Advances in Robotics Research, *Vol. 2, No. 2 (2018) 129-140*
DOI: https://doi.org/10.12989/arr.2018.2.2.129

Robust 2D human upper-body pose estimation with fully convolutional network

Seunghee Lee[1a], Jungmo Koo[1b], Jinki Kim[1c] and Hyun Myung[*1,2]

[1]*Department of Civil and Environmental Engineering, Korean Advanced Institute for Science and Technology, 291 Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea*
[2]*Robotics Program, Korean Advanced Institute for Science and Technology, 291 Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea*

(Received April 28, 2018, Revised May 10, 2018, Accepted May 11, 2018)

Abstract. With the increasing demand for the development of human pose estimation, such as human-computer interaction and human activity recognition, there have been numerous approaches to detect the 2D poses of people in images more efficiently. Despite many years of human pose estimation research, the estimation of human poses with images remains difficult to produce satisfactory results. In this study, we propose a robust 2D human body pose estimation method using an RGB camera sensor. Our pose estimation method is efficient and cost-effective since the use of RGB camera sensor is economically beneficial compared to more commonly used high-priced sensors. For the estimation of upper-body joint positions, semantic segmentation with a fully convolutional network was exploited. From acquired RGB images, joint heatmaps accurately estimate the coordinates of the location of each joint. The network architecture was designed to learn and detect the locations of joints via the sequential prediction processing method. Our proposed method was tested and validated for efficient estimation of the human upper-body pose. The obtained results reveal the potential of a simple RGB camera sensor for human pose estimation applications.

Keywords: human pose estimation; skeleton extraction; fully convolutional network; semantic segmentation; upper-body joint segmentation

1. Introduction

Human body pose estimation is one of the most important techniques that has been studied for decades. There have been extensive efforts to efficiently estimate human body poses along with reliable skeleton extraction results. Such technology allows a higher level of human-computer interaction and the recognition of human activities for various applications (Aggarwal *et al.* 1997, Moeslund *et al.* 2006). Pose estimation is mainly aimed at recognizing the gestures of humans in action; the recognition of human gestures may be adapted for the development of body language or

*Corresponding author, Professor, E-mail: hmyung@kaist.ac.kr
[a]Ph.D. Student, E-mail: seunghee.lee@kaist.ac.kr
[b]Ph.D. Student, E-mail: jungmokoo@kaist.ac.kr
[c]Master Student, E-mail: rlawlsrl@kaist.ac.kr

 ISSN: 2287-4976 (Print), 2287-4984 (Online)

sign language applications. Also, human pose estimation techniques can be useful for sports activities, surveillance systems, and the development of clinical analysis of gait pathologies.

Despite numerous applications of human pose estimation techniques, these techniques still require further investigation because of their poor accuracy and accessibility due to the necessity of using expensive sensors such as motion tracking system.

Automated human body pose estimation methods are categorized into two approaches: graph-based methods and machine learning-based methods. The graph-based human body pose estimation methods use geodesic distance for pose estimation by measuring the geodesic distances between the different points of body parts (Roweis *et al.* 2000). Attempts have been made to develop pose estimation techniques using geodesic distances, such as using anatomical landmarks in a depth geodesic graph and inverse kinematics (Schwarz *et al.* 2012), and using a skeletal tree-like graph to represent the human body (Straka *et al.* 2011). However, inevitable variations in images, e.g., variations in body profiles, clothes, and human joint movements, make pose estimation difficult using the geodesic distance method. Self-articulation-induced partial occlusions, for instance, covering the face with hands or parts of the body, or occlusions by external objects may cause uncertainties in body pose estimation, which may produce insufficient outcomes (Droeschel *et al.* 2011). Compared with the graph-based methods, machine learning-based methods may produce better results by training the system in such various situations.

Moreover, whereas graph-based methods require a calibration procedure for joint detection, machine learning-based pose estimation methods can be applied without a calibration procedure, which makes them suitable for real-time applications (Kim *et al.* 2015). Shotton *et al.* (2013) tested a per-pixel classification of joints and estimated the positions of joints using a random forest algorithm. Hernández-Vela *et al.* (2012) used a graph-cut optimization for image segmentation in depth maps.

The main contributions of this paper are three-fold. Firstly, the proposed human joint estimation algorithm improves upper-body part detection by learning and inferencing the image at pixel-unit and considering adjacent joints using heatmaps. Our method is better than state-of-the-art with respect to accuracy and frame rate. Secondly, the proposed human joint estimation algorithm is suitable to be operated on mobile robots. In order to understand the circumstances and the environment for human-robot interaction, the robot should be able to perceive human beings. Most robots interacting with humans perceive the human closely, which means the robots usually observe the person's upper-body, not a full body. Therefore, in this study, we specifically focus on the development of a human upper-body pose estimation method using an RGB camera sensor. Finally, we made our own segmentation-training datasets for the upper-body, which include various camera views and occlusions of body parts in 12 different scenarios with four different persons. Our datasets can be downloaded from the following website: https://github.com/handale88/Urobot_segmentation_DB.

