Improvement of Human Action Recognition Using 3D Pose Estimation

Kohei Adachi, Paula Lago, Tsuyoshi Okita and Sozo Inoue

Abstract While human action recognition (HAR) using motion capture can perform well with high accuracy, it requires a high computational cost for recording and post-processing. To avoid this, we build a HAR system using 3D pose estimation from single-camera video instead of motion capture. One drawback in this approach is that the performance is considerably dependent on the camera position. This paper investigates how we can use the pose estimate constantly without the effect of camera position even when the camera position in the test data is changed. We augment the data by rotating around the 3D pose estimate to improve the accuracy when using different camera positions in the test data and in the training data. The strategy of augmenting training data shows improvements up to 55.7% in accuracy, compared with the case of 2D pose with no augmentation.

1 Introduction

Human action recognition (HAR) is a task of recognizing different types of activities from sensor or video data. This has become a popular research in ubiquitous computing [1]. HAR is often built with supervised learning which requires to collect labeled data. Takeda et al. [12] proposed a motion capture-based HAR system whose inputs are the position of reflective markers attached to the body. One demerit of

Kohei Adachi
Kyushu Institute of Technology, Fukuoka Japan e-mail: adachi@sozolab.jp

Paula Lago
Kyushu Institute of Technology, Fukuoka Japan e-mail: paula@sozolab.jp

Tsuyoshi Okita
Kyushu Institute of Technology, Fukuoka Japan e-mail: okita@sozolab.jp

Sozo Inoue
Kyushu Institute of Technology, Fukuoka Japan e-mail: sozo@sozolab.jp

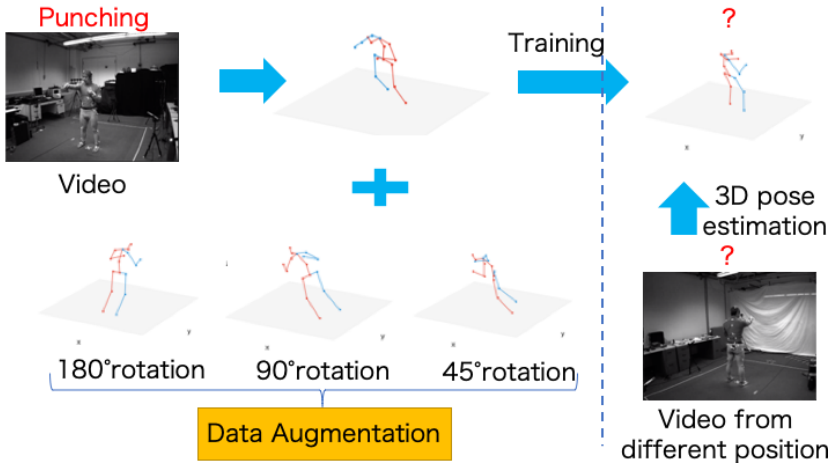
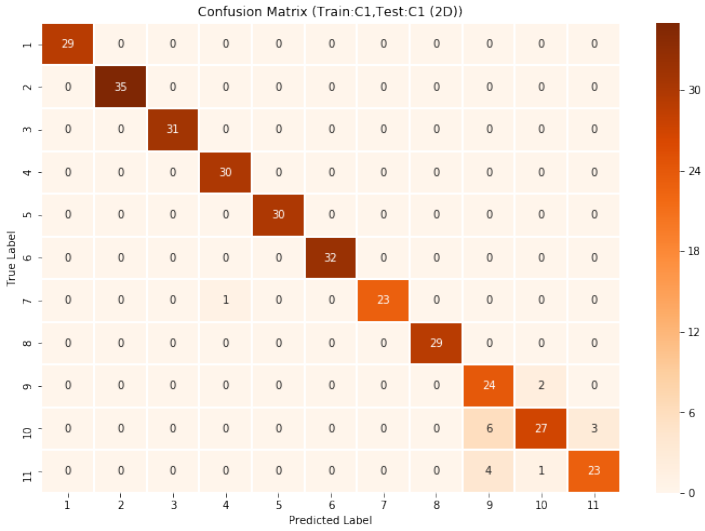
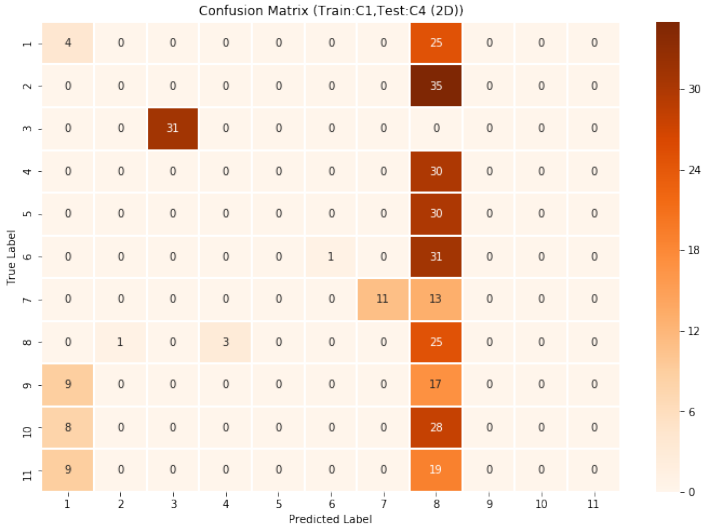


