2024 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE) | 979-8-3503-4501-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/CHASE60773.2024.00030

Accurate Body Pose Matching for Individuals with Stroke using Siamese Networks

Ruslan Gokhman*, Talya Sawdayi*, Rana Khan*, Ashwin Satyanarayana †, Ramana Vinjamuri ‡ and Sai Praveen Kadiyala *
* Yeshiva University, New York, NY 10016, † City Tech, CUNY, New York, NY
‡ University of Maryland Baltimore County, Maryland

***Abstract*—Stroke is one of the major causes of long-term disability in United States. With more than 800,00 people experiencing stroke every year, it is important that efficient means for recovery are presented to support stroke subjects. Exoskeleton and serious game based rehabilitation are some of the state-of-art approaches used in the recovery of stroke subjects. Accurate matching of body poses performed by individuals with stroke is essential in understanding the current state of recovery of the subject and plan further rehabilitation. Established machine learning based approaches fall short in accurately matching the poses of stroke subjects with ground truths. In this work, we present algorithms supported by Siamese architectures to effectively identify the poses performed by the subjects. Our proposed framework involves data pre-processing, extraction, building classification models and validating them using a body pose data set of individuals with stroke. On a considered public database, our proposed pose identification models namely, Siamese based LSTM and Siamese based CNN gave 7.8% and 14.2% better identification accuracy than the traditional LSTM approach.**

I. Introduction

Stroke in general causes motor impairment, systematic rehab helps to restore the lost motor function enabling subject to perform his daily activities. Surface Electromyography (sEMG) based exoskeletons are the choice amongst patients affected with complete stroke [1]. Serious games based rehabilitation is an established approach for providing effective rehabilitation for partial stroke survivors. Immersive serious games, carefully crafted with scoring mechanisms based on movement smoothness and accuracy are effective in systematic and steady motor action recovery of stroke subjects. Serious games are used not only by stroke subjects but also by healthy subjects to keep their motor function active.

Estimation of user pose in Virtual Reality (VR) based games is well explored in literature. Authors in [2], present a low latency approach for estimating the upper body pose of the user. In this, they used a convolutional neural network based architecture for better pose matching, where in the model is trained with 3D joint positions. Usage of camera input image to estimate the body pose in an augmented reality (AR) game application is discussed in [3]. A comparison is made with the pre-built point in cloud data with the camera image to estimate the pose. In order to minimize the errors in body pose estimation authors present a hybrid approach combining the ORB (Oriented Fast and Rotated Brief) descriptor with optical flow that accurately tracks the displacement of key points in consecutive images. A relatively simple method of pose matching is used in [4] for automated designing of game levels for a better user experience. Authors allowed for an error tolerance level for angles of various joint levels, there by matching a set of angles of an obtained pose with the existing ground truths. Main emphasis of this work was to design wide variety of game levels in an automated fashion using Reinforcement Learning. Accurate estimation of camera pose using global features based on rotation consistency and local features based on rotation in-variance is discussed in [5]. Authors pitched this approach for applications in augmented reality and autonomous driving. They tried to optimize the pose estimation model by combining the losses obtained from the global and local features.

Importance of identifying accurate body pose in individuals with stroke for better rehabilitation is discussed in literature. In [6], authors analyze the effect of various body pose detection algorithms for a selected consumer unit (Xbox One Kinect). They focused on the upper-body stroke rehabilitation and observed that an improved tracking of shoulder, elbow and wrist joints along with their temporal information will help improve the pose prediction accuracy. Evaluation of low cost human pose estimators namely *Openpose*, *Detectron 2* has been done in [7]. In this work authors compare the estimates of angles of shoulders and elbows of four different upper body exercises obtained from estimators with the ground truth obtained from RGBD Kinect 2 devices. A numerical comparison based on RMSE and MAE is made between the results obtained from estimators and the ground truth values. Effective estimation of body pose in uncontrolled settings based on videos obtained from single handheld camera is studied in work [8]. Authors trained a convolutional neural network (CNN) model using data from below waist videos of stroke subjects, for predicting clinically relevant gait parameters. Comparison of estimates with results obtained from standard gait estimators using numerical analysis concluded that even with less sophisticated equipment and outside clinical settings deep networks are effective in estimating stroke gait. To assist robotic rehabilitation, authors in [9], have proposed usage of teaching trajectory plan for the bionic motion of robot using body pose estimation. These estimates were collected using Kinect's depth camera and *Openpose*'s deep neural network. After validating the trajectories, it was concluded that they will be of immense help for the rehabilitation doctors to design

2832-2975/24/$31.00 ©2024 IEEE
DOI 10.1109/CHASE60773.2024.00030

effective trajectories for the rehabilitation robots.

