Pattern Recognition 000 (2017) 1–12

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Generation of human depth images with body part labels for complex human pose recognition

K. Nishi*, J. Miura

Department of Computer Science and Engineering, Toyohashi University of Technology, Toyohashi, Aichi, 441–8580, Japan

ARTICLE INFO

Article history:
Received 14 December 2016
Revised 30 April 2017
Accepted 1 June 2017
Available online xxx

Keywords:
Human depth image
Human pose estimation
Convolutional neural network
Care robot

ABSTRACT

This paper describes an efficient generation of large-scale dataset of human depth images with body part labels. The size of image datasets has recently been increasingly important as it is shown to be strongly related to the performance of learning-based classifiers. In human pose recognition, many datasets for ordinary poses like standing, walking, and doing gestures have already been developed and effectively utilized. However, those for unusual ones like lying fainted and crouching do not exist. Pose recognition for such cases may have a large potential applicability to various assistive scenarios. Moreover, locating each body part could also be important for an accurate care and diagnosis or anomaly detection. We therefore develop a method of generating body part-annotated depth images in various body shapes and poses, which are handled by a flexible human body model and a motion capture system, respectively. We constructed a dataset of 10,076 images with eight body types for various sitting poses. The effectiveness of generated dataset is verified by part labeling tasks with a fully convolutional network (FCN) for synthetic and real test data.

1. Introduction

The percentage of the elderly increases every year in the world [1] and developed countries are going to suffer from so-called aging society; it is reported that the population of the over-sixty will be one third of the total population in 2050 and this is recognized as one of the serious social issues. Assistive technologies are expected to support the elderly in various application scenarios. One possible technological application is a robot which takes care of people in care houses. Human pose recognition is one of the necessary functions of such robots, which contributes to an accurate care and diagnosis or anomaly detection.

Many human pose estimation methods have been proposed. Felzenszwalb et al. [2] dealt with an image-based human pose estimation using a pictorial structure representation [3], in which a whole body is represented as a collection of parts with their deformable geometrical relationships. Many improvements to this approach have then been proposed. To achieve a better performance, Ramanan et al. [4] improved the accuracy of part detection and Ferrari et al. [5] limited the search area using GrabCut [6].

As low-cost depth sensors like RGB-D cameras are developed, Shotton et al. [7] developed a method of estimating human poses in a depth image. The method first assigns body part labels to each pixel by using a simple depth difference between two points as a feature, and adopting a random forest classifier. It then estimates the pose of every part based on the assigned labels. Foreground/background separation is easily handled by using depth data. These previous works basically deal with pose estimation in ordinary poses.

In actual applications, unusual poses, such as lying and crouching, must also be considered. Ardiyanto et al. [8] applied a human pose estimation to a fallen person monitoring and rescue scenario. Their system continuously tracks the skeleton of a person using an environmental RGB-D camera and can therefore recognize the pose even after falling; such environmental cameras need to be installed in advance. Suppose a situation that a mobile service robot patrols a residence or a nursing home to see if any emergency situation occurs. Without environmental cameras, the robot has to recognize the human state including his/her pose only using onboard sensors. Therefore a pose estimation method for such applications must be able to estimate unusual poses. This is still a challenging problem which has not been fully solved by existing approaches. Wang et al. [9] improved the method by Felzenszwalb et al. [2] in human region detection to cope with lying person pose estimation using a color image. Although the method shows a good performance, it might be weak in the situation where foreground/background separation is difficult due to, for example, a

* Corresponding author.
E-mail addresses: nishi@aisl.cs.tut.ac.jp (K. Nishi), jun.miura@tut.jp (J. Miura).

http://dx.doi.org/10.1016/j.patcog.2017.06.006
0031-3203/

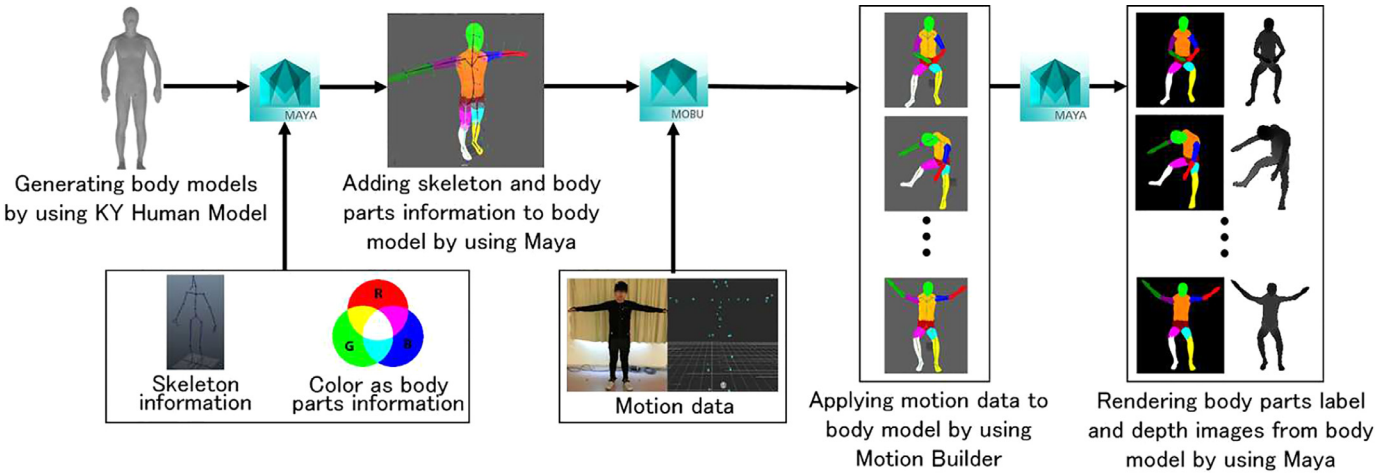

Fig. 1. Outline of generating depth images with body part labels.

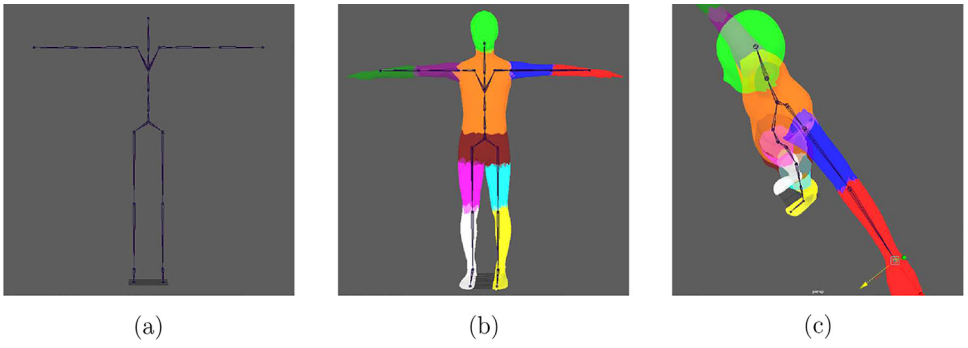

Fig. 2. Attaching skeleton information to the model.

