A Joint Network Based CNN for Yoga Pose Classification and Scoring

Wenxi Yang*

Department of Computer Science University of Wisconsin-Madison Madison, WI, United States, 53715

* Corresponding author: wyang235@wisc.edu

Abstract. Comparing to traditional rehabilitation, rehabilitation at home becomes a need during pandemic. The technique brought up in this paper allows patients and yoga fans exercise at home with low cost and comfort while can also evaluate their postures. Previous works focus either on classifying poses or scoring on the sameness between the two input branches of patients' poses and normative poses, but they ignore the combination of them in one single network. In this study, a residual block based Siamese CNN network with classification and scoring modules is proposed, aiming at providing accurate pose matching scores and classify pose types on yoga postures simultaneously. The Siamese network takes two inputs of learner's pose and standard pose, which are preprocessed skeleton images by OpenPose. With the addition of residual block on the first convolutional module, back propagation is facilitated, which boosts up the process of updating parameters and optimization. The model calculates total loss by summing up cosine embedding loss and cross entropy loss in which the weight parameter lambda could be modified based on need. As for the scoring module, cosine similarity is used to calculate pose resemblance on batch level. The improvement in model performance is obvious when comparing the loss and accuracy between the Siamese network with residual block and VGG-16. Experimental results indicate that the residual block based Siamese network achieves competitive performance compared to the VGG-16 and can provide scoring feedback to learner's yoga poses.

Keywords: Residual block; CNN; Yoga pose; Classification; Scoring.

1. Introduction

Rehabilitation is now benefiting an estimated 2.4 billion people globally [1]. Rehabilitation facilitates people's recovery from injuries in several medical areas, including neuropsychology in which people regain cognitive functions after receiving therapies and physical rehabilitation that recovers patients' functionality and mobility, etc. However, the popularity made rehabilitation paramedic shortage an issue. According to report World Population Prospects 2022, the percentage of people aged 65-year-old or above in the share of global population is predicted to rise from 10 percent in 2022 to 16 percent in 2050 [2]. The report states the number of successive birth cohorts will decline as a result of the loss in fertility [2]. Given that manpower shortages and population aging are unavoidable in the near future, automation in healthcare must be taken into account.

In the traditional physical rehabilitation, therapists would intervene in patients' actions. Although therapists often offer patients professional advice on postures and exercises, humans are also prone to introduce subjective biases. In research investigating low back pain (LBP) and pain beliefs, Rufa et al. argue that clinicians and therapists with higher fear avoidance beliefs (FAB) tend to increase patients' FAB and pain catastrophizing (PC) [3]. More specifically, higher FAB from patients indicates that their estimation of threat will cause avoidance behaviors; while patients are aware of upcoming pain simulation, they tend to bear increasing and exaggerated PC [4, 5]. Additionally, with high association between burnout and physical therapy, various manifestations, including both mentally fatigue, physically tiredness and impatience when facing patients, would lead to negative consequences to patients' rehabilitation and culminate into more medical errors [6]. To prevent subjective factors influencing rehabilitation performance, implementing adaptive scoring metrics with machine learning algorithms can be considered as an effective alternative in this case.

With the emergence of robotics in the area of rehabilitation early as in 1980s, more and more researchers start to embrace high technology into therapies. However, the use of wearable robots in rehabilitation for patients is still under debates as the performance is inferior comparing to having the assistance from therapists. Indeed, the traditional ways of requiring patients to wear sensors to measure motion accuracy raise some concerns over 1) patients' personal preferences in carrying sensors constantly while undergoing rehabilitation, 2) potential lack of therapist's supervision and guidance, 3) decreasing in accuracy of motion detection when the sensor does not attach to the patients well, 4) expensive cost and 5) limited to lab environment [7, 8].

While, with the pose-guided matching model brought up by Qiu et al., patients can easily adopt personal rehabilitation in several low-cost at-home scenarios [7]. In research regarding pose-guided matching on assessing quality of motions in rehabilitation training, Qiu et al. focuses on providing objective and accurate score [7]. The current model Qiu et al. implement is a pair-based Siamese convolutional neural network called ST-AMCNN with the spatial transformer network (STN) to extract multi-scale features [7]. Although they included the scoring function in the network, the classification function for poses was ignored. In fact, it is also crucial to inform patients about their current stage. In this paper, a classification module is also considered to be added into the pairwise Siamese convolutional neural network. The model is implemented on the dataset of yoga postures. Recent research has shown yoga is benefiting people in several perspectives including helping to regulate blood glucose levels and keeping the cardiovascular system healthy [9]. Yoga involved several postures including physical, mental, and spiritual exercises aiming for relaxation, which make it suitable for training the pairwise Siamese CNN model.

