Deep Learning-based Yoga Posture Correction with Dynamically Varying Poses

Sumurth Dixit[1, a)], Vikash Kumar Patel[1, b)], Bharat Kumar Sharma[1, c)], Sushil Kumar[1, d)] and Umang Rastogi[1, e)]

[1]*Department of Computer Science & Engineering, KIET Group of Institutions, Delhi-NCR Ghaziabad, India*

[a)] Corresponding author: sumurth17dixit05@gmail.com
[b)]vikashpateljps1@gmail.com
[c)]iambharatks1@gmail.com
[d)] drsushil.cs@gmail.com
[e)] rastogi225103@gmail.com

Abstract. Since COVID-19 has had such an enormous influence on everyone's lives, Yoga has grown more and more popular every day. This can be explained by Yoga's potentially wide range of physical, mental, and spiritual advantages. Without a teacher's guidance, many people have taken up this trend and practiced Yoga. Nevertheless, practicing Yoga improperly or without the correct direction can harm one's health. There is currently only a handful of pieces of research on this complex topic. The described methods' insufficient precision in similar poses, the fact that there only exist a handful of Yoga positions in the existing datasets, and their demand that the user is under the camera's range of view are the problems. To address the problems, we suggest in this study a deep learning model-based posture correction for Yoga that will provide users and clients with the right instruction. Since earlier models only had a small number of positions and had poor accuracy for comparable postures, the main goal of the proposed work is to include additional positions and strive to provide higher accuracy for similar stances.

INTRODUCTION

Yoga is crucial for good health. Yet, accurately executing Yoga poses can be difficult, especially for beginners. Improper posture execution might result in harm or diminished benefits. Consequently, it's crucial to make sure people are doing Yoga poses correctly. Yoga teachers typically use both verbal instructions and physical corrections to improve posture. Using these techniques, especially in group situations, can be difficult. Moreover, they could be ineffective in correcting positions that alter dynamically while the instructor is changing the pose. In order to successfully correct Yoga postures and manage dynamically changing positions, an automated system is required.

Particularly, convolutional neural networks (CNNs) have made tremendous strides in the field of image recognition. These methods have been used for a variety of purposes, including correcting posture. In this article, we provide a deep learning-based method for correcting Yoga posture using dynamically changing postures. The suggested method analyzes photographs of a person executing Yoga poses using a CNN and then offers suggestions to rectify the stance. The widely recognized convolutional pose machine (CPM) adopted a method that advanced joint recognition over a sequence of network stages. Regarding the pose estimation problem, piled hourglass connections used streams of the oblong structure [23,24]. Techniques that adhere to our notion of thinking holistically about stance have only shown modest results. Practically, in order to transfer the joint locations, Mori and Malik [4] attempt to identify the nearest example from a collection of annotated photos for each test image. Here, we examine related studies on 3D human posture prediction that are most pertinent to our methodology. Most current work that uses the issue as a direct 2D image to 3D pose regression job while using deep features [9].

RELATED WORK

Yoga, which is a physical exercise, has gained tremendous significance in the community of medical researchers. Since the early days of computer vision, the concept of describing articulated objects in a general and human pose in particular as a graph of components has been promoted [1]. The Pictorial Structures (PSs), first developed by Fishler Elschlager were streamlined and made usable by employing the distance transformation approach of Felzenszwalb and Huttenlocher [2,3]. There are therefore many different PS-based models with the immediate application that were afterward created. Pooling and local reaction normalization layers are introduced after an array of convolutional layers, and layers for dropout are used to regularize the layers that are completely interconnected [12]. The person's anterior-body joint coordinates have been regressed using a network that was explicitly conducted to accomplish so. To replace it, there is a softmax loss layer, which exists in the ConvNets in image categorization [13,14]. We modify the dimension and normalize again the markings to a range in the individual joint subject [18].

Li and Chan train a regression model with deep learning to forecast 3D stances as seen in the pictures. The study that comes closest to ours employs CNNs along with neighborhood component analysis to regress toward a point in an embedding representing pose [9, 10]. Unfortunately, a cascade of networks is not used in this work. Nonetheless, face points of deep neural network (DNN) regressors have been employed for localization [11]. Many models showing intricate coupled interactions have been put forth more recently. An amalgam model of parts is employed in [21, 36]. By incorporating a variety of PSs, mixture models have been explored on a full model size, published by Johnson as well as Everingham [22].