2. Related research

2.1 Human pose estimation using depth information

Depth-based automatic human pose estimation systems have attracted much attention following the development of Microsoft's Kinect system (Jain *et al.* 2011). Images acquired from the Kinect camera provide depth information. There have been extensive studies that used geometric

information to detect joints. Haritaoglu *et al*. (1998) divided the blob with geometric information to distinguish body parts such as the head, hands, and feet. Then, Fujiyoshi *et al*. (2004) predicted the blob of a head, hands, and feet without any template model. Similarly, Guo *et al*. (1994) evaluated the location of full body points using the distance as a fitting parameter. To acquire complete positions of the joints of person, genetic algorithms (Takahashi *et al.* 2000) and neural networks (Ohya *et al.* 1994) have also been used.

The recent study by Shotton *et al*. (2013) evaluated the Kinect sensor's performance to classify 20 joints for every pixel image from the sensor in every single image. For the pixel classification procedure, a randomized decision forest was used, which was implemented by building decision trees for training the system and providing the output sources for the classes of body parts. The generation of 3D joint positions was achieved by a weighted Gaussian kernel with a mean shift. The classifier was trained by a massive amount of data for diverse motions and body shapes of large numbers of people.

The OpenNI library is also a widely used tool for the estimation of the skeleton. With a depth-edge counting local descriptors, 15 skeleton joints were estimated by Presti *et al*. (2016). The Canny edge detector was used to extract the depth edge information from the depth map, and static information at edge pixels was evaluated in each patch. Then, the location of skeletal joints was found using an approximate nearest neighbor (ANN) algorithm to match the patch descriptors. Consequently, the patch descriptors were compared to determine the body joint locations.

2.2 Human pose estimation using RGB image information

It was not satisfactory to use only RGB image data for detecting skeletons with changing colors and human body shapes (Wei *et al*. 2016). Recently, to resolve this problem, researchers tried to use GPUs (graphics processing units) to train deep neural networks using a huge number of images. There have been many suggestions for skeleton extraction methods that use GPUs to increase the robustness and frame rates (Droeschel *et al.* 2011, Ganapathi *et al.* 2010, Shotton *et al.* 2013, Zhang *et al.* 2013)

Recently, RGB image-based deep convolutional neural networks (DCNNs) have achieved outstanding performance for the estimation of human pose (Yang *et al.* 2016). Toshev *et al*. (2014) used the DeepPose regression learning system for body joint detection with a convolutional network. Chen *et al*. (2014) suggested a convolutional network dependent on pairwise joint relationships. Tompson *et al.* (2015) used multi-resolution DCNN to accurately detect joints by reducing the pooling effect. These methods perform much better than the conventional methods, but the inaccuracy of estimating joints is still high especially when there are occlusions. Wei *et al*. (2016) proposed a convolutional pose machine (CPM) consisting of six stages using a cascaded regression-based method and achieved state-of-the-art accuracy. However, occluded joints from various human poses are still a challenging problem for this method.

There has been a recent flow of interest in segmentation-based learning for joint estimation. Segmentation-type approaches are simple and fast, and they give a more accurate inference on image at pixel level. Ladicky *et al*. (2013) suggested a joint pixel-wise and part-wise segmentation-based pose estimation using HOG feature component, but it cannot deal with images with large pose variations and some occlusions due to the limitation of the features it uses. Xia *et al.* (2017) suggested to fuse initial joint score maps and part segment score maps through two stages to yield better estimation results. However, inference frame rate of this method is less than 8Hz because of many—more than 101—layers in each stage. In contrast, our model has only one

stage consisting of 8 layers, and it combines FCNs with consideration of adjacent joints' confidence maps. It greatly boosts the accuracy and the speed of handling pose variations especially for the upper-body joints. In this work, we present a fast and accurate estimation method for upper-body joints using segmentation-based learning.

The performance of the pose estimation methods is mostly evaluated with pose estimation datasets that are open to public. There are several benchmark datasets for the skeleton estimation test. Popular upper-body images obtained from Hollywood movies are the Frames Labeled in Cinema (FLIC) dataset, consisting of 4,000 training images and 1,000 test images (Sapp *et al.* 2013). The Leeds Sports Poses (LSP) dataset is another well-known dataset consisting of various sports postures with 14 body joints, containing 11,000 training images and 1,000 testing images (Johnson *et al.* 2010). The MPII Human Pose Dataset (Andriluka *et al.* 2014), the Image Parse (PARSE) dataset (Ramanan 2007), the Buffy dataset (Ferrari *et al.* 2008) are also frequently used for the evaluation of human pose estimation methods. All those datasets only offer RGB images and joint locations. Unfortunately, there are no upper-body segmentation datasets among those open datasets. It is required to acquire segmentation dataset to train segmentation-based network. For this reason, we produced our own upper-body dataset consisting of RGB and segmentation image pairs.

3. Human upper-body pose estimation method

We introduce our human upper-body pose estimation method using a segmentation-based learning classifier. Semantic segmentation understands an image at pixel level, such that they look at human pose more precisely. It takes an RGB image as input and generates a list of joint coordinates. There are seven joints in total: head, left shoulder, left elbow, left hand, right shoulder, right elbow, and right hand.