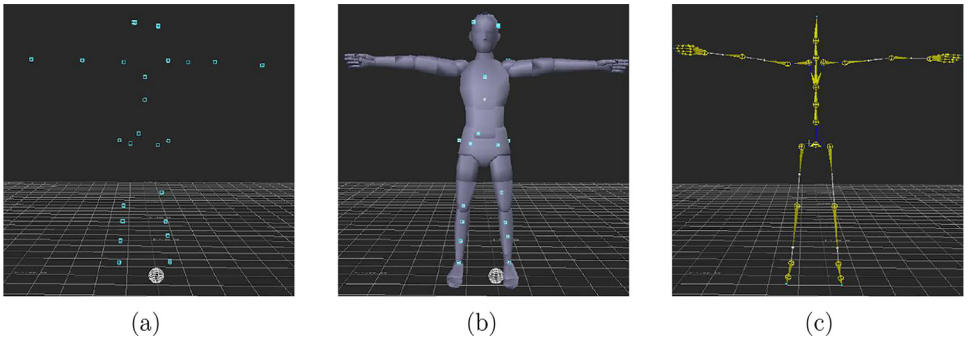

Fig. 3. Converting VICON measurements to skeleton motion data.

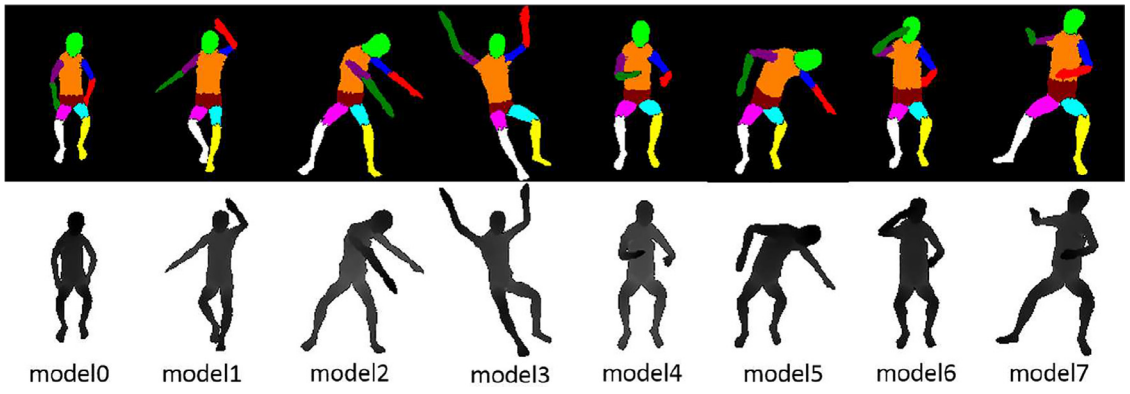

Fig. 4. Generated depth images with body parts labels. **First row:** generated label images. **Second row:** generated depth images.

Please cite this article as: K. Nishi, J. Miura, Generation of human depth images with body part labels for complex human pose recognition, Pattern Recognition (2017), http://dx.doi.org/10.1016/j.patcog.2017.06.006

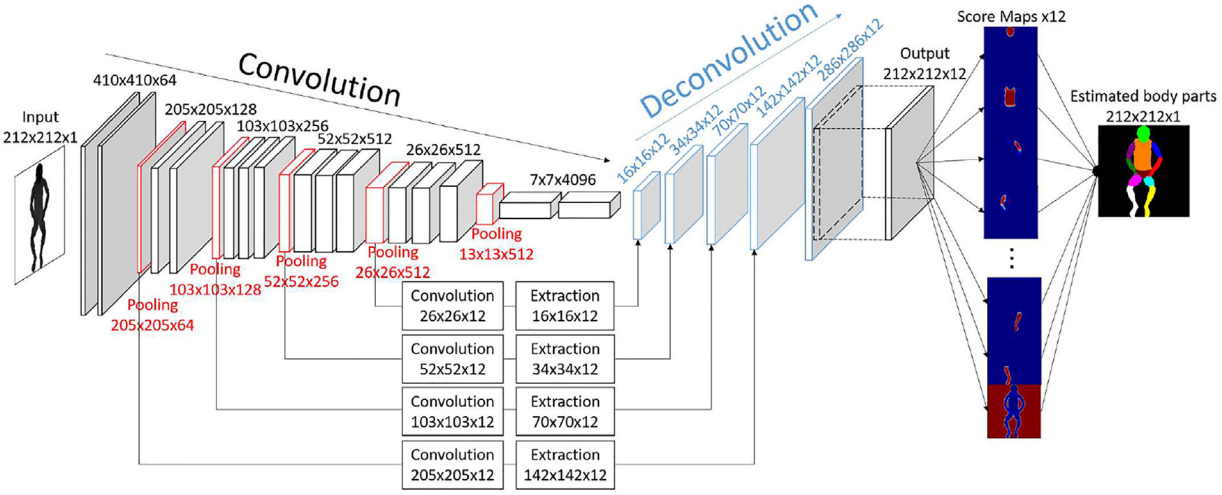

Fig. 5. Network architecture.

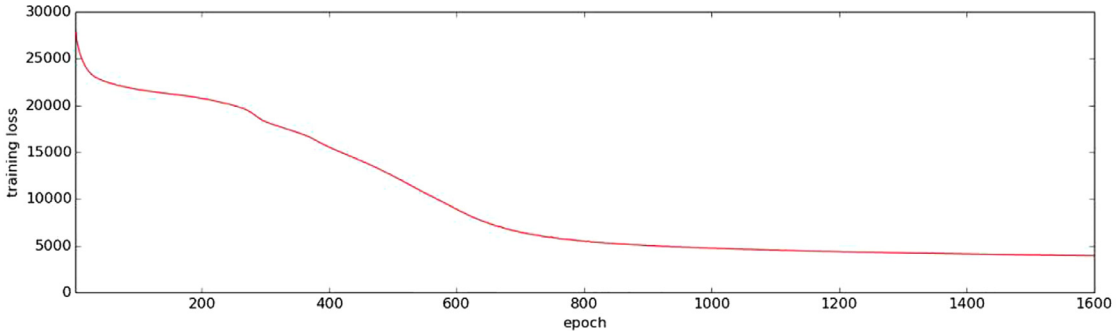

Fig. 6. The change of the training loss.

bad illumination condition and where more complex poses with self-occlusions occur.

Convolutional neural networks (CNNs) [10] have recently been very successful in various recognition tasks including human pose estimation. Pose estimation methods using CNNs are divided into two types. One is to directly estimate hunan joint positions. Toshev et al. [11] take this approach and use FLIC dataset [12] which provides human color images with joint positions. The other is to estimate the pose by classifying each pixels into body parts, as in the case of Shotton et al. [7]. Oliveira et al. [13] take this approach and train a fully convolutional network (FCN) [14] using the PASCAL Parts dataset [15] which provides human color images with body part labels. Although these works exhibit good performances, they rely on color images and may be sensitive to changes illumination, clothing, and skin color. They could also face a privacy issue.

Use of depth images is a promising alternative to solve these problems. However, this leads another big problem, that is, to construct a large dataset of annotated depth images. Nishi and Miura [16] generated a set of depth images with head position annotation for several lying poses from omnidirectional viewpoints using a large rotation table and an RGB-D camera. This approach can be applicable only to a small-sized dataset generation. We can use annotation tools like LabelMe [17,18] for color images, but a similar approach is difficult to apply to depth images. Skeleton tracking techniques [19,20] could be a possible way but these are applicable only to normal poses but not to unusual poses under consideration.

Since the annotating real depth images is difficult, we adopt computer modeling and computer graphics techniques for generating annotated depth images [7]. The issues are then how to construct human models with various body shapes and how to make the models take various poses. Manually producing such variations is extremely hard when constructing a large-scale dataset. Therefore we propose a novel approach that combines a flexible, parameterized body model, a motion capture system, and computer graphics tools in order to generate a large number of body part-annotated depth images efficiently. We evaluate the constructed dataset by conducting body part labeling experiments using an FCN for synthetic and real depth images.

