Presented at 10th Int. Work on Assistive Computer Vision and Robotics (ACVR 2022) held in Conjunction with ECCV 2022.

Depth-based in-bed human pose estimation with synthetic dataset generation and deep keypoint estimation

Shunsuke Ochi and Jun Miura

Department of Computer Science and Engineering, Toyohashi University of Technology

Abstract. This paper describes a method of estimating the pose of a human in bed only from a single depth image. Such estimation is useful for robotic monitoring of the elderly and the disabled, where their lying posture may indicate illness. While it can address privacy and illumination issues, depth images make the pose estimation problem more challenging. We solve this problem by generating training images with cloth simulation and deep keypoint estimation. We evaluated the effectiveness of the dataset using synthetic and real test images. We also show that adding a small number of real training data improves the results.

1 Introduction

Lifestyle support is one of the promising application domains of robotic technologies. Several home service robots (e.g., Toyota's Human Support Robot [33]) have been developed and are expected to work at home in the near future. One of the tasks of such robots is *monitoring*, which is to live with and take care of the elderly or the disabled at home or in care houses, by watching their states frequently. There are many ways of monitoring, for example, activity monitoring [18, 20], health monitoring using dedicated devices [25], and contactless fatigue estimation [12, 11].

Posture is an informative cue of the state of a person, and there is a relationship between sleeping posture and health [3]. Unusual postures, such as crouching and lying with pressing the stomach, might also indicate abnormal health conditions. In robotic monitoring scenarios, persons to monitor are often sleeping in the bed, and the body is mostly or partially occluded by cloth-like objects such as blankets.

Pose estimation techniques can be used for identifying both usual and unusual postures. Image-based pose estimation is a popular research topic in computer vision, and many deep learning-based methods have been developed (e.g., [5]). Some of them use a depth image as input [28]). These methods work well when taking usual postures like standing and walking, but not for unusual postures like crouching or heavily occluded cases.

The use of depth images effectively addresses illumination variations and privacy issues. However, since depth images have less detailed features than RGB images, estimating unusual postures only from depth images is still challenging. Annotating depth images is also a tedious task. We [22, 21] previously proposed a semantic segmentation method of body part labels, which utilizes a large synthetic dataset. However, their method does not work for in-bed pose estimation. Then, We [23] extended this work to generate a depth image dataset of lying persons under blankets using a cloth simulation technique. However, this method works only for synthetic test data and still requires post-processing to convert segmentation results to the posture.

This paper further extends our previous attempts in the following two points. First, we generate real training data and rigorously analyze the effect of utilizing a combined synthetic and real dataset. Second, we adopt joint location estimation instead of body part segmentation to make it easier to estimate poses and generate a real dataset.

The rest of the paper is organized as follows. Section 2 describes related work. Section 3 describes the steps for generating a synthetic dataset using cloth simulation. Section 4 describes the experimental results using synthetic datasets. Section 5 describes the experimental results using a real dataset and analyzes the effect of additional real data for training. Section 6 concludes the paper and discusses future work.

2 Related Work

2.1 RGB Image-based human pose estimation

Human pose estimation has been one of the fundamental problems in computer vision. A large degrees of freedom of human structure and frequent occlusions sometimes make the pose estimation a challenging task. For a robust and reliable estimation, various methods have been proposed [19, 16]. Thanks to recent advances in deep learning techniques, many image-based methods have been proposed, for example, joint position estimation [5, 30] and part segmentation [24]. The joint position estimation task outputs the position of the keypoints of each joint, and the part segmentation outputs pixel-wise classification of a person's body parts, such as head, arms, and legs, from an image.

Toshev et al. proposed a method for estimating the joint positions of a person in a color image [30] by using the FLIC dataset [26] and Alexnet[13]. Oliveira et al. proposed a method that uses the PASCAL Parts dataset [6], which includes pairs of color images and human part labels, for training a Fully Convolutional Network (FCN) [17] for part segmentation.

Liu and Ostadabbas [14] developed a method of image-based in-bed posture classification using a combination of HOG and SVM. They also developed a system that utilizes infrared images with the convolutional pose machine [15].

Although image-based approaches can achieve high performance using a large amount of training data, image-based methods tend to be sensitive to appearance changes. They may also encounter privacy issues at home or in care houses. Using depth images is one way to address them.