Pictorial structural models have been a common component of traditional pose estimation techniques that maximize a partial configuration based on local image evidence for a component and a priority for the relevant components' locations along the human kinematic chain. Poselets are used in an alternative strategy. There have been early examples of posture comparisons utilizing ConvNets for pose estimation [7, 8, 35]. Recently, ConvNet was utilized to directly regress joint coordinates in an AlexNet-like manner, with a cascade of ConvNet regressors to improve accuracy over a single network of pose regressors [15, 16]. Tompson, Jain, and others used ConvNet architectures in a number of studies to directly regress heatmaps for each joint, adding additional layers to build a spatial model based on the markov random field (MRF) [4, 5]. ConvNets were first employed with temporal information from videos for action recognition [17], where the optical flow was utilized as a motion feature for the network's input. In the wake of this research [5, 20] examined the application of temporal similarity, we use flow or RGB from numerous surrounding frames in the posture estimation method predicting joint positions in the current frame using a neural network [6,19].

For certain application areas, such as MS COCO key points [25], MPII human pose database [27], Human3.6M [26], or LSP [28], thoroughly annotated data is readily accessible. These data sources include fully supervised models. Due to many direct labels, methods using these datasets are typically trained without extra priors. Poses have been described using pictorial frameworks [29, 30, 31, 32] and associated uncertainty. For scenarios where a single posture needs to be estimated or numerous positions at once [33, 34], confidence heatmaps are frequently used. In order to learn a posture prediction model, our approach does not employ pre-existing picture annotation.

PROPOSED METHODOLOGY

In this paper, our deep learning paper uses the following methodologies:

- *Data Collection:*

  Assembling video evidence of individuals engaging in numerous acts. A uniform method should be used to identify each video with the appropriate stance.

- *Model Selection:*

  Pick the most relevant deep learning models, and afterward evaluate the output produced by each model to the output produced by our suggested model. In order to process the data, all of the models employ CNN. The entire working of the proposed model is shown in Fig. 1. The working steps of the implementation of the model are shown in Fig. 2.

- *Feature Extraction:*

The model receives input in the form of movies, from which frames are taken at certain intervals and supplied to the technique to create joint vectors. The axis and joints' angles are calculated.

- *Feedback Generation:*

The discrepancies are computed on a regular basis in real-time for each angle, along with figures for the angle and marks indicating that the joint is also being performed appropriately and which practice is being performed incorrectly for each specific exercise.

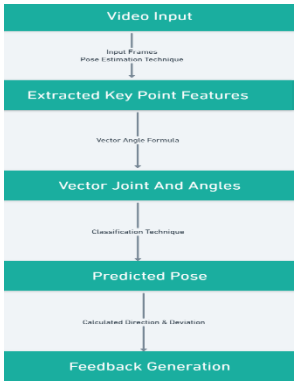


FIGURE 1. Overview of the proposed model

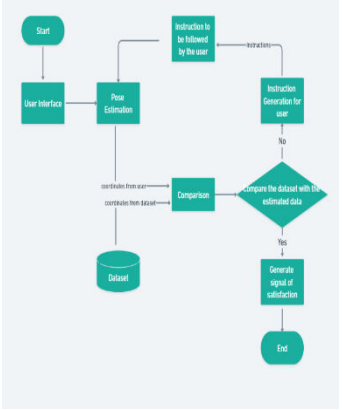


FIGURE 2. Working Steps of the research paper

As the covid-19 has made a huge impact on our lives, the popularity of Yoga is increasing daily. This is due to the fact that Yoga practice has both physical and mental advantages. The major elements of the suggested Yoga posture Correction System are two. It's them,

- Key points detection using OpenCV
- Higher probability prediction and comparision

The following describes the system's general process. The media streaming server is used to record the user's movements and feed them live to the system. The system then uses posture estimation and OpenCV to identify the user's joints or critical locations. The posture estimation algorithms first identify critical spots, which are then transmitted to the Yoga position identification module. Given that the present model can only identify key points, it was feasible to anticipate the user's Yoga posture using the key points information gathered from the preceding components before the user reached the final stage of the asana. Video snapshots of the model output are in Fig. 3.

