Learning People Detection Models from Few Training Samples

Leonid Pishchulin   Arjun Jain   Christian Wojek   Mykhaylo Andriluka
Thorsten Thormählen   Bernt Schiele

MPI Informatics, Saarbrücken, Germany

Abstract

People detection is an important task for a wide range of applications in computer vision. State-of-the-art methods learn appearance based models requiring tedious collection and annotation of large data corpora. Also, obtaining data sets representing all relevant variations with sufficient accuracy for the intended application domain at hand is often a non-trivial task. Therefore this paper investigates how 3D shape models from computer graphics can be leveraged to ease training data generation. In particular we employ a rendering-based reshaping method in order to generate thousands of synthetic training samples from only a few persons and views. We evaluate our data generation method for two different people detection models. Our experiments on a challenging multi-view dataset indicate that the data from as few as eleven persons suffices to achieve good performance. When we additionally combine our synthetic training samples with real data we even outperform existing state-of-the-art methods.

1. Introduction

People detection has been actively researched over the years due to its importance for applications such as mobile robotics, image indexing and surveillance. The most powerful methods for people detection rely on appearance-based features paired with supervised learning techniques. This is true for full-body models such as [8] as well as part-based models such as [1, 13, 15]. Key to best performance for these methods is to collect representative and substantial amounts of training data which is a tedious and time-consuming task and often limits further improvements.

The question we are asking in this paper is if the realism of today's computer graphic models such as [3, 4, 21] can help computer vision to reduce the tedious task of data collection and at the same time improve the quality and the relevant variability of the training data. Even in the early days of computer vision, computer graphics has been seen as a rich source for object models [7, 24, 27]. While these early models lacked realism in appearance more recent rendering techniques have indeed allowed to learn models for objects such as cars using computer graphics models alone [23, 28]. Also in the context of people detection computer graphics models have been used to generate training data. [26], e.g., reports promising results using a game engine to produce training data. While game engines have improved dramatically over the years they are still not as realistic as more elaborate 3D human models such as [3, 4, 21].

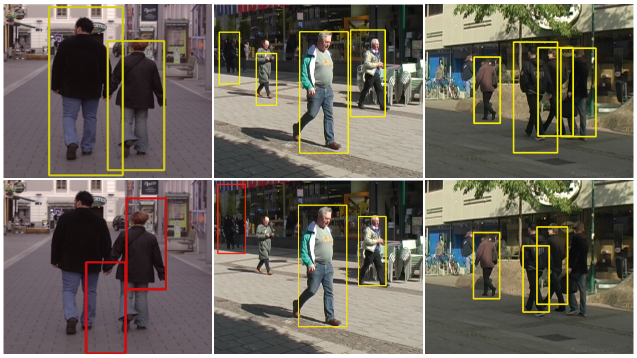


Figure 1: Sample detections at the equal error rate by the model trained on synthetic data generated from 6 people (top row) and on the game engine data of [26] (bottom row). Even training on a subset of data obtained from only 6 different people, we are able to outperform the detector trained on much more variable game engine data. (see Sec. 4.1 for more details)

The first major contribution is to explore the applicability of a state-of-the-art 3D person model from the computer graphics community to learn powerful people detection models. We directly compare to state-of-the-art systems based on the well-known pictorial structures model [1] as well as the Histogram of oriented gradients (HOG) model [8] learned from hundreds of manually labeled training data. Our findings indicate that surprisingly good results can be obtained training from as few as 1 or 2 people only and that comparable results can be obtained already with 11 people. The second main contribution is to compare these results to prior work such as [26]. The third contribution is to analyze different combinations of real and synthetic training data thereby outperforming the current-state-of-the-art using standard training data only. These results are obtained for two prominent people detection methods, namely the pictorial structures model and the HOG model.

Related Work. Using computer graphics to support object modeling in general and human modeling in particular is obviously not a novel idea. A large number of silhouettes rendered with the animation software Poser has been used e.g. to learn a multi-view shape model for humans [18]. A simple 3D human model is used by [6] to generate training data for infrared-based people detection. More recently [26] used a game engine to generate training data for a HOG-based detector. While promising results have been obtained, the employed computer graphics models still lack realism and thus seem suboptimal to train state-of-the-art detection models that rely on appearance based features.

Using an existing pool of training images, another line of research aims to increase training data size by a morphable 2D model based on silhouette and appearance modeling [10]. Improved performance w.r.t. the original pool of training images has been obtained even though a significant part of the improvement can be achieved by simply adding spatial Gaussian noise (often called jittering) to the training data [22]. The reason for this is that the employed morphable model is still inherently 2D and thus limited in generating relevant shape and appearance variations.

