Deep Body-pose Estimation via Synthetic Depth Data: A Case Study

Christopher Pramerdorfer[1,2] and Martin Kampel[2]
[1]*Cogvis, Vienna, Austria*
[2]*Computer Vision Lab, TU Wien, Vienna, Austria*

Keywords: Deep Learning, Synthetic Depth Data, Body-pose Estimation.

Abstract: Computer Vision research is nowadays largely data-driven due to the prevalence of deep learning. This is one reason why depth data have become less popular, as no datasets exist that are comparable to common color datasets in terms of size and quality. However, depth data have advantages in practical applications that involve people, in which case utilizing cameras raises privacy concerns. We consider one such application, namely 3D human pose estimation for a health care application, to study whether the lack of large depth datasets that represent this problem can be overcome via synthetic data, which aspects must be considered to ensure generalization, and how this compares to alternative approaches for obtaining training data. Furthermore, we compare the pose estimation performance of our method on depth data to that of state-of-the-art methods for color images and show that depth data is a suitable alternative to color images in this regard.

1 INTRODUCTION

Current research in Computer Vision is highly data-driven due to the prevalence of deep learning, which has enabled significant performance gains in many fields such as image classification (He et al., 2016) and human pose estimation in color images (Cao et al., 2018). However, large datasets are required to be able to fully utilize the potential of Deep Learning, which are not always available.

This presumably is an important reason why depth data have become less popular in Computer Vision research after a surge in interest following the release of the Kinect depth sensor (Shotton et al., 2011). This is despite depth data having practical advantages over color (or grayscale) images. This applies in particular to practical applications that involve people, where utilizing video cameras raises privacy concerns and the reluctance of users due to feeling monitored. In some practical cases, this effectively precludes technology based on color images for this reason.

One such example that we focus on in this paper is human pose estimation for identifying unhealthy sitting positions at the workplace in order to promote the long-term health of office workers by raising awareness. This task also naturally favors depth data as it requires 3D pose estimation, which is more intuitive in depth data than in color image data. However, there are no large depth datasets available that closely represent this problem and acquiring such a dataset is a considerable effort that involves recruiting a large number of people. This is a common problem that hinders progress in the corresponding research fields.

In this paper, we discuss and compare different approaches to address this problem by obtaining suitable training data. One approach that is the focus of our study is utilizing synthetic depth data created specifically for this purpose. Depth data synthesis allows creating datasets of virtually any size and with accurate labels with comparatively little effort. This not only applies to pose estimation but to most applications that involve human participation and depth data. Yet despite these advantages there are only few synthetic depth datasets and works that utilize such data, particularly works based on deep learning. Whether this is due to issues with generalization of trained models to real data or for other reasons is unclear.

We aim to shed light on this matter by training Convolutional Neural Networks (CNNs) for human pose estimation on synthetic depth data and then study their performance on synthetic validation data as well as real test data. We compare the results to alternative approaches for obtaining training data, namely (i) acquiring a limited amount of real data that reflect the problem, (ii) adapting an existing real dataset for this purpose, and (iii) adapting an exist-

Pramerdorfer, C. and Kampel, M.
Deep Body-pose Estimation via Synthetic Depth Data: A Case Study.
DOI: 10.5220/0008941303180325
In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020) - Volume 5: VISAPP*, pages 318-325
ISBN: 978-989-758-402-2; ISSN: 2184-4321

ing synthetic dataset. This enables us to investigate the applicability of training CNNs on synthetic depth data for solving practical problems in the context of possible alternatives. Our work is based on (Pramerdorfer et al., 2019), which shows that utilizing synthetic depth data for human pose estimation is feasible but lacks studies on how this compares to other approaches, means necessary for generalization across datasets, and the effect of sensor noise simulation.

Furthermore, we compare these results to the state of the art in 2D body pose estimation in color data for an indication of how depth-data-based human pose estimation performs compared to the more popular color-data-based methods. To the best of our knowledge, this is the first study of this kind.