2.2 Depth-based human pose estimation

Shotton et al. [28] developed a human pose estimation method using depth-based features with a random forest classifier. In their method, the difference in depth values between two points on the image is used as a feature value to classify to which human body part each pixel belongs. The region of each part is obtained from the pixel classification result, and then the joint locations are calculated. Vasileiadis et al. [31] proposed a pose estimation method from depth images using an articulated human model and a signed distance function. Although these methods perform well, their applicability to heavily-occluded situations is limited.

We proposed generating human depth images with pixel-wise body part labels using computer graphics and motion capture techniques [22, 21]. We have shown that a deep neural network trained with the generated images can recognize a variety of human poses in real scenes on the condition that the body regions in the depth images are correctly extracted. We [23] extended this approach to pose estimation under cloth-like objects by adopting cloth simulation technology to synthetic data generation. However, the method cannot obtain enough accuracy when applied to real data.

2.3 Sensor-based pose estimation

Pressure sensors installed in a bed can get a pressure distribution of a person lying on the bed. By analyzing the distribution (or *pressure image*), the lying posture is estimated [9, 10, 32, 29]. Deep learning-based approaches have recently been proposed to analyze pressure images. Davoodnia and Etemad [8] developed a CNN-based method for recognizing the user identity and the posture class. Clever et al. [7] developed a physics-based method to simulate human bodies in a bed, generate synthetic pressure images, and train a neural network for predicting human shape and posture. Although pressure image-based methods can also be applied to humans under cloth-like objects or heavy occlusions, a specialized bed or mattress is required.

3 Synthetic Dataset Generation

3.1 Outline of synthetic dataset generation

We use a computer graphics platform, Maya [1], for generating the dataset by following our previous steps [22, 23] and by adopting keypoint detection instead of body parts segmentation. Fig. 1 shows the outline of the data generation. We first construct a model of a human (see Fig. 2) and cloth. The human model has fourteen trackers (see Fig. 3) on its body so that the location of each joint can be extracted.

We make two types of data. One is the depth image to simulate the observation by a depth sensor. This image is generated by visualizing the cloth and rendering depth data. The depth values are normalized to $[0, 1]$. The other is a

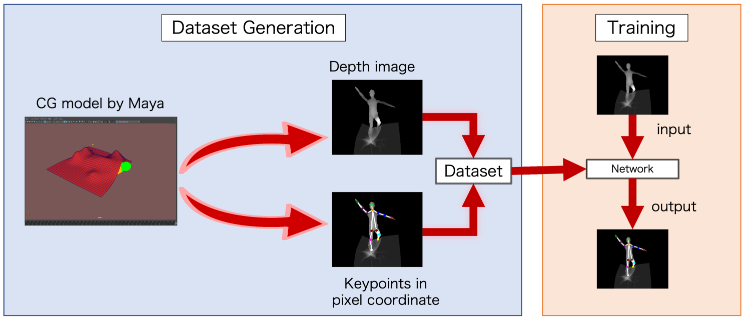


Fig. 1: Outline of dataset generation using computer graphics.

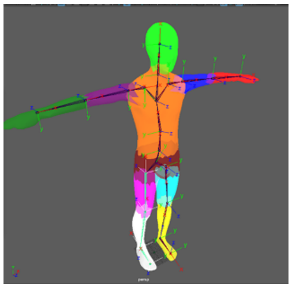


Fig. 2: Human model.

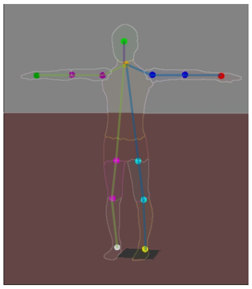


Fig. 3: Joint trackers.

list of keypoint locations. Each pair of a depth and a labeled image is an element of the generated dataset.

We use a human model generated based on [27]. There are fourteen body joints: nose, neck, left/right wrists, left/right elbows, left/right shoulders, left/right hips, left/right knees, left/right ankles. The model has a skeletal structure, and its posture can be modified by specifying joint angles.

3.2 Cloth simulation

We use nCloth [2], the cloth simulation function of Maya, for simulating humans covered by blankets. We use a fixed-sized cloth (150 [cm] × 150 [cm]) with 0 [cm] thickness. An nCloth object is represented as a dynamic mesh, characterized by parameters such as mass, friction, and stretch, compression, and bend resistance. We tested various combinations of the parameters and chose the following: 1.0 for the mass, 0.4 for the friction, and 0.4, 1.0, and 0.3 for the stretch, compression, and bend resistance, respectively. For simulation, we place a cloth 50 [cm] above the human body and make it freely fall while starting the dynamic simulation. We stop the simulation and extract the cloth surface shape when the cloth motion converges.