Multiple studies have been performed to understand the effectiveness of serious game based rehabilitation on stroke subjects' recovery. Many of these works focused on analyzing the immersiveness of the participants, efficiency of the games in rehabilitation etc. Lack of the emphasis on using latest machine learning algorithms in estimating quality of the poses will lead to a superficial analysis and conclusions of the merits and shortcomings of the serious games. Considering the large number of patients to be cared for and the extensive nature of serious games involved in, it is very difficult for physicians to track the accuracy of pose detection. Therefore accurate pose identification is a much needed problem to be addressed for an effective stroke recovery. Usage of advanced technologies to understand the pose of participants is an established practice. Instant understanding of the pose of the subjects can be obtained by the use of sensors or wearable objects obtained. A major drawback with this approach is the low level of comfort of the subject in wearing the device and lack of his willingness to wear the device. The scenario in which, sensors placed else where in the data collection site may need the subject to be stationed in a same place for every data collection episode.

Though there have been multiple works in the domain of pose estimation for general purposes and for stroke rehabilitation, effect of body pose estimation quality in individuals with stroke was never given sufficient emphasis. To address this gap, in this work, we present Siamese network based models which can effectively identify the body pose thereby improving the overall rehabilitation process. The rest of the paper is organized as follows. Section II discusses details of various methods used in this work for body pose matching and the data set considered for validating the methods. Analysis of the results obtained from the methods and a comparative study with other works are presented in Section III. Section IV concludes the paper with a summary of observations and scope for future work.

II. Methods

In this Section, we describe in detail the data set considered in this work and the methods we considered for body pose matching using the pre-processed data.

A. Data Set

Stroke often leads to impairment of movements. To accommodate movements stroke subjects use their strong joints, thus altering the poses. In this work we consider the data set published in [10], which present a set of clinically relevant motions that are considered during the rehabilitation therapy. Comprising of 10 healthy and 9 stroke subjects this data set is obtained using a Microsoft Kinect sensor. All the motions are performed while the participant remains seated, involving only the upper body. A total of 20 motions were performed by all the healthy subjects while stroke subjects could perform motions between 4 to 12, depending on each individual capabilities. For each motion of a given subject, all the 25 joint positions as described in [11], were recorded in world coordinates and were saved in a corresponding csv file. The data is sampled at 2KHz. Data collection setup in shown in Fig. 1.

Fig. 1. Framework for collecting pose data - Adapted from [10]

B. Pose Matching Framework

The considered data set in earlier section is pre-processed for deploying various machine learning models. The pose matching framework used in this paper is shown in Fig. 2. After the pre-processing the data corresponding to a particular motion (series of poses) i of a given subject j is arranged as single input vector $p_{i,j}$. Such input vectors are sent to both Siamese CNN and Siamese LSTM models, where in each vector is compared with another vector to observe the similarity. The percentage of correctly identified input vectors of same motion give the similarity score or accuracy of the model. For LSTM model, we train the model with a training data set extracted from the original data set. Later we test on a testing data set and collect the classification accuracy of the motions. Details of the individual architectures are explained below.