The results show that utilizing synthetic training data outperforms all other approaches in terms of body pose estimation performance on a common realistic test dataset. Acquiring a limited amount of realistic training data, a popular alternative in practice, performs significantly worse despite transfer learning to address the small dataset size. The comparison with color-based methods indicates that pose estimation in depth data is possible with similar accuracy. The results underline that depth data are a viable alternative to color data for human pose estimation and that deep learning from suitable synthetic data can outperform other data acquisition strategies.

This paper is structured as follows. Section 2 covers related works on human pose estimation and synthetic depth data. In Section 3 we discuss the problem considered in this case study and the different approaches for acquiring training data in more detail. Our pose estimation method is explained in Section 4 while Section 5 presents the experiments and results, and Section 6 concludes the paper.

2 RELATED WORK

Human Pose Estimation. Human pose estimation in color images via deep learning is a popular research topic. A seminal work in this field is (Toshev and Szegedy, 2014), in which a CNN is trained for 2D keypoint regression. More recent works such as (Cao et al., 2018) and (Fang et al., 2017) perform dense keypoint (heat-map) prediction for improved performance. A limitation of these methods is that they predict 2D poses. 3D pose estimation from single images is more challenging than the 2D variant due to the larger pose space and ambiguities caused by perspective projection. Methods that perform well at this task have been proposed only recently. A popular approach is to first predict 2D keypoints, which are then mapped to 3D. In (Chen and Ramanan, 2017) this mapping is accomplished using a similarity search in a large dataset of pairs of 2D and 3D keypoints. The authors of (Sun et al., 2018) present an extension of CNN heat-map prediction that supports 3D poses.

In contrast, there are few recent works that utilize depth data. Kinect's pose estimation method (Shotton et al., 2011) is perhaps the most well-known example of utilizing synthetic depth data for this purpose but its performance is no longer competitive (Haque et al., 2016). Moreover the method is based on classification forests, which may generalize from synthetic data to real data differently than CNNs. (Haque et al., 2016) presents a patch-based method for 3D pose estimation in depth data using a combination of a CNN and a recurrent neural network. Two more recent works are (Guo et al., 2017) and (Moon et al., 2018). The former proposes a multi-stage network architecture for 3D pose estimation from depth maps while the latter both processes and predicts keypoints in 3D voxel grids. We utilize a simpler network architecture that processes depth map patches. The work that is most closely related to ours is (Pramerdorfer et al., 2019), which also covers upper-body pose estimation and synthetic training data. However, the paper focuses on pose classification and lacks comparative studies.

Synthetic Depth Data. To our knowledge, (Shotton et al., 2011) was the first work to demonstrate the potential of utilizing synthetic data for 3D pose estimation in depth maps. The work is still one of only few examples and, as mentioned before, not based on deep learning. The most comprehensive public dataset that includes depth maps of people is SURREAL (Varol et al., 2017). We include this dataset in our studies for comparison. (Pramerdorfer et al., 2019) is another example where synthetic data are used successfully for pose estimation purposes.

3 3D POSE ESTIMATION

We consider the task of estimating 3D coordinates of six face and upper-body keypoints, namely the nasion (intersection of the frontal bone and the two nasal bones of the human skull), chin center, front of the throat, manubrium, as well as the left and right shoulders. All keypoints lie on the skin surface. These keypoints were found to be important for identifying unhealthy sitting postures in (Pramerdorfer et al., 2019).

3.1 Test Dataset

We evaluate each training data approach using the same test dataset, which was presented in (Pramer-

dorfer et al., 2019). This dataset consists of 1707 depth maps that were recorded using an Orbbec Astra depth sensor. In each sample, one of 31 people was simulating one of 15 common healthy and unhealthy sitting poses under supervision. Ground-truth 3D keypoint coordinates were obtained using a professional motion capture system. Figure 1 shows a sample from this dataset, highlighting missing and noisy data around object borders and at steep angles.

