A Vision-based Deep Learning Platform for Human Motor Activity Recognition

Mobina Mobaraki[1], Anushree Bannadabhavi[1], Matthew J. Yedlin[1], and Bhushan Gopaluni[2]

[1]Electrical & Computer Engineering, University of British Columbia, Vancouver, BC, Canada
[2]Chemical & Biological Engineering, University of British Columbia, Vancouver, BC, Canada

***Abstract*—To track the body movement of patients with movement disorders, sensors such as Kinect cameras are not easily accessible. Recently-developed deep learning models, as a subset of Artificial Intelligence (AI), can analyze patients' behavior from RGB images of smartphones. The Stacked Hourglass model is a novel pose estimation deep learning model which can accurately determine the location of body joints and a long short-term memory network (LSTM) can determine the corresponding action by analyzing the kinematic behavior of the body joints. This study develops a deep learning model that uses RGB images from the UT-Kinect dataset as input and determines the action performed with 84.14 % accuracy. Specifically, our contributions are: (i) developed the preprocessing pipeline to use stack hourglass model on the UT-kinect dataset (ii) fine-tuning of the model to handle 20 joints (iii) Added a human action recognition component to accurately classify the actions performed. Our method can be an efficient replacement for the hardly-accessible Kinect cameras and can be used to analyze various diseases with movement disorders.**

***Index Terms*—Deep Learning, Human Pose Estimation (HPE), Human Activity Recognition (HAR), Stacked Hourglass model, Long Short Term Memory (LSTM) model, medical application**

I. Introduction

Abnormal behavior in patients' movement would be one of the main factors to evaluate the level of severity of a disease. Early detection and analysis of patient's movement can prevent the progression of the disease [1].

To track the patient's body movement, reliable biomarkers such as a Kinect camera are not easily accessible [2], [3]. Also, doctors' diagnoses are sometimes inconsistent [4]. Deep learning algorithms can help doctors detect abnormal movements more consistently using the commonly-used RGB images from smartphones.

This paper combines novel deep learning models including stacked hourglass and LSTM to understand human body motion using the commonly-used videos from smartphones. This model can be further used to develop a mobile phone application to provide a daily report of the level of severity of the disease by analyzing a video of the patients.

The novelties of this work are as follows: 1) The pre-trained model has already been trained on the MPII dataset[5] to predict 16 joints. This work modifies the model to predict 20 joints of the human body that are more representative of human movement 2) The training dataset includes multiple videos of humans doing specific tasks. This creates more representative data with respect to the previous MPII dataset[5] as the model can learn each action in different possible situations. This results in more generalized and accurate predictions.

II. Background

A. HPE model

A novel subset of deep learning models is Human Pose Estimation (HPE). It forms a skeleton-based representation of a human body to model the likelihood of certain parts (discriminator) and the probability distribution over the part (prior). Therefore, it can determine the location of the special joints (key points) and their connections (pair) from an image of a person. The human Activity Recognition (HAR) model is another deep learning model that uses the joints' predicted location to recognize and classify the action of the human. It extracts the kinematic behaviour of the joints and classifies the unusual movements. The doctor can analyze abnormal movements to determine the disease's severity and limit its progression.

Previous studies on HPE [6], [7], [8], [9] and HAR [10], [11], [12] lack the accuracy and generalization capabilities as they use manual features to predict the key points and recognize the action. A Microsoft Kinect camera is a good example of a classical model (random forest prediction algorithm) to detect body joints from RGB-D images [13]. Earlier HAR models also used traditional machine learning algorithms such as decision trees, support vector machines (SVM), and naïve Bayes. However, these models do not accurately show the landmarks and predict the name of the human's action in the case of hidden joints, differences in human appearance, body proportions, clothing, environment, and different angle of view. Therefore, proposing an alternative to this hardly-accessible poor-quality camera would significantly improve their performance.

The emergence of deep learning in 2014 [14] was a motivation for [15] to successfully replace the Kinect cameras with deep learning models. The complex features extracted by convolutional layers in a deep learning model result in better performance in comparison with the classical models.

