Workshop on Robot Perception of Humans Baden-Baden, Germany,
June 11, 2018. In conjunction with IAS-15

CNN-based Human Body Orientation Estimation for Robotic Attendant

Yoshiki Kohari, Jun Miura, and Shuji Oishi

Department of Computer Science and Engineering,
Toyohashi University of Technology

Abstract. Estimating the state of a person, such as walking and talking with others, is an important function of attendant robot for generating appropriate behaviors. This paper describes an image-based human body orientation estimation using a convolutional neural network. By training the network with a massive SURREAL dataset, the network exhibits a high accuracy while keeping the calculation cost low enough for real-time applications. The evaluation has been done using our precisely annotated dataset.

Keywords: Body orientation estimation, attendant robot, convolutional neural network.

1 Introduction

Attending is one of the tasks of mobile service robots, which requires generating robot actions adaptively to the state (or behavior) of a person, such as walking, sitting, and taking with others. As a key information for the state estimation, we deal with the body orientation. It is related to social relationships to the surrounding people as well as the direction of motion and/or attention.

There are two major approaches to body orientation estimation. One is to use shape information obtained by range sensors like LIDARs [1–3]. The other is to use appearance information obtained by cameras [4–6]. The shape data is useful for numerically estimating orientation but weak to various disturbances such as clothings and items to hold. Image data can provide rich information and image-based CNNs (Convolutional Neural Networks) have been successfully applied to orientation classification but not for orientation estimation.

We pursue the image-based approach for orientation estimation. To increase the accuracy, we design a CNN and use a large dataset of synthetic images with additionally using our original real dataset. The proposed method is evaluated using real data to show its effectiveness.

The rest of the paper is organized as follows. Sec. 2 describes related work. Sec. 3 describes existing and improved CNNs for body orientation estimation. Sec. 4 describes a synthetic dataset and our real dataset. Sec. 5 shows experimental comparison results. Sec. 6 describes preliminary attempts to applying the body orientation estimation to attendant robot. Sec. 7 concludes the paper and discusses future work.

2 Related Work

2.1 LIDAR-based body orientation estimation

The section of waist-level body of a person obtained by a 2D LIDAR placed horizontally changes according to the relative orientation between the person and the robot. Glas et al. [1] estimate the orientation by combining LIDAR data from multiple viewpoints. Matsumoto et al. [2] recognize the person posture by extracting body shape features by multiple LIDARs. These methods require multiple LIDARs put at different positions and are thus not applicable to mobile robots applications.

Shimizu et al. [3] estimate the orientation by matching the scan data from a single LIDAR with a set of scans taken in advance with ten degree-interval of orientations. Since the estimation using only one scan data is sometimes inaccurate, they combines the result with the motion data based on the orientation-motion consistency (i.e., people usually walk forward). However, the method suffers from a low accuracy for stopping or very slow situations as well as is sensitive to body shape and/or bags and clothings.

2.2 Image-based body orientation estimation

Weinrich et al. [4] classify the body orientation using HOG [7] and SVM [8]. Ardiyanto and Miura [5] developed a classification method based on PLS (partial least squares)-based image feature selection. They also used the orientation-motion consistency for improving accuracy. Choi et al. [6] propose to use a lightweight CNN (convolutional neural network) for orientation classification. The classification rate of these works is 81% at best for classification of eight orientations. This is not enough for motion prediction and behavior recognition.

3 CNN-based orientation estimation

Convolutional neural networks (CNNs) are shown to exhibit very high performances in various recognition tasks [9, 10]. Choi et al. [6] use a CNN with two convolutional and three fully-connected layers for orientation classification. In image recognition tasks, much deeper networks have been used [11].

We adopt one of the best networks by Simonyan et al. [10] as a base network. The network has thirteen convolutional and three fully-connected layers. We modify the original network in the input and the output layers, the number of channels in the convolutional layers, and the number of nodes of the fully connected layers.

Fig. 1 shows the configuration of the proposed network. The input is gray-scale images normalized to 100×100 pixels. The output is the motion vector, represented by two elements (i.e., velocity in x and y directions). A linear function is used as an activation function.