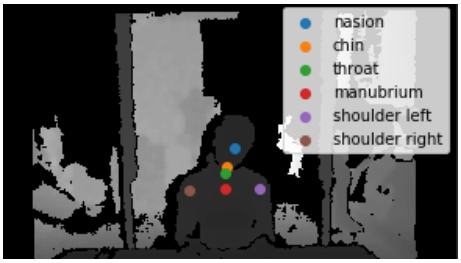

Figure 1: Visualization of a sample from the test dataset. Brighter pixels represent further distances.

3.2 Training Approaches and Datasets

We compare the following approaches to obtaining training data in terms of their performance on the common test set. This allows us to assess the suitability of depth data synthesis for solving practical problems and to compare this approach to alternatives.

Data Synthesis. One approach is to generate synthetic depth data that closely represent the task at hand. This it allows creating datasets of an arbitrary size with comparatively low effort and thus costs. However the resulting depth maps are not realistic in terms of sensor noise, clothing, and background objects, which may impact the generalization performance of trained models to real data.

We implement this approach using the synthetic dataset presented in (Pramerdorfer et al., 2019). The dataset comprises 50,000 depth maps with accurate 3D ground-truth coordinates for all keypoints considered. The depth maps were rendered from 3D models of synthetic humans in various sitting poses. These models were created using the Blender 3D modeling software (15,000 models) and include hair and different facial expressions but no clothing. For increased realism, the depth maps also depict desks, chairs, and a background object. Figure 2 shows an example.

We consider two versions of this dataset, one without sensor noise and one with simulated noise using a method based on (Xu and Cheng, 2013). This enables us to study whether noise simulation can improve the generalization performance.

Data Recording. Another approach is to record and label an own dataset that represents the problem to solve. This represents the standard approach for solving a problem in a data-driven fashion. However, depending on the problem and available resources, the amount of data obtainable this way is limited. On the other hand, the sample quality is higher than with data synthesis as the depth maps are realistic.

Figure 2: Visualization of a sample from the synthetic dataset. Brighter pixels represent further distances.

To represent this approach in our case study, we recorded 17 colleagues for a short period. During this time, the people continued their work in front of their computer screens. The sensor was an Orbbec Astra that was placed on top of the computer screens. 450 random frames were extracted from the resulting recordings and the keypoint image coordinates were marked with the help of color images that were recorded along with the depth data and registered with the depth maps. On this basis, ground-truth coordinates were defined based on depth map lookups at the individual image coordinates. We refer to this dataset as the office dataset.

Adapting a Realistic Dataset. An alternative that might be applicable depending on the task at hand is to adapt an existing realistic dataset to one's needs. In the context of this study, this means taking an existing body-pose estimation dataset and calculating missing ground-truth keypoints based on other available keypoints if possible. This has the advantage of taking less effort than recording and labeling a new dataset and can result in larger amounts of data. On the other hand, the resulting data might not reflect the task at hand accurately as adapting the labels (inferring keypoints) is not always possible without errors.

We utilize the ITOP dataset (Haque et al., 2016) to represent this approach, restricting to the subset of frontal views (22,854 samples) as these samples represent the example problem more closely than the top views. We chose this dataset because it is the largest body pose estimation dataset available. The limitations of this dataset in terms of adaptation to our problem are that it does not include labels for the nasion, chin, and throat keypoints and these keypoints cannot be inferred reliably from other existing keypoints. We represent the nasion keypoint by ITOP's head keypoint but ignore the chin and throat keypoints in the

experiments for this reason. The dataset also lacks a manubrium keypoint but tests showed that its neck keypoint agrees well with this keypoint (the paper does not state how the keypoints are defined exactly).

Adapting a Synthetic Dataset. The last approach we consider is to adapt an existing synthetic dataset for our purposes. This is less effort than creating an own synthetic dataset but has the same disadvantages as adapting existing realistic datasets.