Therefore in this paper we follow a different route by leveraging the latest developments in 3D human shape and appearance modeling pioneered by [3, 4]. This kind of models are truly 3D and as such can – at least in principle – generate all relevant 3D human shape variations. More specifically we employ the model proposed in [21] and train the corresponding appearance from a small set of recordings (in the order of one to eleven people). Models trained on such data are compared both to models trained from a large pool of real training images as well as to models trained from images generated by a game engine [26].

2. People Detection Models

In this section we briefly recapitulate the two prominent people detection models used as the basis for our study. We will start with the pictorial structures model [16] which has been made popular by [1, 14] and then briefly introduce the sliding-window detection model with HOG features [8].

Pictorial structures model. In this model the human body is represented by a flexible configuration $L = \{l_0, l_1, ..., l_N\}$ of N body parts. The state of part i is given by $l_i = (x_i, y_i, \theta_i, s_i)$, where (x_i, y_i) denotes the part position in image coordinates, θ_i the absolute part orientation, and s_i denotes the part scale relative to the part size in the scale normalized training set. Given image evidence E, the posterior of the part configuration L is given by

$$p(L|E) \propto p(E|L)p(L) \tag{1}$$

where $p(L)$ is the kinematic tree prior and $p(E|L)$ corresponds to the likelihood of image evidence E under the particular body part configuration L. The tree prior expresses the dependencies between parts and can be factorized as

$$p(L) = p(l_0) \prod_{(i,j) \in G} p(l_i|l_j) \tag{2}$$

where G is the set of all directed edges in the kinematic tree, l_0 is 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, and pairwise terms are modeled to be Gaussians in the transformed space of part joints [1, 14].

The likelihood term is decomposed into the product of individual part likelihoods:

$$p(E|L) = p(l_0) \prod_{i=0}^{N} p(e_i(l_i)) \tag{3}$$

where $e_i(l_i)$ is the evidence for part i at image location l_i.

As we use the publicly available implementation provided by [1], part likelihoods are computed by boosted part detectors, which use the output of an AdaBoost classifier [17] computed from dense shape context descriptor [5]. Inference is performed by means of sum-product belief propagation to compute marginal posteriors of individual body parts. For pedestrian detection, the marginal distribution of the torso location is used to predict the bounding box, similar to the work of [1].

We slightly adapt the pictorial structures model of [1] to use 6 body parts which are relevant for pedestrian detection: left/right lower and upper legs, torso and head. Also, we use a star prior on the part configuration, as it was shown to perform on par with a tree prior [1] while making the inference much simpler.

In the experiments reported below the part likelihoods as well as the star prior are learned on different training sets ranging from real images as used by [2], over game-engine produced data as used by [26] to images produced from a state-of-the-art 3D human shape model introduced in the section 3.

Sliding-window detection with HOG features. In the sliding-window detection framework the image is scanned over all positions and scales and each window is represented by a feature and classified independently to contain a pedestrian or not. Contrary to the pictorial structures model pedestrians are often represented by a monolithic template without the notion of body parts. In this work we employ HOG features [8]. This feature has been shown to yield state-of-the-art performance for pedestrian detection in a recent benchmark [9]. For a 128×64 detection window HOG features vote the gradient orientation into 8×8 pixel large cell histograms weighted by the gradient's magnitude. To