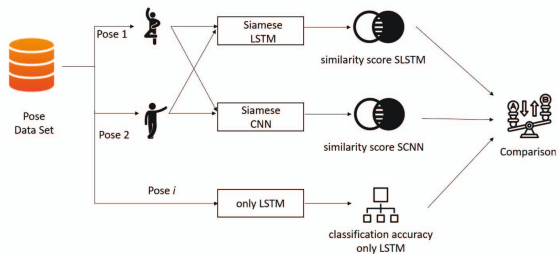

Fig. 2. Framework for Pose Matching/Classification

C. Siamese CNN method

We present the architecture we used for the Siamese CNN model in Fig. 3. The input shape for both sister networks as shown in figure is specified as 75. Next, each of the inputs passes through a 1D convolutional layer with 32 filters and a kernel size of 3. This setup is typically used to extract local features from sequence data, with the small kernel size capturing local dependencies. The output of the convolutional layers is flattened, enabling model to transform the multidimensional convolutional output into a one-dimensional vector. We considered, the first dense layer having 25 neurons, so that the model could capture a moderate level of complexity in the features extracted from the flattened output. The second dense

layer is reduced to a single neuron. This acts as a form of bottleneck feature, where the network is forced to condense the information into the most salient features for the task at hand. Next, the Lambda layer computes the L1 distance between the outputs of the two dense layers corresponding to each input. Here the model focuses on the absolute differences in feature representations to gauge similarity. Finally, we used a dense layer with a sigmoid activation function at the output for a binary classification task, In the context of a Siamese network, this determines whether or not the input pairs are similar.

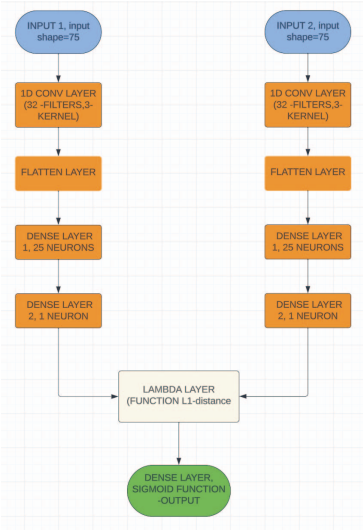

Fig. 3. Architecture of Siamese CNN neural network

D. Siamese LSTM method

In this method, we considered Siamese based LSTM for identifying similarity of two poses. As shown in Fig. 4, any given two poses pass through an LSTM layer configured with 34 units, with a dropout of 0.2, and a recurrent dropout of 0.2. The LSTM's role is to process sequential data, accounting for long-term dependencies within the input sequence which is 75 timesteps long. We used set dropout and recurrent dropout to 0.2 to lay a strong emphasis on preventing over fitting. This will be important in scenarios where the training data is not very large or the sequences contain complex patterns. These are followed by two dense layers, a LAMBDA layer and an output layer with sigmoid activation function as described in earlier Section.

E. only LSTM method

In the final method, we considered only LSTM architecture, which identifies the class of a given poses using a classification method. This is method is performed to compare with the earlier two Siamese based methods and understand their advantages and drawbacks. The architecture is shown in Fig. 5. Here, the input layer is set to accept input sequences of length 75. The LSTM layer has 34 units, and it includes dropout and recurrent dropout, both set at 0.2. This is done so that the

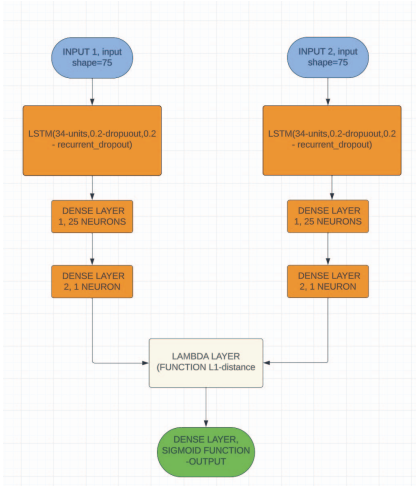

Fig. 4. Architecture of Siamese LSTM neural network

model can process sequences with the capability to remember long-term dependencies. The dropout is used to reduce over fitting by randomly setting a proportion of input units to 0 at each update during training time, which in this case is 20. A fully connected layer with 50 neurons follows the LSTM layer. This layer interprets features extracted from the input sequence by the LSTM layer. The following dense layer has 20 neurons and uses a softmax activation function. This is for classification task where the model is expected to categorize the input into one of 20 possible classes. The softmax function will output a probability distribution over these 20 classes, with the sum of probabilities equaling 1.