We utilize the SURREAL dataset (Varol et al., 2017) for this purpose, which is the largest synthetic dataset available that includes depth maps. We ignore all samples in which the person is depicted from the side or behind according the the shoulder coordinates, resulting in a dataset of 259,417 samples. The dataset lacks nasion, chin, throat, and manubrium keypoints. We thus estimate the throat from the upper and lower neck keypoints and the manubrium from the left and right collarbones. The nasion and chin keypoints are not estimated from other keypoints as this is not possible without significant errors.

4 METHODS

Our pose estimation method predicts 3D camera coordinates for all keypoints considered. The core component is a CNN that predicts image coordinates and distances for all keypoints from depth map patches that depict a person. These image coordinates are then converted to camera coordinates by inverting the geometric transformations applied during patch extraction and using known camera intrinsics.

4.1 Patch Extraction

Given a depth map and a list of ground-truth keypoint coordinates, we first locate the face of the depicted person. For this purpose we estimate the face bounding box center and size based on the nasion and chin keypoints. This bounding box is then extended by a multiplicative factor to obtain a bounding box that captures both the head and upper-body of the person.

Our method thus does not include automatic face or person detection. This is to prevent face detection errors from affecting the pose estimation studies that are the focus of this paper. It would be straightforward to extend our method accordingly though, by replacing the face detection approach described in the previous paragraph with any face detector.

We then compute the median distance in the face region and threshold the depth map on this basis, setting all pixels that differ by more than 75 cm from this distance to zero. This is comparable to the cube-based segmentation approach of (Guo et al., 2017) and (Moon et al., 2018) and removes most background clutter. Finally the person region is extracted and resized to a size of 100×100 pixels.

This procedure introduces consistency between datasets, as shown in Figure 3. It also invalidates projective geometry in the sense that patches depicting people that are further away from the camera are not necessarily smaller. This in turn makes depth-scatter data augmentation (covered below) intuitive.

Figure 3: Visualizations of samples from the test (left), synthetic (center), and ITOP (right) datasets after patch extraction. This results in a similar person scale and alignment across datasets (cf. Figures 1 and 2).

4.2 Network Architecture

Our network architecture is based on ResNet-18 due to its solid performance and efficiency (He et al., 2016). We modify this architecture in two ways. First, we prepend a custom layer that replicates the channels of single-channel inputs as many times as channels expected by the first convolutional layer. This enables compatibility with models that were pre-trained on color images and thus facilitates transfer learning. Second, we replace the final global average-pooling layer with concat-pooling (Howard and Ruder, 2018), i.e. a combination of average- and max-pooling. The network ends with a linear layer with $3k$ neurons, with k being the number of keypoints.

The resulting architecture is simpler than those in related works, which perform e.g. voxel-based (Moon et al., 2018) or dense prediction of keypoint confidence maps (Cao et al., 2018). We choose a simpler architecture as (i) our primary goal is to study relative performance gains rather than outperforming existing methods, and (ii) using a simpler architecture might expose limitations in cross-dataset generalization more than more complex architectures.

4.3 Training and Validation

As in (Pramerdorfer et al., 2019), we train this network to predict keypoint image coordinates and distances as opposed to predicting camera coordinates. This has the advantage of facilitating patch extraction (which entails converting the ground-truth labels accordingly) as well as data augmentation (geometric

transformations such as random crops and image rotations can be applied to image coordinate labels easily, which is contrast to camera coordinates). Predicting distances instead of inferring them from depth maps based on the predicted image coordinates has the potential advantage of allowing the model to become robust to occlusions by e.g. people's arms.

Instead of training from scratch, we employ transfer learning of a model that was trained on ImageNet (Russakovsky et al., 2015), i.e. for classification of color images. This is for two reasons. First, it is a best practice when training on small datasets, which applies to the office dataset. Second, this avoids performance fluctuations due to network parameters being initialized randomly during training, which is important to ensure comparability of the results.

