A Robust Estimation of 2D Human Upper-body Poses using Fully Convolutional Network

Seunghee Lee[1], Jungmo Koo[1], Hyungjin Kim[1], Kwangyik Jung[1],
and Hyun Myung[1,2],

[1] Department of Civil and Environmental Engineering, KAIST,
Daejeon 34141, Korea
{seunghee.lee, jungmokoo, hjkim86, ankh88324,
hmyung}@kaist.ac.kr
http://urobot.kaist.ac.kr
[2] Robotics Program, KAIST,
Daejeon 34141, Korea
{hmyung}@kaist.ac.kr

Abstract. We present an approach to efficiently detect the 2D human upper-body pose in RGB images. Among the system for estimating the joints position, the method using only RGB camera sensor is very cost-effective compared to the system with high-priced sensors such as a motion capture system. In this work, we use semantic segmentation using a fully convolutional network to estimate the upper-body poses of each skeleton and choose the location coordinate using joint heatmaps. The architecture is designed to learn joint locations and their association via the sequential prediction process. We demonstrate the performance of the proposed method using various datasets.

1 Introduction

Human body pose estimation has been actively researched with the aim of fast and reliable skeleton extraction results for decades. It has been applied to gaming, animation, surveillance, medical diagnose, security, and human-computer interaction (HCI) [1] [2]. Pose estimation is mainly used for the gesture recognition. Perceiving human gesture has an important role for worldwide usage without using various languages but with body language and also for deaf people signing communication in a good way. Also, it can be used for the surveillance system for CCTV for the crime or the alarming system for the emergency of elder people living alone. With the accurate skeleton extraction, correcting the posture in a medical way could be possible. Such requires keep the research of the robust and fast skeleton extraction.

Two different approaches for automatic human body pose estimation methods are categorized as follows: graph-based and machine learning-based methods. The first category is the graph-based method where geodesic distance is mostly used. A geodesic distance is a straight line from the curved place [3]. 3D joints are represented easily in the graph-based approach. Plagemann *et al.* [4] find points which are under the maximum geodesic distance, and these points are identified as the body points with body descriptors. Schwarz et al. [5] attempts to find pose estimation using anatomical landmarks in a depth geodesic graph and inverse kinematics. The skeletal graph is also used in the graph-based approach. Straka *et al.* [6] used a skeletal tree-like graph as the human body.

The other category is the machine learning-based method where there is no pre-required information about human. Shotton *et al.* [7] presented a per-pixel classification of the joints and estimation of joint position both using random forests. Hernández-Vela *et al.* [8] used graph cut optimization for the image segmentation over the work per-pixel classification.

However, pose estimation is still considered as a difficult problem since 230 human joints movements are not fully evident. Moreover, clothing and body shape differences cause a variety of situations. Partial occlusions due to self-articulation, for example, hand covering the face or the frontal portion of the body, or occlusions due to external objects may cause ambiguities in the obtained results [9]. Only using graph-based pose estimation requires model calibration before starting the predictions of the joints which mean it is not adequate for the real-time application [10]. However, learning-based pose estimation doesn't require prior calibration and can be trained with vast images for robustness.

Robots should be able to perceive human in order to understand the environment and interact. Hence most interacting robots would look at a person in a similar view of the human, which means that robots would see the person in a close look, so not a full body in a distance, but a part of the person, for example, upper body, or lower body. In this study, we focused on the estimation of the upper-body pose which gives some idea about partly shown body estimation.

In the next section, we review related works on skeleton extraction. The following sections explain pose estimation, introducing our architecture. Finally, we demonstrate pose estimation experiment result and discuss the conclusion.

2 Related research trends

2.1 Skeleton extraction using depth information

Under the development of Kinect system from Microsoft, depth-based automatic human pose estimation system without marker has become a great study [11]. 640 * 480 image at 30 frames per second is obtained from the Kinect camera. Depth information is invariant to color and texture. Haritaoglu et al. [12] divided the blob using geometrical characteristics and decided different joints (head, hands, and feet). Similarly, Fujiyoshi et al. [13] also predict the blob of a head, hands, and feet with the characteristics without template model. Guo et al. [14] determine the location of full body points by using distance function for fitting the model. Neural networks [15] and genetic algorithms [16] have also been used to obtain the complete position of all of the joints of the person.