3.1 Segmentation-based learning using fully convolutional network

Except the output layer configuration, our method is similar to the segmentation network of Long *et al*. (2015), which was the first attempt that used deep segmentation network adapted from deep classification network. Long *et al*. (2015) adopted Alexnet (Krizhevsky *et al.* 2012) deep classification networks and extended it to fully convolutional networks (FCN) and fine-tuned (Donahue *et al.* 2014) it for the segmentation problem by transferring learned representations from whole image inputs. End-to-end training and pixel-wise segmentation predictions were performed in the FCN. Both learning and inference on the whole image at one time were possible by dense feedforward computation and backpropagation.

Fig. 1 illustrates the detailed configuration of the proposed network. It is trained by our upper-body posture dataset to be more specialized in human upper-body pose estimation. There are eight convolutional layers. The first layer has a size equal to the number of pixels times three color channels. Learnable parameters are contained only in convolutional layers and fully connected layers. The filter sizes of the first and second convolutional layers are 11×11 and 5×5, respectively, and the other three have 3×3 filters. After five convolution layers, we put fully connected layers to generate dense pixel-wise prediction map for each class. The difference to FCN-Alexnet (Long *et al.* 2015) is the outputs of our network. Instead of using confidence maps as outputs, we added element-wise (eltwise) layers for more accurate joint estimation. The eltwise

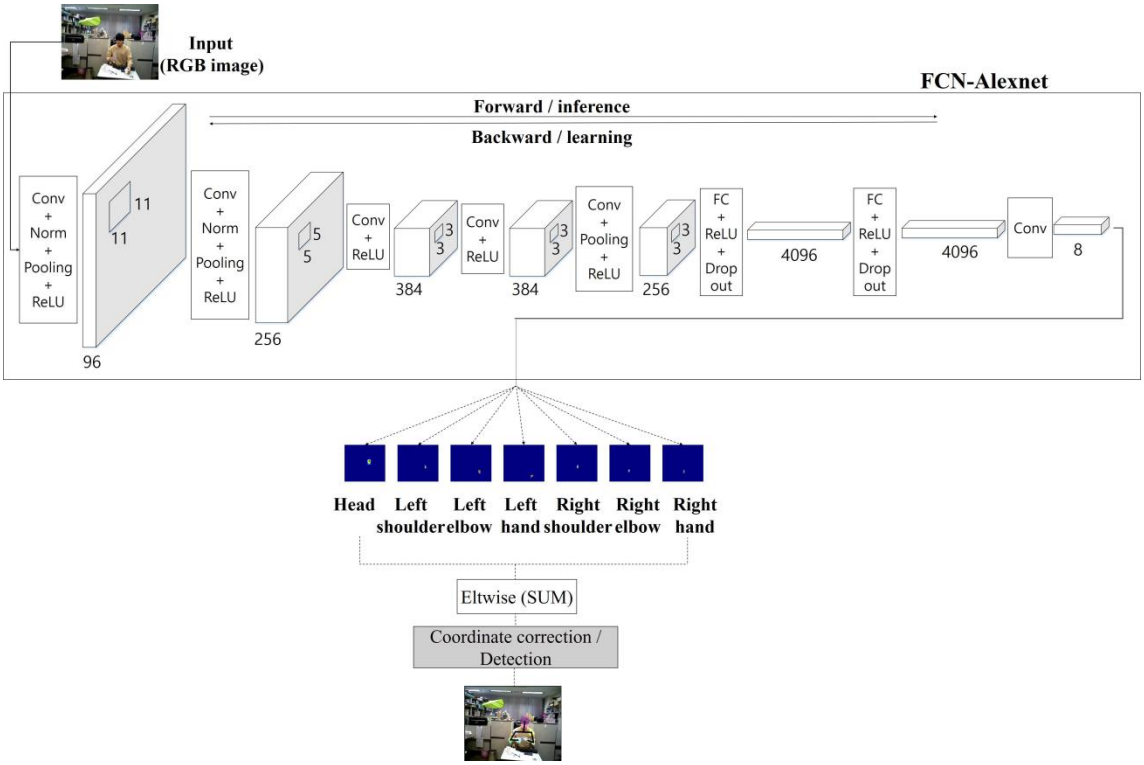


Fig. 1 The overall architecture of our segmentation network. *Conv* denotes a convolutional layer, *Norm* a local response normalization layer, *Pooling* a pooling layer, and *FC* a fully connected layer. The output of each heatmap represents the probability of each joint

layers perform the element-wise summation of different images. By using these layers, adjacent joints are considered when estimating positions of joints. The details of the network parameters and training procedure are elaborated in Section 4.

3.2 Pose estimation using joint confidence map

For pose estimation, the 2D coordinate of each joint is calculated from the 2D confidence map of each joint. Each confidence map represents the likelihood of the position of the body joints.