Fig. 1 After estimating the 3D pose from the video, rotate the 3D pose using affine transformations to augment the data and make it easier to recognize the 3D pose estimated from videos with different camera positions

this system is that it requires additional efforts other than the recording of motion capture, especially in its manual post-processing which requires several extra days. Recently, various deep learning methods have been proposed to estimate pose from video [6, 10], which has great advantages over motion capture in its speed since it doesn't require any manual process. It is straight-forward to build HAR systems based on this pose estimation. Unfortunately, one disadvantage is in its performance since we cannot avoid the inherent errors caused by the pose estimation which propagate to the entire system. For this reason, our previous work [4] targeted at measuring such defects in pose estimation in 2D and 3D individually, and compared their performance in HAR. One additional big obstacle for using pose estimation from video for HAR is the change in camera position which made the performance suddenly drop (Fig.2). In this paper, we have identified the cause of this problem. To solve it, we increased the training data size with data augmentation by rotating 3D pose using affine transformations (Fig.1). Data augmentation significantly improves the recognition accuracy. Our results show improvements up to 55.7% compared with the case of using 2D pose estimation as input for HAR.



(a) F1-Score:94.8%



(b) F1-Score:18.2%

Fig. 2 The confusion matrix when performing action recognition with the 2D pose estimated from video. If the test data is at the same camera position as the training data, it can be recognized with high accuracy (a), but if the camera position is different, the recognition accuracy will be low (b)

2 Related Work

HAR using 2D pose estimation is studied well. Takasaki et al. [11] aim to analyze activity in real time from video. When one or ten images are acquired from video, 2D pose information extracted by OpenPose is input as explanatory variables, and action is recognized using various machine learning methods.

Okita and Inoue [9] proposed a method to translate across multiple modalities (accelerometer, mocap, and camera) using an RNN encoder-decoder model. they used OpenPose to extract 2D pose from camera. In this work, we focus only on pose data from video without mocap nor accelerometer data.

Video-based HAR using different viewpoints has also been studied previously. Varol et al. [14] responded to the problem that the recognition accuracy deteriorated when changing the direction of the person with respect to the camera in video-based action recognition. They propose a method to improve recognition accuracy by creating artificially created videos and augmenting training data. In contrast, we augmented data by rotating 3D pose without increasing the image data.

In addition, several data augmentation methods for HAR using sensors have been proposed [8, 13]. Data augmentation for sensor data has shown to be effective for increasing the robustness of HAR systems to sensor position variability. In this work, we evaluate data augmentation for video data.

In this research, we focus on improving the robustness of video-based action recognition to variations in camera shooting position. To this end, we estimate the 3D pose from the video and augment the data by rotating the estimated 3D pose using affine transformations. This process will be explained in the next section.

3 Method

In this paper, we propose to use data augmentation to improve the accuracy of action recognition using pose estimation from video data as input. In this section, we describe our proposed method of data augmentation and action recognition from video-based pose estimation. Fig.3 shows the overview of our method. In short, we use video from one camera to estimate the body pose for each frame and use these poses as input for training an activity recognition model. Then we evaluate the recognition accuracy *when the test data is recorded using a different camera position*.

We evaluate the proposed method by comparing the results of action recognition using 2D and 3D poses estimated from video and using 3D poses combined with the proposed data augmentation method.

In what follows, we describe the the methods used for pose estimation, data augmentation, feature extraction and classification model.