The model was pre-trained on images whose pixel values were mapped to $[0,1]$ via division by $v = 255$. This must be replicated for the depth data, which requires setting v carefully as in this cases pixel values encode distances. One consideration is that we study the cross-dataset performance and these datasets differ significantly in terms of depicted people's distances, as visible in Figure 4. One approach would be to set $v = 10$ m, which covers all datasets considered. However this lowers the contrast unnecessarily for datasets that do not cover the full distance range, which is common. We thus instead set v based on the distribution of the target datasets. Specifically, we set $v = \max(l_{99}, t_{99}) + 1$ m, with l_p and t_p denoting the p% percentiles of the training and test dataset. This ensures that all body parts of at least 99% of samples are mapped to $[0,1]$ while maximizing the possible contrast. Afterwards we normalize the samples by subtracting the mean and dividing by the standard deviation of ImageNet, as done during pre-training.

The loss function minimized during training is a weighted sum of two Huber losses (Huber, 1992) that penalize image coordinate and distance prediction errors, respectively. The weights are set such that both losses contribute roughly equally to the overall loss. We minimize this loss using the Adam optimizer with weight decay set as in (Loshchilov and Hutter, 2017).

4.4 Data Augmentation

Figure 4 shows the distance distributions of people according to their ground-truth nasion keypoints for all datasets, highlighting that these distributions are very different and that there is little overlap. On this basis, we cannot expect models to generalize well across datasets and particularly to the test set.

To overcome this problem, we put forward a training data augmentation technique called *depth-scattering*. During training and for each sample, this technique samples a random scalar from $[s_0, s_1]$, which it adds to all non-zero inputs and ground-truth distances. This ensures that the distances of samples seen during training capture both the training (and validation) set and the test set if s_0 and s_1 are set accordingly, thereby enabling the models to generalize to the latter. We set s_0 and s_1 similarly to v, namely $s_0 = \min(0, t_1 - l_5)$ and $s_1 = \max(0, t_{99} - l_{95})$.

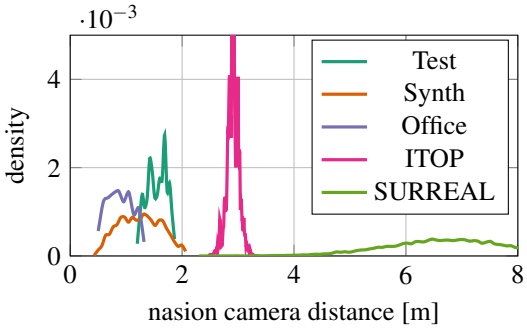

Figure 4: Person distance distributions of all datasets in terms of the ground-truth nasion keypoint (head and throat for ITOP and SURREAL, respectively).

In addition we augment the training data via random crops to 88×88 pixels and random rotation at angles up to ± 10 degrees.

5 EXPERIMENTS

All models are trained twice and the reported results are averages of both runs in order to limit the impact of random data augmentation on the results.

5.1 Impact of Training Data

We first study how the training data approach affects the test performance, with a focus on how much the performance decreases by transitioning from the different training sets to the common test dataset. The datasets are abbreviated as follows: SY is the synthetic dataset without simulated noise, SN is the same dataset with simulated noise, OF is the office dataset, IT is the ITOP dataset, and SU is SURREAL.

We reiterate that the individual datasets do not define keypoints identically, with some including no detailed definitions as all. This leads to systematic errors that cannot be avoided or corrected in a principled way. This is a compromise that often cannot be avoided when adapting existing datasets to the task at hand, as in this study. Heuristics such as subtracting offset vectors based on validation data would also

mask systematic errors due to other causes, which is why we do not apply them in this analysis.

Validation Performance. In order to establish baselines, we first compare the results on the individual validation sets and independently for each keypoint. The results are reported as the median error when predicting 3D coordinates (Euclidean distances between predictions and ground-truths) in cm.

Figure 5 visualizes these results. Missing entries are due to some datasets missing certain keypoints. The results on SY and SN are comparable, with differences attributable to randomness during training. Those on IT are better than on OF despite the former dataset being more challenging, possibly due to the small size of the latter. SU shows the worst performance despite being a synthetic dataset and the largest in size, suggesting that it is more challenging in terms of poses. The shoulders are the hardest do predict accurately, for all datasets considered.