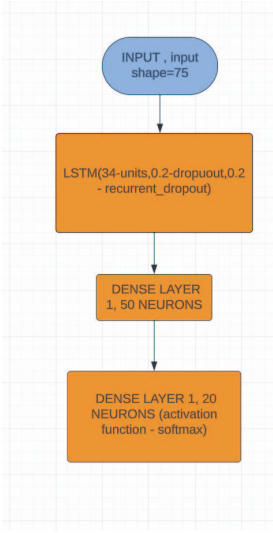

Fig. 5. Architecture of LSTM

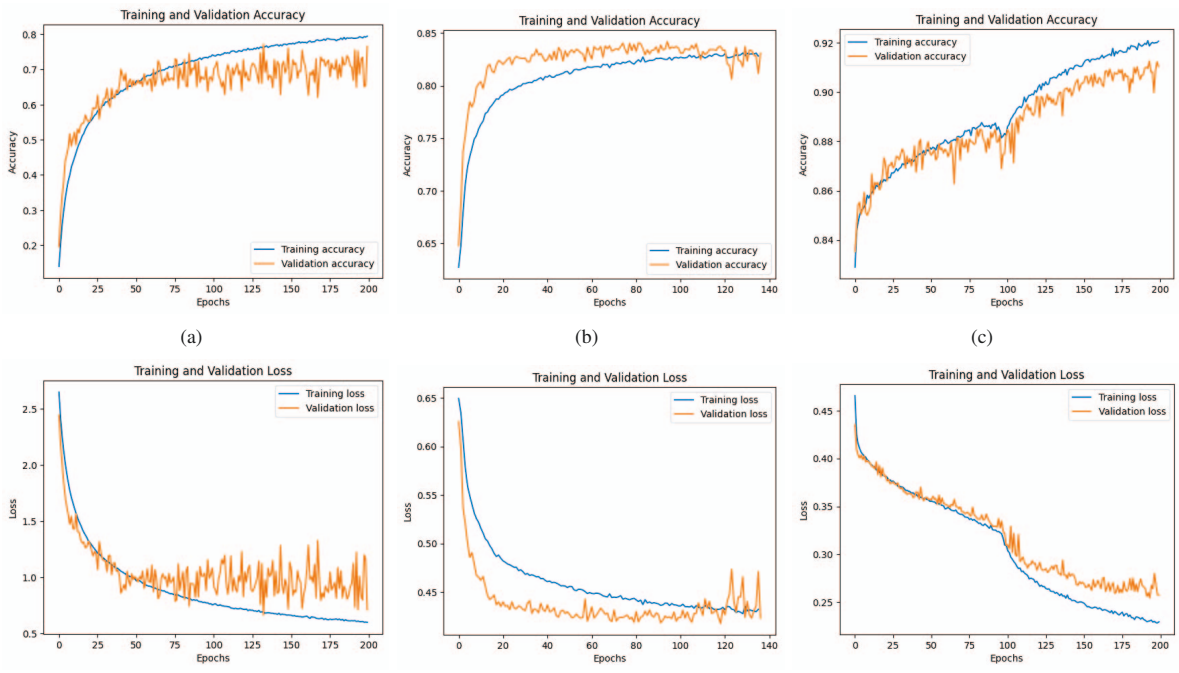

Fig. 6. a.), b.) c.) Training and Validation accuracy and d.) e.) f.) loss of only LSTM, Siamese LSTM, Siamese CNN respectively. Siamese CNN outperforms Siamese LSTM both in accuracy and loss. This can be due to the fact that CNN better exploits spatial sequences and our data has both spatial and temporal sequences.

III. Results and Discussion

In this Section, we present the results obtained from applying three different methods described in earlier Section on the body pose of individuals with stroke data set [10]. Further, we analyse the results and present a comparative study with other works.