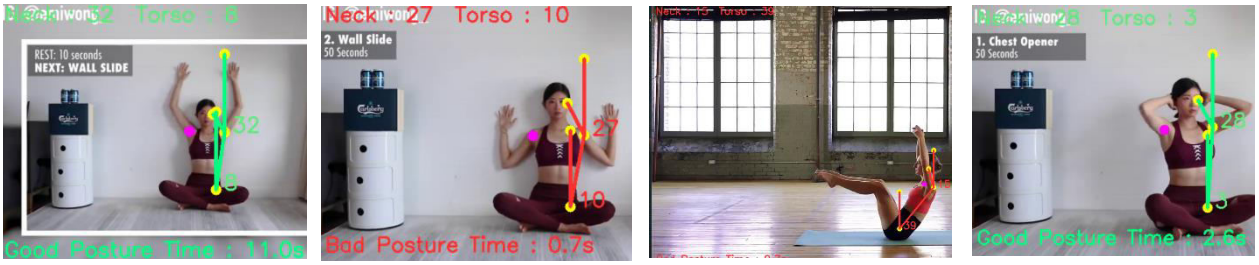


FIGURE 3. Working examples of the model on different people doing Yoga and exercise including different postures in the gym. The model analyzes the body joints of the people and captures them to show good posture time and bad posture time. Good posture times appear with green lines and bad posture times appear with red lines.

RESULT ANALYSIS AND DISCUSSION

Input, hidden, and output layers are the three different sorts of layers that make up CNNs' structure. Depending on how complicated the training data is, different numbers of hidden layers may be best; too few hidden layers may cause overfitting, while too many may result in underfitting. Each node in the subsequent layers of the CNN is connected to every other node, which is referred to as being completely connected. By CNN, supervised training is frequently used, where each input data point is associated with a particular output label or class. CNNs are used to classify human poses by computing the angles between important points, which are subsequently fed into the CNN played by computing angles between key points, which are then used as input for the CNN.

- *Model Accuracy and Model Loss:* The model's accuracy on the training dataset was 98.5782% after 60 iterations, and its accuracy on the validation dataset was 94.2101%. The model's accuracy was predicted using the formula given in Eq. (1).

$$Model\ Accuracy = \frac{Total\ Number\ of\ green\ lines}{Total\ number\ of\ green\ plus\ red\ lines} \quad (1)$$

- *Precision:* It is employed to assess the model's effectiveness. Given that all the samples were accurately identified as positive, it is calculated as the ratio of those samples. Our model's accuracy is 0.92813. It is given by Eq. (2).

$$Model\ Precision = \frac{TP}{TP+FP} \quad (2)$$

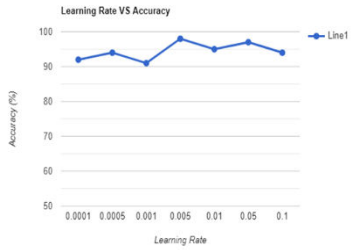
- *Recall:* It is used to assess the model's effectiveness. Given that all samples are expected to be positive, it is calculated as the proportion of samples that were correctly identified as positive. Our model has a 0.90926 recall rate. In mathematics, it is represented by Eq. (3).

$$Model\ Recall = \frac{TP}{TP+FN} \quad (3)$$

TP = True Positive, FP = False Positive, FN = False Negative

- *Impact of Learning Rate on Model Performance:* It is defined as the magnitude of the model's correction in response to the mistake shown each time the model's weights are adjusted. Although a very fast learning rate can considerably shorten the learning period, accuracy suffers Finding the ideal learning rate is a laboriously difficult procedure that necessitates trial and error. The impact of learning rate vs. accuracy is shown in Fig. 4.

The table 1 shows the accuracy of the model in 6 Yoga compared on different library implemented models with the proposed model of this research paper along with graph in Fig. 5 of the data obtained from table 1.