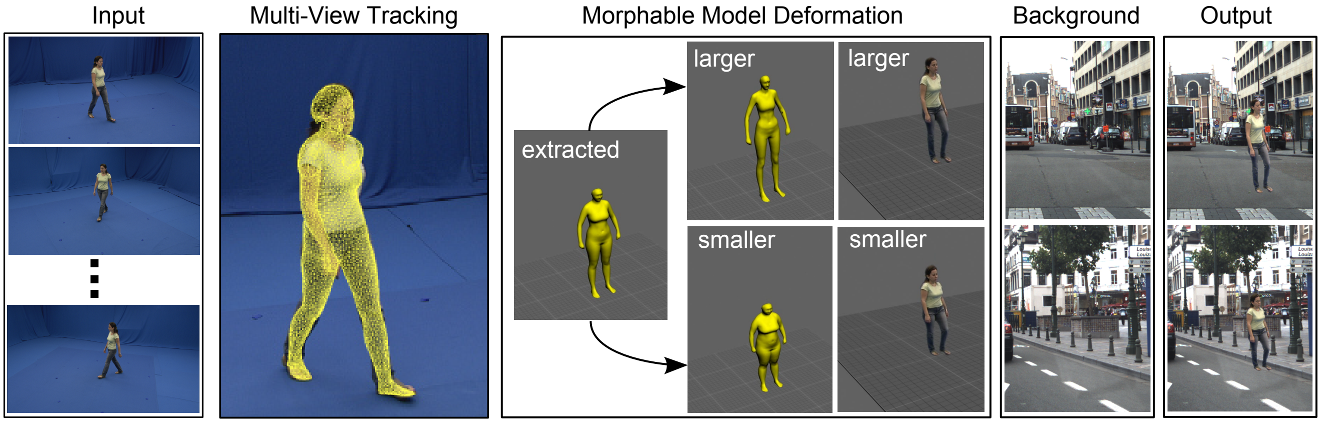


Figure 2: Overview of the approach to generate training data from real examples using a morphable 3D body model that drives a 2D image deformation.

tolerate slight variations in position and scale the responses are interpolated with respect to orientation and location and distributed into neighboring bins and cells. More robustness with respect to lighting conditions is achieved by normalization over 2 × 2 groups of cells. As a classifier we employ a histogram intersection kernel SVM which can be computed efficiently at test time [25]. To merge nearby detections on the same object we employ mean-shift mode search as a non-maximum suppression step.

3. MovieReshape: 3D Human Shape Model

In order to generate synthetic training data for the people detection models, we adopt an approach to reshape humans in videos [21]. The core component of this work is a morphable 3D body model that represents pose and shape variations of human bodies. Starting from an image sequence of an individual this model allows to generate large amounts of synthetic training data representing 3D shape and pose variations of the recorded individual. Fig. 2 shows an overview of the approach. To generate the required input data, we ask subjects to perform movements in front of a uniformly colored background in our motion capture studio. Each person is captured with 8 HD cameras with a resolution of 1296 × 972 pixels. First, the subject is segmented from the background and the extracted silhouettes are used to automatically fit the morphable 3D body model to the input sequences. We then randomly sample from the space of possible 3D shape variations that is defined by the morphable body model. These shape parameters drive a 2D deformation of the image of the subject. In the last step, an arbitrary background is selected and is composited with the image of the deformed subject. To generate large amounts of training data for each subject, the random selection of the 3D shape parameters and the background is repeated several times resulting in an arbitrary number of composited training images with different body shapes for all subjects, all performed poses, and all camera views.

Morphable 3D body model. The morphable body model is generated from a database of 3D laser scans of humans (114 subjects in a subset of 35 poses). Additionally, body weight, gender, age, and several other biometric measures of the subjects are recorded [20]. From this data a morphable 3D body model is built, similar to the well known SCAPE model [3]. This morphable model is capable of representing almost all 3D pose and shape variations available in the database. The pose variations are driven by a skeleton in combination with linear blend skinning that is defined once manually for the template mesh fitted to all 3D scans in the database. The shape variations across individuals are analyzed and represented via principal component analysis (PCA). The first 20 PCA components are used capturing 97% of the variations in the observed body shapes.

Markerless motion capture. Given the segmented input images, we employ a particle filter-based estimator [21] to fit the parameters of the morphable body model to the extracted silhouettes. The estimated parameters are the 28 joint angles of the skeleton and the 20 PCA coefficients. The approach selects those particles whose parameters produce the lowest silhouette error in all camera views.

Image deformation. Once we know the parameters of the subject in the video this defines our deformation source. The corresponding deformation target is defined by randomly selecting different shape parameters from our database. Thereby, we allow samples from 3 times the standard deviations that was observed in the 3D shape database of scanned subjects (corresponding to a 99% confidence interval). The difference between the 3D source and target model defines 3D offset vectors for all the vertices of the morphable model template mesh. As detailed in [21], a subset of these 3D offset vectors can be used to drive a 2D deformation in the image plane. This 2D deformation is consequently motivated by the knowledge about the shape variations of subjects in the database and the results are different from simple image transformations (like non-uniform scal-

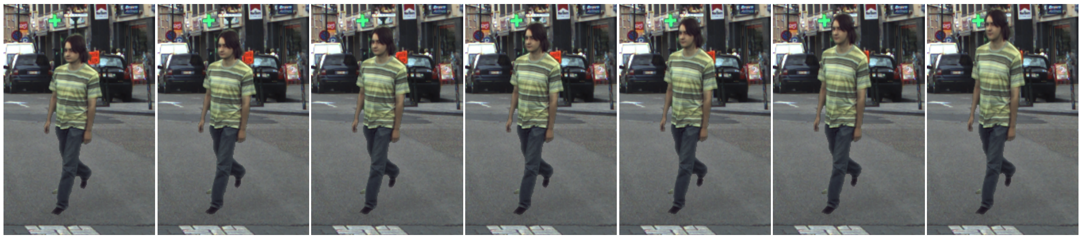


Figure 3: Sample Reshape images of a person with modified height. The leftmost and the rightmost images represent extreme deviations and the middle image corresponds to the original height; the 2nd and 6th images show deviations of 2σ, while the 3rd and 5th images correspond to the deviations of 1σ from the original height.

ing or shearing). It is, e.g., possible that the depicted subject only becomes bigger at the belly, or gets shorter legs, or enjoys more muscular arms. The image deformation is repeated multiple times with randomly sampled body shapes.