Our body orientation estimation task is simpler than the original, a 1000-category classification task. We therefore reduce the size of the parameters from

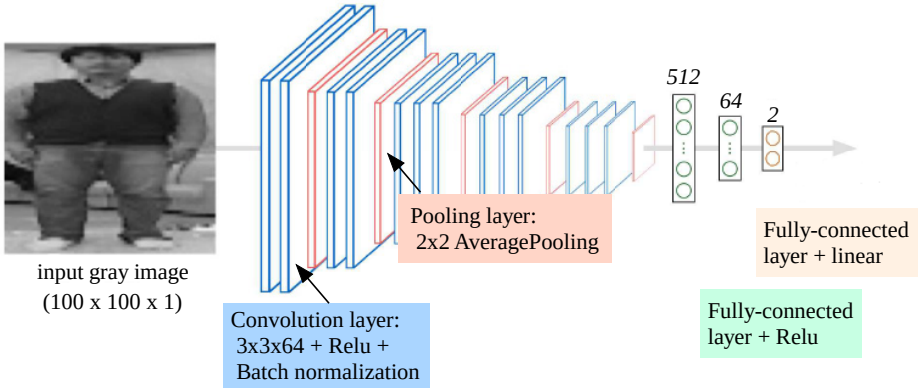


Fig. 1. Proposed network configuration.

the original one. The number of channels of all convolutional layers is set to 64 and the numbers of nodes of the fully-connected layers are set to 512, 64, and 2 from the input side to the output one. We also apply a batch normalization step [12] just after every convolutional layer.

4 Datasets

4.1 Dataset for training

The amount and the variety of the dataset used for training are crucial for effective deep learning. We used Synthetic hUmans foR REAL tasks (SURREAL) dataset [13]. This dataset includes about six million images, which are generated by combining synthetic person images with real background images. The dataset provides a set of an RGB image, a depth image, a part labels image, and 2D- and 3D joint locations, but does not provide body orientation nor a bounding box. We thus generate these additional data as follows.

The body orientation vector is calculated by (see Fig. 2):

1. Extract body-to-left shoulder and body-to-right shoulder vectors to form a basis.
2. Calculate the 3D body orientation vector as an outer product of these vectors.
3. Project the 3D vector to the ground plane (the X-Y plane).

The bounding box for each data is determined as the circumscribing rectangle of pixels with any body part labels (see Fig. 4). The number of images in the final dataset is 2,336,851.

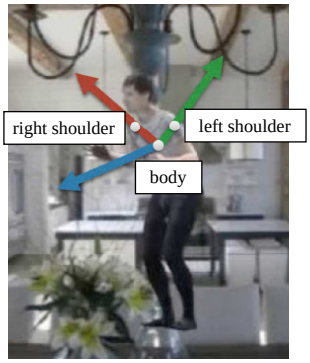


Fig. 2. Calculate the orientation vector in 2D.

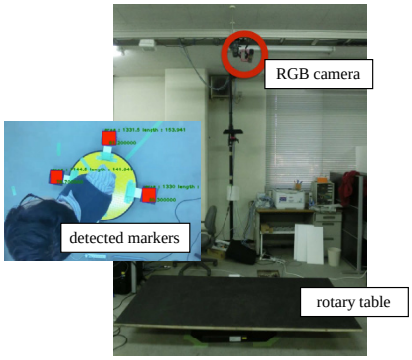


Fig. 3. Data collection system.

(a) Part label image. (b) Binary image. (c) Bounding box.

Fig. 4. Calculating a bounding box.

4.2 Data for testing

There are few data for human images with body orientation annotations. So we collected such data by ourselves using the system with a rotary table developed by Nishi and Miura [14] (see Fig. 3). The table can measure the rotation angle by observing the markers on it from the top. Sensors used for estimation are set horizontally to observe the person on the table, supposing the use by a mobile robot. Considering various application scenarios, we use the following three sensors: an RGB camera, a 2D LIDAR (Hokuyo UST-20LX), and a 3D LIDAR (Velodyne HDL-32e). We obtained data for thirteen male subjects.