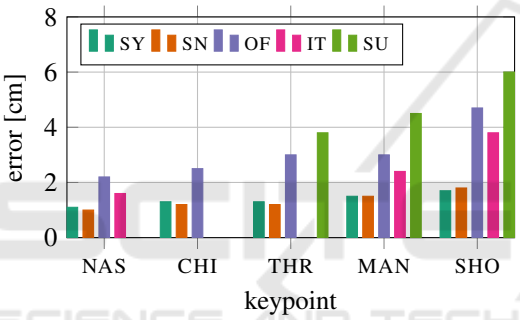

Figure 5: Validation error medians for all datasets and keypoints. NAS is the nasion, CHI the chin, THR the throat, MAN the manubrium, and SHO are the shoulders.

Test Performance. Figure 6 shows the same performance numbers for the test data, highlighting how well models trained on each dataset perform on the common test dataset. This in turn shows how suitable each approach for obtaining training data is in the context of this study. Models trained on SY and SN achieve the best overall performance on the test set. This confirms that training on synthetic depth data that accurately represents the task at hand is superior to the alternatives considered.

The models trained on SY perform consistently better than those trained on SN, i.e. the same data but with simulated sensor noise. This shows that noise simulation was ineffective in this case, either because the models are able to handle sensor noise themselves or because the simulations did not reflect the actual sensor noise characteristics well enough. We investigate this matter in more detail below.

Training on a small but realistic dataset (OF), a common approach if only limited data are available, performs significantly worse than utilizing synthetic data apart from the manubrium keypoint. A possible reason for this exception is inaccurate marker placement, as detailed below.

Training on IT, and thus the approach of adapting a larger existing dataset to the task at hand, results in a performance between the two aforementioned approaches for the manubrium and shoulder keypoints. On the other hand, the models trained on this data are unable to predict the nasion reliably. This was expected as the IT dataset has only a head keypoint, which was used in approximation in this case, and is an example for systematic errors due to limits in adapting existing datasets to other tasks.

Models trained on SU perform the worst on the test data for all keypoints available. This is again due to differences in the keypoint definitions (internal vs. surface points) but also due to prediction errors that are also apparent in the validation results.

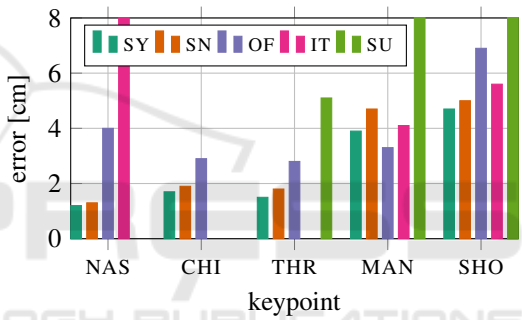

Figure 6: Test error medians for all training datasets and keypoints. NAS is the nasion, CHI the chin, THR the throat, MAN the manubrium, and SHO are the shoulders.

Generalization Gap. Comparing Figures 5 and 6 shows that the performance loss on the test data varies significantly depending on the training set, with the two synthetic datasets generalizing the best on average. This is unexpected as we assumed that transitioning from synthetic to real data would incur an additional performance penalty. The version without simulated noise (SY) generalizes consistently better than the version with simulated noise (SN), which indicates that the noise simulation method utilized does not capture the sensor noise characteristics properly.

On both synthetic datasets, the generalization gap for the manubrium and shoulder keypoints is much larger than for the other keypoints. This is likely due to a combination of the following reasons. First, realistic conditions, namely clothing and sensor noise, might affect these keypoints more than the others. Second, a visual inspection indicates that the ground-truth coordinates for these keypoints are not always perfectly accurate. This is because the markers used

for obtaining these coordinates were glued to the skin or clothing of the test subjects, which was harder to do accurately for the manubrium and shoulder keypoints than for the other keypoints. We will investigate this circumstance more closely in the future.

In summary, the results show that using synthetic depth data for training convolutional neural networks is a promising alternative to the other approaches considered, namely to collecting a limited amount of real training data as well as to adapt existing datasets that do not closely reflect the task at hand.