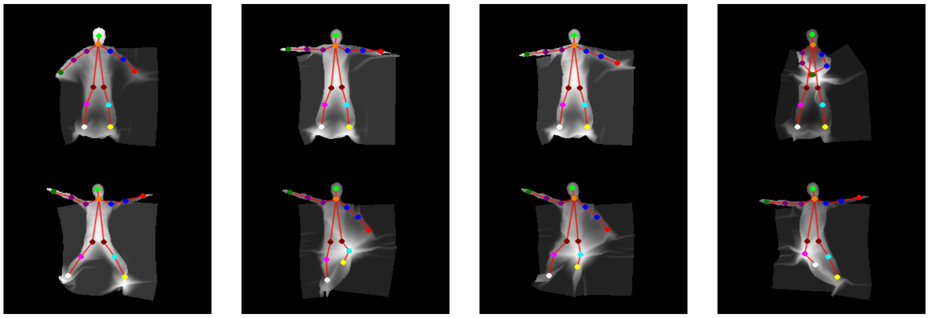


Fig. 4: Examples of human postures with cloth in dataset.

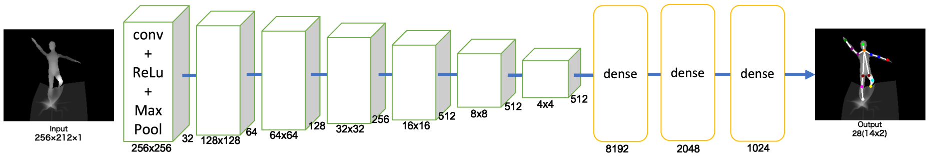


Fig. 5: Network architecture for human pose estimation.

3.3 Dataset details

The camera is set on the ceiling, looking right downward from 250 [cm] above for rendering. To make images of various lying orientations, we rotate the human body with the cloth around the vertical axis at 5 [deg] intervals. The dataset is generated with 67 human poses and 72 different angles. Fig. 4 shows the examples of a human model with the cloth. The size of the images is scaled to 212 × 256. Keypoints are specified by their normalized pixel coordinate values, where the upper-left corner is (0,0), and the bottom-right corner is (1,1). We split the dataset into 4,248 and 576 for training and testing, respectively. We also generated another dataset without the cloth for comparison purposes.

4 Exepriments with Synthetic Dataset

4.1 Training

Fig. 5 depicts the CNN-based network architecture used. The network takes a single 256 × 212 × 1 depth image as input and outputs a 28-dimensional tensor, which is composed of the locations of fourteen joint keypoints in the pixel coordinates. The network is trained with two different datasets, with-cloth and without-cloth, to examine the effect of cloth on the estimation accuracy. The training condition is as follows: GPU: Nvidia Titan X, framework: TensorFlow, optimizer: Adam, learning rate: 10^{-4}, batch size: 32, epochs: 100.

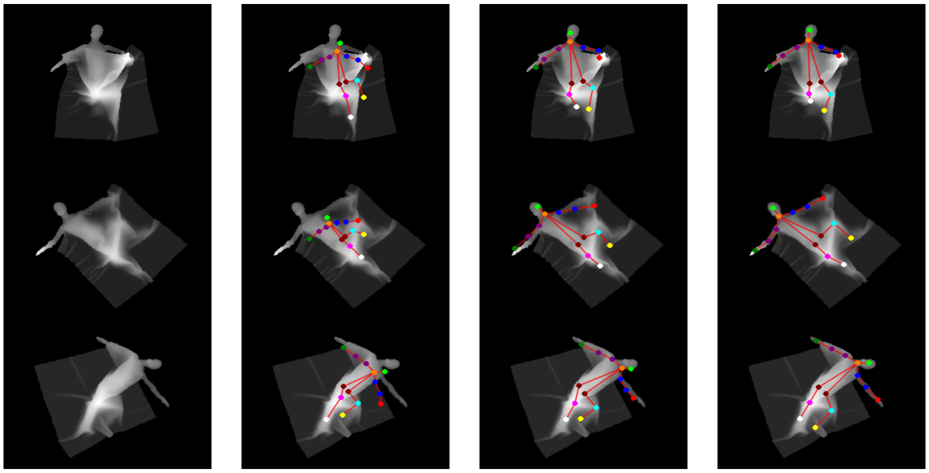

(a) input (b) output (W/O) (c) output (W) (d) ground truth

Fig. 6: Estimation results for synthetic data. (a) Input. (b) Output with the model trained with the without-cloth (W/O) dataset. (c) Ouput with the model trained with the with-cloth (W) dataset. (d) Ground truth.