Background compositing. In the final step, we sample randomly from a database of backgrounds containing images of urban scenes without pedestrians. We blend the segmentation masks of persons with a Gaussian with σ of 2 pixels. Then, we composite the background with the deformed images of subjects by adding weighted background and foreground pixel values together. See Fig. 3 for sample outputs of the system varying the height of the person.

4. Results

This section experimentally evaluates the applicability of training data obtained by the 3D human shape model described in section 3. These results are compared to training data obtained from real images [2] as well as from a game engine [26]. First, we briefly introduce the different datasets used for training and evaluation. Then, we show that already a small number of people in our training dataset allows to achieve performance almost on par with the detector trained on real data containing hundreds of different people. We also show that combining detectors trained on real and synthetic data allows to outperform the detectors trained only on real data.

Reshape training dataset. In order to obtain synthetic training data, we collected a dataset of 11 subjects each depicted in 6–9 different poses corresponding to a walking cycle. Each pose is seen from 8 different viewpoints separated by 45 degrees apart from each other. Synthetic images were obtained as described in section 3. For each original image we generated 30 gradual changes of height: 15 modifications making a person shorter and 15 making a person taller, which results in almost 2000 images per person and 20400 positive training samples in total (see for samples Fig. 3). We note that the applied transformation is non-linear and therefore different from simply scaling the original image. The MovieReshape model also allows to automatically obtain bounding box as well as body part annotations which are required for the pictorial structures model. The annotations for the unmodified image are obtained by backprojecting the morphable 3D model to the image plane. For the reshaped images we apply the same inverse mapping to these annotations which is used to morph appearance. This is one of the key advantages which facilitates the generation of large amounts of data without the need to manually annotate each image. All persons are rescaled to 200 pixel in height and embedded in background images of driving sequences containing no pedestrians. To record the background sequences a calibrated camera has been used and thus synthetically generated pedestrians can be easily embedded at geometrically plausible positions on the ground plane. Some sample images are shown in Fig. 4 (top row). We additionally perform smoothing along the shape boundaries separating persons from background in order to get more realistic gradients for the shape context descriptor. Finally, we adjust the luminance of the embedded pedestrians such that their mean approximately matches the backgrounds' mean luminance.

CVC training dataset. The second dataset contains synthetic images produced by a game engine which were kindly provided by the authors of [26]. These images of virtual pedestrians are generated by driving through virtual cities in the computer game Half-Life 2. The CVC dataset which we were provided with consists of 1716 pedestrians shown from arbitrary views with annotated bounding boxes. In comparison to our Reshape dataset, the appearance variability of the CVC dataset is significantly larger (*c.f*. Fig. 4, middle row). We manually annotated the body parts of people and also rescaled the images so that all subjects have the same height of 200 pixels. Finally, we mirrored all images in order to obtain more training data, resulting in 3432 images in total. This data is complemented by a negative set of 2047 images of the same virtual urban scene environment without pedestrians.

Multi-viewpoint dataset. The third dataset we used in our experiments is the challenging multi-viewpoint dataset [2] consisting of real images of hundreds of pedestrians shown from arbitrary views. The dataset comes with 1486 part-annotated pedestrians for training, 248 for testing and 248 for validation. The images from the training set were mirrored in order to increase the amount of training data. Sample images for different viewpoints can be seen in Fig. 4