Each data has the bounding box and body orientation annotations. Bounding boxes are determined by using SSD (single shot multibox detector) [15]. The number of images in the test dataset is 15,142.

5 Experiments

5.1 Training of the CNN

We use 90% of the data for training and the rest for validation. Table 1 summarizes the parameters used for training. Fig. 5 shows the training curve of the

Table 1. Training parameters

Loss function	MSE
Batch size	64
Learning rate	0.0001
Optimizer	Adam
Graphic card	GeForce Titan X Pascal

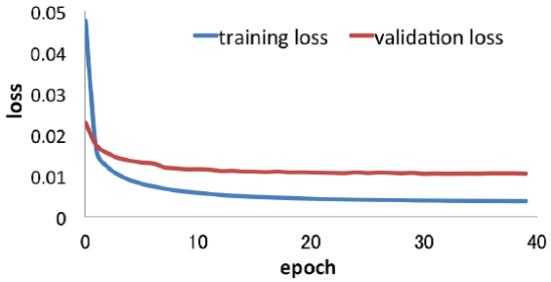


Fig. 5. Training curve of the network.

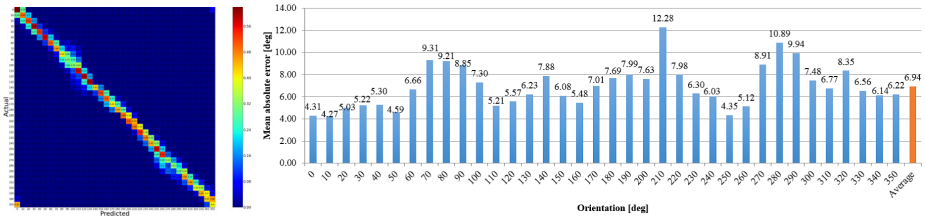
proposed network. The total training time was 82 hours (40 epochs) and the average training time per epoch was two hours. The validation loss decreases rapidly until epoch 3 and does gradually afterwards.

5.2 Performance evaluation

Performance evaluations have been done using the test (real) dataset. Fig. 6 shows the confusion matrix and mean absolute errors of the proposed method. The confusion matrix is generated by discretizing the estimation result into bins with the ten-degree interval. The accuracies with zero-degree, ten-degree, and twenty-degree allowance are 47.7%, 89.7%, and 97.5%, respectively. The averaged absolute error is about 6.9 [deg], which is accurate enough for orientation estimation. The averaged processing time is 0.00873 [sec].

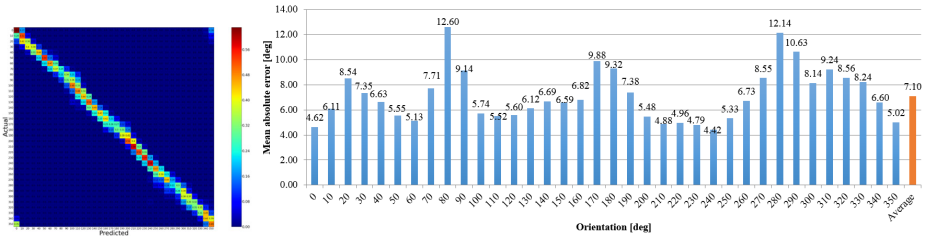
Fig. 7 shows the result of a modified network whose inputs are RGB images instead of gray images. Fig. 7 shows the result. The accuracies with zero-degree, ten-degree, and twenty-degree allowance are 42.9%, 89.3%, and 98.1%, respectively. The averaged absolute error is about 7.1 [deg]. The performance is slightly worse than that of the gray image-based. The averaged processing time is 0.00911 [sec].