The rest of the paper is organized as follows. Section 2 describes the detailed procedure of dataset generation. Section 3 explains the FCN that we used for evaluation. Section 4 describes experimental results to show the effectiveness of the dataset. Section 5 concludes the paper and discusses future work.

2. Data generation

Fig. 1 shows the outline of the proposed dataset generation method. The first step is to generate human body models. We use KY Human Model [21] that can deal with various body shapes. Other human models [22,23] can be used. Since this model has only shape information, at the second step, we attach part labels and skeleton information, for generating annotated depth images and for controlling the pose with joint angles, respectively. The third step is to collect human motion data using a VICON motion capture system [24] and to apply them to the body models. The last step is to generate human depth images with body part labels.

We use Maya [25] for the second and the last step, and MotionBuilder [26] for the third step.

Please cite this article as: K. Nishi, J. Miura, Generation of human depth images with body part labels for complex human pose recognition, Pattern Recognition (2017), http://dx.doi.org/10.1016/j.patcog.2017.06.006

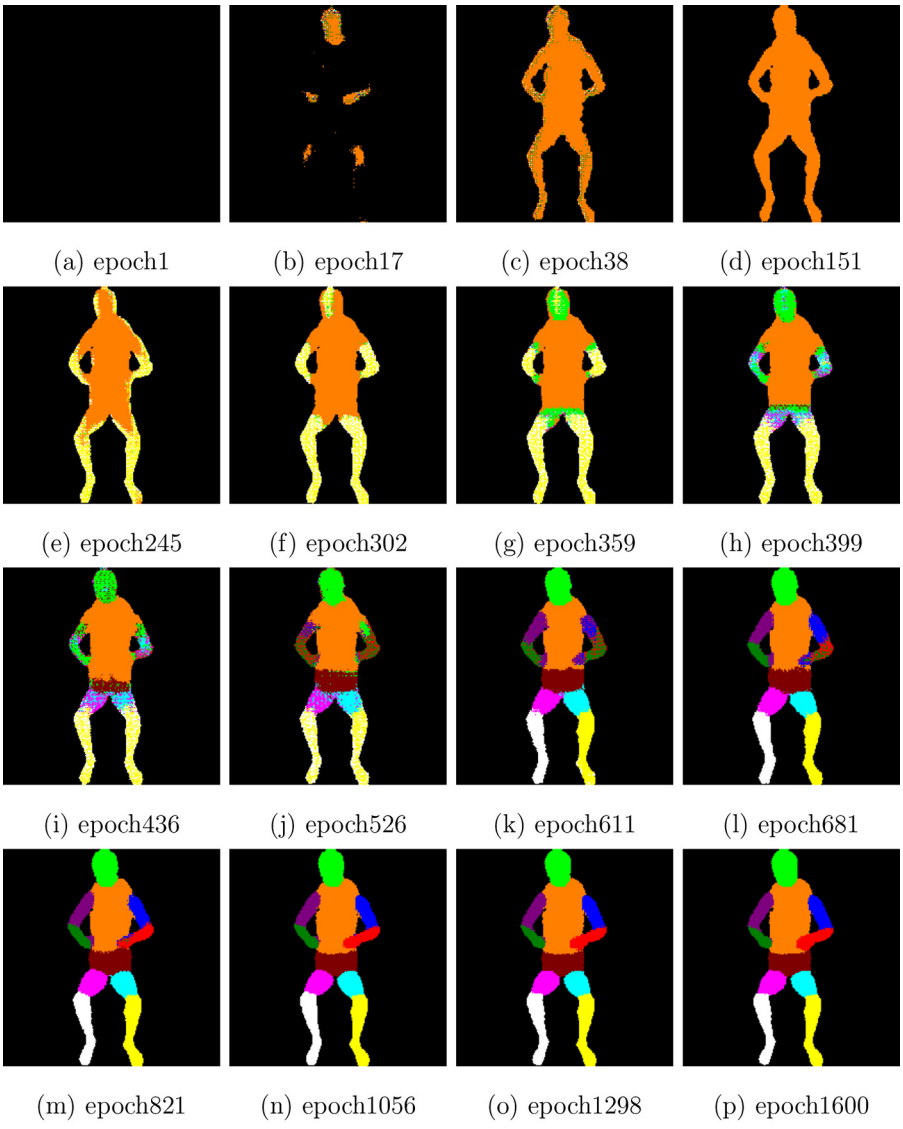

Fig. 7. Change of the part labeling results through training.

Table 1
Parameters of the generated human body models.

Model No.	Height (cm)	Weight (kg)	Waist breadth (cm)	Param 1st	Param 4th
0	149.2	36.6	22.7	120.0	0.0
1	155.3	45.0	24.2	80.0	0.0
2	172.6	51.9	23.7	0.0	40.0
3	171.4	56.8	25.2	0.0	20.0
4	169.6	61.7	26.6	0.0	0.0
5	167.7	66.6	27.7	0.0	-20.0
6	166.1	71.5	29.0	0.0	-40.0
7	182.8	78.4	28.8	-80.0	0.0

2.1. Human model generation

The variety of data is crucial for learning-based approaches. We used KY Human Model [21] for generating a variety of human body models. KY Human Model was constructed by choosing 17 out of 49 human body data in the AIST/HQL database [27], and analyzing them using PCA (principal component analysis). The constructed KY Human Model has eleven parameters to adjust for changing the body shape. In this paper, we adjusted the first and the fourth parameter which mainly affect the height and the width of the body, respectively, and leave the other nine parameters be zero (i.e., the mean value). Table 1 shows the parameter pairs and the corresponding body dimensions for the eight models used in this paper.

2.2. Adding body part labels and skeleton information to the generated shape model

The eight models mentioned above only have shape data. We thus attach part labels and skeleton information to them. Part labels are attached as follows. We consider the following eleven parts: head, torso, left/right upper arm, left/right forearm, hip, left/right upper leg, left/right lower leg. These labels are repre-

Please cite this article as: K. Nishi, J. Miura, Generation of human depth images with body part labels for complex human pose recognition, Pattern Recognition (2017), http://dx.doi.org/10.1016/j.patcog.2017.06.006

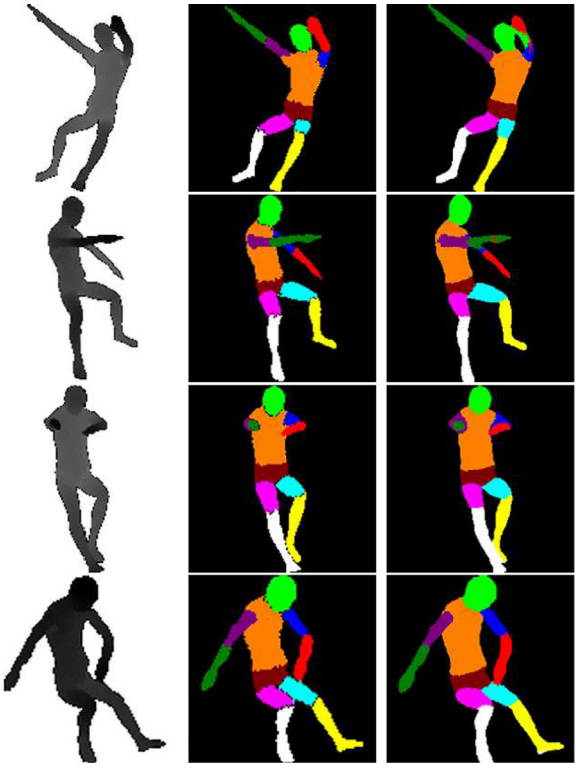

Fig. 8. Result for synthetic data. **First column:** generated synthetic depth images for test. **Second column:** generated label images. **Third column:** labeling results.

sented by respective distinctive colors in the model. When attaching a color to a body part using Maya, we choose SurfaceShader material whose rendered appearance is constant irrespective of illumination conditions. These colors in a rendered image are converted to label ID's to generate a labeled image.