To select the proper HPE model, there are some surveys [16], [17], [18], [19], [20], [21] which classify the current deep learning-based pose estimation models from different points of view. They can mainly be divided into top-down and bottom-up approaches. An example of a top-down algorithm is the AlphaPose model [22] which first localizes the human

and then calculates the pose. However, this approach raises errors as the predicted bounding box in the first step may not include the required poses for the second step. As for the other approach, a good example of bottom-up algorithms is the OpenPose model proposed by Zhe Cae et al. [23] which first detects the key points by predicting the confidence map and then forms the appropriate pairs by predicting the part affinity fields. However, grouping the key points into individual poses would be a challenge for real-time usage. To this end, [23], [24], [25] tried to tackle the issue using greedy parsing algorithms. As a novel solution, a recently developed "stacked hourglass network" [26] tackled the problems of the previous HPE models. Each hourglass is an extensive residual module that changes the resolution of an image by passing it through a pooling and subsequent upsampling layer. This allows the model to extract deeper features from both local and global scales. Also, the subsequent hourglass reassesses the previous output, making the final result more accurate. Stacking multiple hourglasses produces a symmetric architecture of end-to-end bottom-up and top-down interference. They continue to reach the output resolution and they are terminated by two successive 1*1 convolutions to predict the heat maps of each joint's probability at each pixel as output.

B. HAR model

To select the appropriate HAR model, a standard deep learning-based model such as Convolutional Neural Network (CNN) would lack the required feedback connections from a data sequence to accurately classify different activities. HAR needs these feedback connections from the current and previous frames to accurately classify the action in the presence of lags of unknown duration between starting and ending points. This may happen when two people do the same action at different speeds. In this regard, the Long Short-Term Memory (LSTM) model would be an ideal solution for processing sequences of data due to its feedback connections.

III. Methodology

The general road map of this study is shown in Figure 1. The details of the steps are discussed as follows.

A. Dataset

This study uses the open-source "UT-Kinect" dataset [27]. It includes 20 videos of 10 people doing ten different actions twice. In addition, for each RGB-D image in the dataset, the (x,y,z) location of the 20 body joints (including hip center, spine, shoulder center, head, left/ right shoulder, left/ right elbow, left/ right wrist, left/ right hand, left/ right hip, left/ right knee, left/ right ankle and left/ right foot), as well as the name of the action are provided as labels of each image. The size of the train, validation, and test datasets are 10787, 2697, and 1576, respectively.

B. The AI Model

In this subsection, the structure of the model and the modifications are discussed.

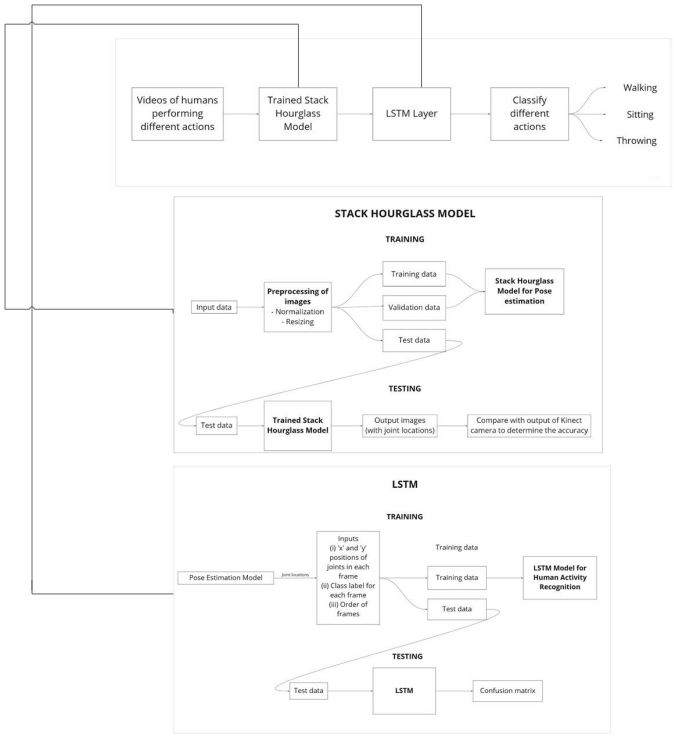

Fig. 1. General steps to develop an AI model to analyze human's movements

1) Model's Structure: The HPE model consists of stacks of hourglasses. Each hourglass gets an image of the person as input. The images first undergo the encoder section of the model. It consists of convolutional, batch normalization, ReLU activation, max pooling, and residual layers which are respectively responsible for extracting the features, avoiding gradient vanishing by confining the changes, adding nonlinearity, and reducing the memory by extracting the most important features.

The encoder is followed by a bottleneck layer to create a deeper network. Finally, the model will end up with the decoder section to extract the spatial features. It includes upsampling layers which enlarge the images and convert the features back into the images.