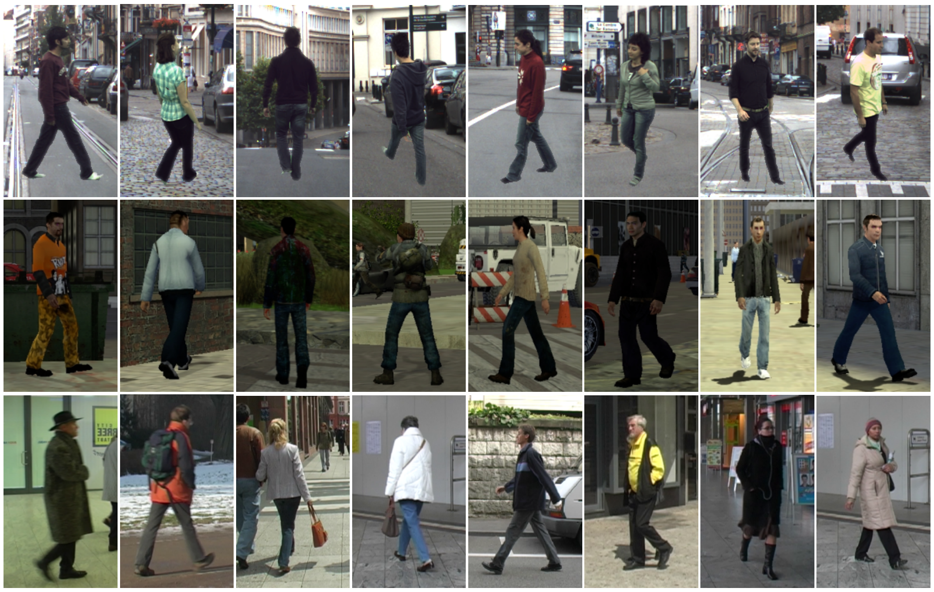


Figure 4: Samples from the training data used in our experiments: Reshape images (top row), CVC virtual pedestrians (middle row) and multi-viewpoint dataset (bottom row). Synthetic Reshape images look similar to the real ones while being much more realistic than CVC pedestrians. Real images often contain persons wearing long or wide clothes and caring a bag, which does not occur in the synthetic data.

(bottom row).

Experimental setup. To evaluate all trained models we use the multi-viewpoint dataset's test data and are thus directly comparable to the state-of-the-art on this dataset [2]. Thus, whenever we use the multi-viewpoint training data we refer to the experiment as *Andriluka*. For our experiments which use the Reshape dataset we use the multi-viewpoint dataset's negative training data. Experiments on the CVC data showed minor performance differences between using the negative data provided with the CVC data or the negative data provided by the multi-viewpoint data. In the following we thus only report results obtained with the CVC negative dataset. All results are provided as precision vs. recall curves and throughout this chapter we use the equal error rate (EER) to compare results. EERs for each experiment are also reported in the respective plots' legend. To match ground truth annotations to objects detections we use the PASCAL criterion [12], which demands at least 50% overlap of ground truth bounding box and detection.

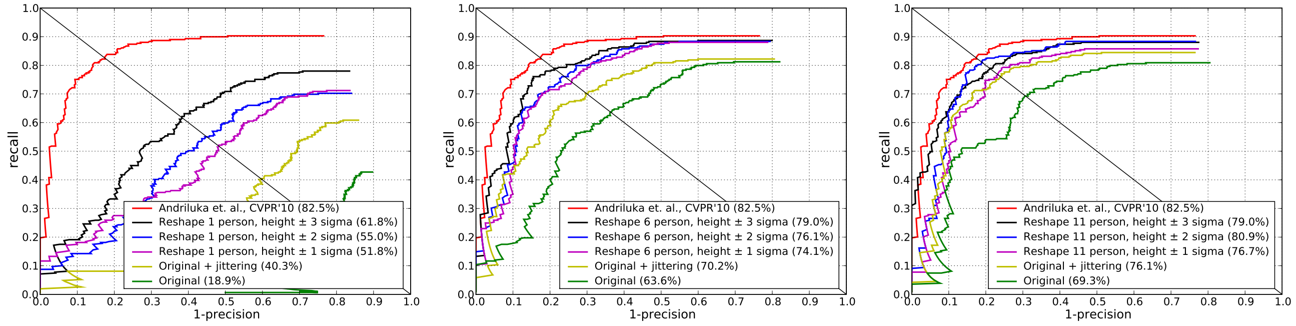
4.1. Results using the Reshape data

We start by evaluating the pictorial structures model's performance when it is trained on the Reshape data and compare its performance to training on the multi-viewpoint training dataset and the CVC training dataset.

Fig 5(a)-(c) show the results obtained using one, six, and eleven people to train a generic pictorial structures model. To understand the influence of different parameters of the model we vary the employed subset of the Reshape data. The green lines in figure Fig 5(a)-(c) show the results obtained using the original training sequences acquired from one, six and eleven people without applying the human reshape model of section 3. While the performance increases with more people the maximum performance obtained with eleven people is only 69.2% EER (equal error rate).