5.2 Ablation Studies

We next assess the importance of simulating sensor noise as well as depth-scattering in more detail.

Sensor Noise Simulation. The previous results suggest that the method for simulating sensor noise is ineffective as the test errors of models trained on SY (without simulated noise) are lower than those trained on SN (with noise). Figure 7 shows more detailed test results in the form of average precisions at 3D error thresholds up to 10 cm. Training on simulated noise increases the test errors significantly for the throat and manubrium keypoints across the threshold range, while the other changes are explainable by randomness in training. This confirms our earlier findings.

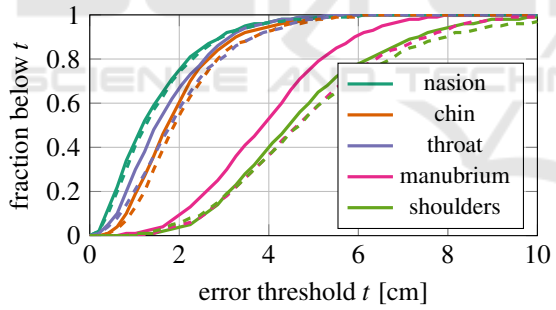

Figure 7: Average 3D keypoint precisions of models trained on SY (solid lines) and SN (dashed lines).

Depth-scattering. In order to assess the impact of depth-scattering on the test errors, we retrain the models on the SY, OF, and IT datasets without this form of data augmentation and compare mean 3D keypoint prediction errors over all keypoints.

In case of SY, disabling depth-scattering has no significant effect on the errors while for OF and IT, doing so increases the errors by over 500% in both cases. This confirms that depth-scattering is mandatory for generalization across datasets with different distance ranges but is neither beneficial nor detrimental otherwise (cf. Figure 4).

5.3 Color-based Methods

We next compare the keypoint image coordinate prediction accuracy of our best models (trained on SY) to state-of-the-art 2D pose estimators for color images, namely OpenPose (Cao et al., 2018) and AlphaPose (Fang et al., 2017). To do so, we run these detectors at their default settings on the color frames of the test set, which were recorded alongside the depth data. The depth and color frames are registered, enabling us to convert the available 3D ground-truth coordinates to image coordinates for evaluation purposes. We report PCKh scores (Andriluka et al., 2014), which measure the fraction of predictions with an error below t times the head size. The head size is estimated as the distance between the nasion and manubrium, and t is varied between 0 and 0.4.

For a fair comparison, the correct person is selected manually if multiple persons are incorrectly detected. Neither OpenPose nor AlphaPose locate the nasion, which we calculate as the center between both eye detections. OpenPose does not predict the manubrium but tests showed that its neck keypoint aligns closely with it. AlphaPose does not provide either keypoint, so we estimate the manubrium as the center between the shoulders. On this basis we restrict our comparison to the nasion, manubrium, and shoulder keypoints as these are available in all cases or can be estimated from other keypoints.

Figure 8 summarizes the results. OpenPose and AlphaPose perform almost identically. This is in contrast to findings in the literature and might be because the test dataset does not include particularly challenging poses. Our method outperforms both color-based methods at predicting nasion and manubrium keypoints. For the manubrium, this may be in part due to the keypoint estimation process. For the nasion, this is not the case as the manubrium is defined as the region between the eyes, as used for estimation. On the other hand, the color-based method achieve significantly higher scores for the shoulders.

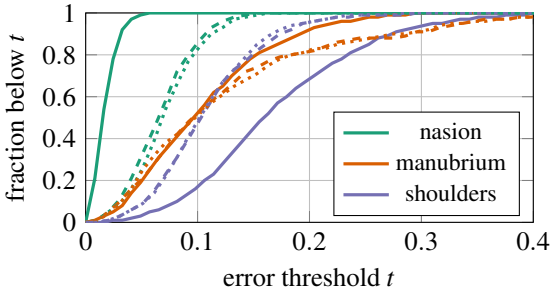

Figure 8: Keypoint image coordinate prediction performance on the test set of our method (solid lines) as well as OpenPose (dashed lines) and AlphaPose (dotted lines).