The most well-known depth using method is from Shotton et al. [7] inferring 20 joints for every pixel classification in every single depth image from Kinect sensor. A randomized decision forest is performed for the pixel classification. A random decision forest is operated by building decision trees when training and giving the output for the body part classes. A weighted Gaussian kernel with mean shift is used to generate 3D joint position proposals. The classifier has been trained by using a vast amount of database for the variety of motions and different body shape-people.

OpenNI library is also widely used for skeleton estimation. 15 skeleton joints are estimated by depth-edge counting local descriptors [17]. Canny edge detector extracts depth edges from the depth map, and statics about edge pixels are computed in each patch. Then the locations of skeletal joints are searched by the approximate nearest neighbors (ANN) algorithm for matching the patch descriptors. By the time all the patch descriptors are compared, the body joint locations are determined one-by-one.

2.2 Skeleton extraction using RGB image information

Using image data for inferring human joints in the image has been suspicious for variation of color difference and human body shape. However, emerging of the advanced hardware has made it possible to train a vast number of datasets using the neural network. Many skeleton extraction methods use graphic processing units for the increase of robustness and frame rate [8, 18-20].

Recently, Deep Convolutional Neural Networks (DCNNs) using RGB images have achieved excellent performance on human pose estimation [21]. DeepPose [22] used regression learning for the body joint locations with the convolutional network. Tompson et al. [23, 24] suggested multi-resolution DCNN for the better accuracy by reducing the pooling effect. Chen et al. [25] proposed convolutional network dependent on pairwise spatial joints relationships. Chu et al. [26] used the relationship among body joints at the feature level.

Evaluation of the graphical methods is mostly held with renowned public pose estimation data sets. There are various benchmarks for human pose estimation. The Frames Labeled in Cinema (FLIC) dataset is upper body images obtained from Hollywood movies, and it consists of 4000 training images and 1000 test images [27]. The Leeds Sports Poses (LSP) dataset consists of sports images with 14 body joints and contains 11000 training images and 1000 testing images [28]. The Image Parse (PARSE) dataset [29], the MPII Human Pose Dataset [30], the Buffy dataset [31], and so on are also popularly used.

3 Pose Estimation Algorithm

In this study, segmentation based learning classifier is used for the estimation of the upper-body human pose. The total number of joints is seven, and the joints are a head, left shoulder, left elbow, left hand, right shoulder, right elbow, and right hand.

3.1 Fully Convolutional Network

For human upper-body segmentation, we deploy a network architecture that is based on and initialized by segmentation network of Long et al. [32]. They cast the problem of deep classification. They adapt contemporary classification networks such as Alexnet [33] into fully convolutional networks and fine-tune [34] to the segmentation task so they can pass their learned representations. FCN is trained end-to-end and predicts pixel-wise segmentation result. Dense feedforward and backpropagation computation enable both learning and inference on the whole image at one time.

Our segmentation network is focused on the human-body oriented and trained on our upper-body pose dataset. Details on the network architecture and its training procedure are provided in chapter 4.2.

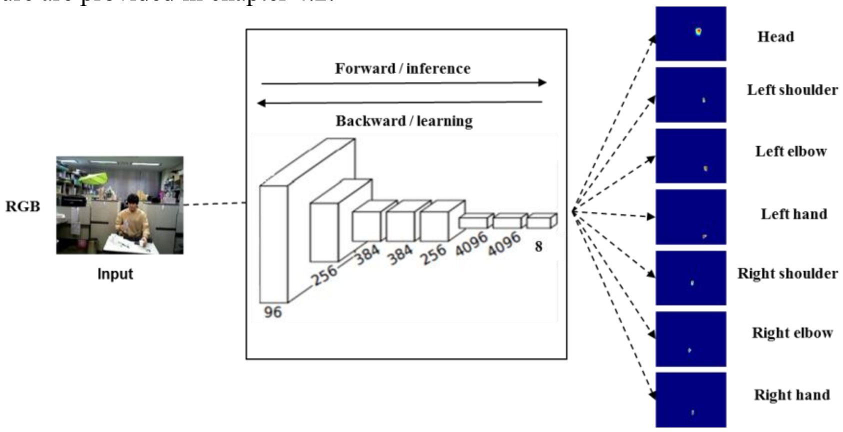

Fig. 1. The architecture of the proposed method. We show a convolutional architecture based on segmentation training and learning. The network results in heatmaps of each joint which are the probability maps of each joint existence.

3.2 Pose estimation using joint confidence map

For human pose estimation, we formulate localization of 2D keypoints as estimation of 2D confidence maps, where each map contains information about the likelihood of a certain body joint is present at a location.