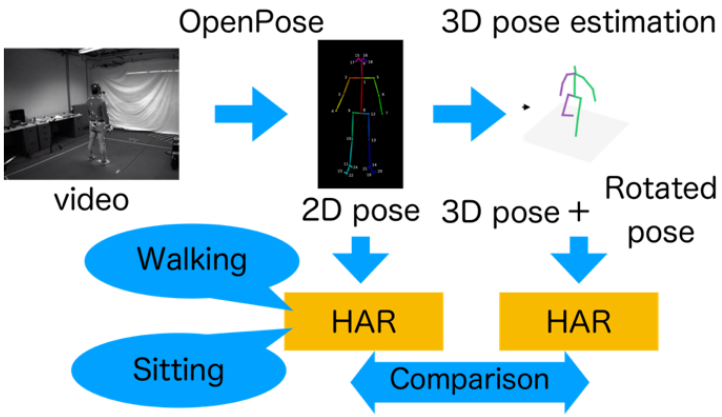


Fig. 3 Overview of our method. We estimate the 2D pose from the video using OpnePose, and further estimate the 3D pose from the 2D pose. After that, action recognition is performed using 2D pose and 3D pose and the 3D pose with data enhancement, and the recognition accuracy is compared.

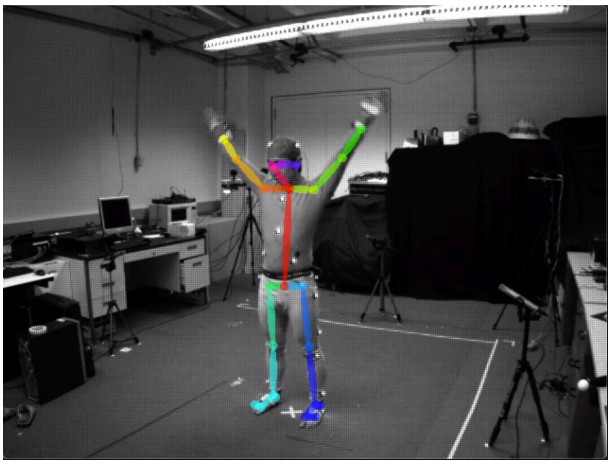


Fig. 4 2D keypoints extracted by OpenPose. There are total of 25 key points.

3.1 Extracting 2D pose from video

We used OpenPose [2] to get 2D key-points for each frame of each video. OpenPose is an opensource library for multi-person keypoint detection, which detects human body, feet hands, and facial key-points(135 key-points in total) In this work, we estimates only 25 key-points in the whole body using Openpose. Therefore, OpenPose returns the (X,Y) positions of 25 key-points, and there are 50 signals in total. An example of the extracted key points is shown in Fig.4.

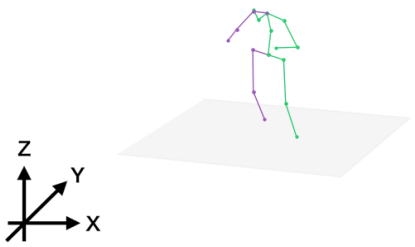


Fig. 5 3D keypoints estimated by 3d-pose-baseline

3.2 Estimating 3D pose from 2D pose

We estimated 3D pose from 2D keypoints extracted as explained in Sec.3.1. For the 3D pose estimation, we used the implementation of 3d-pose-baseline [6] pre-trained with the Human3.6M [3] dataset. This method estimates a total of 32 key-points consisting of (X, Y, Z), and there are 96 signals in total. An example of the estimated key points is shown in Fig.5.

3.3 Data Augmentation Using 3D Affine Transformation

To augment the training data, we rotate the 3D pose estimation around the vertical Z axis using affine transformation (Eq. (1)). The rotation matrix used is shown in Eq. (2). An example of the rotated 3D pose is shown in Fig. 6. We evaluate six different rotation patterns for data augmentation as shown in Table 1.

$$\begin{pmatrix} x_i' \\ y_i' \\ z_i' \end{pmatrix} = R_z(\theta) \begin{pmatrix} x_i \\ y_i \\ z_i \end{pmatrix}, i \in \{1...32\} \tag{1}$$

$$R_z(\theta) = \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{2}$$

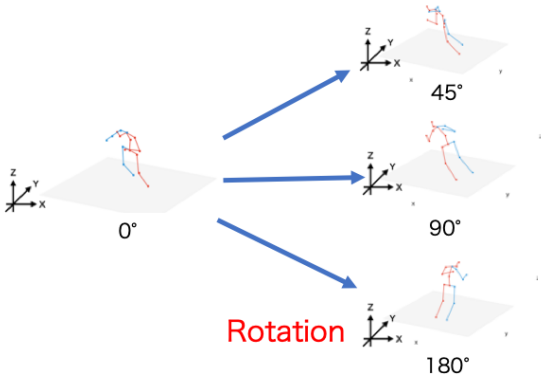


Fig. 6 Rotating the estimated 3D pose around the Z axis using affine transformation