The outputs of the pre-trained model are 16 heat maps of the position of body joints (including right ankle, right knee, right hip, left hip, left knee, left ankle, pelvis, thorax, upper neck, head top, right wrist, left wrist, right shoulder, left shoulder, right elbow, left elbow, respectively). The heat map is a 64*64 Gaussian distribution of the probability of each joint in each pixel of an image. For each joint two heat maps from the original and flipped version of the image are created and the result is averaged so it improves the validation performance by 1%. Figure 2 shows one output of the HPE model. The output of the pre-trained HPE model is a heatmap and can be converted into coordinates. The coordinates are combined and fed into the LSTM layer to predict the name of the action that the person is taking.

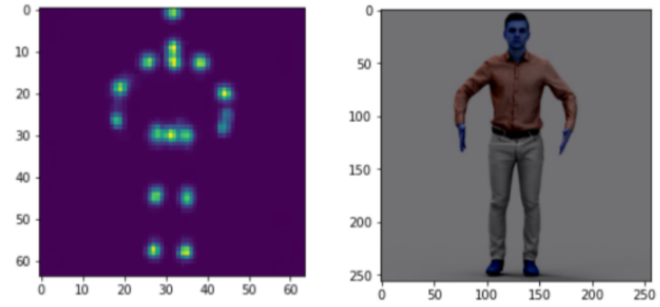

Fig. 2. Sample output of the pre-trained HPE model; The Gaussian probability of the joints is shown as a heatmap in the left figure. They are converted into coordinates and shown in light blue in the right picture

C. Training process

As a first step of the training process, the number of outputs was increased from 16 to 20. Then the parameters of the last layer of the stack hourglass model were updated to find the optimal parameters of the last layer which result in the most accurate prediction of the joint's location. The model was trained with a learning rate of 2.5 e-4, an epoch of 100, a batch size of 32, and MSE loss.

The HAR model was also trained with the learning rate of 5e-4, the epoch of 1e5, a batch size of 64, and cross-entropy loss.

IV. Results and Discussion

A. Qualitative Results

Figures 3 and 4 show the predicted heatmaps and the predicted landmarks after training the HPE model. It is worth mentioning that in Fig. 4, the ankles and the feet are hidden so only 15 joints are visible. Consequently, the corresponding heatmaps in Fig. 3 also do not include any color dots. The model can detect the key points with 67.5 % accuracy on a new dataset based on the "Probability of Correct Keypoint (PCK)" method.

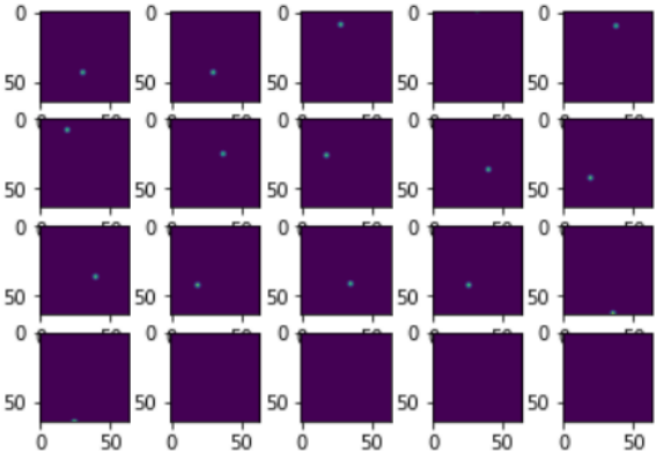

Fig. 3. Predicted heatmaps with 20 body joints after training the HPE model

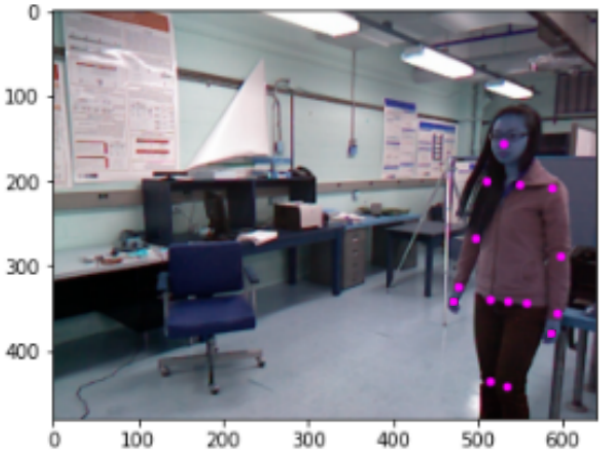

Fig. 4. Predicted landmarks with 20 body joints after training the HPE model

Figure 5 shows the confusion matrix after training the HAR model. It can determine the name of the action that the person is taking in a video with 84.18 % accuracy.