Given the image feature representation produced by the encoder, an initial score map is predicted and calculated the central moment of the region. A complete overview over the network architecture is located in Chapter 4.

4 Evaluation

4.1 Datasets for experiments

For training dataset, there is no available dataset that provides segmentation image of human upper-body. Therefore, we complement them with a new dataset by ourselves. Our dataset is built upon 3 different persons performing 12 different scenarios including cleaning the table, making a cereal in a bowl, reading a book, etc.

For each new scenario, we randomly sample a new camera location, which is roughly located at the probable place where a camera of the humanoid robot Mybot, developed in the Robot Intelligence Technology Laboratory at KAIST, would take place.

In total, our dataset provides 433 images for training and 58 images for evaluation with a resolution of 640 x 480 pixels. All samples come with full annotation of 7 joints: head, left shoulder, left elbow, left hand, right shoulder, right elbow, and right hand. Fig. 2 shows a sample of the dataset.

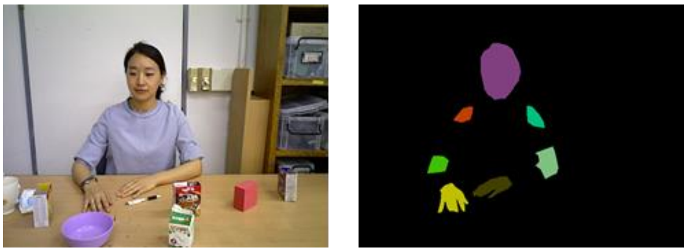

Fig. 2. Our new dataset provides segmentation image with 7 classes: head, left shoulder, left elbow, left hand, right shoulder, right elbow, and right hand.

4.2 Training and Testing method

We used an ASUS Xtion pro to estimate the joints poses. The camera is fixed at the height of Mybot, and the movement of the camera downward or above is about ±15. For the computation, Intel i5 3.0 GHz quad-core CPU is used, and two GTX1080 were used for training and testing. We define and implement our model using the Caffe[35] libraries and DIGITS for deep learning.

It was trained for upper-body joints segmentation with a batch size of 1 and using Adam solver [36]. The network architecture is initialized using weights of FCN-AlexNet and fine-tuned with a pre-trained model by PASCAL-VOC dataset [37]. The learning rate was $1 \cdot 10^{-4}$, and sigmoid decay was used for the learning rate policy.

4.3 Experimental results

In this section, we present our numerical results with our dataset. We trained the network for 30 epochs, which takes an average of 0.5 hours on the Nvidia DIGITS.

We show in Fig 3 our results on the upper-body image. Each heatmap gives the probability map of each joint's existence. The red is the highest confidence, and the blue is the lowest confidence. With the segmentation result according to heatmaps of each joint, final pose estimation result is obtained as shown in Fig 4. Different View and environment dataset were tested as the test.

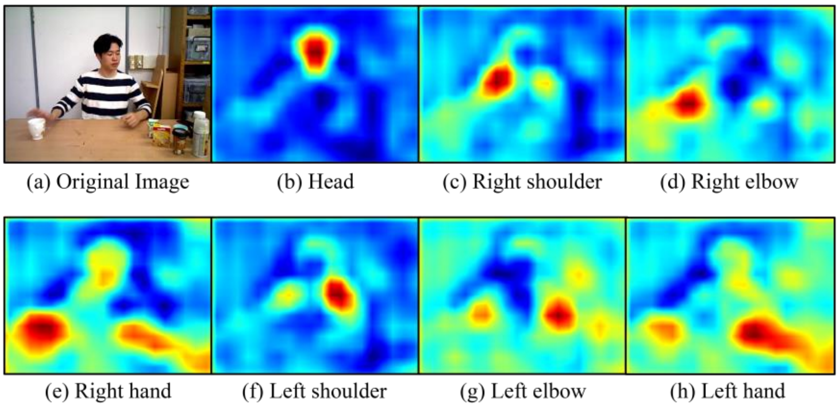

Fig. 3. Heatmap results of our method on each joint. We see that the method is able to provide the confidence map of each joint. Red color in the map represents the higher confidence, and blue color represents the lower confidence. [38]