A. Analysis of Results Obtained

As discussed in II-A, not all types of poses are preformed by all the considered subjects. Out of the 20 different poses, some poses like reach side to side and back using left hand (Rch_sd2sd_Bck_L), reach forward backward using left hand (Rch_Fwd_Bck_L) are more prominently performed by the subjects than poses like reach forward backward using right hand (Rch_Fwd_Bck_R). A distribution of number of observations for each pose category, collected from 19 different subjects is shown in Fig. 7. After pre-processing, the pose data was sent to each of the three models as described in Section II.

From Fig. 6, we can observe the training and validation accuracies and losses for the three different methods in identifying the right class for each pose. The Siamese LSTM model has an AUC of approximately 0.89, indicating its effectiveness distinguish between the similar and non similar classes. The Siamese CNN model gave a higher AUC of around 0.95, can be seen in Fig. 8. This proves the CNN model's excellent ability to grasp patterns from successive

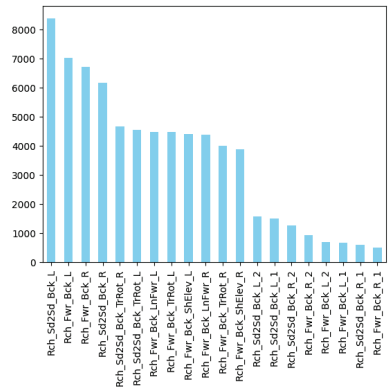

Fig. 7. Distribution of samples by pose category in data set [10]. Not all the 20 poses are performed by all the 19 subjects considered, due to individual limitations.

inputs. It has an improved result than the Siamese LSTM model, and the curve suggests that it is particularly effective across various thresholds, maintaining a high true positive rate and a low false positive rate. In conclusion, while both models exhibit strong performance, the Siamese CNN model is more effective in identifying body pose accurately, as indicated by the higher AUC value. The only LSTM model, is less effective than other two methods in this comparison. Its predictive performance could be more applicable to different types of

TABLE I
COMPARISON TABLE

Model	Avg. Score (Right Pairs)	Avg. Score (Wrong pairs)
ST-AM-CNN [12]	0.9686	0.7077
ST-VGG-16 [12]	0.9959	0.9827
Siamese LSTM	0.750480	0.280108
Siamese CNN	0.872325	0.148438

data or scenarios where understanding temporal dynamics is important. The pink dotted line in Fig. 8, indicates the micro-average ROC curve for the only LSTM model with an AUC of 0.46, is counter intuitive since an AUC less than 0.5 is worse than random guessing. It could possibly imply severe class imbalance and the model is performing poorly on the majority class.

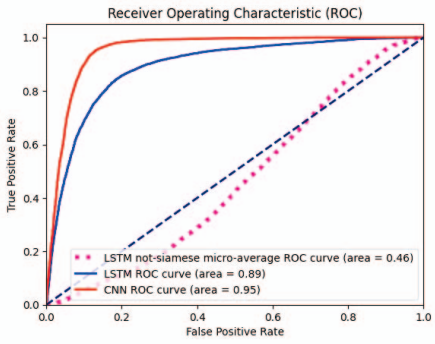

Fig. 8. Performance Comparison of only LSTM, Siamese LSTM and Siamese CNN. Siamese CNN has better AUC, due to its greater ability in identifying the patterns in poses.

B. Comparison with Other Work

In work [12], methods for pose evaluation and matching are described. This work considered skeleton like images as input to their Siamese based models. For comparison with our work, we choose models by their best scores from their work. In Table I, we present the average score obtained for identifying the right pairs (i.e. two poses of same category are identified as same, two poses of different category are identified as different) and the wrong pairs (i.e. two poses of different category are identified as same, two poses of same category are identified as different). The higher score for right pairs and lower score for wrong pairs is the ideal scenario.

The Siamese CNN model exhibits a strong preference for precision, as evidenced by its superior performance in minimizing 'wrong pairs', signifying its efficiency in reducing the errors. This trait is particularly advantageous in settings where the consequences of incorrect matches are significant. Although its score for 'right pairs' is not the highest, it suggests a deliberate calibration towards precision at the expense of capturing every potential correct match, a trade-off that may be beneficial in high-stakes applications requiring utmost accuracy. In stark contrast, the ST-VGG-16 model, despite its excellent score for 'right pairs', indicating a high sensitivity to correct matches, performs poorly on 'wrong pairs'. This implies a tendency to make more erroneous matches, which could be problematic in scenarios where such mistakes are costly. The Siamese CNN model, therefore, presents a more prudent choice over the ST-VGG-16 for tasks where the precision of the match is more critical than the sheer number of matches identified.