4.2 Evaluation metrics

We use two evaluation metrics: Root Mean Squared Error (RMSE) and Percentage of Correct Keypoints (PCK) [4]. RMSE is a metric that indicates the distance between the estimated and the ground truth keypoint locations, defined by:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \|\mu_i - \bar{\mu}_i\|^2}, \tag{1}$$

where N is the number of keypoints (i.e., fourteen), μ_i and $\bar{\mu}_i$ are the ground truth and the estimated location of the ith keypoint, respectively.

PCK is a metric that indicates the percentage of correctly estimated keypoints, defined by:

$$PCK = CEK/N, \tag{2}$$

$$CEK = \sum_{i=1}^{N} K, \quad \begin{cases} K = 1, \text{where} \sqrt{\|\mu_i - \bar{\mu}_i\|^2} < \ \epsilon \\ K = 0, \text{where} \sqrt{\|\mu_i - \bar{\mu}_i\|^2} \geq \ \epsilon \end{cases} \quad , \tag{3}$$

where ϵ is the threshold to judge the correctness. A half of the diagonal length of the ground-truth head bouding box is commonly used as the threshold; the metric using that threshold is called PCKh@0.5.

4.3 Experimental results

Fig. 6 shows the estimation results when the with-cloth test dataset is supplied to the two models; one is trained with the *without-cloth dataset* and the other with the *with-cloth dataset*. The latter exhibits better results than the former and outputs results close to the ground truth even though the cloth occludes most of the body surface. We also compare the models in terms of RMSE and PCKh@0.5. The averaged metrics for the *without-cloth* model and the *with-cloth* model are 16.58 [pix] and 4.55 [pix] in RMSE and 0.260 and 0.929 in PCKh@0.5, respectively. These results show the effectiveness of the dataset generated with cloth simulation.

5 Experiments wit Real Scene Dataset

5.1 Aquisition of real scene dataset

Fig. 7 shows an overview of the real scene experiment. The data for the real-world evaluation was obtained by an Azure Kinect RGB-D sensor installed on the ceiling. For the cloth, we used a curtain cloth with a thickness of 0.008 [cm]. A person lay on the floor and took various poses. We took a pair of images with and without the cloth for a pose. We applied OpenPose v1.7.0 to the RGB images taken without the cloth to obtain keypoints as ground truth. Since a raw depth image includes many noise pixels, we preprocess the images so that only depth data within the correct range (between 0 [m] and 2.45 [m] (camera height)) exist. Fifty pairs of depth images were taken while a person was changing posture. We scaled and cropped the captured images to 256 × 212 so that they have the same angle of view and the image size as the synthetic images. The dataset was then augmented by rotating the images by 360 [deg] with 5 [deg] intervals, which provides 72 images for each posture. We have thus 3,600 images in total. The images are split into 1,440 and 2,160 for training and testing. Fig. 8 shows four example pairs of preprocessed depth images. We also modified the keypoint locations of synthetic data so that they match those of OpenPose outputs.

5.2 Experimental results

Testing the models trained only with synthetic data Fig. 9 shows pose estimation results for the model trained with the without-cloth synthetic dataset tested against without-cloth real test data. The results look reasonable, although some joints, such as wrists and ankles, suffer from significant errors. The averaged metrics are 11.565 [pix] in RMSE and 0.464 in PCKh@0.5. Even though the model is trained only with synthetic data, it can robustly estimate the pose when a cloth does not cover the human body.

Fig. 10 shows pose estimation results for the model trained with the with-cloth dataset tested against with-cloth real data. The averaged metrics are 25.622 [pix] in RMSE and 0.124 in PCKh@0.5. The estimation accuracy is very low, possibly because the shape of the cloth is significantly different between synthetic and real data.