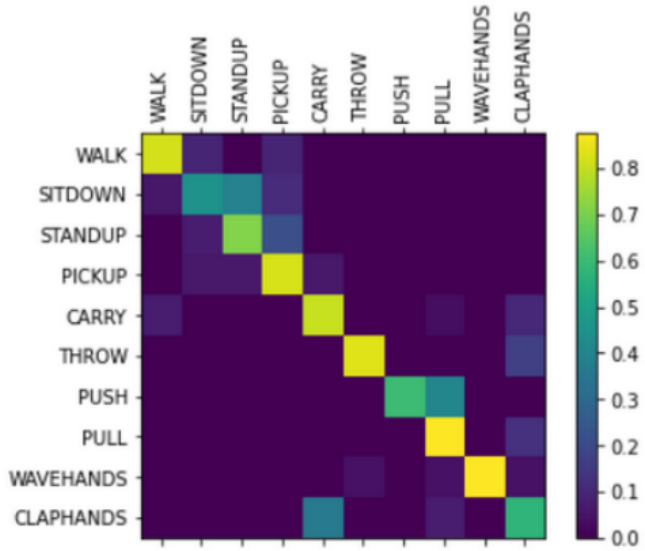

Fig. 5. Confusion matrix after training the HAR model

B. Quantitative result analysis

Results of our HPE model are presented in Table I. Other popular HPE models like AlphaPose[22] and OpenPose[23] report performance on MPII[5] and MSCOCO datasets. Since our experiments are performed on the UT-Kinect dataset, comparing accuracies directly would not be fair. Additionally the base model, Stacked Hourglass model that we use reports PCKh scores on the MPII dataset[5]. We mainly ablate on the number of hourglasses that are most suitable for pose estimation for the UT-Kinect dataset.

Table I shows a comprehensive analysis of the performance of our HPE model in case of changes in the number of

TABLE I
QUANTITATIVE RESULTS: PERFORMANCE OF THE HPE MODEL FOR DIFFERENT NUMBER OF HOURGLASSES IN THE MODEL

# hourglasses	# parameters	Accuracy (%)	Model size (MB)
1	3,586,960	33.76	482.71
2	6,730,912	67.50	742.99
8	25,594,624	73.71	2640.64

hourglasses of the model. As the table shows, increasing the number of hourglasses from 1 to 2, significantly improves the accuracy but further increase it to 8 doesn't have significant impact on the accuracy while increasing the number of parameters in the model 4 times and the model's volume more than 3 times. Therefore, hourglass 2 is the optimal number for our application.

V. CONCLUSION

This study develops a system consisting of the Stacked Hourglass model and LSTM for accurate pose estimation and action recognition with the detected poses. This is greatly beneficial for healthcare applications like diagnosis of Parkinson's disease that require human gate, posture, and intentional movement analysis. Specifically, it can be used to assist clinicians in the diagnosis of Dyskinesia (an involuntary movement of the head, arm, leg, or entire body) or Bradykinesia (slowness of movement) in Parkinson's patients by analysing videos captured by normal smartphone cameras making the process more convenient compared to conventional Kinect-camera based techniques. Our model achieves 84.14% accuracy for action recognition on the UT-Kinect dataset. A more representative, large dataset that includes cases of occlusion, hidden joints and noisy background is worth exploring in the future to determine the robustness and efficacy of our model.

REFERENCES

[1] Flavio Nobili, Eric Westman, Rosalie V Kogan, Joana B Pereira, Federico Massa, Matteo Grazzini, Sanne K Meles, and Klaus L Leenders. Clinical utility and research frontiers of neuroimaging in movement disorders. *The Quarterly Journal of Nuclear Medicine and Molecular Imaging: Official Publication of the Italian Association of Nuclear Medicine (AIMN)[and] the International Association of Radiopharmacology (IAR),[and] Section of the Society of..*, 61(4):372–385, 2017.

[2] Kenneth Marek, Danna Jennings, Shirley Lasch, Andrew Siderowf, Caroline Tanner, Tanya Simuni, Chris Coffey, Karl Kieburtz, Emily Flagg, Sohini Chowdhury, et al. The parkinson progression marker initiative (ppmi). *Progress in neurobiology*, 95(4):629–635, 2011.