Upon generation of heatmap images at the end of segmentation network, the central moment of each region is used as the position of each joint (Lee *et al.* 2017). However, this central moment may not be the exact location of each joint since it is not robust enough in occlusions. Instead, after the heatmap of each joint is generated, the position of each joint is estimated by considering an adjacent joint's confidence map using eltwise layer. Fig. 2 describes the steps for joint prediction with an example of left shoulder area. The adjacent joints of the left shoulder are the head and the left elbow (Fig. 2(a)). The confidence maps of those joints are combined into an image, and this combined image results in high probability in the area near the head or the left elbow (Fig. 2(b)). In this case, the area near the left elbow has the maximum probability. Whereas the joint confidence map is expected to have the maximum probability at one point, it sometimes

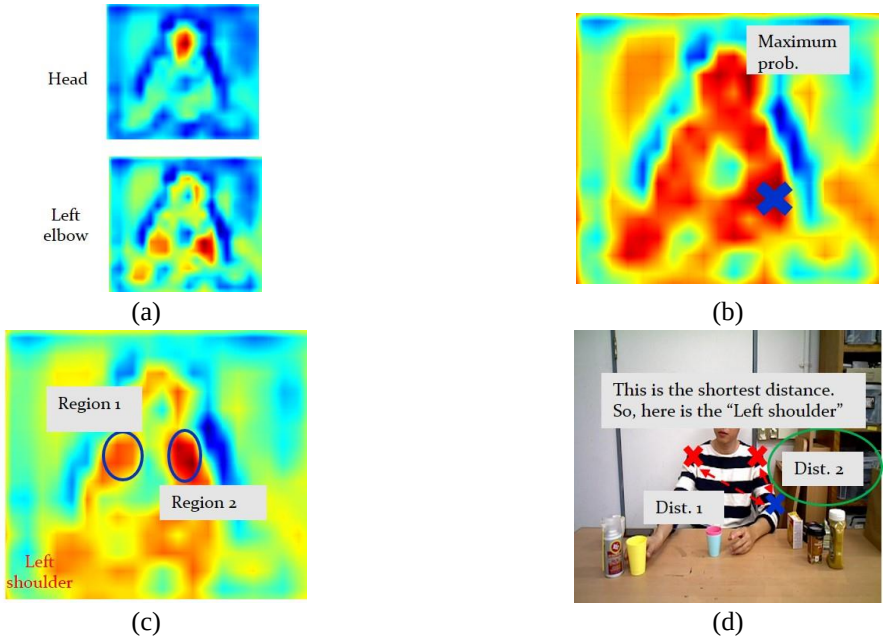


Fig. 2 Example of detecting left shoulder. (a) Confidence maps of head and left elbow. (b) Confidence map of left shoulder. (c) The resulting confidence map combining head and left elbow. (d) Comparing the distance between the maximum probability point and the candidates of left shoulder. Finally, Region 2 is selected because Dist.2 is shorter than Dist 1

shows multiple local maxima in probability distribution which may produce ambiguity. In this case, the distances from the body parts to the coordinate of the maximum probability in the combined image are compared, and the joint with shorter distance from the maximum probability point is selected (Fig. 2(d)). Therefore, the left elbow joint is finally selected in this example.

4. Experiments

4.1 Datasets for training and testing

Since there is no dataset available that provides pairs of RGB and segmentation images of the human upper-body, we supplemented ourselves with a new dataset. Our dataset consists of RGB and segmentation images that depict 12 different scenarios, such as cleaning the table, preparing cereal in a bowl, reading a book, etc., with four different persons. For each new scenario, a new camera location is randomly sampled at the probable place where the camera of the humanoid robot called Mybot (Kim *et al.* 2017) is likely to be located.

There are a total of 433 images for training and 58 images for evaluation with a 640×480 pixel resolution. All the samples had full annotation of seven joints; head, left shoulder, left elbow, left hand, right shoulder, right elbow, and right hand. A sample of the dataset is shown in Fig. 3.