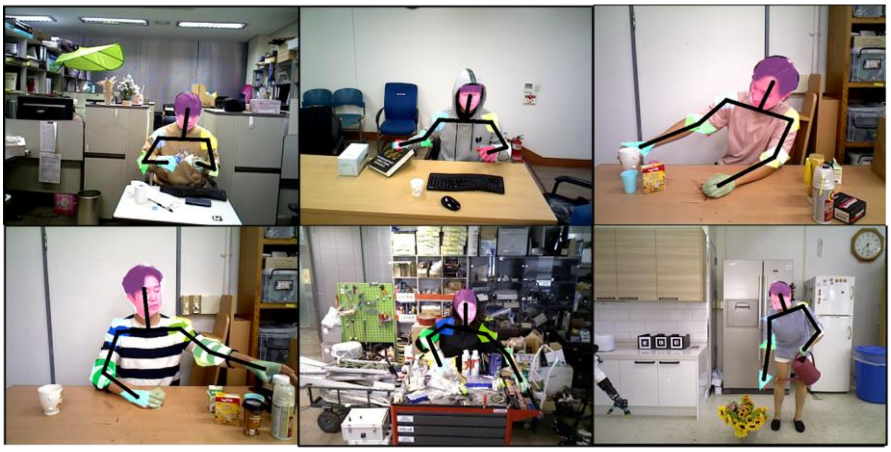

Fig. 4. Upper-body pose estimation results on different environment and person. [38]

For the accuracy evaluation, we have set the ground truth manually for the random test images as we don't have ground truth dataset. As the Table 1 describes, the average of error length of all the joints with the proposed algorithm is about 45.25mm, and all errors are below 72mm. The error length of the left elbow was the highest. One reason might be because there were many datasets which are occluded the left elbow by working on the desk.

Table 1. Experimental results of the distance errors for each joint for the test dataset.

Head	Left Shoulder	Left El-bow	Left Hand	Right Shoulder	Right Elbow	Right Hand
6.35	43.15	56.43	50.05	33.11	72.09	55.57

5 Conclusion

In this paper, we proposed a method of recognizing the pose attitude through the segmentation-based articulated body joints estimation for upper-body RGB images. In addition, we made upper-body image dataset on our own consisting of mostly working and cleaning situation on the table. We trained the joint features with a fully convolutional network to get a part confidence map and infer the body pose by calculating the central moment of each joint's confidence map. This method is expected to be applicable to a real robot including Mybot. In the future, a further research is needed to enable joint position estimation even when harsh articulation of the body joints and radical change of the human orientation appear.

6 Acknowledgement

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). 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

[1] Aggarwal J, Cai Q (1999) Human Motion Analysis: A Review. Computer Vision and Image Understanding 73:428-440.

[2] Moeslund T, Hilton A, Krüger V (2006) A survey of advances in vision-based human motion capture and analysis. Computer Vision and Image Understanding 104:90-126.

[3] Oxford dictionaries. http://oxforddictionaries.com/definition/english/VAR. Accessed 16 Oct 2017

[4] Plagemann, C., Ganapathi, V., Koller, D., & Thrun, S. (2010, May). Real-time identification and localization of body parts from depth images. In Robotics and Automation (ICRA), 2010 IEEE International Conference 3108-3113

[5] Schwarz, L. A., Mkhitaryan, A., Mateus, D., & Navab, N. (2012). Human skeleton tracking from depth data using geodesic distances and optical flow. Image and Vision Computing, 30: 217-226.

[6] Straka, M., Hauswiesner, S., Rüther, M., & Bischof, H. (2011). Skeletal Graph Based Human Pose Estimation in Real-Time. BMVC 1-12.

[7] Shotton J, Sharp T, Kipman A et al. (2013) Real-time human pose recognition in parts from single depth images. Communications of the ACM 56:116.

[8] Hernández-Vela, A., Zlateva, N., Marinov, A., Reyes, M., Radeva, P., Dimov, D., & Escalera, S. (2012). Graph cuts optimization for multi-limb human segmentation in depth maps. Computer Vision and Pattern Recognition (CVPR) 726-732.

[9] Droeschel, D., & Behnke, S. (2011). 3D body pose estimation using an adaptive person model for articulated ICP. Intelligent Robotics and Applications. 157-167.

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

[11] Jain, H., Subramanian, A., Das, S., & Mittal, A. (2011). Real-time upper-body human pose estimation using a depth camera. Computer Vision/Computer Graphics Collaboration Techniques. 227-238.

[12] HARITAOGALU, I. (1998). W4S: A real-time system for detecting and tracking people in 2 1/2-D. European Conference on Computer Vision.

[13] Fujiyoshi, H., Lipton, A. J., & Kanade, T. (2004). Real-time human motion analysis by image skeletonization. IEICE TRANSACTIONS on Information and Systems. 87(1): 113-120.