Skeleton information is attached as follows. For changing the pose, Maya needs the sizes and the positions of the following parts as a skeleton: head, neck, spine, hip, shoulders, upper arms, forearms, wrists, upper legs, lower legs. These parts are defined by the joint positions. We first define a rough skeleton model (see Fig. 2(a)) and put it on the human model (see Fig. 2(b)). Then, we manually adjust joint positions so that they matches with those in the human model. Fig. 2(c) shows a scene of adjusting the left wrist position.

2.3. Adding recorded motion data

Generating natural pose data by manually adjusting joint angles is very difficult. We thus take a more intuitive approach. That is, we use VICON motion capture system [24] to collect a large number of natural poses and give them to the human body model. Our VICON system can track ball markers at 50 fps and export the motion of each marker as a sequence of 3D positions. Fig. 3(a) shows a snapshot of the sequence.

We use MotionBuilder [26] to convert marker positions at a time to joint angles in the human model. For this conversion, we give the marker positions in the human body to MotionBuilder. Fig. 3(b) and (c) show the marker positions on the human model and the corresponding skeleton model, respectively. Since this skeleton model is already attached to the human model, we can generate any pose data by actually taking that pose.

2.4. Generating depth images with part labels

The steps explained above produce a set of labeled human models with various poses. We then render the models also using Maya. The viewpoint is set at the pose of a real camera on the top of our robot. Fig. 4 shows examples of generated images; each pair of a color-labeled image and a depth image corresponds to a human model and a pose. Each model number corresponds to that in Table 1. We can see that an enough variety of models and poses can be generated.

In this paper, the input and the output of a pose estimation system are a depth image and an ID-labeled image, respectively. Therefore we generate a large number of the pairs of depth and ID-labeled images to construct a dataset.

3. Recognition of human parts using FCN

3.1. Network architecture

We use a fully convolutional network (FCN) [14] for the part labeling task for depth images. FCNs do not have fully-connected layers, unlike usual convolutional neural networks (CNNs). Oliveira et al. [13] applied an FCN to a part labeling task for color images and showed an outstanding performance. The network has fifteen

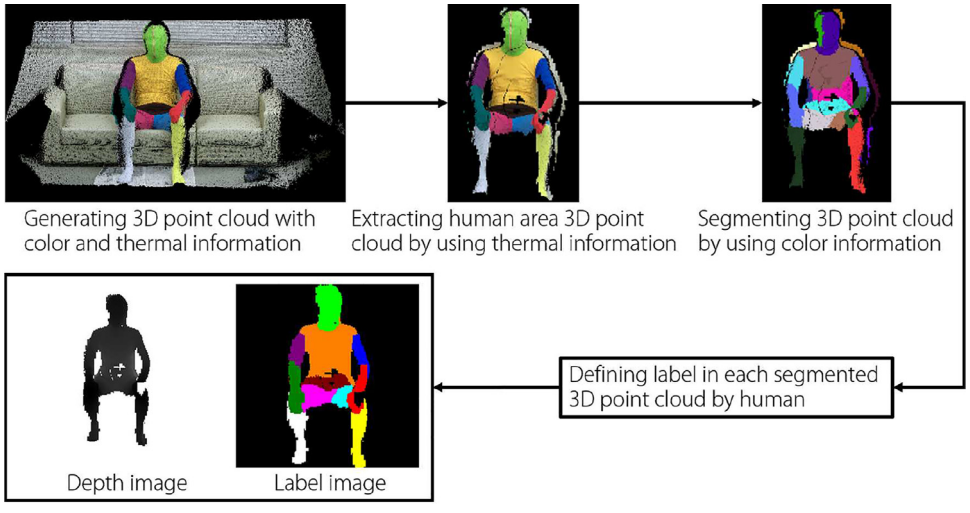

Fig. 9. Generating flow for depth and body label images in real scene.

Please cite this article as: K. Nishi, J. Miura, Generation of human depth images with body part labels for complex human pose recognition, Pattern Recognition (2017), http://dx.doi.org/10.1016/j.patcog.2017.06.006

Table 2
Confusion matrix for synthetic data.

Actual \ Predicted	Head	Torso	LU arm	RU arm	LF arm	RF arm	Hip	LU leg	RU leg	LL leg	RL leg	BG
Head	0.92	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04
Torso	0.00	0.95	0.01	0.02	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01
LU arm	0.00	0.10	0.83	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.05
RU arm	0.00	0.08	0.00	0.86	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.03
LF arm	0.00	0.01	0.09	0.01	0.70	0.07	0.01	0.01	0.00	0.00	0.00	0.10
RF arm	0.00	0.00	0.00	0.10	0.03	0.79	0.00	0.00	0.01	0.00	0.00	0.07
Hip	0.00	0.12	0.00	0.02	0.01	0.00	0.81	0.01	0.02	0.00	0.00	0.01
LU leg	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.91	0.00	0.03	0.00	0.02
RU leg	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.01	0.88	0.00	0.04	0.05
LL leg	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.92	0.01	0.05
RL leg	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.01	0.93	0.04
BG	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

LU = Left upper, RU = Right upper, LF = Left fore, RF = Right fore.
LL = Left lower, RL = Right lower, BG = Background.

Table 3
Confusion matrix for real data.

Actual \ Predicted	Head	Torso	LU arm	RU arm	LF arm	RF arm	Hip	LU leg	RU leg	LL leg	RL leg	BG
Head	0.69	0.22	0.05	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.03
Torso	0.00	0.83	0.05	0.06	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.03
LU arm	0.00	0.09	0.74	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.07
RU arm	0.00	0.06	0.00	0.80	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.04
LF arm	0.02	0.03	0.14	0.00	0.59	0.01	0.01	0.05	0.00	0.05	0.00	0.10
RF arm	0.01	0.02	0.01	0.11	0.04	0.65	0.01	0.00	0.07	0.00	0.03	0.05
Hip	0.00	0.25	0.00	0.00	0.03	0.02	0.65	0.01	0.02	0.00	0.00	0.02
LU leg	0.00	0.00	0.00	0.00	0.06	0.00	0.16	0.66	0.02	0.05	0.00	0.05
RU leg	0.00	0.00	0.00	0.00	0.00	0.02	0.16	0.02	0.71	0.00	0.06	0.03
LL leg	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.03	0.00	0.86	0.02	0.07
RL leg	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.04	0.01	0.87	0.07
BG	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

LU = Left upper, RU = Right upper, LF = Left fore, RF = Right fore.
LL = Left lower, RL = Right lower, BG = Background.

convolution layers and five deconvolution layers. Convolution layers extract features and the size of image decreases as they go through the layers. Deconvolution layers are then applied to the final output of convolution layers to have a labeled image with the same size as the input one. To compensate for the missing details, each deconvolution layer additionally uses the output from the corresponding pooling layer. Before adding a pooling layer output, it is convoluted and extracted for making the layer be the same size as the corresponding deconvolution layer.