To give a general view, the contributions of this paper can be summarized as 1) a deep learning-based method with high performance was proposed to classify various yoga poses 2) develop a joint network that combines classification and scoring module simultaneously. The experiment results proved that the proposed model achieved the promising results.

2. Methodology

2.1. Data description and preprocessing

The source of the dataset used in this study is Yoga Poses Dataset from Kaggle [10]. The original dataset divides data into train/test subdirectories for easy access. It contains 1,551 RGB images in total with 1,081 in subdirectory train and 470 in subdirectory test. The dataset includes five yoga poses: down dog, goddess, plank, tree, and warrior. Figure 1 presents the sample data in the collected dataset.

Figure 1. Original sample images.

In this project, *OpenPose* is used to extract skeleton information of the Yoga Poses as shown in Figure 2. It is a real-time multi-person system for detecting human body, hand, foot, and facial key points in single images [11], it helps eliminate potential uncertainty in the training process brought up, such as height. With only skeleton information, the model is trained on useful features and provides higher accuracy. After converting all five categories of yoga poses into skeleton images, the training result shows that down dog and plank fail to give successful learn as the images picture human body from side instead of front, making it harder to learn the poses. As a result, down dog and plank folders are removed for overall performance while goddess, tree, and warrior poses are kept for subsequent training.

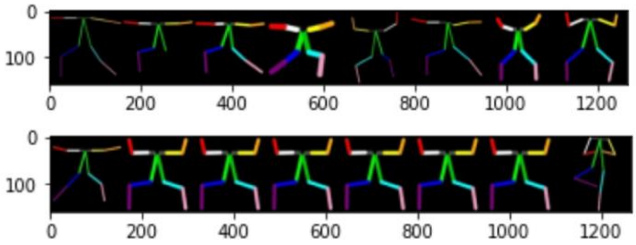

Figure 2. Sample skeleton images.

In this work, the learner and standard pose both consists of 544 skeleton images of single human figure. Standard images are made by taking one of each pose from the learner images and making multiple duplications. To smooth the process of learning, images of same type of poses are put together in the order of goddess, tree, warrior. Additionally, resizing input images is an important step of preprocessing. Each image is made uniformed and cropped to size of 156×156. Then, images are converted into tensor objects for training and testing. The size of the training dataset is 70% of all 544 images, and the size of the testing dataset is the rest 30%. Lastly, the model also shuffles tensor objects for randomness and normalized them.

2.2. Proposed method

Convolutional Neural Network (CNN) which consists of convolutional layers, pooling layers and full-connect layers, is developed from Artificial Neural Network (ANN) and has strengths in making the network more suitable for image-focused tasks [12]. In a CNN, convolutional layers focus on feature extraction following an activation function; a pooling layer then performs down-sampling along the spatial dimensionality of the given input and further reduces the number of parameters; lastly, a full-connected layer attempts to produce class scores from the activations [12].

A Siamese network architecture takes two branches of input with shared weights and parameters [13]. In this study, a residual block based Siamese Network with joint functions of scoring and classification is implemented. The network learns from positive image pairs which the pose in learner images can match with those of standard images match; conversely, if the poses in learner images do not match with those of standard images, the network learns them as negative image pairs.

The network shown above is inspired by Siamese network with two identical CNN. A slight change on the CNN is addition of shortcut layer, also considered as a residual block from ResNet. The purpose of adding residual block is to help with back propagation, facilitating the process of updating parameters and optimization. Figure 3 gives a high-level overview of this network. The paired input poses were obtained from the Yoga Pose Dataset. The image above, named Ip, is the skeleton learner image extracted by *OpenPose* from a patient's pose; the image below, denoted Is, comparatively, is the standard pose. A higher similarity between Ip and Is gives a high score on patient's posture, showing that poses are matched successfully. Both Ip and Is are thrown into the network and go through five convolutional modules. The residual block-based CNN of Ip gives scoring output as a term in metric to evaluate the similarity between learn pose and standard pose. A cosine embedding loss shown in formula (1),