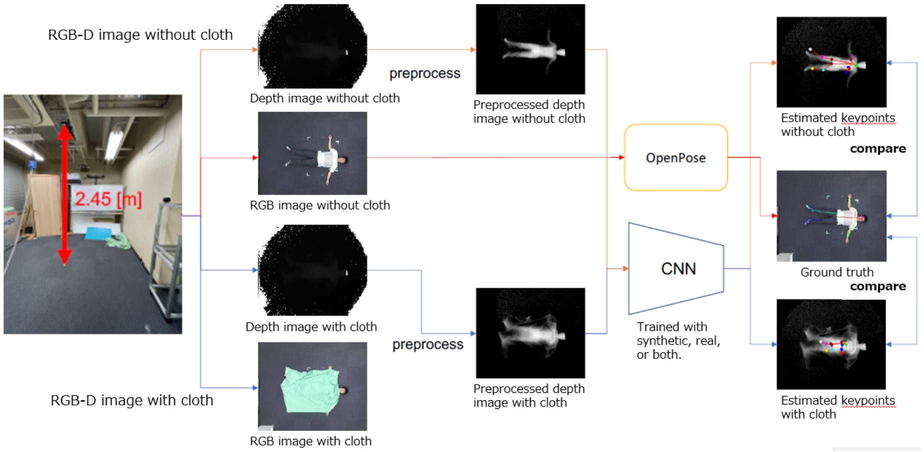

RGB-D image without cloth
2.45 [m]
Depth image without cloth
preprocess
Preprocessed depth image without cloth
RGB image without cloth
OpenPose
Depth image with cloth
preprocess
Preprocessed depth image with cloth
RGB-D image with cloth
RGB image with cloth
CNN
Trained with synthetic, real, or both.
Estimated keypoints without cloth
compare
Ground truth
compare
Estimated keypoints with cloth

Fig. 7: Overview of human pose estimation experiments in real scene.

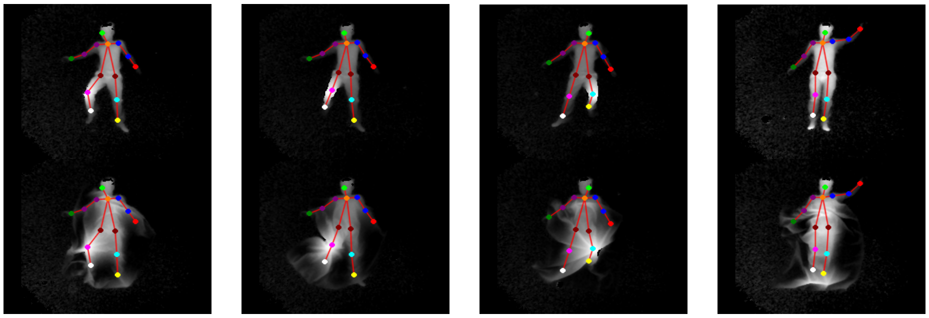
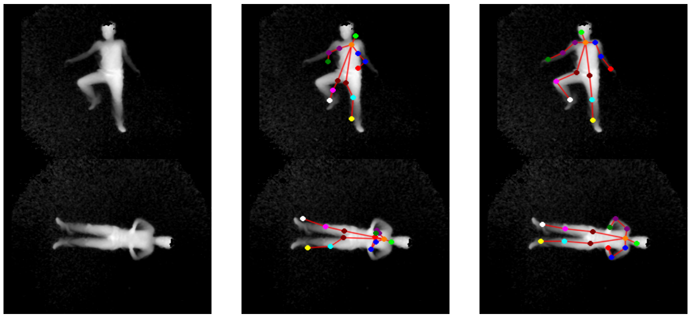


Fig. 8: Examples of preprocessed real scene depth image. Top row: data without cloth, bottom row: data with cloth. Keypoints are superimposed as ground truth.

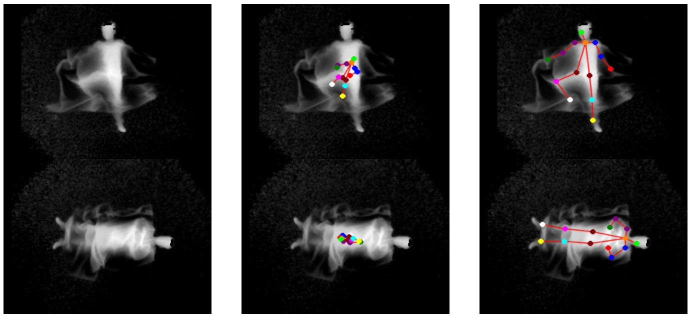
Training with both synthetic and real data The model trained only with synthetic data cannot robustly estimate the human pose in with-cloth situations due to the reality gap. On the other hand, obtaining lots of real data is costly, and the variety of poses may be restricted. Combining synthetic and real data would be a promising way of solving those issues. Thus, we investigate the effectiveness of such a combination in the training data.

Fig. 11 shows the comparison results for the model without and with additional synthetic data for training. The former model uses only real data for five poses, while the latter uses those data and additional synthetic data for 59 poses. The figure shows that adding synthetic data improves the estimation accuracy by supplementing the lack of pose variations. Fig. 12 shows the results with different numbers of additional real data. As the number of real data increases, the estimation results are improved.



