model we are able to generate thousands of synthetic instances having a relevant and complementary shape distribution, covering a wide range of human shapes and poses. We show that careful design of the rendering procedure where model instances are rendered from various viewpoints into non-photorealistic edge images allows to compute low level feature representations which generalize to unseen real test data. Experimental results revealed significant improvements in detection performance when training a detector on joint synthetic and real data over training on real data alone, supporting our claim about complementarity of our synthetically generated data to the real data alone. Finally, we show in the generic as well as in the multi-viewpoint setting that the combination of our detector trained on the joint data with other detectors trained on real data alone allows to improve the state-of-the-art performance on a challenging multi-viewpoint dataset.



(a) Joint vs. Andriluka (b) Our combined detector vs. Andriluka

Figure 7: Typical failures of Andriluka's detector at EER and its comparison to (a) generic detector trained on joint rendered and real data and (b) combination of generic detectors. Andriluka's detector (every second picture) fails due to missing edge evidence while our detector trained on joint data relies on partial shape evidence taken from external edges. Edge images better seen on the screen when zoomed. Combination of our detector with *Andriluka* and *HOG/HIKSVM* helps to detect pedestrians despite unusual shapes, clutter, image blur and poor contrast.

In the future we will aim to further improve our rendering procedure to also explicitly model currently missing internal edges and learning the 3D human motion model from mocap data. Given the promising results reported in this paper such extensions will help to further improve detection performance.

Acknowledgements. We would like to thank Michael Stark for providing rendering software and Mykhaylo Andriluka for making detector code available.

References

[1] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *CVPR*, 2009.

[2] M. Andriluka, S. Roth, and B. Schiele. Monocular 3d pose estimation and tracking by detection. In *CVPR*, 2010.

[3] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. SCAPE: Shape completion and animation of people. In *ACM TOG (Proc. SIGGRAPH)*, 2005.

[4] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *PAMI*, 24(4):509–522, 2002.

[5] A. Broggi, A. Fascioli, P. Grisleri, T. Graf, and M. Meinecke. Model-based validation approaches and matching techniques for automotive vision based pedestrian detection. In *CVPR*, 2005.

[6] J Canny. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8:679–698, November 1986.

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[8] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian Detection: A Benchmark. In *CVPR*, 2009.

[9] M. Enzweiler and D. M. Gavrila. Monocular pedestrian detection: Survey and experiments. *PAMI*, 2009.

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

[11] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL visual object classes (VOC) challenge. *IJCV*, 88(2):303–338, June 2010.

[12] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 32:1627–1645, 2010.

[13] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61:55–79, January 2005.

[14] M. A. Fischler and R. A. Elschlager. The representation and matching of pictorial structures. *IEEE Trans. on Computer*, 22(1):67–92, January 1973.

[15] Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[16] K. Grauman, G. Shakhnarovich, and Trevor Darrell. Inferring 3d structure with a statistical image-based shape model. In *ICCV*, 2003.

[17] N. Hasler, C. Stoll, M. Sunkel, B. Rosenhahn, and H.-P. Seidel. A statistical model of human pose and body shape. In *CGF (Proc. Eurographics 2008)*, volume 2, 2009.

[18] A. Jain, T. Thormählen, H.-P. Seidel, and C. Theobalt. Moviereshape: Tracking and reshaping of humans in videos. *ACM Trans. Graph. (Proc. SIGGRAPH Asia)*, 29(5), 2010.

[19] J. Liebelt and C. Schmid. Multi-view object class detection with a 3d geometric model. In *CVPR*, 2010.

[20] J. Liebelt, C. Schmid, and K. Schertler. Viewpoint-independent object class detection using 3D feature maps. In *CVPR*, 2008.

[21] S. Maji, A.C. Berg, and J. Malik. Classification using intersection kernel SVMs is efficient. In *CVPR*, 2008.

[22] J. Marin, D. Vazquez, D. Geronimo, and A.M. Lopez. Learning appearance in virtual scenarios for pedestrian detection. In *CVPR*, 2010.

[23] L. Pishchulin, A. Jain, C. Wojek, M. Andriluka, T. Thormählen, and B. Schiele. Learning people detection models from few training samples. In *CVPR*, 2011.

[24] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from a single depth image. In *CVPR*, 2011.

[25] M. Stark, M. Goesele, and B. Schiele. Back to the future: Learning shape models from 3D CAD data. In *BMVC*, 2010.