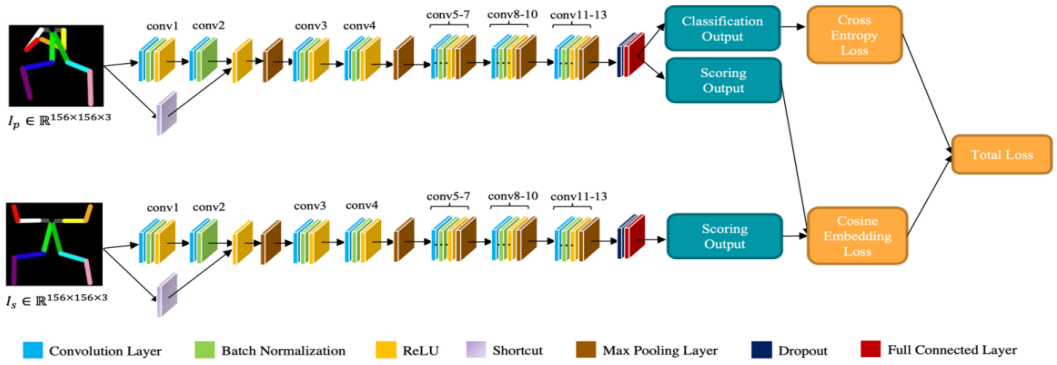

Figure 3. Flowchart of residual block based Siamese Network.

$$loss(x, y) = \begin{cases} 1 - \cos(x_1, x_2), & \text{if } y = 1 \\ \max(0, \cos(x_1, x_2) - \text{margin}), & \text{if } y = -1 \end{cases} \quad (1)$$

is calculated where x_1 and x_2 are two poses from learner and standard branch, y denotes the ground true label, and margin refers to the minimum distance between correct and wrong poses sample pairs. The cosine embedding loss works as a direct evaluation on the network's performance on matching poses between learner and standard, assuring that matched images in a pair are pulled closer and unmatched images are pushed far away [13]. Specifically, when $y = 1$, two images are match; conversely, $y = -1$ when two images fail to match. The network returns a classification output for learner branch as a function to classify whether the pose is goddess, tree, or warrior. A cross entropy loss shown in formula (2),

$$loss = -\sum_{i=1}^{size}{}^{output} \; y_i \cdot \log \hat{y}_i \quad (2)$$

is calculated and then added to the cosine embedding loss to generate total loss for $I_P$and I_s. The total loss is given by formula (3),

$$total\ loss = \lambda_s \begin{cases} 1 - \cos(x_1, x_2) & \text{if } y = 1 \\ \max(0, \cos(x_1, x_2) - \text{margin}) & \text{if } y = -1 \end{cases} - \lambda_c \sum_{i=1}^{size}{}^{output} \; y_i \cdot \log \hat{y}_i \quad (3)$$

where λ_s indicates the weight of cosine embedding loss and λ_c indicates the weight of cross entropy loss. In this paper, both λ_s and λ_c are set to 1, but the total loss is subject to change when lambda changes.

2.3. Implementation details

This paper uses Pytorch on the model with Apple M1 Pro chip. The residual-block based CNN network is trained on Yoga Pose Dataset with $batch\ size = 8$, $learning\ rate = 0.0001$, $Adam\ optimizer$, $epoch = 15$. The classification accuracy is calculated by iterating through all images and doing division in the end; whenever the tensor objects from learner and standard images equal to each other, the count of match increases by one. The accuracy of whether the learner pose is close to standard pose is calculated by using

$$CosineSimilarity = \frac{x_1 \cdot x_2}{\max(\|x_1\|_2 \cdot \|x_2\|_2, \ \epsilon)} \quad (4)$$

When the paired images have $label = 1$, indicating a good learner pose, a cosine similarity is calculated and added to the result array; otherwise, -1 is appended to indicate a bad learner pose. Besides, VGG-16 is the model alongside with the residual-block based CNN brought up in this paper. Table I provides the architecture of the proposed residual-block based CNN and comparative model (i.e., VGG16).

3. Results and discussion

Table 1. Classification performance of various models

Model	Performance			
	Training Loss	**Training Accuracy**	**Testing Loss**	**Testing Accuracy**
Proposed Residual block based CNN	0.3218	0.9604	0.1039	0.9937
VGG-16	0.3235	0.9608	0.1430	0.9812

As shown in Table II, the proposed residual block based CNN discussed in this paper gives a loss of 10.39% and an accuracy of 99.37% on the testing dataset, which are better than the results from VGG-16 which is a typical backbone in various computer vision tasks. Both VGG-16 and the residual block based CNN show good classification, but the addition of residual block in CNN slightly improves overall performance.