Although the wide range of height modifications allows to cover 99% of data variability spanned in this direction, having extremely short and tall pedestrians in the training set can be unnecessary, since they are quite rare in real world data. This consideration motivates to subsample the Reshape data w.r.t. maximal and minimal height of subjects. For that purpose we train pictorial structures model on subsets of images corresponding to no modification, $\pm 1, 2$ and 3σ (standard deviation) from the original mean height of people. The results for 1, 6 and 11 persons are again shown in Fig. 5. It can be observed that in all cases including the images with increasing number of height modifications helps to improve performance.

In order to understand whether the improvement comes from the increased variability of data rather than from the increased amount of positive samples, we also train the model on the set of original images enriched by jittering [22]. The results are shown in Fig. 5(a)-(c) in yellow. As expected, the performance of the model in the latter case is worse, as nonlinear data transformation due to height modifications allows to capture more realistic variability of the data than simple 2D jittering. The largest difference can be observed when using a single person only for training (*c.f.* Fig. 5(a)). In this case, jittering helps to improve the performance from 18.7% to 40.1% EER, while training on the data with height modification $\pm 1\sigma$ results in an EER of 51.8%. When using six and eleven people the difference between using ± 2 or $\pm 3\sigma$ becomes less pronounced. For



(a) 1 person

(b) 6 people

(c) 11 people

Figure 5: Results using Reshape data. Shown are results using 1 (a), 6 (b) and 11 (c) people to train a generic pictorial structures model. Each plot show results obtained by [1] on real data (red), training on the unmodified training data (green), the reshape model with different variations of σ (violet, blue, black) and results using jittering (yellow)

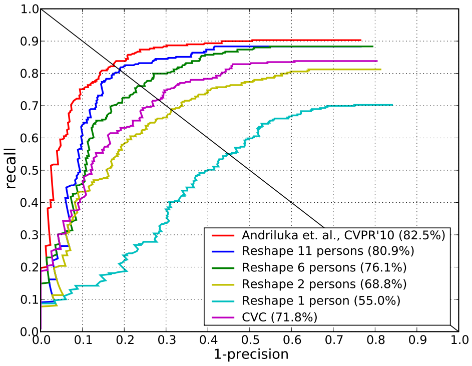


Figure 6: Detection rate w.r.t. the number of different persons represented in the training data. Already one person is enough to provide reasonable variability in the Reshape data. The increasing number of persons results in significant improvement which allows to achieve performance almost on par with the detector trained on the real data. The model trained on CVC data performs well, but noticeably worse than ours.

instance, for six people this difference constitutes 3.9% on EER. As $\pm 2\sigma$ corresponds to faster training times due to less data we use this setting for the remainder of the paper.

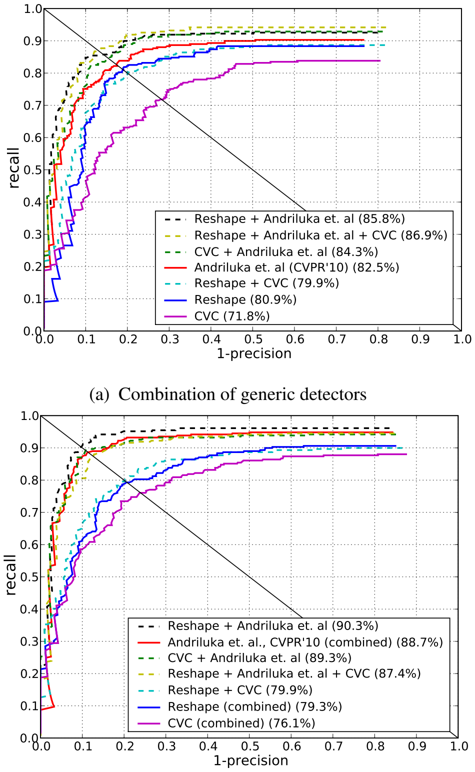
Fig. 6 summarizes how the number of different persons contained in the Reshape dataset affects performance. Surprisingly, already training data from a single person obtains an EER of 55.0% suggesting that this data already covers a reasonable variability (this performance can be further improved using $\pm 3\sigma$ as shown in Fig 5(a)). Not surprisingly, increasing the number of people improves performance considerably. More interesting however is the fact that with as few as 11 people we are able to achieve performance of 80.9% EER, which is almost on par with the model trained on the real multi-viewpoint data (red curve, 82.5% EER) containing hundreds of different people.

Fig. 6 also contains a curve (in violet) for the model trained on the CVC dataset. As expected, the model trained on the virtual people and thus less realistic data performs worse achieving 71.8% EER, despite much larger number of different appearances contained in the dataset. For comparison, the model trained on a subset of our data from just six persons achieves 76.1% on EER. We also provide some sample detections obtained in this case which are shown in Fig. 1. These results clearly show the advantage of using our Reshape data for training.