IV. CONCLUSION

In this work we present a novel models for body pose matching in individuals with stroke, based on Siamese LSTM and Siamese CNN architectures. This improved pose matching will help in providing efficient rehabilitation methods. Our Siamese based models outperformed their only LSTM counterpart significantly. Compared to existing works on pose matching our work minimizes the errors reported pose matching. We would like to extend this work by validating the models on our own data set and also verify its direct affect on stroke subject rehabilitation.

REFERENCES

[1] K. Sai Praveen, C. Ke, G. Ziyang, O. Parthan, C. Andrew, S. Ashwin, and V. Ramana, "Novel hand gesture classification based on empirical fourier decomposition of semg signals," in *IEEE EMBS International Conference on Data Science and Engineering in Healthcare, Medicine Biology, St. Julius, Malta*, 2023.

[2] T. Anvari, K. Park, and G. Kim, "Upper body pose estimation using deep learning for a virtual reality avatar," *Applied Sciences*, vol. 13, no. 4, p. 2460, 2023.

[3] J. Bang, D. Lee, Y. Kim, and H. Lee, "Camera pose estimation using optical flow and orb descriptor in slam-based mobile ar game," in *2017 International Conference on Platform Technology and Service (PlatCon)*. IEEE, 2017, pp. 1–4.

[4] Y. Zhang, B. Xie, H. Huang, E. Ogawa, T. You, and L.-F. Yu, "Pose-guided level design," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–12.

[5] M. Xu, Z. Zhang, Y. Gong, and S. Poslad, "Regression-based camera pose estimation through multi-level local features and global features," *Sensors*, vol. 23, no. 8, p. 4063, 2023.

[6] J. Sarsfield, D. Brown, N. Sherkat, C. Langensiepen, J. Lewis, M. Taheri, C. McCollin, C. Barnett, L. Selwood, P. Standen *et al.*, "Clinical assessment of depth sensor based pose estimation algorithms for technology supervised rehabilitation applications," *International journal of medical informatics*, vol. 121, pp. 30–38, 2019.

[7] Ó. G. Hernández, V. Morell, J. L. Ramon, and C. A. Jara, "Human pose detection for robotic-assisted and rehabilitation environments," *Applied Sciences*, vol. 11, no. 9, p. 4183, 2021.

[8] L. Lonini, Y. Moon, K. Embry, R. J. Cotton, K. McKenzie, S. Jenz, and A. Jayaraman, "Video-based pose estimation for gait analysis in stroke survivors during clinical assessments: a proof-of-concept study," *Digital Biomarkers*, vol. 6, no. 1, pp. 9–18, 2022.

[9] T. Tao, X. Yang, J. Xu, W. Wang, S. Zhang, M. Li, and G. Xu, "Trajectory planning of upper limb rehabilitation robot based on human pose estimation," in *2020 17th International Conference on Ubiquitous Robots (UR)*. IEEE, 2020, pp. 333–338.

[10] E. Dolatabadi, Y. X. Zhi, B. Ye, M. Coahran, G. Lupinacci, A. Mihailidis, R. Wang, and B. Taati, "The toronto rehab stroke pose dataset to detect compensation during stroke rehabilitation therapy," in *Proceedings of the 11th EAI international conference on pervasive computing technologies for healthcare*, 2017, pp. 375–381.

[11] X. Xu and R. W. McGorry, "The validity of the first and second generation microsoft kinect™ for identifying joint center locations during static postures," *Applied ergonomics*, vol. 49, pp. 47–54, 2015.

[12] Y. Qiu, J. Wang, Z. Jin, H. Chen, M. Zhang, and L. Guo, "Pose-guided matching based on deep learning for assessing quality of action on rehabilitation training," *Biomedical Signal Processing and Control*, vol. 72, p. 103323, 2022.