4.2 Training and testing method

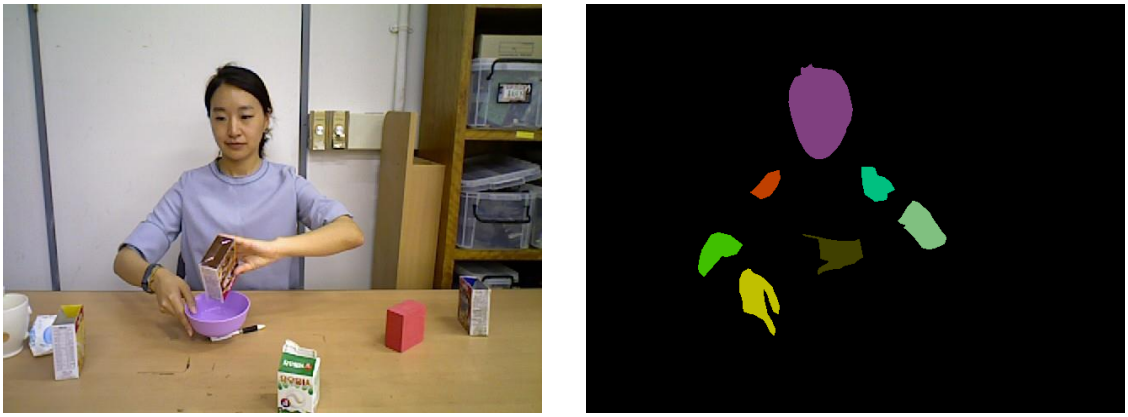
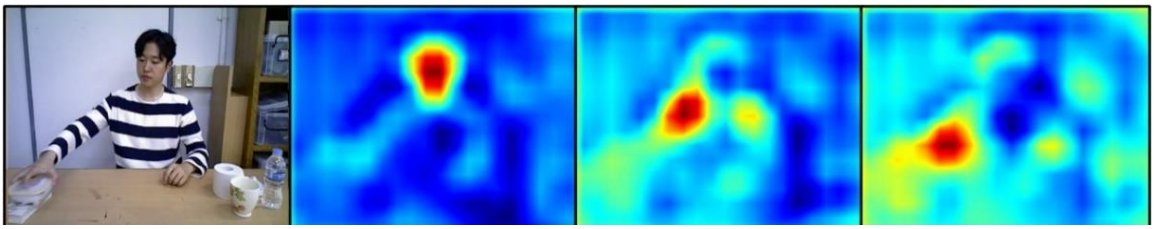
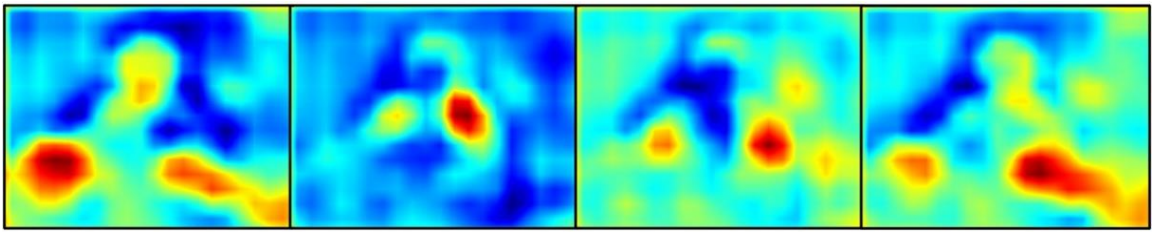


Fig. 3 An exemplary scene of our segmented image dataset with seven classes; head, left shoulder, left elbow, left hand, right shoulder, right elbow and right hand



(a) Original image (b) Head (c) Right shoulder (d) Right elbow



(e) Right hand (f) Left shoulder (g) Left elbow (h) Left hand

Fig. 4 An example of heatmap results on each joint. The more reddish in the map is, the higher the confidence of existence is. The bluer in the map is, the lower the confidence of existence is

For constructing the human upper-body pose dataset, RGB images from ASUS Xtion pro (Asus, 2018) was used. The position of this camera was fixed at the height of Mybot's camera, and the movement of the camera is set approximately ±15 degrees up and down. The training and testing of the network were conducted using a laptop computer with one GTX1080 GPU (NVIDIA, 2018a). The Caffe (Jia *et al.* 2014) libraries and DIGITS (NVIDIA 2018b) for deep learning were used to define and implement our model.

To train the network for upper-body joint segmentation, a batch size was set to one for on-line learning and the Adam solver (Kingma *et al.* 2014) was used for the optimization. The network architecture was first initialized with the weights of FCN-AlexNet and fine-tuned with a pre-trained model by PASCAL-VOC dataset (Everingham *et al.* 2015). The learning rate was set to 1.0×10^{-4}, and sigmoid decay was used for the learning rate policy.

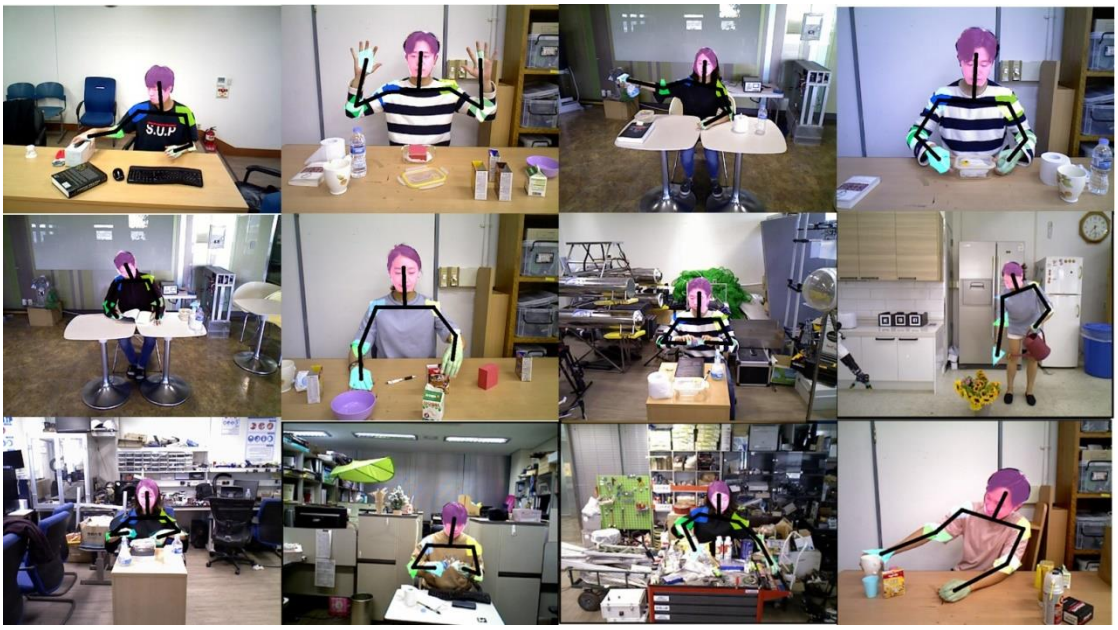


Fig. 5 Examples of upper-body joint estimation results for different persons and environments. Each color represents a different joint