We constructed our FCN based on the one by Oliveira et al. Fig. 5 shows the architecture of our network, which differs from theirs only in the size of inputs and output layers. We use the depth images with 212 × 212 pixels and twelve classes (eleven parts and one background). We therefore use 212 × 212 × 1 nodes for the input and 212 × 212 × 12 nodes for the output. We obtain twelve score maps, each of which indicates pixel-wise scores for the corresponding class. The final result (that is, a labeled image) is given by choosing the class with the highest score at each pixel.

Please cite this article as: K. Nishi, J. Miura, Generation of human depth images with body part labels for complex human pose recognition, Pattern Recognition (2017), http://dx.doi.org/10.1016/j.patcog.2017.06.006

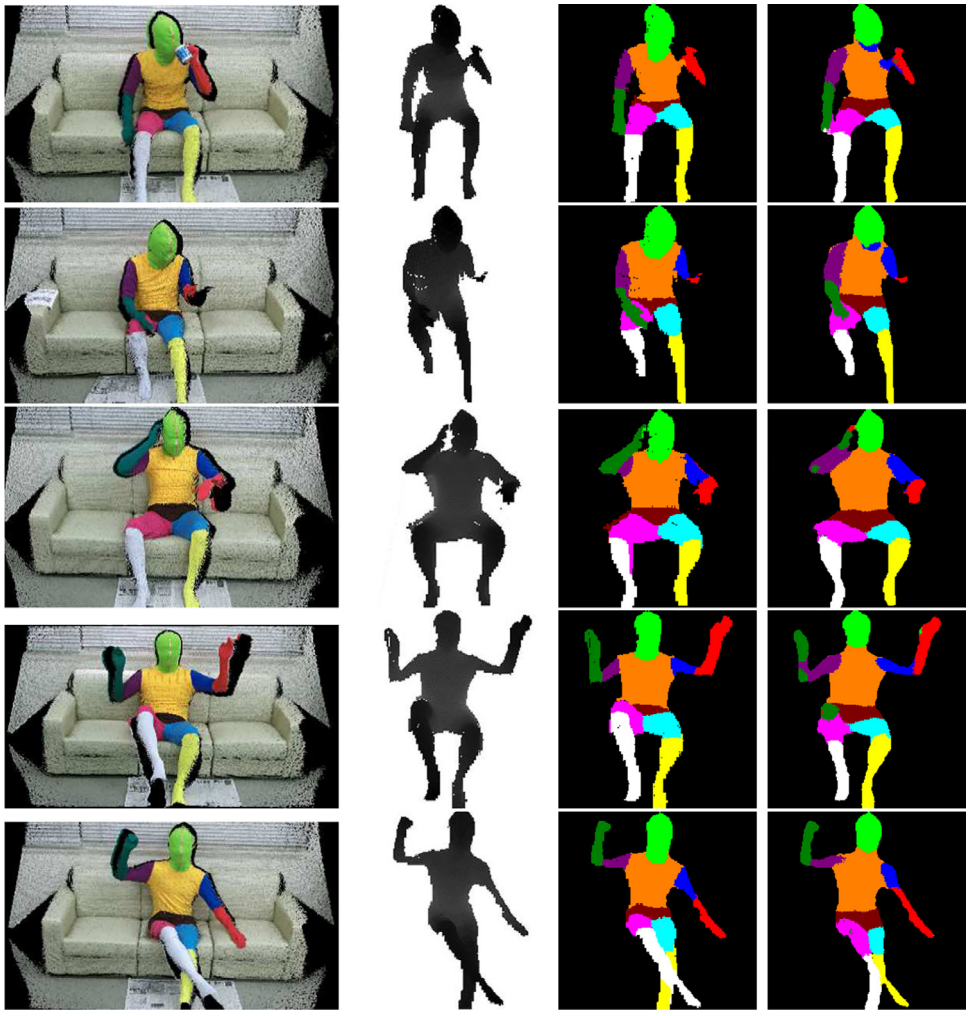

Fig. 10. Result on real data by using a color-coded clothing. **First column:** generated 3D point cloud with color and thermal information. **Second column:** Extracted depth data fort testing. **Third column:** target part labels generated from the colored point cloud. **Fourth column:** labeling results.

3.2. Network training

We deal with scenes where persons sit on a sofa with various poses. 10,076 depth images with part labels are generated as a dataset from the eight human models with various sitting poses. The dataset contains relatively static pose data such as drinking, reading a book, and using a gadget, and more dynamic pose data such as swinging the body, arms and legs. To apply the images to the FCN, we extract the person region in the depth images, which are randomly flipped horizontally, and normalize the size to 212 × 212 pixels. We also normalize the depth value, from the range [0 mm, 2000 mm] to [0, 1]. We applied the stochastic gradient descent (SGD) optimizer with momentum [28] for training. The learning rate and the momentum are set to 10^{-10} and 0.99, respectively; these values are the same with the one used in [13]. Each mini-batch consists of twelve images. We implemented the network using Chainer [29] and ran it on a single GeForce GTX TITAN X for 22 days.

Fig. 6 shows the change of the training loss during training. Although the training loss monotonically decreases at each epoch, the improvement is saturated around epoch 800 (11 days). Fig. 7 shows how the discriminative power increases as the training proceeds. The figure shows the labeling result at each selected epoch. Premature networks classify the body region poorly, but as the training proceeds, the network is refined gradually so that a more correct classification is performed. We here give some conjecture about the training process based on the classification results. Between epoch 0 and 50, the network learned that pixels with depth data constitute the body region. Since the torso is the largest body part, all body regions are classified as torso. Between epoch 50 and 300, the arms and the legs region are labeled as mixtures of yellow (left lower leg) and white (right lower leg). It means the networks learned that rod-like region are recognized as either of the legs. Between epoch 300 and 400, the network learned that all regions adjacent to the torso region are head, but this is partially corrected between epoch 400 and 600 such that the regions above the torso become head and those under the torso become hip. As the training proceeds, the other parts are also learned correctly, basically from the central parts to the peripheral ones. As shown in Fig. 6, the results after about epoch 800 are almost identical as the learning is considered saturated.

4. Experiments

4.1. Experiments using synthetic data

We collected motion data and generated another set of 4984 annotated depth images of the sit-on-a-sofa scene for testing. Fig. 8 shows example recognition results; parts labels are mostly correctly assigned. Table 2 shows the confusion matrix, summarizing the pixel-wise comparison results for the assigned and the

Please cite this article as: K. Nishi, J. Miura, Generation of human depth images with body part labels for complex human pose recognition, Pattern Recognition (2017), http://dx.doi.org/10.1016/j.patcog.2017.06.006

