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

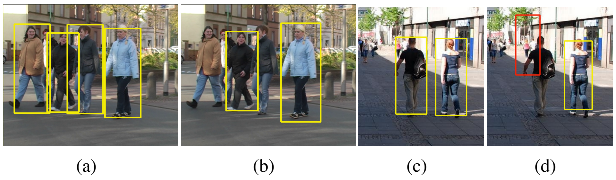

Figure 1: Performance of our detector without (b,d) and with (a,c) complementary shape information

Motivation. 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 [3] from computer graphics to add relevant shape information to learn more powerful people detection models. Contrary to existing works [4, 5, 6] 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*. 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. See Fig. 1 for sample detections by our detector with and without complementary shape information.

Data generation. In order to generate non-photorealistic 2D edge images of pedestrians we employ a statistical model of 3D human shape [3] which is learned from a publicly available database of 3D laser scans of humans and describes plausible shape and pose variations of human bodies. The overview of our approach is given in Fig. 2. First, we uniformly sample shape parameters of six semantic attributes. Then we sample 3D joint angles from a set of various walking poses. After shape and pose changes are applied, the 3D model is rendered from an arbitrary viewpoint into a 2D edge image. Prior to training of people detectors we combine the rendered edges with the background edges and use the result as a direct input to shape context descriptors. All annotations needed for training of people detector are produced automatically from known 3D joint positions of the body model. 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.

Results. For our experiments, we generate 15,000 non-photorealisticly rendered edge images where training samples are seen from a large range of viewpoints. As second dataset we use multi-view data [2] containing images of real pedestrians from arbitrary viewpoints. We train the publicly available detector [1] on different types of data and report the results on the challenging test set of the multi-viewpoint dataset. The results are shown in Fig. 3. It can be seen that training on synthetic data alone achieves reasonable performance of 53.3% EER, which is still less than 80.8% EER obtained by training on real data, due to missing internal edges and clothes in the synthetic samples. Hovewer, by jointly training on synthetic and real data we are able to improve the detection rate from 80.8% (real data alone) to a remarkable 86.1% EER (combined synthetic with real data - *Joint*). 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 detectors. Our method outperforms the results of Andriluka et.al [2] (82% EER) who trained the detector on real data alone. As it can be seen from Fig. 4 the detector of [2] (middle) fails when edge evidence for several body parts is missing or only partial e.g. due to poor contrast, while our detector (left) 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, and clearly shows the advantage of our method.

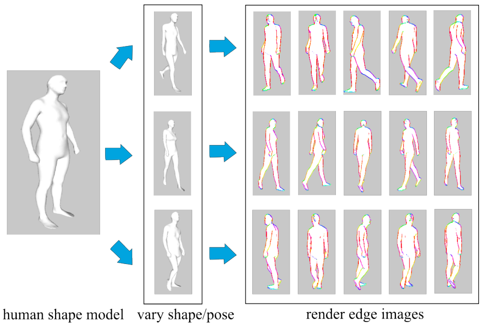

Figure 2: Overview of our method.

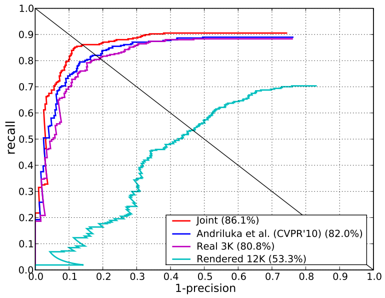

Figure 3: Results using rendered data alone and jointly with real data. Our detector trained on joint data outperforms the one trained on real data alone.

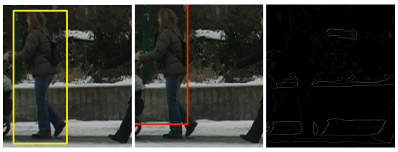

Figure 4: Our detector (left) trained on joint data relies on partial shape evidence taken from external edges and thus more robust to missing edges compared to the detector of [2] (middle) trained on real data alone. Right: edge image obtained by application of Canny edge detector.

[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] 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.

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

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

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