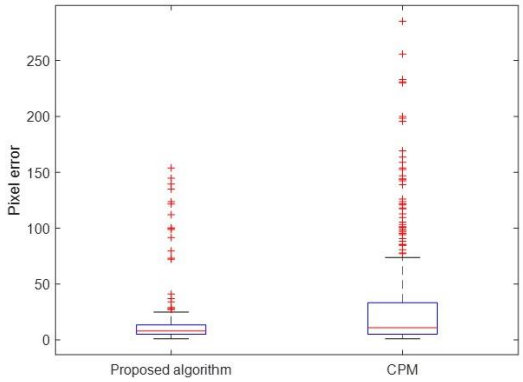


Fig. 6 Boxplots for joint estimation pixel error using the proposed and CPM algorithms ($p < 0.001$ in a paired *t*-test)

Table 1 The average joint estimation pixel errors for the proposed and CPM algorithms (unit: pixel)

	Head	Right shoulder	Right elbow	Right hand	Left shoulder	Left elbow	Left hand	Total average
Proposed method	13.05	7.08	11.76	19.32	6.44	13.18	29.18	14.29
CPM	24.46	12.29	31.20	42.55	16.43	38.06	50.36	30.76

Table 2 Average frame rates of joint estimation (unit: frame per second)

	Proposed method	CPM
Frame rate (Hz)	19.2433	14.5134

4.3 Experimental results

In this section, numerical results are presented with our datasets. We trained the network for 30 epochs, which took an average of 0.5 hours with NVIDIA DIGITS running on a GTX1080 GPU.

Fig. 4 shows the results of the upper-body image estimation. Each heatmap provides the confidence map of each joint, which gives the existence likelihood of each joint. The red color represents the highest confidence of existence, while the blue color represents the lowest confidence of existence. We obtained the final pose estimation results with the segmentation results from the joint heatmaps as described in Fig. 5. It presents 18 examples of joint estimation results with 10 different environments, 14 different views, and 4 different persons. All the estimated joint locations are overlaid on the original RGB images. Each color represents a different joint.

To evaluate the performance of our proposed method, the average of joint distance errors of the proposed algorithm are compared. We manually provided the ground truth for the random test images because there is no ground truth dataset. Fig. 6 shows the comparison results between the proposed joint estimation algorithm and the CPM algorithm (Wei *et al*. 2016). Median and the third quartile of the proposed algorithm are all smaller than those of the CPM algorithm. We performed a hypothesis testing using a paired *t*-test to determine whether there is a statistically significant difference between the proposed algorithm and CPM method. We found that the *p*-value is 1.5891×10^{-10} (< 0.001), and we can ensure the superiority of our proposed method. 40% of outliers in the proposed method are caused by left hand. Also, 25% and 24% of outliers in CPM method are caused by left and right hands, respectively. This can be also seen in Table 1. The highest errors in both methods were resulted from the left hand. This is due to many datasets having occluded hands holding something. The average error for all joints using the proposed algorithm is approximately 13 pixels. Table 2 shows the average frame rates of joint estimation for the proposed method and CPM while testing with the test datasets. The proposed method is 1.33 times faster than CPM on average.

5. Conclusions

This paper proposed a method to estimate the human pose using segmentation-based skeleton extraction for upper-body RGB images. We built our own upper-body image dataset, which mostly consists of working and cleaning scenarios on a table similar view of working robots. We trained an FCN with the joint features for getting a joint confidence map and inferred the articulated pose with the consideration of adjacent joints' confidence maps. We have evaluated our method through several experiments. Our method outperforms the state-of-the-art method in accuracy and frame rate suggesting that it is applicable to a real robot.

However, our proposed method still has a few limitations. The performance of the proposed algorithm can vary when there are occlusions and radical changes in human orientation. In addition, the current training data are generated manually and it takes a long time due to the time required for segmenting the images with pixel-unit.

As a future work, dealing with the following topics can improve our current method. Firstly, considering the arrangement of their limbs and recognizing face in the network can improve the performance with radical changes in human orientation. Secondly, auxiliary information such as depth, texture, and thermal data can be used for human pose estimation. Lastly, adding a great

amount of various upper-body training datasets for training the network will greatly increase the performance in occlusions especially for hands.

Acknowledgements

This work was supported by the Technology Innovation Program, 10045252, Development of robot task intelligence technology, supported by the Ministry of Trade, Industry, and Energy (MOTIE, Korea). This work was also supported by the ICT R&D program of MSIP/IITP (2016-0-00563, Research on Adaptive Machine Learning Technology Development for Intelligent Autonomous Digital Companion). The students are supported by Korea Minister of Ministry of Land, Infrastructure and Transport (MOLIT) as U-City Master and Doctor Course Grant Program.