Table 1 Rotation patterns for data augmentation

Pattern id.	Rotation Angles
1	45°
2	90°
3	180°
4	45°, 90°
5	45°, 180°
6	90°, 180°

3.4 Feature Extraction

For action recognition, we extracted features from the estimated poses. For each axis of each key-point, average, standard deviation, maximum value, and minimum value were extracted as features. Therefore, the number of features extracted from each pose is 200 in the 2D pose scenario and 384 in the 3D pose case.

3.5 Classification Model

The features obtained in Sect.3.4 are input as explanatory variables for an action classification model. We used the machine learning algorithm RandomForest [5] to create the action classification model. Table 2 shows that parameters used in the classification model of RandomForest.

Table 2 Parameters used in the classification model of RandomForest

Parameter	Value
Bootstrap	True
Critetion	gini
Max_Depth	None
Min_samples_leaf	1
Min_samples_split	2
N_estimators	500

Table 3 action labels: 11 classes

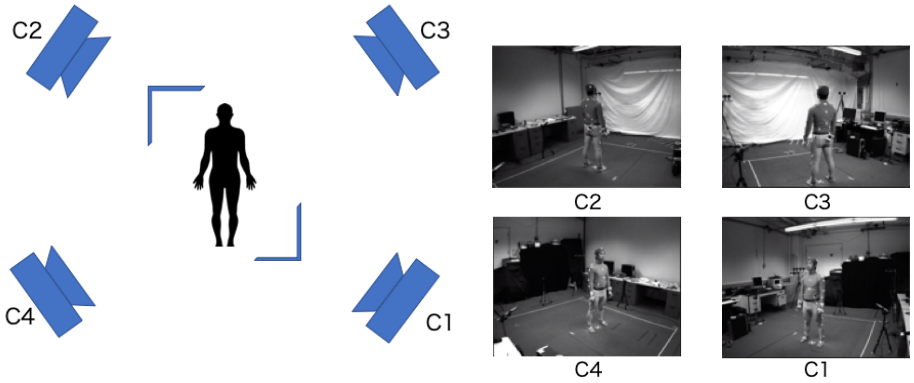
Class label	action
1	Jumping in place
2	Jumping jacks
3	Bending - hands up all the way down
4	Punching(boxing)
5	Waving - two hands
6	Waving - one hand(right)
7	Clapping hands
8	Throwing a ball
9	Sit down then stand up
10	Sit down
11	Stand up

4 Evaluation method

In this section, we describe the data set and evaluation method used for this experiment. The goal of the experiments is to understand if and when does data augmentation increase the robustness to camera shooting position variability.

4.1 Data Set

We used the Berkeley Multimodal Human Action Database (HMAD) [7] for this experiment. HMAD data set contains 11 actions, as listed in Table 3. These actions, performed by 12 subjects (7 males and 5 females), were recorded using audio, video, accelerometers, motion capture, and kinect. For this work, we used the video data. These data were taken from cameras placed surrounding the subject. Fig.7 shows the layout of cameras and an example image taken from each camera.



(a) Layout of cameras (b) Example taken from each camera

Fig. 7 Camera layout. Four cameras are installed around the subject. Subject performed actions look at C1 and C4 cameras

4.2 Evaluation

To evaluate the models, we split data into 50 percent for training and 50 percent for test randomly. Note that activities recorded at the same timing at different camera positions were not included in both the training data and the test data. For comparison, the evaluation was performed when the training data and test data were set at the same camera position. We use the F1-score of the test data as an evaluation measure. We evaluate using features from 2D poses, 3D poses, and 3D poses with augmented data for all camera position combinations, for a total of 128 test cases.

5 Result

In this section, we show results of this experiment. The following settings are compared between training data and test data:

- When using video from the same camera positions
- When using video from different camera positions
  - Without data augmentation
  - With data augmentation

Table 4 and Table 5 show the results of HAR for 2D pose and 3D pose. Table 6 shows the results of HAR for 2D pose, 3D pose and 3D pose with data augmentation.

Table 4 F1-Score results for test data when performing HAR for 2D pose. C1, C2, C3, and C4 indicate the camera positions. Bold text represents the best score for each combination.

		test data			
		C1	C2	C3	C4
training data	C1	**94.8**	39.1	46.4	18.2
	C2	58.0	**93.5**	57.3	41.8
	C3	40.0	47.4	**94.3**	27.1
	C4	54.4	26.7	11.4	**93.0**

Table 5 F1-Score results for test data when performing HAR for 3D pose. C1, C2, C3, and C4 indicate the camera positions. Bold text represents the best score for each combination

		test data			
		C1	C2	C3	C4
training data	C1	**96.4**	20.0	28.1	33.5
	C2	26.1	**93.2**	34.0	20.5
	C3	1.37	28.2	**94.8**	1.29
	C4	53.3	7.34	4.82	**96.6**

5.1 When Using Video from the same Camera Positions

From the results of using videos from the same camera position for training and test (diagonal of Tables 4 and 5), it was found that actions can be recognized with high accuracy from both the estimated 2D and 3D poses in this scenario.

5.2 When Using Video from different Camera Positions without Data Augmentation

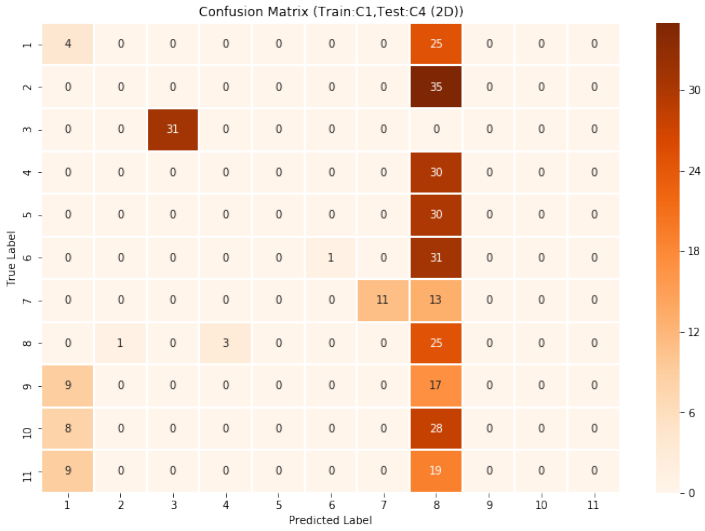
From the result of using videos from different camera positions, such as training data from C1 and test data from C2 (non-diagonal cells of Tables 4 and 5), we observe that the recognition accuracy drops significantly compared to that of using videos from the same camera position. We also observe that the recognition accuracy when using features from the 2D pose was higher than that of the method using features from the 3D pose. There were 11 cases out of 12 cases in which the 2D pose had better performance. The recognition accuracy of the 3D pose was higher than that of the 2D pose only when the training data was from C1 and the test data was from C4. In that case, the recognition accuracy was improved by 15.3%.

On the other hand, the recognition accuracy of the 3D pose had the the biggest drop from that of the 2D pose when the training data was C3 and the test data was C1. In this case, the recognition accuracy deteriorated by 38.6%.

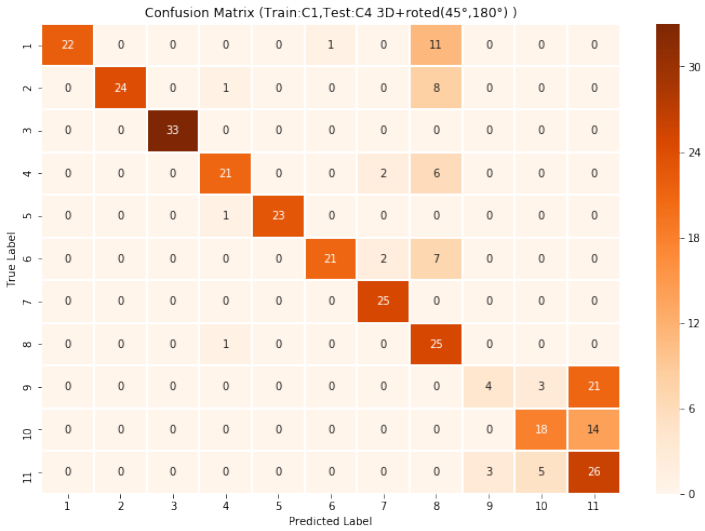
5.3 When Using Video from different Camera Positions with Data Augmentation

Recognition accuracy was improved in 10 out of 12 cases when data augmentation was applied to training data and the training and testing data come from different camera positions. The only combinations that did not benefit from data augmentation were C1-C2 and C2-C1 (Table 5). Fig.5.3 shows an example in which the recognition accuracy is most improved. This is when training data is from camera C1 and test data from C4 and the training data is augmented by rotating the 3D pose by 45° and 180°. In this case, the recognition accuracy is improved by 55.7% compared to the model using 2D pose features.

On the other hand, Fig.5.3 shows an example in which the recognition accuracy has deteriorated. This is the case when training data is from camera C3 and test data from camera C1 and the 3D pose of the training data was rotated by 45° and included as augmented training data. In this case, the recognition accuracy was deteriorated by 22.8%.

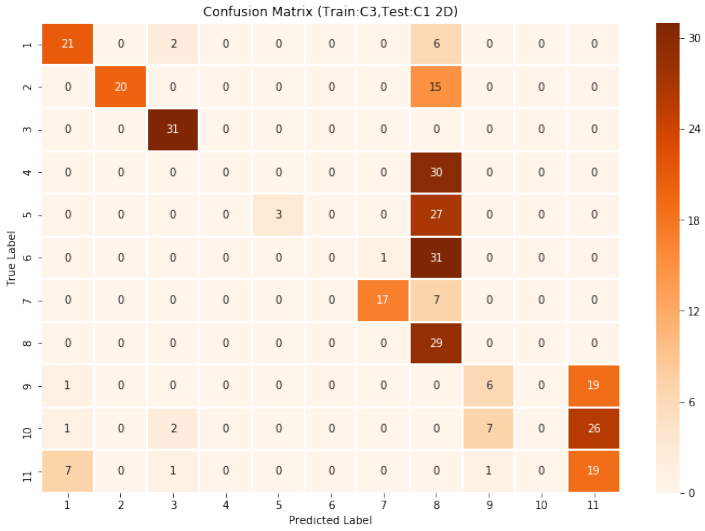


(a) F1-Score:18.2%)

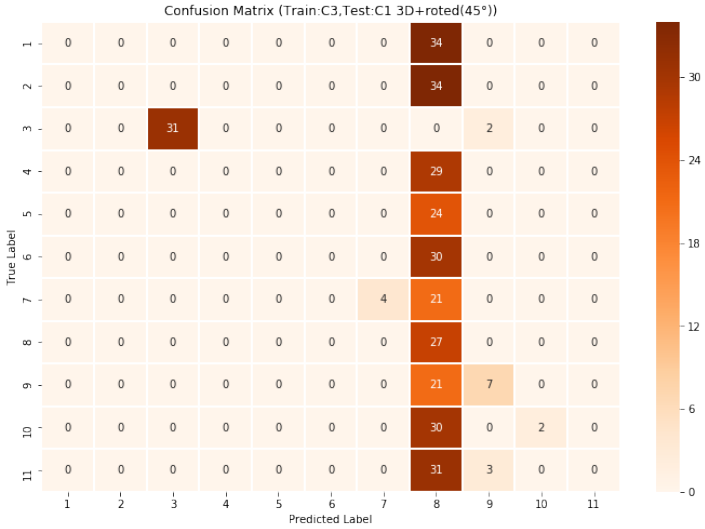


(b) F1-Score:73.9%

Fig. 8 An example of improvement. Training data is C1, test data is C4 camera. (a) Using 2D pose. (b) Using 3D pose with data augmentation(rotated by 45° and 180°)



(a) F1-Score:40.0%



(b) F1-Score:17.2%

Fig. 9 An example of depreciation. Training data is C3, test data is C1 camera. (a) Using 2D pose. (b) Using 3D pose with data augmentation(rotated by 45°)

Table 6 F1-Score results for test data when performing HAR for 2D pose, 3D pose, and 3D pose with data augmentation(%). The vertical axis represents the camera used for training, and the horizontal axis represents the camera used for testing. C1, C2, C3, and C4 indicate the camera positions. The cameras C1 and C4 take the front of the person, and the cameras C2 and C3 take the back of the person. Bold text represents the best score for each combination of training and testing camera

	Test data			
Training data	C1	C2	C3	C4
C1(2D)	94.8	**39.1**	46.4	18.2
C1(3D)	**96.4**	20.0	28.1	33.5
C1(+45°)	95.3	32.7	41.1	62.7
C1(+90°)	96.4	27.5	40.3	62.8
C1(+180°)	93.9	37.4	**47.5**	68.7
C1(+45°, +90°)	93.6	30.3	40.4	41.0
C1(+45°, +180°)	94.8	35.7	42.6	**73.9**
C1(+90°, +180°)	95.6	31.4	44.2	72.5
C2(2D)	**58.0**	**93.5**	57.3	41.8
C2(3D)	26.1	93.2	34.0	20.5
C2(+45°)	43.7	91.5	65.5	39.1
C2(+90°)	45.6	90.7	64.7	36.2
C2(+180°)	56.2	91.2	**77.8**	**51.6**
C2(+45°, +90°)	46.4	90.0	69.1	41.8
C2(+45°, +180°)	45.7	89.8	70.9	44.9
C2(+90°, +180°)	46.4	90.7	69.1	41.8
C3(2D)	40.0	47.4	94.3	27.1
C3(3D)	1.37	28.2	**94.8**	1.29
C3(+45°)	17.2	42.5	94.5	7.4
C3(+90°)	34.5	45.8	93.3	32.4
C3(+180°)	**40.5**	44.5	93.0	**40.8**
C3(+45°, +90°)	36.6	47.0	93.3	34.0
C3(+45°, +180°)	36.1	**49.6**	93.3	37.2
C3(+90°, +180°)	35.4	47.4	93.9	39.8
C4(2D)	54.4	26.7	11.4	93.0
C4(3D)	53.3	7.34	4.82	**96.6**
C4(+45°)	71.4	15.1	11.5	**96.6**
C4(+90°)	78.1	45.4	46.1	95.9
C4(+180°)	72.7	45.7	47.6	94.4
C4(+45°, +90°)	**86.8**	47.2	44.2	95.0
C4(+45°, +180°)	77.1	42.3	46.1	95.9
C4(+90°, +180°)	84.1	**47.8**	**49.1**	95.2

6 Discussion

In this section, we discuss the following two points:

- Improvement of recognition accuracy by data augmentation.
- Deterioration of recognition accuracy by data augmentation.

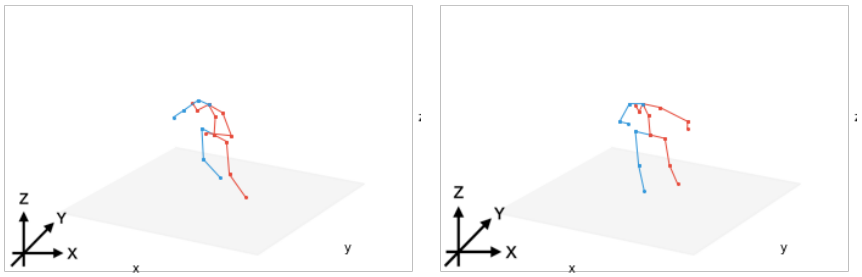
6.1 Improvement of Recognition Accuracy by Data Augmentation

When using the C1 camera for training data and data augmentation (+45°, +180°) and the test data was from the C4 camera, recognition accuracy improved by 54.8% compared to using features from the 2D pose. Therefore, we succeeded in creating a classification model that can correctly classify actions from the estimated 3D pose from videos at different camera positions by rotating the 3D pose. Particularly, the recognition accuracy was improved in the cases when the training data comes from camera C1 and test data comes from camera C4 (or vice-versa) and any of the cases for data augmentation was used . The reason for this improvement is that C1 and C4 are shooting the front of the person (Fig.7), and there is less error when estimating the 3D pose than when shooting from the back. It appears that the rotating data augmentation technique worked well without being affected by errors or false key-point detections (Fig.10).

6.2 Deterioration of Recognition Accuracy by Data Augmentation

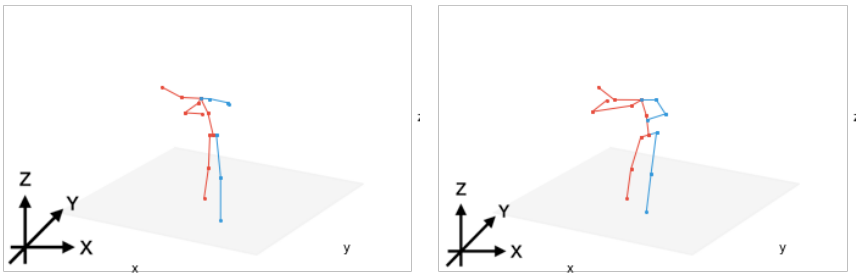
When using the C3 camera for training data and data augmentation (+45°) and the test data was from the C1 camera, recognition accuracy deteriorated by 22.8% compared to the model using features from the 2D pose. In addition, when the training data comes from cameras that shoot from behind the human actor (C3, C2), the improvement in recognition accuracy when data is augmented in the training data is lower than when training data comes from cameras that capture the front of the human actor (C1, C4).

The reason is that the 3D pose estimated from a camera that captures a person from the back is more influenced by errors and false key-point detections than a camera that captures a person from the front (Fig.11). Furthermore, the recognition accuracy is improved when the training data and test data are augmented between the front and back surfaces, but otherwise the recognition accuracy is poorly improved. It is possible that these reasons are also due to the 3D pose error and detection estimated from the camera that shows the person from the back. Therefore, it is necessary to consider the optimal rotation angle and the error of the estimated 3D pose when augmenting the data using the 3D pose.



(a) C1 (b) C4

Fig. 10 3D pose estimated from videos captured by cameras C1 and C4



(a) C2 (b) C3

Fig. 11 3D pose estimated from videos captured by cameras C2 and C3

7 Conclusion

In this paper, we studied the problem of action recognition using the 3D pose estimated from single-camera video. We used data augmentation by rotating the 3D pose using affine transformation to improve the recognition accuracy in the case of changing the shooting camera position in the test data.

As a result, it was found that recognition accuracy at different camera positions was improved by up to 55.7%. However, it was also found that the accuracy decreased depending on the rotation angle of the data augmentation.

As future work we would like to study the following:

- Proposing the optimal rotation angle for data augmentation
- Selecting augmented data so as not to adversely affect the classification model
- Feature extraction to enhance robustness for different camera positions

References

1. Bulling, A., Blanke, U., Schiele, B.: A tutorial on human activity recognition using body-worn inertial sensors. ACM Comput. Surv. **46**(3) (2014). DOI 10.1145/2499621. URL https://doi.org/10.1145/2499621
2. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR (2017)
3. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(7), 1325–1339 (2014)
4. Kohei, A., Tsuyoshi, O., Sozo, I.: Human action recognition using 3d pose estimation. Proceedings of the 21st SOFT Kyushu Chapter Annual Conference pp. 40–43 (2019)
5. Liaw, A., Wiener, M.: Classification and regression by randomforest. Forest **23** (2001)
6. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3d human pose estimation. In: ICCV (2017)
7. Ofli, F., Chaudhry, R., Kurillo, G., Vidal, R., Bajcsy, R.: Berkeley mhad: A comprehensive multimodal human action database. In: WACV, pp. 53–60. IEEE Computer Society (2013). URL http://dblp.uni-trier.de/db/conf/wacv/wacv2013.html#OfliCKVB13
8. Ohashi, H., Al-Naser, M.O.A., Ahmed, S., Akiyama, T., Sato, T., Nguyen, P., Nakamura, K., Dengel, A.: Augmenting wearable sensor data with physical constraint for dnn-based human-action recognition. In: Time Series Workshop. Time Series Workshop @ ICML, located at ICML 2017, August 11-11, Sydney, Australia (2017)
9. Okita, T., Inoue, S.: Activity recognition: Translation across sensor modalities using deep learning. In: Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers, UbiComp '18, pp. 1462–1471. Association for Computing Machinery, New York, NY, USA (2018). DOI 10.1145/3267305.3267512. URL https://doi.org/10.1145/3267305.3267512
10. Pavllo, D., Feichtenhofer, C., Grangier, D., Auli, M.: 3d human pose estimation in video with temporal convolutions and semi-supervised training. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
11. Takasaki, C., Takefusa, A., Nakada, H., Oguchi, M.: A study of action recognition using pose data toward distributed processing over edge and cloud. In: 2019 IEEE International Conference on Cloud Computing Technology and Science (CloudCom), pp. 111–118 (2019). DOI 10.1109/CloudCom.2019.00027
12. Takeda, S., Lago, P., Okita, T., Inoue, S.: Reduction of marker-body matching work in activity recognition using motion capture. In: Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers, UbiComp/ISWC '19 Adjunct, pp. 835–842. Association for Computing Machinery, New York, NY, USA (2019). DOI 10.1145/3341162.3345591. URL https://doi.org/10.1145/3341162.3345591
13. Um, T.T., Pfister, F.M.J., Pichler, D., Endo, S., Lang, M., Hirche, S., Fietzek, U., Kuliundefined, D.: Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction, ICMI '17, pp. 216–220. Association for Computing Machinery, New York, NY, USA (2017). DOI 10.1145/3136755.3136817. URL https://doi.org/10.1145/3136755.3136817
14. Varol, G., Laptev, I., Schmid, C., Zisserman, A.: Synthetic humans for action recognition from unseen viewpoints (2019)