(a) input (b) output (c) ground truth

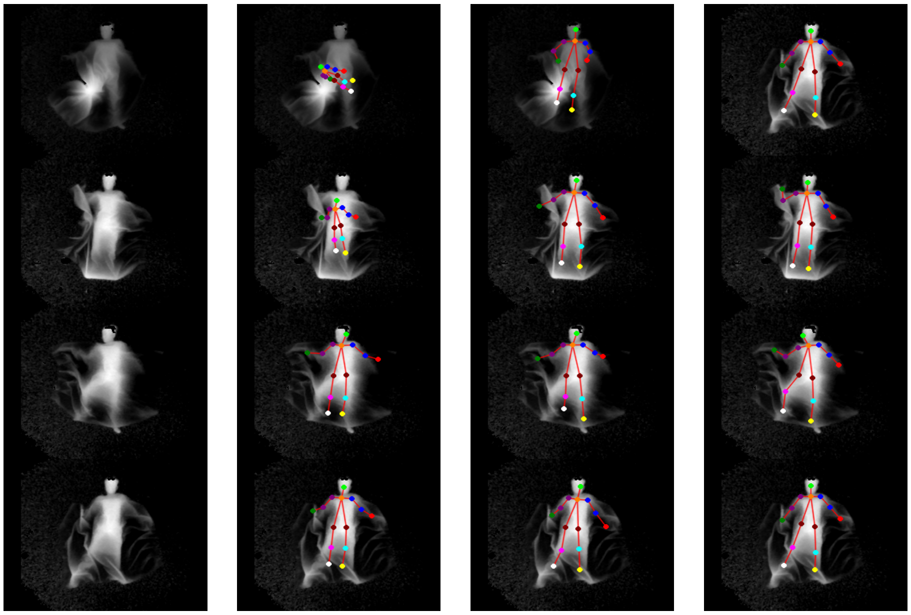
Fig. 9: Estimation results in real scene (without-cloth). Training dataset: Synthetic dataset without cloth. Test dataset: Real data without cloth.



(a) input (b) output (c) ground truth

Fig. 10: Estimation results in real scene (with cloth). Training dataset: Synthetic dataset with cloth. Test dataset: Real data with cloth.

Table 1 summarizes the RMSEs and PCKs for various training datasets. From the table, we can see that introducing or increasing the number of real data in the training dataset improves the performance. For example, from lines 1 to 4, real data are effective compared to synthetic data, even if the number of real data is relatively small; this is probably because the variation of inputs is not very large in our current setting. On the other hand, from pairs of real data only and real plus synthetic (lines 4 and 7, for example), synthetic data are also useful when combined with real data. An interesting observation is that the combination of a small number of real data and a large synthetic dataset (line 5) shows comparable performance to a large number of real data (line 4); the synthetic dataset seems to supplement the lack of pose variations in the real



(a) input (b) real data only (c) real + synthetic (d) ground truth

Fig. 11: Estimation results of models with and without synthetic data against the real test data.

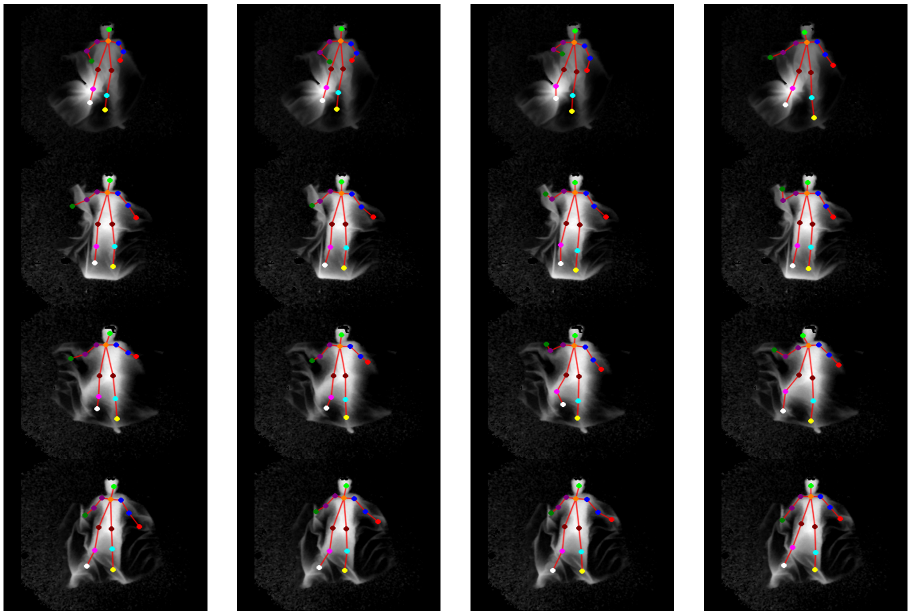
data. This result suggests an approach to reducing the cost of generating a real dataset by effectively utilizing synthetic data.