[3] Yefei He, Tao Yang, Cheng Yang, and Hong Zhou. Integrated equipment for parkinsonrsquo;s disease early detection using graph convolution network. *Electronics*, 11(7), 2022.

[4] Jing Zhang. Mining imaging and clinical data with machine learning approaches for the diagnosis and early detection of parkinson's disease. *npj Parkinson's Disease*, 8(1):1–15, 2022.

[5] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele.

[6] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *2009 IEEE conference on computer vision and pattern recognition*, pages 1014–1021. IEEE, 2009.

[7] Pedro F Felzenszwalb and Daniel P Huttenlocher. Pictorial structures for object recognition. *International journal of computer vision*, 61(1):55–79, 2005.

[8] Yi Yang and Deva Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR 2011*, pages 1385–1392. IEEE, 2011.

[9] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *CVPR 2011*, pages 1465–1472. IEEE, 2011.

[10] Sameh Neili Boualia and Najoua Essoukri Ben Amara. Pose-based human activity recognition: a review. In *2019 15th International Wireless Communications Mobile Computing Conference (IWCMC)*, pages 1468–1475, 2019.

[11] Ankita Jain and Vivek Kanhangad. Human activity classification in smartphones using accelerometer and gyroscope sensors. *IEEE Sensors Journal*, 18:1169–1177, 2018.

[12] Qin Ni, Lei Zhang, and Luqun Li. A heterogeneous ensemble approach for activity recognition with integration of change point-based data segmentation. *Applied Sciences*, 8(9), 2018.

[13] Ditte Rudå, Gudmundur Einarsson, Anne Sofie Schott Andersen, Jannik Boll Matthiassen, Christoph U. Correll, Kristian Winge, Line K. H. Clemmensen, Rasmus R. Paulsen, Anne Katrine Pagsberg, and Anders Fink-Jensen. Exploring movement impairments in patients with parkinson's disease using the microsoft kinect sensor: A feasibility study. *Frontiers in Neurology*, 11, 2021.

[14] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1653–1660, 2014.

[15] Elise Klæbo Vonstad, Xiaomeng Su, Beatrix Vereijken, Kerstin Bach, and Jan Harald Nilsen. Comparison of a deep learning-based pose estimation system to marker-based and kinect systems in exergaming for balance training. *Sensors*, 20(23):6940, 2020.

[16] Yucheng Chen, Yingli Tian, and Mingyi He. Monocular human pose estimation: A survey of deep learning-based methods. *Computer Vision and Image Understanding*, 192:102897, 2020.

[17] Qi Dang, Jianqin Yin, Bin Wang, and Wenqing Zheng. Deep learning based 2d human pose estimation: A survey. *Tsinghua Science and Technology*, 24(6):663–676, 2019.

[18] Wenjuan Gong, Xuena Zhang, Jordi Gonzàlez, Andrews Sobral, Thierry Bouwmans, Changhe Tu, and El-hadi Zahzah. Human pose estimation from monocular images: A comprehensive survey. *Sensors*, 16(12):1966, 2016.

[19] Zhao Liu, Jianke Zhu, Jiajun Bu, and Chun Chen. A survey of human pose estimation: the body parts parsing based methods. *Journal of Visual Communication and Image Representation*, 32:10–19, 2015.

[20] Pichao Wang, Wanqing Li, Philip Ogunbona, Jun Wan, and Sergio Escalera. Rgb-d-based human motion recognition with deep learning: A survey. *Computer Vision and Image Understanding*, 171:118–139, 2018.

[21] Nikolaos Sarafianos, Bogdan Boteanu, Bogdan Ionescu, and Ioannis A Kakadiaris. 3d human pose estimation: A review of the literature and analysis of covariates. *Computer Vision and Image Understanding*, 152:1–20, 2016.

[22] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2334–2343, 2017.

[23] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

[24] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin Murphy. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In *Proceedings of the European conference on computer vision (ECCV)*, pages 269–286, 2018.

[25] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. Pifpaf: Composite fields for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[26] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation, 2016.

[27] Lu Xia, Chia-Chih Chen, and Jake K Aggarwal. View invariant human action recognition using histograms of 3d joints. In *2012 IEEE computer society conference on computer vision and pattern recognition workshops*, pages 20–27. IEEE, 2012.