Table 2. Comparative Network architecture

Type	**Residual-block based CNN**	**VGG16**
Convolution	$\times 2 + shortcut$	$[3 \times 3, 4] \times 2$
Max pooling	2×2	2×2
Convolution	$[3 \times 3, 8] \times 2$	$[3 \times 3, 8] \times 2$
Max pooling	2×2	2×2
Convolution	$[3 \times 3, 16] \times 3$	$[3 \times 3, 16] \times 3$
Max pooling	2×2	2×2
Convolution	$[3 \times 3, 32] \times 3$	$[3 \times 3, 32] \times 3$
Max pooling	2×2	2×2
Convolution	$[3 \times 3, 64] \times 3$	$[3 \times 3, 64] \times 3$
Max pooling	2×2	2×2
Full connected	64	64

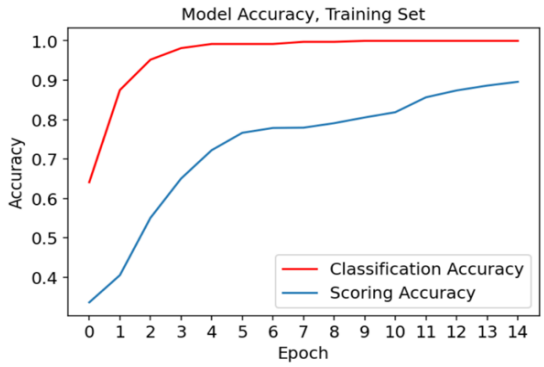

Figure 4. Model accuracy on training set.

Both classification accuracy and scoring accuracy increase smoothly in training set as expected. Figure 4 shows that the red curve reaches 1.0 as soon as it is around 4/15 epoch. The scoring accuracy keeps growing from 0.3359 to 0.8958 from epoch 0 to epoch 14.

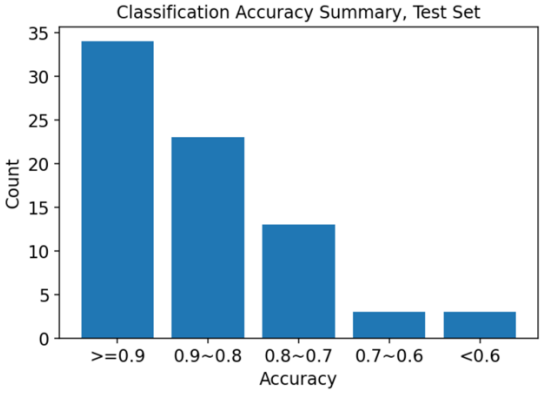

Figure 5. Model accuracy on training set.

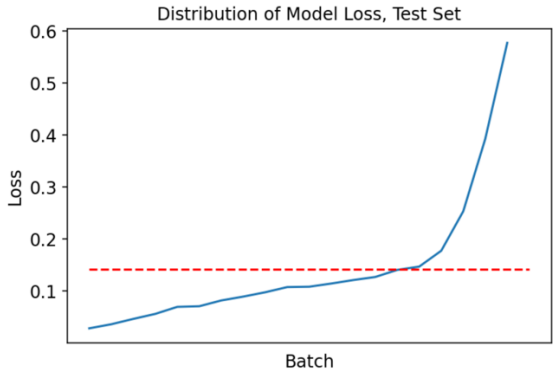

Figure 6. Distribution of model loss on test set.

A good, trained model gives more accuracies greater than 0.9 and the portion of accuracies less than 0.6 should be as small as possible. In Figure 6, there is a clear decreasing trend on the portion of accuracies in five categories. It is also shown in Figure 6 that the amount of loss less than and greater than average (shown as the red dashed horizontal line) are half and half. Both Figure 4 and 5 indicate a good performance of classification model after training.

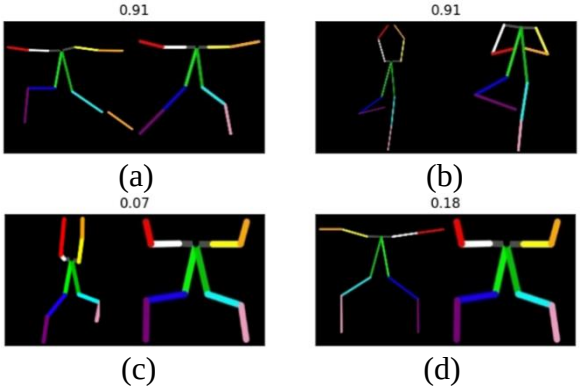

Figure 7. Sample output with high and low scoring accuracy.