4.2. Combining different datasets for training

In the previous section good results have been obtained using the Reshape data from as few as eleven different people as well as using training data from real images. Therefore this section explores the possibility to combine models trained on different types of data in order to boost performance further. In order to combine detectors, we follow a detector stacking strategy (also used in [2]). More precisely we train detectors on different datasets first and then combine them by an another SVM that is trained using the vectors of detector outputs as features (normalized by mean/variance). For SVM training, we use the validation set provided with the multi-viewpoint dataset.

We consider two different settings. First, we consider the combinations of the models trained on all viewpoints of the corresponding data, as it is done in the previous section. The results are shown in Fig. 7(a), where single detectors are denoted by solid lines, and combined ones are marked by dotted lines. The combination *Andriluka+CVC* (84.1% EER) improves performance slightly over *Andriluka* alone (82.5% EER) whereas the combination *Reshape+CVC* (79.9%) does not improve performance w.r.t. *Reshape* (80.9%). The combination *Reshape+Andriluka* (85.8%) does improve both over *Andriluka* alone as well as *Reshape* alone. Further adding CVC (*Reshape+Andriluka+CVC*) slightly improves the performance achieving 87% EER. Overall this combination obtains the best performance reported in the literature for this setting (multi-viewpoint pictorial structures model). The combination *Reshape+CVC* performs similarly to *Reshape* data alone. This might be due to the fact that in both types of data subjects wear tight clothes such as trousers, jackets and T-shirts, but no coats or dresses which sometimes occur in the test data. Additionally this combination suffers from less realistic appearance



(a) Combination of generic detectors

(b) Combination of viewpoint specific detectors

Figure 7: Combination of generic detectors (a) and viewpoint specific detectors (b). In both cases, the combination of our detector with the one trained on the real data helps to improve detection performance.

of the virtual pedestrians. Hence, the additional *CVC* samples are not complementary to the Reshape samples. This intuition is also confirmed by a noticeable improvement obtained by combining the detector trained on Reshape data with the one of [2] trained on real multi-viewpoint data. As quite a few images in the real multi-viewpoint training set contain persons wearing long clothes the training data and thus the detectors are more complementary. For the same reason, the combination of the CVC detector and the Andriluka detector performs better than Andriluka's detector, though the combination's performance is slightly worse than the combination with the Reshape data.

The second setting explored in this section is to combine not only one detector trained on each dataset but to first train viewpoint-specific detectors on appropriate subsets of the different data and then train a stacked classifier on combinations thereof. The main advantage is that the part detectors as well as the kinematic tree prior are more specific for each view and thus more discriminative. The results are shown in Fig. 7(b). First, the combination of 8 viewpoint-specific detectors trained on Reshape data clearly outperforms those trained on CVC virtual pedestrians (79.3% against 76.1% EER) which again shows the advantage of training on our synthetic Reshape data. However, the performance achieved is still below the results provided by [2] (red curve) who combined 8 viewpoint-specific detectors, 2 sideview detectors that contain feet and one generic detector trained on all views. By enriching this set of detectors by 8 viewpoint-specific and one generic detector trained on human Reshape data, we are able to outperform the results of [2] increasing the detection rate from 88.6% to 90.3% EER. The 8 CVC viewpoint-specific detectors are not complementary enough to further boost performance w.r.t. the combinations mentioned above.

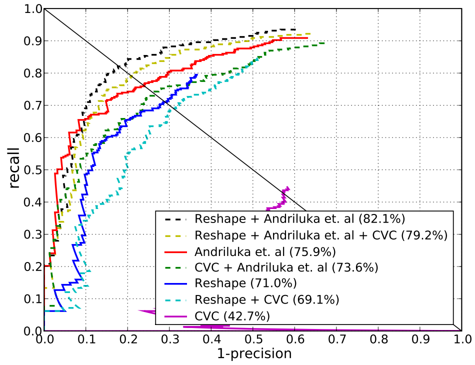


Figure 8: Performance of the sliding window detector of on different types of data. Similar to pictorial structures model, the best performance achieved when trained on human reshape together with real data. Detector trained on CVC pedestrians again performs worst.