To investigate the effectiveness of convolutional layers, we examine the performance of the network with three convolutional layers; this is a modification of [6], which outputs the 2D orientation vector. The accuracies with zero-degree, ten-degree, and twenty-degree allowance are 31.4%, 71.3%, and 86.9%, respectively. The averaged absolute error is about 15.5 [deg]. The averaged processing



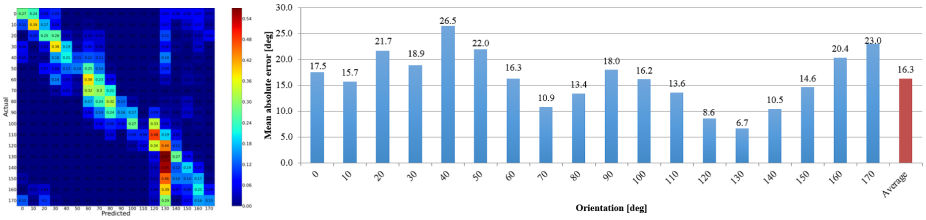
(a) Confusion matrix. (b) Mean absolute errors.

Fig. 6. Accuracy of the proposed method.



(a) Confusion matrix. (b) Mean absolute errors.

Fig. 7. Accuracy of the proposed network with RGB inputs.



(a) Confusion matrix. (b) Mean absolute errors.

Fig. 8. Accuracy of a LIDAR-based method [3].

time is 0.00434 [sec]. By increasing the number of convolutional layers (from three to thirteen), the accuracy increases with a small amount of extra computation.

We then compare the proposed image-based method with our LIDAR-based method [3] using the same test dataset. Since the body shape, which is usually ellipse-like, inherently has an ambiguity between orientations with 180 [deg] interval, we evaluated the method for the orientation range between $0 \sim 180$ [deg]. The accuracies with zero-degree, ten-degree, and twenty-degree allowance are 26.4%, 63.4%, and 78.9%, respectively. The averaged absolute error is about 16.3 [deg]. The averaged processing time is 0.00035 [sec].

Table 2. Comparison between body orientation estimation methods.

method	accuracy [%] (±0 [deg])	accuracy [%] (±10 [deg])	accuracy [%] (±20 [deg])	averaged error [deg]	averaged time [msec]
proposed (gray)	**47.7**	**89.7**	97.5	**6.94**	8.73
proposed (color)	42.9	89.3	**98.1**	7.10	9.11
modified Choi's	31.4	71.3	86.9	15.5	4.34
LIDAR-based	26.4	63.4	78.9	16.3	**0.35**

Table 2 summarizes and compares these results. From the table, the proposed method exhibits a nice performance with a reasonable processing time for real-time applications.

6 Application to Attendant Robot

This section shows an example of applying the proposed body orientation estimation to generating a robotic attending behavior. We have proposed a method of generating adaptive attending behaviors of a robot according to the user's states such as walking and sitting [16]. We here show the case for the user's sitting state.

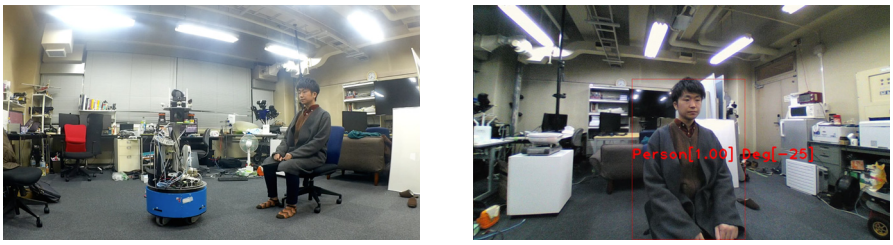
After recognizing that state, the robot plans its waiting position considering the user's comfort as well as collision avoidance. Reliably estimating the body orientation is an important factor for the user's comfort. Fig. 9 shows snapshots of the robot's attending behavior to a sitting person. The robot first detects a sitting person using our human detector [17] and estimate the body orientation using the proposed method (see Fig. 9(a)). The right image is the view from the robot camera and the numbers inside indicate the confidence of the person detection (left) and the estimated body orientation (right). The robot approaches to the waiting position while estimating the person position and the body orientation, thereby localizing itself with respect to the person (see Fig. 9(b)) and finally reaches to the waiting position (see Fig. 9(c)).