In summary, these results indicate that 2D pose estimation in depth data is possible at an accuracy similar to 2D pose estimation in color images, suggesting that depth data are suitable for this purpose in terms of the achievable accuracy. However, given the diverging results and the limited number of keypoints that are consistent across detectors, we aim to carry out more studies in the future to confirm this.

6 CONCLUSIONS

We have presented a case study on how utilizing synthetic depth data for solving a practical problem via deep learning, namely 3D human pose estimation for health care applications, compares to alternative means for acquiring training data. The results show that synthetic training data are a promising alternative particularly to acquiring own realistic data if this results in a dataset that is small by deep learning standards, despite using transfer learning. We presume that this applies for related problems such as face and person detection in depth data as well as these tasks are similar in terms of data characteristics. For the future we plan to verify this empirically and to investigate why the sensor noise simulation method employed did not lead to conclusive results. On this basis we hope to be able to develop an improved noise simulation method that helps to further reduce the generalization gap from synthetic to real data.

ACKNOWLEDGEMENTS

This work was supported by the Austrian Research Promotion Agency (FFG-855696).

REFERENCES

Andriluka, M., Pishchulin, L., Gehler, P., and Schiele, B. (2014). 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3686–3693.

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., and Sheikh, Y. (2018). OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *arXiv preprint arXiv:1812.08008*.

Chen, C.-H. and Ramanan, D. (2017). 3D Human Pose Estimation = 2D Pose Estimation + Matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7035–7043.

Fang, H.-S., Xie, S., Tai, Y.-W., and Lu, C. (2017). RMPE: Regional Multi-Person Pose Estimation. In *International Conference on Computer Vision*.

Guo, H., Wang, G., Chen, X., and Zhang, C. (2017). Towards Good Practices for Deep 3D Hand Pose Estimation. *arXiv preprint arXiv:1707.07248*.

Haque, A., Peng, B., Luo, Z., Alahi, A., Yeung, S., and Fei-Fei, L. (2016). Towards Viewpoint Invariant 3D Human Pose Estimation. In *European Conference on Computer Vision*, pages 160–177.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Howard, J. and Ruder, S. (2018). Universal Language Model Fine-Tuning for Text Classification. *arXiv preprint arXiv:1801.06146*.

Huber, P. J. (1992). Robust Estimation of a Location Parameter. In *Breakthroughs in statistics*, pages 492–518. Springer.

Loshchilov, I. and Hutter, F. (2017). Fixing Weight Decay Regularization in Adam. *CoRR*, abs/1711.05101.

Moon, G., Yong Chang, J., and Mu Lee, K. (2018). V2V-PoseNet: Voxel-to-Voxel Prediction Network for Accurate 3D Hand and Human Pose Estimation from a Single Depth Map. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5079–5088.

Pramerdorfer, C., Kampel, M., and Heering, J. (2019). 3D Upper-Body Pose Estimation and Classification for Detecting Unhealthy Sitting Postures at the Workplace. In *International Conference on Informatics and Assistive Technologies for Health-Care, Medical Support and Wellbeing*.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252.

Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., and Blake, A. (2011). Real-Time Human Pose Recognition in Parts from Single Depth Images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1297–1304. Ieee.

Sun, X., Xiao, B., Wei, F., Liang, S., and Wei, Y. (2018). Integral Human Pose Regression. In *European Conference on Computer Vision*, pages 529–545.

Toshev, A. and Szegedy, C. (2014). DeepPose: Human Pose Estimation via Deep Neural Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1653–1660.

Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M. J., Laptev, I., and Schmid, C. (2017). Learning from Synthetic Humans. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 109–117.

Xu, C. and Cheng, L. (2013). Efficient Hand Pose Estimation from a Single Depth Image. In *International Conference on Computer Vision*, pages 3456–3462.