This study verifies that using the proposed residual block based CNN model to match and score yoga poses is feasible since the model can successfully differentiate learner's pose and standard pose. To have a direct view of how the residual block based CNN performs, a piece of output is extracted and shown in Figure 7. Within all paired images possessing high accuracy score, some paired-poses physically match perfectly, such as Figure 7a, but some do not tend to be perfectly matched in human eyes, such as Figure 7b which is the reason of implementing neural network to help with human eyes. As shown in Figure 7c and 7d, the model also gives reasonable scores on poses that do not tend to similar. The addition of scoring module does not seemingly drag down the performance of classification model. One potential reason is both modules work on the skeleton information and learn the difference between two skeleton poses. Furthermore, in the training process, both modules may extract similar features, and the optimizations of loss distribution are possibly subjected to the same direction as well. In this way, the combination of scoring and classification modules work well. Moreover, the network can also be used on other images other than yoga poses if paired learner's images and standard images in skeleton form are fed into it.

4. Conclusion

In this paper, a Siamese CNN model with residual block is proposed with scoring and classification modules on yoga pose dataset. Comparison between the residual block based Siamese Network and VGG-16 is done and the testing result shows that classification accuracy reaches 99.37% and scoring accuracy reaches 87.85%, indicating that the residual block based Siamese CNN model is reliable and trustworthy. Also, experimental results show that this model performs better than VGG-16. One limitation of current study is that current model is only used on Yoga Pose dataset. In fact, pose

matching and scoring could be implemented on several areas as well, such as rehabilitation at clinics or fitness training. In future, it is expected to train, test, and modify the proposed model on various topics, while also keep trying new combination of various neural networks for better performance.

References

[1] W. H. O. (WHO), "Rehabilitation," 11 November 2021. [Online]. Available:https://www.who.int/news-room/fact-sheets/detail/rehabilitation

[2] P. D. United Nations Department of Economic and Social Affairs, "World Population Prospects 2022: Summary of Results," United Nations Publication, New York, 2022.

[3] A. Rufa, M. J. Kolber, J. Rodeghero and J. Cleland, "Musculoskeletal Science and Practice 55 (2021) 102425 Available online 7 July 2021 2468-7812/© 2021 Elsevier Ltd. All rights reserved.Original article The impact of physical therapist attitudes and beliefs on the outcomes of patients with low back pain," Musculoskeletal Science and Practice, vol. 55, 2021.

[4] R. J. Gatchel, R. Neblett, N. Kishino and C. T. Ray, "Fear-Avoidance Beliefs and Chronic Pain," Journal of Orthopaedic & Sports Physical Therapy, vol. 46, no. 2, pp. 38-43, 2016.

[5] P. J. Quartana, C. M. Campbell and R. R. Edwards, "Pain catastrophizing: a critical review," Expert Review of Neurotherapeutics, vol. 9, no. 5, pp. 745-758, 2009.

[6] S. D. Burri, K. M. Smyrk, M. S. Melegy, M. M. Kessler, N. I. Hussein, B. D. Tuttle and D. J. Clewley, "Risk factors associated with physical therapist burnout: a systematic review," Physiotherapy, vol. 9, no. 24, 2022.

[7] 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, 2021.

[8] A. Elkholy, M. E. Hussein, W. Gomaa, D. Damen and E. Saba, "Efficient and Robust Skeleton-Based Quality Assessment and Abnormality Detection in Human Action Performance," IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 1, pp. 280-291, January 2020

[9] I. Stephens, "Medical Yoga Therapy," Children, vol. 4, no. 2, 2017

[10] N. Pandit, "Yoga Poses Dataset," 2020. Retrieved July 27, 2022 from https://www.kaggle.com/datasets/niharika41298/yoga-poses-dataset

[11] Z. Cao, T. Simon, S. Wei, Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," IEEE Conference on Computer Vision and Pattern Recognition, pp. 7291-7299, 2017

[12] K. O'Shea, R. Nash, "An Introduction to Convolutional Neural Networks," arXiv: 1511.08458[cs.NE]

[13] I. Melekhov, J. Kannala, E. Rahtu, "Siamese network features for image matching," 2016 23rd International Conference on Pattern Recognition (ICPR), 2016, pp. 378-383, doi: 10.1109/ICPR.2016.7899663