4.3. Sliding-window detection using HOG

For the combination of different datasets we additionally verified our findings for a sliding-window detector framework (see Fig. 8). For this experiment we trained a generic detector for all viewpoints consisting of a monolithic HOG feature representation [8] combined with a fast histogram intersection kernel as classifier [25]. We used the exact same training data as for the experiments reported above. Overall the results obtained are slightly below the pictorial structure model's results in Fig. 7(a). This may be explained by the test set's difficulty, which contains people seen from all viewpoints and under all poses for which a part-based representation is favorable. As for the pictorial structures model the combination of the Reshape data with the multi-viewpoint data provided by Andriluka obtains best performance with an EER of 82.1%. When we additionally add the CVC virtual samples the performance drops to an EER of 79.2% which can be explained by the less realistic appearance of these samples. However, both combinations outperform the detector which is only trained on data by [2] (EER 75.9%). Consistent with our finding for the pictorial structures model, the performance drops to an EER of 73.6% when the CVC data is added. Also the detector trained only on the Reshape data (EER 71.0%) performs worse than the detector trained on real data. Similarly to the real multi-viewpoint data, the combination of Reshape with CVC data decreases performance (EER 69.1%). The performance with the detector only trained on the CVC virtual samples is substantially worse. Interestingly Marin *et al*. [26] have reported equal performance of virtual samples and their real data when a sliding-window size of 48×96 pixels is used. This might be explained by the fact that real data and virtual data appear more similar on the lower resolution DaimlerDB [11] automotive test data (pedestrian median height is 47 pixels), while for higher resolution (median pedestrian height on the multi-viewpoint test data is 184 pixel) the classifier might loose performance due to unrealistic appearance. Overall we find that the results for the sliding-window detector framework to be consistent with the results obtained by the pictorial structures model leading to the same conclusions. We would also like to highlight that a detector trained on the combination of multi-viewpoint and Reshape data clearly outperforms a detector which is only trained on real multi-viewpoint data.

5. Conclusion

This paper explored the possibility to generate synthetic training data from a state-of-the-art computer graphics 3D human body model (called Reshape data in the paper). Learning people detection models from as few as 11 people enabled to achieve performance nearly on par with state-of-the-art systems trained on hundreds of manually labeled images. This result has been obtained for two of the best known people detection models, namely the pictorial structures model (in two different settings) as well as the HOG-detector. Using less realistic training data generated from a game engine [26] has led to far less compelling results. Combining the detectors trained on the Reshape data with detectors trained on the manually labeled data has allowed to outperform the state-of-the-art for challenging multi-viewpoint data introduced by [2].

Considering the fact that only 11 people have been recorded and used to train the respective appearance models the results reported in this paper are indeed promising. In fact, using recordings from several hundreds of people should allow to reach performance levels that are beyond what can be reached with today's manually and tediously labeled data. To further increase the variability in appearance we also envision the combination of the Reshape data generation with an additional model for clothing generation such as Eigen Clothing [19].

Acknowledgements. We would like to thank Javier Marin Tur for provided dataset of virtual pedestrians.

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] A. Balan, L. Sigal, M. Black, J. Davis, and H. Haussecker. Detailed human shape and pose from images. In *CVPR*, 2007.

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

[6] 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.

[7] R. Brooks, R. Creiner, and T. Binford. The acronym model-based vision system. In *Intern. Joint Conference on Artiticial Intelligence*, pages 105–113, 1979.

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

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

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

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

[12] 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.

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

[14] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61:55–79, Jan. 2005.

[15] V. Ferrari, M. Marin, and A. Zisserman. Progressive search space reduction for human pose estimation. In *CVPR*, 2008.

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

[17] 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.

[18] K. Grauman, G. Shakhnarovich, and T. Darrell. Inferring 3d structure with a statistical image-based shape model. In *ICCV*, pages 641–648, 2003.

[19] P. Guan, O. Freifeld, and M. J. Black. A 2D Human Body Model Dressed in Eigen Clothing. In *ECCV*, 2010.

[20] 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.

[21] 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.

[22] I. Laptev. Improving object detection with boosted histograms. *Image Vision Comput.*, 27(5):535–544, 2009.

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

[24] D. Lowe. Three-dimensional object recognition from single two-dimensional images. *Artificial Intelligence*, 31:355–395, 1987.

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

[26] J. Marin, D. Vazquez, D. Geronimo, and A. Lopez. Learning appearance in virtual scenarios for pedestrian detection. In *CVPR*, pages 137–144, 2010.

[27] D. Marr and H. Nishihara. Representation and recognition of the spatial organization of three-dimensional shapes. *Proc. Roy. Soc. London B 200*, pages 269–194, 1978.

[28] M. Stark, M. Goesele, and B. Schiele. Back to the future: Learning shape models from 3D CAD data. In *British Machine Vision Conference (BMVC)*, 2010.