7 Conclusions and Future Work

This paper has described an image-based human body orientation estimation using a convolutional neural network. The input to the network is a cropped gray image of human region and the output is the 2D body orientation vector. By training the network with a massive SURREAL dataset, the network exhibits a high accuracy while keeping the calculation cost low enough for real-time applications. Especially it achieves a better accuracy than a LIDAR-based method that uses metric information directly. We have also shown an application of the body orientation estimation to an attendant robot.

The current network design is somehow heuristic and further investigation would be necessary to evaluate the trade-off between training/calculation costs

(a) Detect a sitting person.



(b) Approach to the waiting position.

(c) Reach to the waiting position.

Fig. 9. Attending behavior to a sitting person.

and estimation accuracy. We are also applying the method to recognizing a more variety of human behaviors.

Acknowledgment

This work is in part supported by JSPS KAKENHI Grant Number 17H01799.

References

1. D.F. Glas, T. Miyashita, H. Ishiguro, and N. Hagita. Laser-Based Tracking of Human Position and Orientation using Parametric Shape Modeling. *Advanced Robotics*, Vol. 23, No. 4, pp. 405–428, 2009.

2. T. Matsumoto, M. Shimosaka, H. Noguchi, T. Sato, and T. Mori. Pose Estimation of Multiple People using Contour Features from Multiple Laser Range Finders. In *Proceedings of 2009 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pp. 2190–2196, 2009.
3. M. Shimizu, K. Koide, I. Ardiyanto, J. Miura, and S. Oishi. LIDAR-based Body Orientation Estimation by Integrating Shape and Motion Information. In *Proceedings of 2016 IEEE Int. Conf. on Robotics and Biomimetics*, pp. 1948–1953, 2016.
4. C. Weinrich, C. Vollmer, and H.-M. Gross. Estimation of Human Upper Body Orientation Estimation for Mobile Robotics using an SVM Decision Tree on Monocular Images. In *Proceedings of 2012 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pp. 2147–2152, 2012.
5. I. Ardiyanto and J. Miura. Partial Least Squares-based Human Upper Body Orientation Estimation with Combined Detection and Tracking. *Image and Vision Computing*, Vol. 32, No. 11, pp. 904–915, 2014.
6. J. Choi, B.-J. Lee, and B.-T. Zhang. Human Body Orientation Estimation using Convolutional Neural Network. *CoRR*, Vol. abs/1609.01984, , 2016.
7. N. Dalal and B. Briggs. Histograms of Oriented Gradients for Human Detection. In *Proceedings of 2005 IEEE Conf. on Computer Vision and Patttern Recognition*, pp. 886–893, 2005.
8. C.J.C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, Vol. 2, No. 2, pp. 121–167, 1998.
9. A. Krizhevsky, I. Sutskever, and G.E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pp. 1097–1105. 2012.
10. K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
11. K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *Proceedings of 2016 IEEE Conf. on Computer Vision and Pattern Recognition*, 2016.
12. S. Ioffe and C. Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of 2015 Int. Conf. on Machine Learning*, pp. 448–456, 2015.
13. G. Varol, J. Romero, X. Martin, N. Mahmood, M.J. Black, I. Laptev, and C. Schmid. Learning from Synthetic Humans. In *Proceedings of 2017 IEEE Conf. on Computer Vision and Pattern Recognition*, 2017.
14. K. Nishi and J. Miura. A Head Position Estimation Method for a Variety of Recumbent Positions for a Care Robot. In *Proceedings of 2015 Int. Conf. on Advanced Mechatronics*, pp. 157–158, 2015.
15. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A.C. Berg. SSD: Single Shot Multibox Detector. In *Proceedings of 2016 European Conf. on Computer Vision*, pp. 21–37, 2016.
16. S. Oishi, Y. Kohari, and J. Miura. Toward a Robotic Attendant Adaptively Behaving according to Human State. In *Proceedings of 2016 IEEE Int. Symp. on Robot and Human Interactive Communication*, pp. 1038–1043, 2016.
17. K. Koide and J. Miura. Identification of a Specific Person using Color, Height, and Gait Features for a Person Following Robot. *Robotics and Autonomous Systems*, Vol. 84, No. 10, pp. 76–87, 2016.