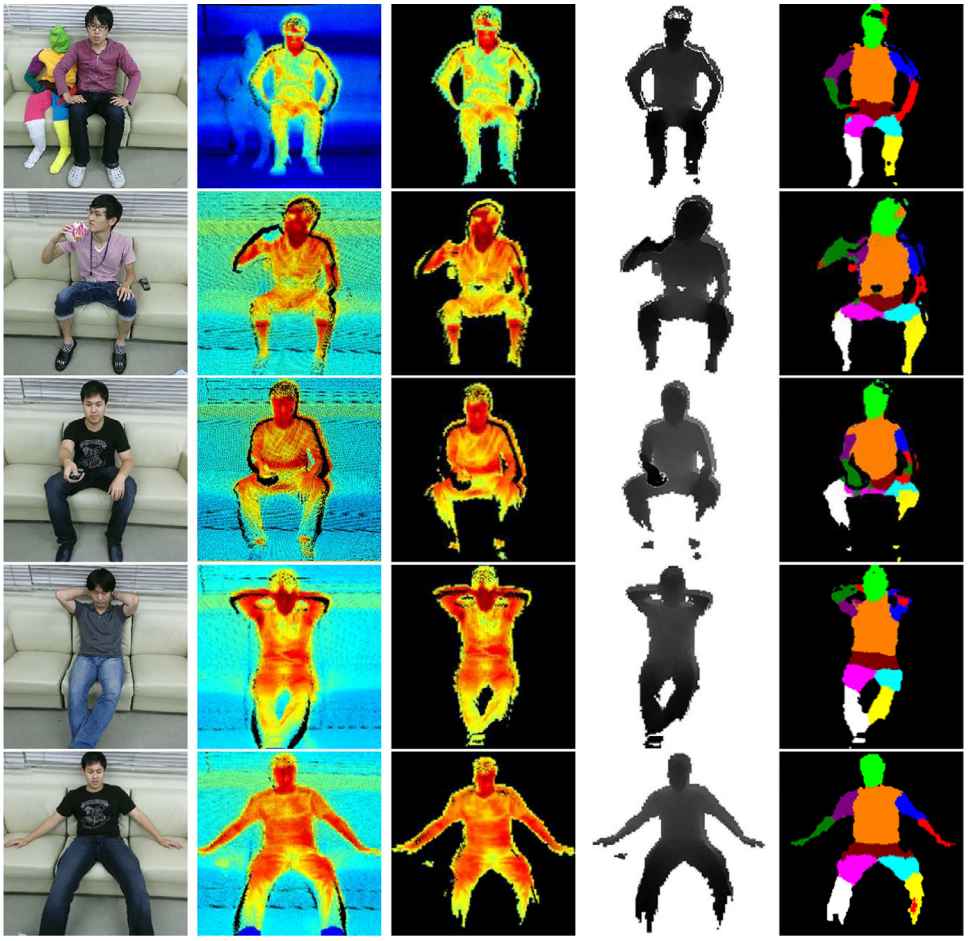

Fig. 11. Good labeling results on real data. **First column:** test scenes. **Second column:** generated thermal point clouds. **Third column:** extracted human region using thermal information. **Fourth column:** generated depth data for testing. **Fifth column:** labeling results.

correct labels. The recognition rates are relatively low for the forearms probably because the variations of their positions are larger than the other parts.

4.2. Evaluation using a color-coded clothing

4.2.1. Color-coded clothing

The previous subsection quantitatively evaluated the effectiveness of the proposed dataset and the FCN using simulated test data. We would like to do the quantitative evaluation also for a real scene. For this purpose, we need to have a dataset of annotated depth images constructed from real data. In the case of a person with normal clothing, however, the labeling should be done manually and is tedious and time consuming. To make this process much easier, we used a color-coded tight-fit clothing so that each part of the body can easily be distinguished using a color image.

In this paper, we use a 3D point cloud with thermal data for extracting the region of a person in a real scene. Point cloud data are obtained by a pair of a depth camera (KinectV2, Microsoft) and a far infrared (FIR) camera (PI200, Optris). The relative pose between the cameras is calibrated in advance [30,31]. In a room with a normal temperature, we can extract a person region relatively easily by extracting pixels with a temperature within some predetermined range (currently, 25 °C ∼ 35 °C).

Fig. 9 shows the developed clothing and the process of generating annotated depth images. First, a 3D point cloud with color and thermal information is generated. Next, the 3D points of the human body are extracted using thermal data, and they are then segmented using color. Since the segmentation is not complete, the segmented regions are manually corrected as body parts and put to the corresponding depth image as the annotation. We collected 145 annotated depth images for eight persons. It takes about one minute to make one annotated depth image.

4.2.2. Result of evaluation

We conducted experiments for the real test data using the FCN trained for the simulated dataset. Fig. 10 shows example labeling results. The labeling is basically acceptable but some parts are incorrectly labeled, mostly when they have a large positional deviation from the training dataset.

Table 3 is the confusion matrix showing a quantitative evaluation result. By examining the confusion states, we can see that when a part is incorrectly labeled, it is usually labeled as its neighboring part. For example, 22% of the head and 25% of the hip are misclassified as the torso. This shows that an overall structure of parts is correctly recovered but precise boundaries are not. In body pose estimation, relative poses between parts are more important than the precise boundary information. We thus conclude the combination of our dataset and the FCN is effective for human pose estimation.

4.3. Experiments using real data

Fig. 11 shows example labeling results. The columns indicate input scenes, thermal images, extracted person regions in the thermal images, extracted depth image regions, and part labeling results, respectively, from left to right. Using thermal data, person

Please cite this article as: K. Nishi, J. Miura, Generation of human depth images with body part labels for complex human pose recognition, Pattern Recognition (2017), http://dx.doi.org/10.1016/j.patcog.2017.06.006

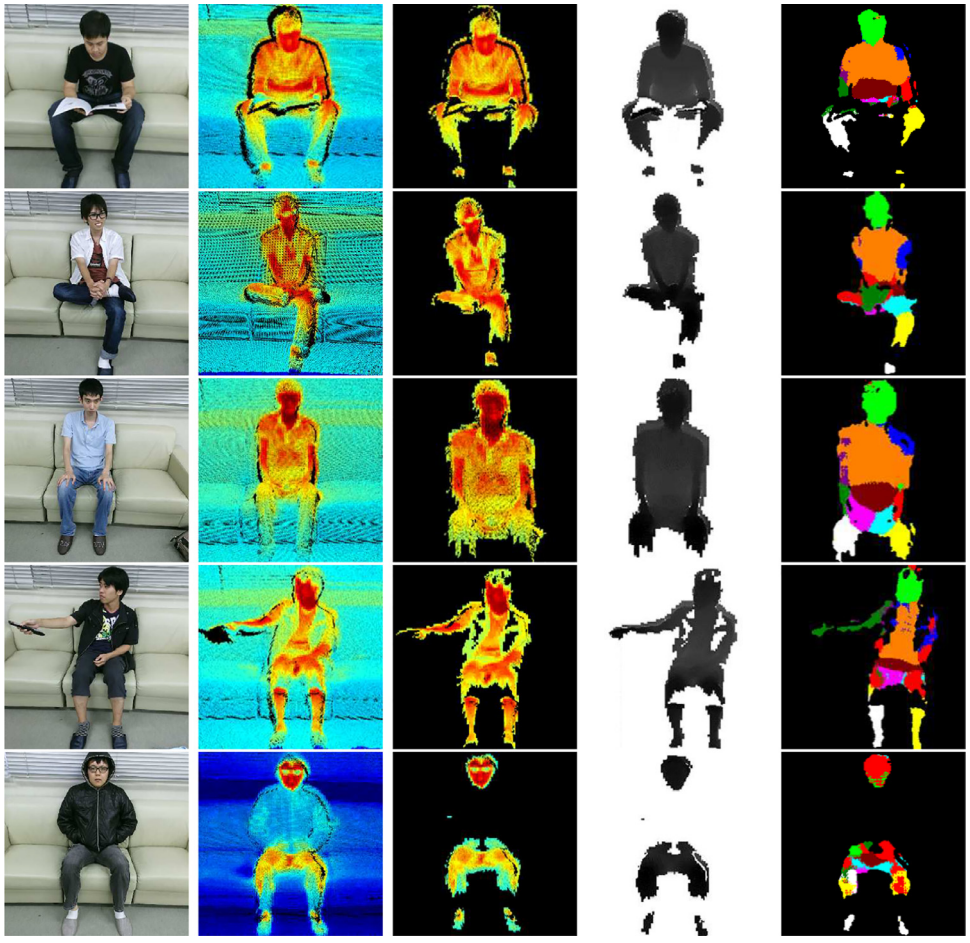

Fig. 12. Labeling results with poor human region extraction. **First column:** test scenes. **Second column:** generated thermal point clouds. **Third column:** Extracted human region using thermal information. **Fourth column:** generated depth data for testing. **Fifth column:** labeling results.

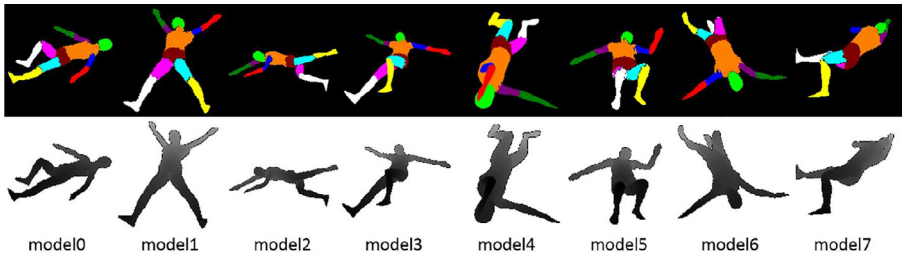

Fig. 13. Generated depth images with body parts labels for recumbent poses. **First row:** generated label images. **Second row:** generated depth images.

regions can be extracted even when they touch surrounding objects.

However, when the person region extraction fails to some extent, the recognition results will be degraded as shown in Fig. 12. In these cases, some parts of the region are missing due to, for example, occlusions and unexpected surface temperatures. We think this degradation in recognition is due to the lack of training data with incomplete region extractions. Adding such data to the dataset could increase the robustness of the recognition.

4.4. Experiments for unusual poses

We deal with a scene where persons are in recumbent positions. We took pose data for various recumbent positions including supine and lateral ones, applied them to the human models, and generated 326,984 depth images with part labels with 360-degree viewing direction. Fig. 13 shows examples of generated images; each pair of a color-labeled image and a depth image corresponds to a human model and a pose. We trained the same FCN using the generated dataset. We used one GeForce GTX TITAN X and two NVIDIA TITAN X's for training for 21 days.

Fig. 14 shows example of labeling results on real data. Estimated parts labels are mostly correctly assigned.

5. Conclusions and discussion

This paper presented an efficient procedure of generating a dataset of human body depth images with part labels, which is suitable for training convolutional neural networks (CNNs) in a depth image-based human pose estimation scenario. To generate data for various body shapes and poses, we first generate a variety of body shape models and then add two types information: skeleton and part labels. The former is to easily generate arbitrary model poses using the joint angle data obtained by a motion cap-

Please cite this article as: K. Nishi, J. Miura, Generation of human depth images with body part labels for complex human pose recognition, Pattern Recognition (2017), http://dx.doi.org/10.1016/j.patcog.2017.06.006

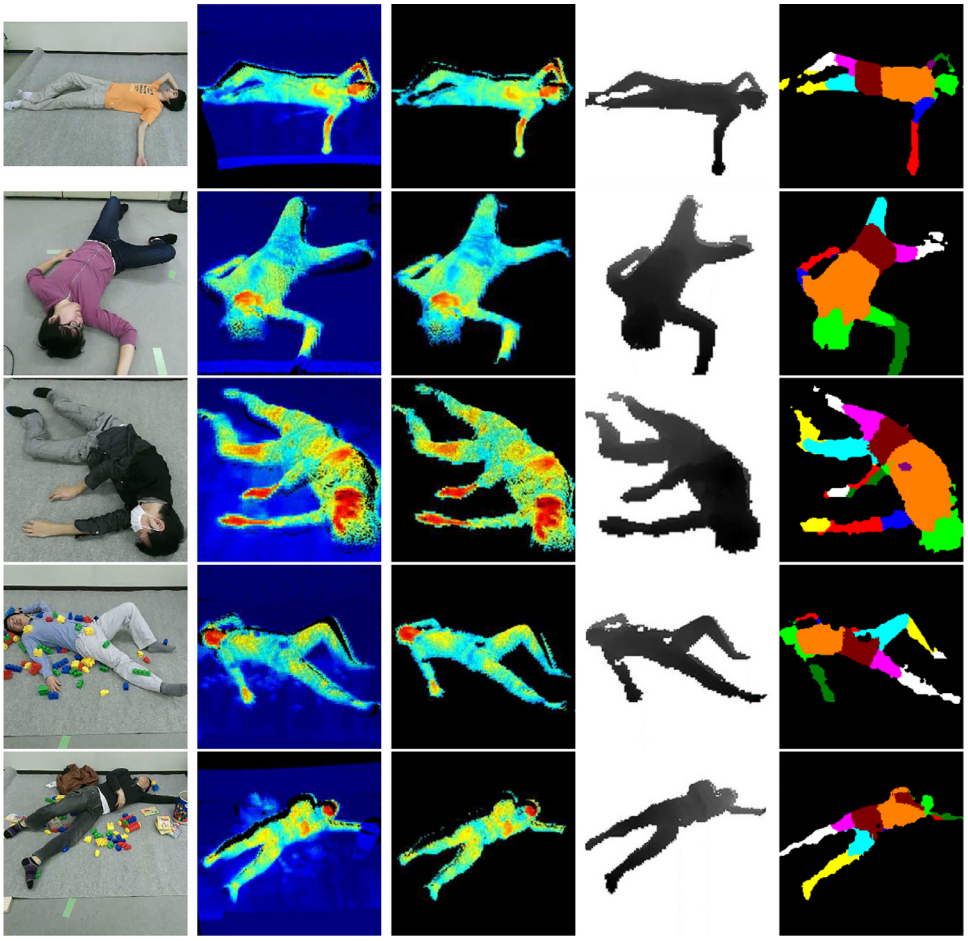

Fig. 14. Results on real data for recumbent poses. **First column:** test scenes. **Second column:** generated thermal point clouds. **Third column:** extracted human region using thermal information. **Fourth column:** generated depth data for testing. **Fifth column:** labeling results.

ture system. The latter is to render body depth images with part label annotations. A dataset is generated and evaluated for a sitting scenario. A fully-convolutional network (FCN) was trained using the dataset and applied to part labeling tasks for both synthetic and real data. Another dataset was generated for recumbent poses, as examples of unusual poses, and evaluated using the same FCN. Evaluation results show the effectiveness of the datasets.

We are interested in developing a human support robot that can recognize a human state and perform appropriate assistive actions. Pose recognition is one of the necessary functions of such a robot and must be able to handle a large variety of poses including unusual ones such as falling and crouching. We are now extending the dataset to include various possible poses. Moreover, since a part of the body region may be missing in a depth image due to self-occlusion and occlusion by other objects, adding data for such cases is also planned.

Evaluation of the dataset is done by the recognition performance of an FCN trained using the dataset. Although the FCN has been shown to be very effective in part labeling tasks for both color and depth images, we need to seek other network architectures which could achieve a better performance.

We are now working on making the results of the research be publicly available as follows. One is the dataset itself [32]. The datasets described in the paper is already available at this website, and we are planning to add more datasets as they are generated. These datasets can be used for train any network architectures for the part labeling task. The other type of data we are planning to make available is a set of human models with skeleton and part label information. This makes it possible for the users to generate a new dataset with poses they want to deal with.

Acknowledgments

The authors would like to thank Dr. Shuji Oishi for helping them to use KY Human Model. This work is in part supported by JSPS Kakenhi No. 25280093.

References

[1] United Nations, Department of Economic and Social Affairs, Population Division, World population ageing 2015 (st/esa/ser.a/390).
[2] P.F. Felzenszwalb, D.P. Huttenlocher, Pictorial structures for object recognition, Int. J. Comput. Vis. 61 (1) (2005) 55–79.
[3] M.A. Fischler, R.A. Elschlager, The representation and matching of pictorial structures, IEEE Trans. Comput. 22 (1) (1973) 67–92.
[4] D. Ramanan, Learning to parse images of articulated bodies, in: Advances in Neural Information Processing Systems, 2006, pp. 1129–1136.
[5] V. Ferrari, M. Marin-Jimenez, A. Zisserman, Progressive search space reduction for human pose estimation, in: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, IEEE, 2008, pp. 1–8.
[6] C. Rother, V. Kolmogorov, A. Blake, Grabcut: Interactive foreground extraction using iterated graph cuts, in: ACM Transactions on Graphics (TOG), vol. 23, ACM, 2004, pp. 309–314.
[7] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, R. Moore, Real-time human pose recognition in parts from single depth images, Commun. ACM 56 (1) (2013) 116–124.
[8] I. Ardiyanto, J. Satake, J. Miura, Autonomous monitoring framework with fallen person pose estimation and vital sign detection, in: Information Technology and Electrical Engineering (ICITEE), 2014 6th International Conference on, IEEE, 2014, pp. 1–6.
[9] S. Wang, S. Zabir, B. Leibe, Lying pose recognition for elderly fall detection, Robotics (1) (2012) 345–353.

Please cite this article as: K. Nishi, J. Miura, Generation of human depth images with body part labels for complex human pose recognition, Pattern Recognition (2017), http://dx.doi.org/10.1016/j.patcog.2017.06.006

[10] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems (NIPS), 2012, pp. 1097–1105.
[11] A. Toshev, C. Szegedy, Deeppose: human pose estimation via deep neural networks, in: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1653–1660.
[12] B. Sapp, B. Taskar, Modec: multimodal decomposable models for human pose estimation, in: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2013.
[13] G. Oliveira, A. Valada, C. Bollen, W. Burgard, T. Brox, Deep learning for human part discovery in images, in: Proceedings of the IEEE Int. Conf. on Robotics and Automation (ICRA), 2016.
[14] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3431–3440.
[15] X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun, A. Yuille, 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.
[16] K. Nishi, J. Miura, A head position estimation method for a variety of recumbent positions for a care robot, in: Proceedings of the Int. Conf. on Advanced Mechatronics (ICAM), 2015.
[17] B.C. Russell, A. Torralba, K.P. Murphy, W.T. Freeman, Labelme: a database and web-based tool for image annotation, Int. J. Comput. Vis. 77 (1–3) (2008) 157–173.
[18] S. Song, S.P. Lichtenberg, J. Xiao, Sun rgb-d: a rgb-d scene understanding benchmark suite, in: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 567–576.
[19] J. Sung, C. Ponce, B. Selman, A. Saxena, Unstructured human activity detection from rgbd images, in: Proceedings of the IEEE Int. Conf. on Robotics and Automation (ICRA), IEEE, 2012, pp. 842–849.
[20] L. Xia, C. Chen, J. Aggarwal, View invariant human action recognition using histograms of 3d joints, in: Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, 2012, pp. 20–27.
[21] M. Shinzaki, Y. Iwashita, R. Kurazume, K. Ogawara, Gait-based person identification method using shadow biometrics for robustness to changes in the walking direction, in: 2015 IEEE Winter Conference on Applications of Computer Vision, IEEE, 2015, pp. 670–677.
[22] N. Hasler, T. Thormählen, B. Rosenhahn, H.-P. Seidel, Learning skeletons for shape and pose, in: Proceedings of the 2010 ACM SIGGRAPH symposium on Interactive 3D Graphics and Games, ACM, 2010, pp. 23–30.
[23] L. Pishchulin, S. Wuhrer, T. Helten, C. Theobalt, B. Schiele, Building statistical shape spaces for 3d human modeling, Pattern Recognit. 67 (2017) 276–286.
[24] VICON. (accessed 17.04.29) URL https://www.vicon.com/.
[25] Maya,. (accessed 17.04.29) URL http://www.autodesk.com/products/maya/overview.
[26] MotionBuilder. (accessed 17.04.29) URL http://www.autodesk.com/products/motionbuilder/overview.
[27] M. Kouchi, M. Mochimaru, AIST/HQL database of human body dimension and shape 2003. (accessed 17.04.29) URL https://www.dh.aist.go.jp/database/fbodyDB/index.html.
[28] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back-propagating errors, Cognit. Model. 5(3) 1.
[29] S. Tokui, K. Oono, S. Hido, J. Clayton, Chainer: a next-generation open source framework for deep learning, in: Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Twenty-ninth Annual Conference on Neural Information Processing Systems (NIPS), 2015.
[30] S. Vidas, P. Moghadam, M. Bosse, 3d thermal mapping of building interiors using an rgb-d and thermal camera, in: Robotics and Automation (ICRA), 2013 IEEE International Conference on, IEEE, 2013, pp. 2311–2318.
[31] J. Rangel, S. Soldan, A. Kroll, 3d thermal imaging: fusion of thermography and depth cameras, in: International Conference on Quantitative InfraRed Thermography, 2014.
[32] AISL HDIBPL (Human Depth Images with Body Part Labels) Database. (accessed 17.04.29) URL http://www.aisl.cs.tut.ac.jp/database_HDIBPL.html.

Please cite this article as: K. Nishi, J. Miura, Generation of human depth images with body part labels for complex human pose recognition, Pattern Recognition (2017), http://dx.doi.org/10.1016/j.patcog.2017.06.006

Kaichiro Nishi received the B.Eng. and the M.Eng. degree in computer science and engineering in 2015 and 2017, respectively, from Toyohashi University of Technology, Japan. His research interests include human pose estimation and care robotics.

Jun Miura received the B.Eng. degree in mechanical engineering in 1984, the M.Eng. and the Dr.Eng. degree in information engineering in 1986 and 1989, respectively, all from the University of Tokyo, Tokyo, Japan. In 1989, he joined Department of Computer-Controlled Mechanical Systems, Osaka University, Suita, Japan. Since April 2007, he has been a Professor at Department of Computer Science and Engineering, Toyohashi University of Technology, Toyohashi, Japan. From March 1994 to February 1995, he was a Visiting Scientist at Computer Science Department, Carnegie Mellon University, Pittsburgh, PA. He received several awards including Best Paper Award from the Robotics Society of Japan in 1997, Best Paper Award Finalist at ICRA-1995, and Best Service Robotics Paper Award Finalist at ICRA-2013. Prof. Miura published over 180 papers in international journals and conferences in the areas of intelligent robotics, mobile service robots, robot vision, and artificial intelligence.

Please cite this article as: K. Nishi, J. Miura, Generation of human depth images with body part labels for complex human pose recognition, Pattern Recognition (2017), http://dx.doi.org/10.1016/j.patcog.2017.06.006