References

Aggarwal, J.K. and Park, S. (2004), "Human motion: Modeling and recognition of actions and interactions", *Proceedings of the 2nd International Symposium on 3D Data Processing, Visualization and Transmission* (*3DPVT*), Thessaloniki, Greece, September.

Andriluka, M., Pishchulin, L., Gehler, P. and Schiele, B. (2014), "2D human pose estimation: New benchmark and state of the art analysis", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Washington, D.C., U.S.A., June.

Asus (2015), *Xtion Pro Live*, <https://www.asus.com/3D-Sensor/Xtion_PRO_LIVE/>.

Chen, X. and Yuille, A.L. (2014), "Articulated pose estimation by a graphical model with image dependent pairwise relations", *Proceedings of the Advances in Neural Information Processing Systems Conference* (*NIPS*), Montreal, Canada, December.

Chu, X., Ouyang, W., Li, H. and Wang, X. (2016), "Structured feature learning for pose estimation", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, Nevada, U.S.A.,

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E. and Darrell, T. (2014), "Decaf: A deep convolutional activation feature for generic visual recognition", *Proceedings of the International Conference on Machine Learning* (*ICML*), Beijing, China, June.

Droeschel, D. and Behnke, S. (2011), "3D body pose estimation using an adaptive person model for articulated ICP", *Proceedings of the International Conference on Intelligent Robotics and Applications* (*ICIRA*), Aachen, Germany, December.

Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J. and Zisserman, A. (2015), "The pascal visual object classes challenge: A retrospective", *J. Comput. Vis.*, **111**(1), 98-136.

Ferrari, V., Marin-Jimenez, M., and Zisserman, A. (2008), "Progressive search space reduction for human pose estimation", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Anchorage, Alaska, U.S.A., June.

Fujiyoshi, H., Lipton, A.J. and Kanade, T. (2004), "Real-time human motion analysis by image skeletonization", *IEICE Trans. Inf. Syst.*, **87**(1), 113-120.

Ganapathi, V., Plagemann, C., Koller, D. and Thrun, S. (2010), "Real time motion capture using a single time-of-flight camera", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), San Francisco, California, U.S.A., June.

Guo, Y., Xu, G., and Tsuji, S. (1994), "Tracking human body motion based on a stick figure model", *J. Vis. Commun. Image R.*, **5**(1), 1-9.

Haritaogalu, I. (1998), "W4S: A real-time system for detecting and tracking people in 2 1/2-D", *Proceedings of the 5th European Conference on Computer Vision* (*ECCV*), Freiburg, Germany, June.

Hernández-Vela, A., Zlateva, N., Marinov, A., Reyes, M., Radeva, P., Dimov, D. and Escalera, S. (2012), "Graph cuts optimization for multi-limb human segmentation in depth maps", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Providence, Rhode Island, U.S.A., June.

Jain, H.P., Subramanian, A., Das, S. and Mittal, A. (2011), "Real-time upper-body human pose estimation using a depth camera", *Proceedings of the International Conference on Computer Vision/Computer Graphics Collaboration Techniques and Applications*, Rocquencourt, France, October.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S. and Darrell, T. (2014), "Caffe: Convolutional architecture for fast feature embedding", *Proceedings of the 22nd ACM International Conference on Multimedia*, Orlando, Florida, U.S.A., November.

Johnson, S., and Everingham, M. (2010), "Clustered pose and nonlinear appearance models for human pose estimation", *Proceedings of the British Machine Vision Conference* (*BMVC*).

Kim, D.H., Park, G.M., Yoo, Y.H., Ryu, S.J., Jeong, I.B. and Kim, J.H. (2017), "Realization of task intelligence for service robots in an unstructured environment", *Ann. Rev. Control*, **44**, 9-18.

Kim, H., Lee, S., Kim, Y., Lee, D., Ju, J. and Myung, H. (2015), "Human pose estimation algorithm for low-cost computing platform using depth information only", *Proceedings of the International Conference on Robot Intelligence Technology and Applications* (*RiTA*), Bucheon, Korea, December.

Kim, H., Lee, S., Kim, Y., Lee, S., Lee, D., Ju, J. and Myung, H. (2016), "Weighted joint-based human behavior recognition algorithm using only depth information for low-cost intelligent video-surveillance system", *Expert Syst. Appl.*, **45**, 131-141.

Kim, H., Lee, S., Lee, D., Choi, S., Ju, J. and Myung, H. (2015), "Real-time human pose estimation and gesture recognition from depth images using superpixels and SVM classifier", *Sensors*, **15**(6), 12410-12427.

Kingma, D.P. and Ba, J. (2015), *Adam: A Method for Stochastic Optimization*, arXiv preprint arXiv:1412.6980.

Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012), "Imagenet classification with deep convolutional neural networks", *Proceedings of the Advances in Neural Information Processing Systems* (*NIPS*), Lake Tahoe, Nevada, U.S.A., December

Ladicky, L., Torr, P.H. and Zisserman, A. (2013), "Human pose estimation using a joint pixel-wise and part-wise formulation", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Portland, Oregon, U.S.A., June.