[14] Guo, Y., Xu, G., & Tsuji, S. (1994). Tracking human body motion based on a stick figure model. Journal of Visual Communication and Image Representation. 5(1): 1-9.

[15] Ohya, J., & Kishino, F. (1994). Human posture estimation from multiple images using genetic algorithm. Pattern Recognition, 1994. Vol. 1-Conference A: Computer Vision & Image Processing. Proceedings of the 12th IAPR International Conference. Vol. 1. 750-753

[16] Takahashi, K., Uemura, T., & Ohya, J. (2000). Neural-network-based real-time human body posture estimation. In Neural Networks for Signal Processing X, 2000. Proceedings of the 2000 IEEE Signal Processing Society Workshop Vol. 2. 477-486

[17] Presti, L. L., & La Cascia, M. (2016). 3D skeleton-based human action classification: A survey. Pattern Recognition. 53: 130-147.

[18] Zhang, Z., Seah, H. S., Quah, C. K., & Sun, J. (2013). GPU-accelerated real-time tracking of full-body motion with multi-layer search. IEEE Transactions on Multimedia. 15(1): 106-119.

[19] Shotton, J., Girshick, R., Fitzgibbon, A., Sharp, T., Cook, M., Finocchio, M., ... & Blake, A. (2013). Efficient human pose estimation from single depth images. IEEE Transactions on Pattern Analysis and Machine Intelligence. 35(12): 2821-2840.

[20] Ganapathi, V., Plagemann, C., Koller, D., & Thrun, S. (2010). Real time motion capture using a single time-of-flight camera. Computer Vision and Pattern Recognition (CVPR). 755-762

[21] Yang, W., Ouyang, W., Li, H., & Wang, X. (2016). End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation. IEEE Conference on Computer Vision and Pattern Recognition. 3073-3082.

[22] Toshev, A., & Szegedy, C. (2014). Deeppose: Human pose estimation via deep neural networks. IEEE Conference on Computer Vision and Pattern Recognition. 1653-1660.

[23] Tompson, J., Goroshin, R., Jain, A., LeCun, Y., & Bregler, C. (2015). Efficient object localization using convolutional networks. IEEE Conference on Computer Vision and Pattern Recognition. 648-656.

[24] Tompson, J. J., Jain, A., LeCun, Y., & Bregler, C. (2014). Joint training of a convolutional network and a graphical model for human pose estimation. Advances in neural information processing systems. 1799-1807.

[25] Chen, X., & Yuille, A. L. (2014). Articulated pose estimation by a graphical model with image dependent pairwise relations. Advances in Neural Information Processing Systems 1736-1744.

[26] Chu, X., Ouyang, W., Li, H., & Wang, X. (2016). Structured feature learning for pose estimation. IEEE Conference on Computer Vision and Pattern Recognition. 4715-4723.

[27] Sapp, B., & Taskar, B. (2013). Modec: Multimodal decomposable models for human pose estimation. IEEE Conference on Computer Vision and Pattern Recognition. 3674-3681.

[28] Johnson, S., & Everingham, M. (2010). Clustered pose and nonlinear appearance models for human pose estimation.

[29] Ramanan, D. (2007). Learning to parse images of articulated bodies. Advances in neural information processing systems. 1129-1136.

[30] Andriluka, M., Pishchulin, L., Gehler, P., & Schiele, B. (2014). 2d human pose estimation: New benchmark and state of the art analysis. IEEE Conference on computer Vision and Pattern Recognition. 3686-3693.

[31] Ferrari, V., Marin-Jimenez, M., & Zisserman, A. (2008). Progressive search space reduction for human pose estimation. Computer Vision and Pattern. 1-8.

[32] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556.

[33] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems. 1097-1105.

[34] Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., & Darrell, T. (2014, January). Decaf: A deep convolutional activation feature for generic visual recognition. International conference on machine learning. 647-655.

[35] Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., ... & Darrell, T. (2014, November). Caffe: Convolutional architecture for fast feature embedding. the 22nd ACM international conference on Multimedia. 675-678

[36] Kingma, D., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv:1412.6980.

[37] Everingham, M., Eslami, S. A., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2015). The pascal visual object classes challenge: A retrospective. International journal of computer vision. 111(1): 98-136.

[38] J. Koo, S. Lee, H. Kim, K. Jung, T. Oh, H. Myung (2017). Human upper-body pose estimation using fully convolutional network and joint heatmap. IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems.