2020 International Conference on Machine Vision and Image Processing (MVIP)
Faculty of Engineering, College of Farabi, University of Tehran, Iran, 19 & 20 Feb. 2020

An Ensemble Model For Human Posture Recognition

Bardia Esmaeili
Faculty of Computer Engineering
Shahid Rajaee Teacher Training University
Tehran, Iran
bardia.esm@gmail.com

Alireza AkhavanPour
Shenasa AI
Tehran, Iran
Akhavan@Shenasa.ai

Alireza Bosaghzadeh
Faculty of Computer Engineering
Shahid Rajaee Teacher Training University
Tehran, Iran
a.bosaghzadeh@srttu.edu

***Abstract*—Human Body Pose Estimation (HBPE) and Human Body Posture Recognition(HBPR) have improved significantly in the past decade. Gaining access to huge amounts of data, Kinect camera, neural networks and specifically deep convolutional neural networks (deep convnets) have led to fascinating success in these fields. In this paper we propose an ensemble model for human body posture recognition. Deep convnets are the main building block and fundamental aspect of our proposed model. We leverage deep convnets in two variations to classify postures. First, we use them for an end-to-end training scenario. We perform transfer learning with Imagenet weights on deep convnets with our gathered dataset of RGB images to classify five different postures. Second, we use a pre-trained deep convnet[1] (pose estimator) for estimating human body joints in RGB images. The pre-trained pose estimator has been trained to calculate a total of 17 2D joints coordinates and we utilize these coordinates to train a decision tree-based classifier for classification among five classes. Both variations are examined with different settings. The best settings for both variations are combined together to create our proposed model. More specifically, the classification layers of both variations are stacked together and fed to a logistic regression unit for a better classification result. Transfer learning, training and experiments in this paper are based on only RGB images from our gathered dataset and human body joints coordinates extracted from these images, which conveys that our proposed model does not require depth images or any sensor. Eventually, experimental results on the images show that the proposed model has higher performance than fundamental variations. Specifically, our model is able to correctly recognize the human posture in the majority of the images that one of the two fundamental variations fails to classify. The code for the proposed model and our gathered dataset are available on github[1].**

Keywords—Deep Convolutional Neural Networks, Decision Tree, XGBoost, Human Pose, Human Posture, Proposed model

I. INTRODUCTION

There has been magnificent improvements in the field of HBPE and HBPR. HBPE refers to estimating the human body joints and the connections between pairs of adjacent joints in an image to form the body pose skeleton, while HBPR is concerned with classification of human postures (i.e., sitting or standing). These two fields could be used in a huge amount of applications. They could be utilized to understand human emotions[2], which is called body language. Body language recognition could lead to significant improvements in the interaction between machine and human. Furthermore, body hand movements could be another source for interaction between human and robot[3]. Head pose estimation[4] can be used for driver assistance systems or eye-gaze estimation systems. Nowadays, HBPE and HBPR are used for robotics[5], action recognition[6], video games, animations, augmented reality or examining the performance of an exercise or a dancing move[7].

The reminder of this article is organized as follows. In Section II we review some of the recent works published in HBPR area. Section III is dedicated to our proposed model and its fundamental aspects. Section IV is concerned with introducing our gathered dataset (including train set, test set 1 and test set 2) and the results of various experiments on the dataset. Finally, in Section V, a conclusion of our proposed model, its results and our future work are mentioned.

II. RELEATED WORK

In [8] RGB, depth images as well as skeleton data were provided by Kinect camera as input data. With the help of the available input, coordinates and angles for informational joints were extracted and fed to a SVM classifier. Authors in [9] used depth image from Kinect camera to extract the silhouette of the human body. A specified vector of features was calculated for each image based on the extracted silhouette, its center of gravity and its contour. The feature vector was fed to a neural network for classification. Similar to [8] and [9], in [10] the authors used, Kinect camera to obtain 2D RGB image and depth image. With the help of depth images and 2D joint coordinates from 2D images, 3D coordinates were calculated. 3D coordinates and angles were fed to a SVM classifier. In [11] two main approaches were presented for posture recognition. The first approach was acquiring joints coordinates from 2D and depth images with the help of Kinect camera to produce the input for a SVM classifier. The second approach involved transfer learning on the AlexNet deep convnet. The AlexNet was utilized as an end-to-end approach for 2D images and depth images, separately. Using Kinect Camera helps with providing more information such as depth image, which can be used for calculating 3D coordinates. However, there are disadvantages including the necessity of using Kinect camera and its limited range.

Moreover, the ideas for body posture recognition can be applied to hand posture recognition. In [12] A multi-channel

[1] https://github.com/BaRdia-eSm/An-Ensemble-Model-For-Human-Posture-Recognition

978-1-7281-6832-6/20/$31.00 ©2020 IEEE

convolutional network was used to classify hand postures. One channel used gray image and two other channels used sobel operators to diversify the input. The input was ran through cubic and 2D convolutions. Channels were then connected to fully connected layers and eventually to a logistic regression layer. Another work [13] took advantage of a two stage hand posture recognition system with the help of a Kinect camera for sign language recognition. In the first stage, hand detection and tracking were implemented using both color and depth image from the Kinect camera. In the second stage, hand images were passed through a deep convnet to automatically extract features.

In [14] a standing posture recognition system was proposed by the means of an intelligent standing surface. The pressure data from the standing surface was transformed into pressure image, then the pressure image was used for training a convolutional neural network.

Authors in [15] proposed a synthetic approach involving depth data, skeleton data, knowledge of anthropometry and a neural network. It was a conditional process starting with the calculation of the ratio of the height of the human head to the height of the body posture. Depending on the ratio, the 3D spatial relations of crucial points and the neural network, which takes extracted and transformed feature vectors as input, might be used throughout the recognition process.

Our proposed model does not require the usage of Kinect camera or any sensor attached to the human body, it operates only on RGB images and there is no need for images to be taken in a specific angle, environment or from a specific distance.

III. METHODS

As mentioned, our proposed model is dependent on body joint coordinates estimated from a pre-trained deep convnet as the pose estimator (body joint approach) and other deep convnets for end-to-end training (end-to-end approach). Both variations were inspected separately with various settings. The best setting of both variations were combined together to form the proposed model. Since the images in the dataset have different sizes, images are resized to a unique size for the end-to-end approach. In body joint approaches, for utilizing joints distances or coordinates, they were normalized to be size-invariant. In subsection A, the end-to-end approach, different body joint approaches are explained to further understand how the proposed model works. Also, it is worth noting that test set 1 is completely composed of images with detected body joints, while in test set 2 the majority of images have no detected body joint.

A. Periliminary Methods

a) Body joint approach: As for body joint approach, three settings were inspected. Normalized body joint coordinate (normalized joint coordinate), the angle among adjacent joints (joint angle) and the normalized distance between all possible pairs of joints (normalized joint distance). In all three methods, the input data is acquired by preprocessing on joints coordinates extracted from the pose estimator.

Data from the body joint approach is a structured form of data and XGBoost[16] is one of the best choices for classification on structured data. XGBoost is a decision tree-based algorithm, which uses gradient boosting for training on an ensemble of machine learning models. It trains a hierarchy of decision trees. In each step of the training, the samples which have been predicted falsely, will get higher weights in the next step. Sequentially, the next prediction will be according to all of the previous decision trees. This is one of the main advantages of XGBoost. Basically, XGBoost uses a number of weak learners to create a strong learner. Moreover, XGBoost can handle missing value. This feature is fundamental for posture recognition with body joints, as it is not always possible to find all 17 human body joints. XGBoost also uses parallel processing and tree-pruning and it preforms regularization to avoid overfitting.

The first setting for body joint approach is the normalized joint coordinates approach. Each feature is a value of a 2D coordinate. In different images with different sizes, humans are standing in various distances and positions relative to the camera, which conveys that the extracted coordinates and calculated line (link) distances between adjacent coordinates in different images are not in the same scale. This could hurt the classification based on normalized joint coordinate approach. A normalization preprocess is proposed as a solution.

Fig.1 depicts the stages of normalization preprocess for two extracted skeletons in different images with different sizes. In the first stage the skeleton extracted from the pose estimator is shown. In the second stage, all images are transferred to the top left corner of the image in order to have a transition-invariant normalization. In a typical human body, the ratio of different links to one another is roughly fixed, therefore in the third stage, for every image, we divided all body joints coordinates in that particular image by the length of a specified link in the image. As a result, the normalization becomes size-invariant. Here, we used the distance between nose and neck coordinates as our link, since these joints were recognized in most of the images. The link between nose and neck in Fig. 1 is shown in red, while other links are shown in white.

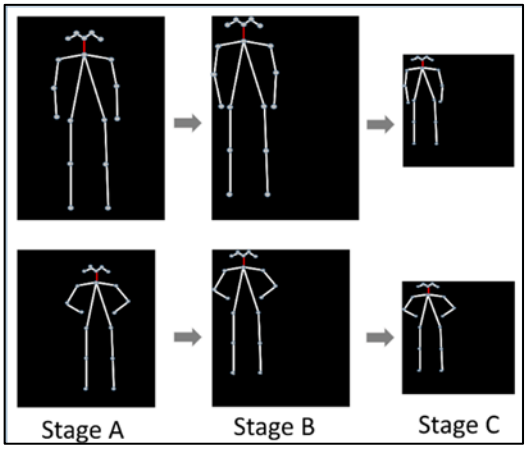

Fig. 1. Different stages of normalization for normalized joint coordinate approach

As our desired link for division must me unique across distinct images, we need to calculate the nose and neck coordinates and accordingly, the link length between them for images in which the pose estimator was not able to find the two coordinates. Towards that end, we used 30 images from the

neutral class, with all 17 coordinates being detected correctly, to calculate the average ratio of nose-neck link length to other link lengths. In images which nose or neck coordinates were not recognized, if any two adjacent joints were detected, then their link could be calculated and sequentially, nose-neck link length could be concluded with the help of calculated average ratios.

The second setting is the angles between adjacent links. For the purpose of posture recognition, using angles between adjacent links with a common joint coordinate as vertex, is more convenient than using normalized joint coordinates themselves, since angles are constant in different scales. In order to calculate angles, we used inner product of the two adjacent links.

The third setting is the normalized joint distance. Body joints can be preprocessed to acquire the normalized distance between pairs of joints. In this approach we used link lengths between all possible pairs of joints. In another word, our input data comprised 136 possible combinations of pairs of joints from 17 joints. Needless to say, this data have to be normalized. We use the same normalization term for division as for normalized joint coordinates approach, which means that all calculated distances were divided by the link length between the nose and neck.

Another valuable feature of XGBoost is verifying more informative and influential features of the input data, as it is a decision tree based classifier and more influential features are higher up the decision trees. We tried combining the best features of joint angle approach and normalized joint distance approach (angle-distance feature selection). We also examined the combination of the best features of all three mentioned body joint approaches (coordinate-angle-distance feature selection), which yielded the best result among body joint approaches and was chosen as a building block of our proposed model. Specifically, 22 features were picked from normalized joint coordinate approach, (11 2D coordinates). From joint angle approach, 8 features were selected and 17 features were chosen from normalized distance approach.

b) End-to-end approach: For end-to-end classification, two deep convnets have been used to perform transfer learning. MobileNet[17] and VGG16[18]. In order to perform transfer learning, the classification layer of these networks was popped, then a dense layer of neurons was added and at last, a classification layer with five neurons was added, as our dataset has five classes. To use the weights of the pre-trained networks, the weights of the layers before the new dense layer were untrainable and the weights did not change, on the other hand the weights for the dense layer and the classification layer were trained.

On test set 1, body joint approaches have higher f1-score than end-to-end approach. However, this does not imply that body joint approaches are always the better option. For majority of images in test set 2, the pose estimator was not able to find any joint at all and XGBoost is dependent on this information to recognize a posture. Fortunately, the end-to-end approach does not need any joint coordinate and only needs the image to perform prediction. In such scenarios, end-to-end approach is the answer.

B. *Proposed model*

In our proposed model we used VGG16 as the end-to-end approach and coordinate-angle-distance feature selection approach as the body joint approach. To combine the two approaches, the classification layer of both approaches, are stacked together to form a new set of features as input for a logistic regression unit, which is trained and validated on the new input. The logistic regression unit is trained only on the subset of dataset with detected joints coordinates. For test set 2 in which the pose estimator was unable to find any coordinates, for most of the images, the prediction is exclusively based on the end-to-end approach. In Fig. 2 an abstract architecture of the proposed model is depicted. The whole figure presents the proposed model on test set 1, while the black dashed rectangle presents the proposed model on test set 2.

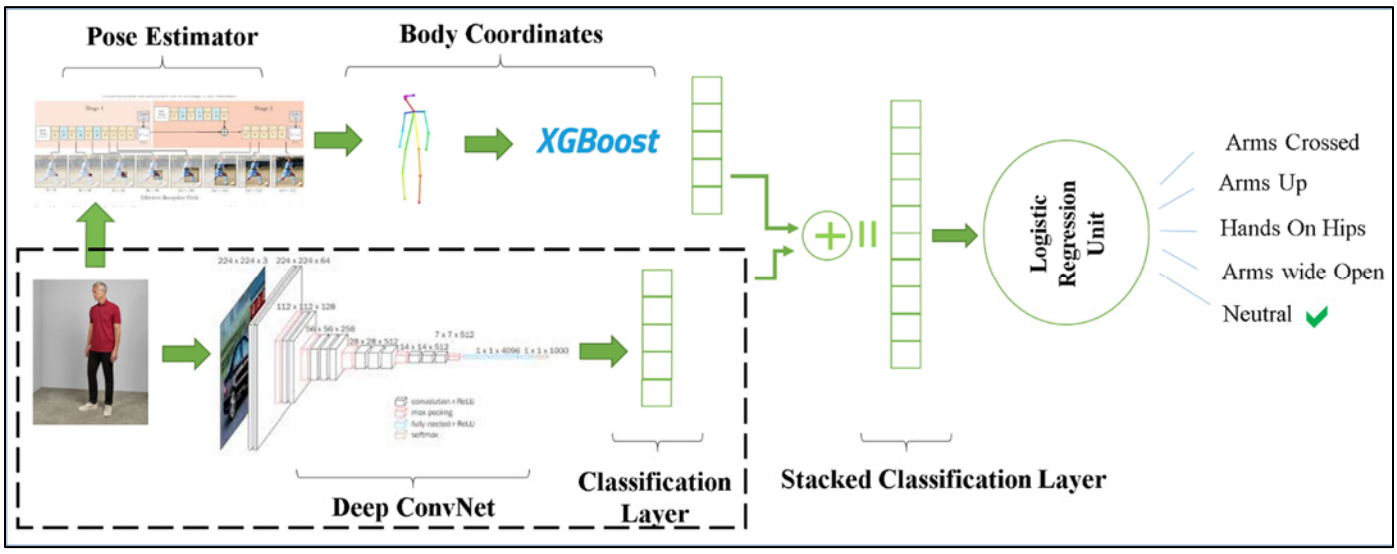

Fig. 2. Architecture of the proposed model

IV. Experiments

A. Dataset

Our dataset was acquired from google image search and images of volunteers. Dataset classes comprises arms crossed, arms up, arms wide open and neutral. A series of samples for each class are visualized in Fig. 3. There are a total of 2583 images in the dataset with the following distribution: 480 in arms crossed class, 535 in arms up class, 570 in arms wide open class, 495 in hands on hips class and 505 in neutral class. Images are taken in various sizes, backgrounds, illuminations, perspectives and from different distances relative to the camera.

Fig. 3. Dataset image samples

Since it was not possible for the pose estimator to extract joints in all images, our dataset consists of images in which joints were detected and images in which no joints were detected. Our train data and first test set (test set 1) are consisted of images with detected joints. The small remaining part of the images with detected joints and the images with no detected joints, are our second test set (test set 2). Test set 1 have balanced distribution across all classes, however for test set 2 the number of images in different classes are not the same and the test set is mostly consisted of image with no detected joints. For the end-to-end approach 20% of the train set is used as validation set. In TABLE I the number of images for train and test sets are shown. In test set 1 both body joint approach and end-to-end approach were used, yet in test set 2, as the majority of images do not have any detected joints, body joint approach could not be used.

TABLE I. Dataset Distribution

	Arms Crossed	Arms Up	Arms Wide Open	Hands on Hips	Neutral	Total
Train Data	329	329	329	329	329	1645
Test Set 1	82	82	82	82	82	410
Test Set 2	69	124	159	82	94	528

B. Results

In this section, the results of body joint approaches, end-to-end approaches as well as the proposed model are described with precision, recall and f1-score. These metrics provide a more detailed understanding of different approaches in each class. Additionally, test set 2 is imbalanced and these metrics are appropriate tools for examining the performance on the test set. Additionally, the confusion matrices for the proposed model on both test sets are presented.

The formula for calculating precision, recall and f1-score are presented in (1)~(3). For calculating these metrics, we use true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). Since we have five classes, we compute these values in a one-vs-all setting. One-vs-all means that for calculating these four values for each class, we think of the data samples as whether belonging to that class or not.

$$Recall = \frac{TP}{TP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3}$$

In order to find the best body joint approach, all mentioned body joint approaches were examined on test set 1 and the results are depicted in Table II with average precision, average recall and average f1-score. Based on Table II, it is clear that coordinate-angle-distance approach has the highest score in all three metrics among body joint approaches.

TABLE II. Average Precision, Recall and F1-Score For Examined Body Joint Approaches On Test Set 1

	Average Precision	Average Recall	Average F1-Score
Normalized Joint Distance	0.91	0.91	0.91
Joint Angles	0.86	0.86	0.86
Normalized Joint Coordinates	0.89	0.89	0.89
Angle-Distance Feature Selection	0.92	0.92	0.92
Coordinate-Angle-Distance Feature Selection	**0.93**	**0.93**	**0.93**

For end-to-end approach, transfer learning on two deep convnets, including VGG16 and MobileNet, was examined. Table III-V show precision, recall and f1-score for coordinate-angle-distance approach, VGG16, MobileNet and our proposed model on test set 1.

In Table III, the proposed model has the highest precision score in four classes in test set 1 except in hands on hips class. coordinate-angle-distance feature selection approach has the best score in hands on hips class and end-to-end approaches have the lowest scores in this class and neutral class.

TABLE III. Precision Table For Test Set 1

	Arms Crossed	Arms Up	Arms Wide Open	Hands on Hips	Neutral
Coordinate-Angle-Distance Feature Selection	0.93	0.93	0.91	**0.99**	0.93
VGG16	0.93	0.94	0.93	0.85	0.88
MobileNet	0.9	0.94	0.9	0.79	0.76
Proposed Model	**0.99**	**0.96**	**0.99**	0.92	**0.94**

In Table IV, the proposed model has the highest recall score in four classes in test set 1 except in arms crossed class. coordinate-angle-distance feature selection approach has the

best score in this class and end-to-end approaches have the lowest scores in neutral class.

TABLE IV. RECALL TABLE FOR TEST SET 1

	Arms Crossed	Arms Up	Arms Wide Open	Hands on Hips	Neutral
Coordinate-Angle-Distance Feature Selection	**0.98**	0.9	0.91	0.93	0.96
VGG16	0.91	0.91	0.91	0.9	0.87
MobileNet	0.8	0.88	0.91	0.88	0.79
Proposed Model	0.96	**0.96**	**0.95**	**0.94**	**0.98**

The proposed model has higher precision, and recall in three classes, hence it has higher f1-score in these classes in Table V. In arms crossed class, the sum and multiplication of precision and recall for the proposed model are higher than coordinate-angle-distance feature selection approach, therefore the f1-score for this class is higher for the proposed model as well. However, in hands on hips class, the proposed model has both lower precision, recall and as a result, lower f1-score than coordinate-angle-distance feature selection approach. It is worth to mention that coordinate-angle-distance feature selection approach has higher f1-score than VGG16 in three classes. In two other classes VGG16 has higher f1-score than coordinate-angle-distance feature selection approach, however its score is not higher than the proposed model. VGG16 is better than MobileNet in all classes. In addition, end-to-end approaches have the lowest precision, recall and f1-score in neutral class by a considerable margin, which conveys their weak performance in this class.

TABLE V. F1-SCORE TABLE FOR TEST SET 1

	Arms Crossed	Arms Up	Arms Wide Open	Hands on Hips	Neutral
Coordinate-Angle-Distance Feature Selection	0.95	0.91	0.91	**0.96**	0.95
VGG16	0.92	0.93	0.92	0.88	0.87
MobileNet	0.85	0.91	0.91	0.83	0.77
Proposed Model	**0.98**	**0.96**	**0.97**	0.93	**0.96**

Table VI-VIII present the same information as Table III-V, however the results are reported on test set 2 instead of test set 1. Since test set 2 majorly contains images without any detected joint, coordinate-angle-distance feature selection approach cannot be examined on this test set and no numbers were reported for this approach in Table VI-VIII. On this test set, our proposed model is basically VGG16 and relies on the results from VGG16. As a result, the numbers for the proposed model on test set 2 are the same numbers for VGG16.

In Table VI, MobileNet has the same precision score as the proposed model in arms crossed class, however in every other class the proposed model has a higher score with a considerable margin. The scores of different classes for the proposed model are in near proximity of each other. On the other hand, MobileNet, has distinguishable score reduction in hands on hips and neutral class in comparison with its score in other classes.

TABLE VI. PRECISION TABLE FOR TEST SET 2

	Arms Crossed	Arms Up	Arms Wide Open	Hands on Hips	Neutral
MobileNet	**0.84**	0.85	0.87	0.68	0.77
Proposed Model(VGG16)	**0.84**	**0.9**	**0.92**	**0.85**	**0.82**

In Table VII, MobileNet has the same recall score as the proposed model in hands on hips class, however in every other class the proposed model has a considerable higher score. Similar to Table VI, the scores of different classes for the proposed model are in near proximity of each other, while MobileNet, has distinguishable score reduction in arms crossed and arms up class relative to its score in other classes.

TABLE VII. RECALL TABLE FOR TEST SET 2

	Arms Crossed	Arms Up	Arms Wide Open	Hands on Hips	Neutral
MobileNet	0.71	0.75	0.89	**0.82**	0.82
Proposed Model(VGG16)	**0.91**	**0.84**	**0.9**	**0.82**	**0.89**

In Table VIII, the proposed model has higher f1-score in every class and MobileNet performs weakly in hands on hips and arms crossed class.

TABLE VIII. F1-SCORE TABLE FOR TEST SET 2

	Arms Crossed	Arms Up	Arms Wide Open	Hands on Hips	Neutral
MobileNet	0.77	0.8	0.88	0.74	0.79
Proposed Model(VGG16)	**0.87**	**0.87**	**0.91**	**0.83**	**0.86**

Base on Table III-VIII, coordinate-angle-distance feature selection approach mostly performs better than end-to-end approaches on test set 1 and the ensemble of these two approaches can yield better overall performance on the test set. On test set 1 and 2, VGG16 performs considerably better than MobileNet. As there is no body joint approach for test set 2, VGG16 is chosen as the proposed model on the test set. In the following TABLE IX-X the confusion matrix for test set 1 and test set 2 are shown for the proposed model.

In Table IX, the most errors in test set 1 are related to images which were incorrectly predicted to be a part of hands on hips class (6 images) and neutral class (5 images). The major cause of the errors is the angle of the person in the image relative to the camera, especially if the person is standing sideways. In the most of these images, body joint approach could extract the human skeleton and classify correctly, however the end-to-end approach performs poorly in this kind of images and classifies incorrectly, sequentially stacking the classification layer of end-to-end approach with the body joint approach, causes the proposed model to classify incorrectly as well. Here, it could be concluded that the combination of the end-to-end approach with the body joint approach, when one of the approaches have classified correctly and the other has not, does not always mean that the proposed model will classify correctly. Another major cause of error was wrong estimation of human pose skeleton (i.e., estimating an angle joint in the same location as an ear),

which leads to wrong classification. Minor errors were related to occlusion, background and illumination, which mostly affected the end-to-end approach rather than the body joint approach and as a result affected the proposed model.

TABLE IX. CONFUSION MATRIX FOR THE PROPOSED MODEL ON TEST SET 1

		Predicted Label				
		Arms Crossed	*Arms Up*	*Arms Wide Open*	*Hands on Hips*	*Neutral*
True Label	*Arms Crossed*	**79**	0	0	2	1
	Arms Up	0	**79**	1	2	0
	Arms Wide Open	1	0	**78**	2	1
	Hands on Hips	0	2	0	**77**	3
	Neutral	0	1	0	1	**80**

In Table X for test set 2, 12 images were incorrectly classified in arms crossed class, 12 images in arms up class, 13 images in arms wide open class, 12 images in hands on hips class and 18 images in neutral class. One of the highest misclassification numbers was due to the misclassification between arm up and arms wide open classes. As mentioned, end-to-end approach is sensitive to illumination, distance, background and the angle of the human relative to the camera, which was the most common cause of error. The other mentioned causes roughly affected the same considerable number of images.

TABLE X. CONFUSION MATRIX FOR THE PROPOSED MODEL ON TEST SET 2

		Predicted Label				
		Arms Crossed	*Arms Up*	*Arms Wide Open*	*Hands on Hips*	*Neutral*
True Label	*Arms Crossed*	**63**	0	1	3	2
	Arms Up	4	**103**	6	3	7
	Arms Wide Open	2	7	**143**	3	4
	Hands on Hips	4	3	3	**67**	5
	Neutral	2	2	3	3	**84**

Utilizing the proposed model reduced the number of false detection noticeably on test set 1. There were a total of 40 misclassifications for VGG16 and 27 for the coordinate-angle-distance approach. However, this number was brought down to only 17 for the proposed model. Specifically, 24 errors of VGG16 and 20 errors of the coordinate-angle-distance approach were classified correctly with the proposed model. Fig. 4 shows some of the images that were classified incorrectly with VGG16 and correctly with the coordinate-angle-distance approach.

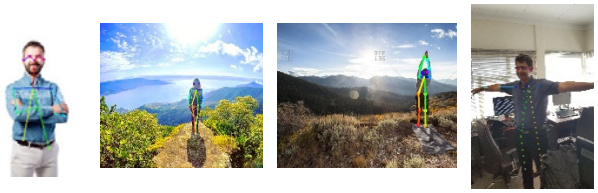

Fig. 4. Samples with their skeleton classified incorrectly with VGG16 and correctly with coordinate-angle-distance approach.

Moreover, Fig. 5 shows images that were classified correctly with VGG16 and incorrectly with the coordinate-angle-distance approach.

Fig. 5. Samples with their skeleton classified incorrectly with coordinate-angle-distance approach and correctly with VGG16.

On the other hand, in Fig. 6 we can observe 4 samples out of a total of 6, which both approaches were unable to classify correctly and as a result, the proposed model could not perform classification correctly as well.

Fig. 6. Samples with their skeleton classified incorrectly both with coordinate-angle-distance approach and VGG16.

V. CONCLUSION

In this paper we proposed an ensemble model for Human Body Posture Recognition based on RGB images. The proposed model was designed by stacking the classification layer of two main approaches together, which were both implemented by deep convnets. One was based on transfer learning on a deep convnet (end-to-end approach) and the other was based on extracting the body joint coordinates from a deep convnet (the pose estimator) and then feeding them to a decision tree-based classifier (body joint approach). As for the end-to-end approach, we performed transfer learning with Imagenet[19] weights on VGG16 and MobileNet. In body joint approach, we used a gradient boosting library called XGBoost.

We also introduced our gathered dataset which includes a train set, test set1 and test set 2. Test set 1 was a series of images with detected body joints, however in the majority of images on test set 2, there was not any detected body joint and body joint approach could not be used on this test data. Therefore, our results on test set 2 for the proposed model were the same as the results for VGG16 on the test set. Experimental results on test set 1 conveyed that the combination of both approaches could correct a considerable subset of the wrong classifications when one approach classified correctly and the other did not. Regardless, the case was not true for all misclassifications of individual approaches. Considering using only raw image as input, the proposed model did not need sensors or a Kinect camera for gathering depth image or skeleton data. Moreover, our model did not require input images to be taken in a certain angle, from a certain distance relative to the camera or in a

specific environment. The most significant aspect of our proposed model was the ability to work with body joint approach and end-to-end approach to use the advantages of both approaches and to reduce their individual disadvantages.

In our future work, we intend to use the pose estimator in an end-to-end approach. As it has valuable information of human body pose, it is likely to be advantageous for transfer learning. Moreover, body joint coordinates could be used with videos and recurrent neural networks for action recognition as well.

ACKNOWLEDGMENT

The authors gratefully acknowledge the voluntary contributions of Shenasa AI group in gathering our dataset.

REFERENCES

[1] Z. Cao, T. Simon, S. E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," CVPR, 2017.

[2] T. Randhavane, U. Bhattacharya, K. Kapsaskis, K. Gray, A. Bera, and D. Manocha, "Identifying Emotions from Walking using Affective and Deep Features," arXiv:1906.11884, 2019.

[3] P. Halarnkar, S. Shah, H. Shah, and J. Shah, "Gesture Recognition Technology: A Review." International Journal of Engineering Science and Technology, 2014, vol 4, pp. 4648-4654.

[4] M. Ariz, A. Villanueva, and R. Cabeza, "Robust and accurate 2D-tracking-based 3D positioning method: Application to head pose estimation," Comput. Vis. Image Underst., vol. 180, pp. 13–22, Mar. 2019.

[5] C. Zimmermann, T. Welschehold, C. Dornhege, W. Burgard, and T. Brox, "3D Human Pose Estimation in RGBD Images for Robotic Task Learning," in Proceedings - IEEE International Conference on Robotics and Automation, 2018, pp. 1986–1992.

[6] H. Kim, S. Lee, D. Lee, S. Choi, J. Ju, and H. Myung, "Real-time human pose estimation and gesture recognition from depth images using superpixels and SVM classifier," Sensors, vol. 15, no. 6, pp. 2410–12427, 2015.

[7] S. Deb, A. Sharan, S. Chaturvedi, A. Arun, and A. Gupta, "Interactive Dance Lessons through Human Body Pose Estimation and Skeletal Topographies Matching." International Journal of Computational Intelligence & IoT 2.4, 2018.

[8] T. L. Le, M. Q. Nguyen, and T. T. M. Nguyen, "Human posture recognition using human skeleton provided by Kinect," International Conference on Computing, Management and Telecommunications, ComManTel 2013, 2013, pp. 340–345.

[9] W.-J. Wang, J.-W. Chang, S.-F. Haung, and R.-J. Wang, "Human Posture Recognition Based on Images Captured by the Kinect Sensor," Int. J. Adv. Robot. Syst., vol. 13, no. 2, p. 54, Mar. 2016.

[10] B. Cao, S. Bi, J. Zheng, and D. Yang, "Human Posture Recognition Using Skeleton and Depth Information," WRC Symposium on Advanced Robotics and Automation, WRC SARA 2018 - Proceeding, 2018, pp. 28–33.

[11] M. El Amine Elforaici, I. Chaaraoui, W. Bouachir, Y. Ouakrim, and N. Mezghani, "Posture recognition using an rgb-d camera: exploring 3d body modeling and deep learning approaches," IEEE Life Sciences Conference, 2018, pp. 69–72.

[12] P. Barros, S. Magg, C. Weber, and S. Wermter, "A multichannel convolutional neural network for hand posture recognition," in Artificial Neural Networks and Machine Learning – ICANN 2014, S. Wermter, C. Weber, W. Duch, T. Honkela, P. Koprinkova-Hristova, S. Magg, G. Palm, and A. E. P. Villa, Eds., ser. Lecture Notes in Computer Science 8681. Springer International Publishing, Sep. 15, 2014, pp. 403–410.

[13] A. Tang, K. Lu, Y. Wang, J. Huang, and H. Li, "A real-time hand posture recognition system using deep neural networks," ACM Trans. Intell. Syst. Technol., vol. 6, no. 2, Mar. 2015.

[14] G. Li, Z. Liu, L. Cai, and J. Yan, "Human standing posture recognition based on CNN and pressure floor," J. Comput. Methods Sci. Eng., pp. 1–10, Oct. 2019.

[15] B. Li, C. Han, and B. Bai, "Hybrid approach for human posture recognition using anthropometry and BP neural network based on Kinect V2," EURASIP J. Image Video Process., vol. 2019, no. 1, p. 8, 2019.

[16] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., vol. 13-17-Augu, pp. 785–794, 2016.

[17] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Marco Andreetto, H. Adam "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications", arXiv:1704.04861, 2017.

[18] A. Krizhevsky, I. Sutskever and G. E. Hinton, "ImageNet Classification with Deep Convolutional Networks". Neural Information Processing Systems, 2012.

[19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database". Computer Vision and Pattern Recognition, 2009.