Lee, S., Kim, H., Lee, S., Kim, Y., Lee, D., Ju, J. and Myung, H. (2014), "Detection of a suicide by hanging based on a 3-D image analysis", *IEEE Sens. J.*, **14**(9), 2934-2935.

Lee, S., Koo, J., Kim, H., Jung, K. and Myung, H. (2017), "A robust estimation of 2D human upper-body poses using fully convolutional network", *Proceedings of the International Conference on Robot Intelligence Technology and Applications* (*RiTA*), Daejeon, Korea, December.

Long, J., Shelhamer, E. and Darrell, T. (2015), "Fully convolutional networks for semantic segmentation", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Boston, Massachusetts, U.S.A., June.

Moeslund, T.B., Hilton, A., and Krüger, V. (2006), "A survey of advances in vision-based human motion capture and analysis", *Comput. Vis. Image Und.*, **104**(2-3), 90-126.

NVIDIA (2018a), *GTX 1080*, <https://www.nvidia.com/en-us/geforce/products/10series/geforce-gtx-1080/>.

NVIDIA (2018b), *DIGITS*, <https://github.com/NVIDIA/DIGITS>.

Ohya, J. and Kishino, F. (1994), "Human posture estimation from multiple images using genetic algorithm", *Proceedings of the 12th IAPR International Conference on Pattern Recognition*, Jerusalem, Israel, October.

Plagemann, C., Ganapathi, V., Koller, D. and Thrun, S. (2010), "Real-time identification and localization of body parts from depth images", *Proceedings of the IEEE International Conference on Robotics and Automation* (*ICRA*), Anchorage, Alaska, U.S.A., May.

Presti, L.L. and La Cascia, M. (2016), "3D skeleton-based human action classification: A survey", *Pattern Recogn.*, **53**, 130-147.

Ramanan, D. (2007), "Learning to parse images of articulated bodies", *Proceedings of the Advances in Neural Information Processing Systems* (*NIPS*), Vancouver, British Columbia, Canada.

Roweis, S.T. and Saul, L.K. (2000), "Nonlinear dimensionality reduction by locally linear embedding", *Science*, **290**(5500), 2323-2326.

Sapp, B. and Taskar, B. (2013), "Modec: Multimodal decomposable models for human pose estimation", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Portland, Oregon, U.S.A., June.

Schwarz, L.A., Mkhitaryan, A., Mateus, D. and Navab, N. (2012), "Human skeleton tracking from depth data using geodesic distances and optical flow", *Image Vis. Comput.*, **30**(3), 217-226.

Shotton, J., Girshick, R., Fitzgibbon, A., Sharp, T., Cook, M., Finocchio, M., Moore, R., Kohli, P., Criminisi, A., Kipman, A. and Blake, A. (2013), "Efficient human pose estimation from single depth images", *IEEE Trans. Pattern Anal. Mach. Intell.*, **35**(12), 2821-2840.

Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., Cook, M. and Moore, R. (2013), "Real-time human pose recognition in parts from single depth images", *Commun. ACM*, **56**(1), 116-124.

Simonyan, K. and Zisserman, A. (2014), *Very Deep Convolutional Networks for Large-Scale Image Recognition*, arXiv preprint arXiv:1409.1556.

Straka, M., Hauswiesner, S., Rüther, M. and Bischof, H. (2011), "Skeletal graph based human pose estimation in real-time", *Proceedings of the British Machine Vision Conference* (*BMVC*), Dundee, Scotland, U.K., August-September.

Takahashi, K., Uemura, T. and Ohya, J. (2000), "Neural-network-based real-time human body posture estimation", *Proceedings of the 2000 IEEE Signal Processing Society Workshop*, Lafayette, Louisiana U.S.A.

Tompson, J., Goroshin, R., Jain, A., LeCun, Y. and Bregler, C. (2015), "Efficient object localization using convolutional networks", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Boston, Massachusetts, U.S.A., June.

Tompson, J.J., Jain, A., LeCun, Y. and Bregler, C. (2014), "Joint training of a convolutional network and a graphical model for human pose estimation", *Proceedings of the Advances in Neural Information Processing Systems* (*NIPS*), Montreal, Canada, December.

Toshev, A. and Szegedy, C. (2014), "DeepPose: Human pose estimation via deep neural networks", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Columbus, Ohio, U.S.A., June.

Wei, S.E., Ramakrishna, V., Kanade, T. and Sheikh, Y. (2016), "Convolutional pose machines", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, Nevada, U.S.A., June-July.

Xia, F., Wang, P., Chen, X. and Yuille, A. (2017), *Joint Multi-Person Pose Estimation and Semantic Part Segmentation*, arXiv preprint arXiv:1708.03383.

Yang, W., Ouyang, W., Li, H. and Wang, X. (2016), "End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, Nevada, U.S.A., June-July.

Zhang, Z., Seah, H.S., Quah, C.K. and Sun, J. (2013), "GPU-accelerated real-time tracking of full-body motion with multi-layer search", *IEEE Trans. Multimedia*, **15**(1), 106-119.

HM