Table 1: RMSE and PCKh@0.5 for various training data.

No.	Training Dataset	RMSE	PCKh@0.5
1	synthetic dataset only	26.976	0.167
2	real data only (5 poses)	15.773	0.508
3	real data only (10 poses)	13.225	0.636
4	real data only (20 poses)	8.570	0.789
5	real data (5 poses) + synthetic dataset	8.760	0.738
6	real data (10 poses) + synthetic dataset	7.729	0.800
7	real data (20 poses) + synthetic dataset	**7.061**	**0.841**

6 Conclusions and Discussion

This paper described a method of estimating the pose of humans under cloth-like objects such as blankets. We use depth images to avoid sensitivity to illumination



(a) add 5 real poses (b) add 10 real poses (c) add 20 real poses (d) ground truth

Fig. 12: Estimation results of models with synthetic data and different numbers of real data against the real test data with cloth.

conditions and privacy concerns. We need to have a large dataset for training to adopt depth images for pose estimation. We thus utilize a cloth deformation simulation for generating pairs of the depth image of a human under a blanket and the list of joint locations. We showed the usefulness of cloth simulation-based data generation for pose estimation using synthetic test data. However, using only synthetic data for training is not enough for pose estimation in real scenes. Therefore, we analyzed the effect of combining real and synthetic data. The analysis shows that the combination is better than real data-only or synthetic data-only cases. We also showed that a small number of real data combined with a large synthetic dataset provides a good balance of the data generation cost and the estimation performance.

Further improvements are needed to apply the proposed approach to real application scenarios. It is necessary to increase the variety of synthetic data to cope with a more variety of scenes. Possible ways to increase the variation are: using human models of various body shapes and dimensions, using various types of cloth objects with different cloth parameters such as thickness and stiffness, and adding more postures. Several data augmentation techniques can also be adopted. It is also necessary to develop a method of abnormal posture detection, as our ultimate goal is to develop a monitoring robot that can detect persons in physically critical situations. Therefore, mapping from a posture to a physical state will be necessary. Not a single posture data but a time series of postures could be more informative for that purpose.

References

1. Maya, http://www.autodesk.com/products/maya/overview/
2. Maya ncloth, https://knowledge.autodesk.com/support/maya/learn-explore/caas/CloudHelp/cloudhelp/2018/ENU/Maya-CharEffEnvBuild/files/GUID-ED791F1C-8412-4785-829F-9925F2604E8A-htm.html
3. Nation sleep foundation, http://sleepfoundation.org
4. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
5. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of 2017 IEEE Conf. on Computer Vision and Pattern Recognition (2017)
6. Chen, X., Mottaghi, R., Liu, X., Fidler, S., Urtasun, R., Yuille, A.: Detect what you can: Detecting and representing objects using holistic models and body parts. In: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2014)
7. Clever, H., Erickson, Z., Kapusta, A., Turk, G., Liu, C., Kemp, C.: Bodies at rest: 3d human pose and shape estimation from a pressure image using synthetic data. In: Proceedings of 2020 IEEE Conference on Computer Vision and Pattern Recognition (2020)
8. Davoodnia, V., Etemad, A.: Identity and posture recognition in smart beds with deep multitask learning. In: Proceedings of 2019 IEEE Int. Conf. on Systems, Man, and Cybernetics (2019)
9. Harada, T., Mori, T., Nishida, Y., Yoshimi, T., Sato, T.: Body parts positions and posture estimation system based on pressure distribution image. In: Proceedings of 1999 IEEE Int. Conf. on Robotics and Automation (1999)
10. Harada, T., Sato, T., Mori, T.: Pressure distribution image based human motion tracking system using skeleton and surface integration model. In: Proceedings of 2001 IEEE Int. Conf. on Robotics and Automation (2001)
11. Hasegawa, M., Hayashi, K., Miura, J.: Fatigue estimation using facial expression features and remote-ppg signal. In: Proceedings of 2019 IEEE Int. Conf. on Robot and Human Interactive Communication (2019)
12. Huang, R.Y., Dung, L.R.: Measurement of heart rate variability using off-the-shelf smart phones. Biomedical Engineering Online **15**(1) (2016), 16 pages
13. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (NIPS) (2012)
14. Liu, S., Ostadabbas, S.: A vision-based system for in-bed posture tracking. In: Proceedings of the 5th Int. Workshop on Assistive Computer Vision and Robotics (2017)
15. Liu, S., Yin, Y., Ostadabbas, S.: In-bed pose estimation: Deep learning with shallow dataset. IEEE J. of Translational Engineering in Health and Medicine (2019)
16. Liu, Z., Zhu, J., Bu, J., Chen, C.: A survey of human pose estimation: the body parts parsing based methods. J. of Visual Communication and Image Representation **32**, 10–19 (2015)

17. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2015)
18. Mikic, I., Huang, K., Trivedi, M.: Activity monitoring and summarization for an intelligent meeting room. In: Proceedings of IEEE Workshop on Human Motion (2000)
19. Moeslund, T., Hilton, A., Krüger, V.: A survey of advances in vision-based human motion capture and analysis. Computer Vision and Image Understanding **104**, 90–126 (2006)
20. Mori, T., Tominaga, S., Noguchi, H., Shimoasaka, M., Fukui, R., Sato, T.: Behavior prediction from trajectories in a house by estimating transition model using stay points. In: Proceedings of IEEE/RSJ Int. Conf. on Intelligent Robots and Systems. pp. 3419–3425 (2011)
21. Nishi, K., Demura, M., Miura, J., Oishi, S.: Use of thermal point cloud for thermal comfort measurement and human pose estimation in robotic monitoring. In: Proceedings of 5th Int. Workshop on Assistive Computer Vision and Robotics (2017)
22. Nishi, K., Miura, J.: Generation of human depth images with body part labels for complex human pose recognition. Pattern Recognition **71**, 402–413 (2017)
23. Ochi, S., Miura, J.: Human pose recognition uder cloth-like objects from depth images using a synthetic image dataset with cloth simulation. In: Proceedings of 2021 IEEE/SICE Int. Symp. on System Integration (2021)
24. Oliveira, G., Valada, A., Bollen, C., Burgard, W., Brox, T.: Deep learning for human part discovery in images. In: Proceedings of 2016 IEEE Int. Conf. on Robotics and Automation (2016)
25. Pantelopoulos, A., Bourbakis, N.: A survey on wearable sensor-based systems for health monitoring and prognosis. IEEE Trans. on Systems, Man, and Cybernetics Part C: Applications and Reviews **40**(1), 1–12 (2010)
26. Sapp, B., Taskar, B.: Modec: Multimodal decomposable models for human pose estimation. In: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2013)
27. Shinzaki, M., Iwashita, Y., Kurazume, R., Ogawara, K.: Gait-based person identification method using shaodow biometrics for robustness to changes in the walking direction. In: Proceedings of 2015 IEEE Winter Conf. on Applications of Computer Vision. pp. 670–677 (2015)
28. Shotton, J., Sharp, T., Kipman, A., Fitzgiboon, A., Finocchio, M., Blake, A., Cook, M., Moore, R.: Real-time human pose recognition in parts from single depth images. Communications of the ACM **56**(1), 116–124 (2013)
29. Sun, Q., Gonzalez, E., Sun, Y.: On bed posture recognition with pressure sensor array system. In: 2016 IEEE SENSORS (2016)
30. Toshev, Z., Szegedy, C.: Deeppose: Human pose estimation via deep neural networks. In: Proceedings of 2014 IEEE Conf. on Computer Vision and Pattern Recognition. pp. 1653–1660 (2014)
31. Vasileiadis, M., Malassiotis, S., Giakoumis, D., Bouganis, C.S., Tzovaras, D.: Robust human pose tracking for realistic service robot applications. In: Proceedings of the 5th Int. Workshop on Assistive Computer Vision and Robotics (2017)
32. Xu, X., Lin, F., Wang, A., Song, C., Hu, Y., Xu, W.: On-bed sleep posture recognition based on body-earth mover's distance. In: Proceedings of 2015 Biomedical Circuits and Systems Conf. (2015)
33. Yamamoto, T., Terada, K., Ochiai, A., Saito, F., Asahara, A., Murase, K.: Development of human support robot as the research platform of a domestic mobile manipulator. ROBOMECH Journal **6**(1) (2019)