Exercises	EpiplorPos	OpenPose	PoseNet	MediaPipe	Paper
Balasana	37.53	72.22	75.01	78.78	88.49
Marjaryasana	40.24	61.45	69.52	78.53	85.54
Tadasana	51.25	55.55	70.55	87.59	90.92
Sirsasana	48.25	58.75	66.66	81.25	85.75
Veerbhadrasan	62.50	69.51	72.22	81.28	81.81
Padmasna	56.25	69.25	77.77	87.52	88.89

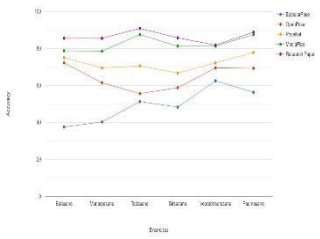


FIGURE 4 (LEFT). The learning rate vs. accuracy of our model which signifies that stable and quick learning happens at 97% accuracy between 0.001 and 0.005 learning rate
TABLE 1 (MID). The accuracy of the mentioned 6 Yoga when run on different algorithms and compared to our algorithm.
FIGURE 5 (RIGHT). The comparison of the accuracy of the models in estimating 6 Yoga poses.

CONCLUSION

Using a typical RGB camera, we propose a Yoga tracking system in this paper. In order to accurately represent Yoga practitioners, human pose estimate is necessary (HPE). In a standard RGB image, HPE proposes to replicate all skeletal connected components of a given individual. It has significant applications in motion graphics, software system, and human detection and recognition. For HPE to achieve high accuracy, representations must be both locally geometrically accurate and globally semantically distinguishing. So, the estimation of the correctness of the multiple Yoga exercises carried out by people varies depending on wide ranging multi-scale information. unobstructed pose is employed to observe the individual's attention to recognize the essential characteristics. Videos that have been previously recorded are an opportunity for people to access the model. In order for the model to achieve the intended outcome, the research has taken the monitoring angles from the users' exercises and utilized them as a feature. Angles between the joints and the ground are taken into account. Any changes to the joint angles have an impact on the result. The end-to-end deep learning-based system does away with the requirement for manually creating the features, making it possible to add new asanas by just keeping the model. We used the LSTM to remember the pattern observed in recent frames and the time-distributed CNN layer to find patterns between important points in a single frame. The process is made even more reliable by using LSTM to store the memories of prior frames, eliminating mistakes brought on by erroneous key point identification. When compared to past approaches like OpenCV, PoseNet, and MediaPipe, the model we developed shows promising results. With the added bonus that the approach still uses simple processing, it may be used in daily life to identify bad postures and assist people in avoiding significant joint and anterior-related disorders.

REFERENCES

Using a typical RGB camera, we propose a Yoga tracking system in this paper. In order to accurately represent Yoga practitioners, human pose estimate is necessary (HPE). In a standard RGB image, HPE proposes to replicate all skeletal connected components of a given individual. It has

1. M. A. Fischler and R. A. Elschlager. The representation and matching of pictorial structures. Computers, IEEE Transactions on, 100(1):67–92, 1973
2. R. Nevatia and T. O. Binford. Description and recognition of curved objects. Artificial Intelligence, 8(1):77–98, 1977.
3. P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. International Journal of Computer Vision, 61(1):55–79, 2005.
4. . G. Mori and J. Malik. Estimating human body configurations using shape context matching. In ECCV, 2002.
5. A. Jain, J. Tompson, Y. LeCun, and C. Bregler. MoDeep: A deep learning framework using motion features for human pose estimation. Proc. ACCV, 2014.
6. J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. Proc. NIPS, 2014.
7. P. Buehler, M. Everingham, D. P. Huttenlocher, and A. Zisserman. Upper body detection and tracking in extended signing sequences. IJCV, 2011.
8. M. Eichner, M. Marin Jimenez, A. Zisserman, and V. Ferrari. 2d articulated human estimation and retrieval in (almost) unconstrained still images. IJCV, 2012.
9. S. Li and A. B. Chan. 3d human poses estimation from monocular images with a deep convolutional neural network. In ACCV, 2014.
10. G. W. Taylor, R. Fergus, G. Williams, I. Spiro, and C. Bregler. Pose-sensitive embedding by nonlinear nca regression.In NIPS, 2010.
11. Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on pages 3476–3483. IEEE, 2013.
12. Pfister, T., Charles, J., Everingham, M., Zisserman, A.: Automatic and efficient long term arm and hand tracking for continuous sign language TV broadcasts. BMVC (2012)
13. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition videos. NIPS (2014)
14. Tompson, J., Jain, A., LeCun, Y., Bregler, C.: Joint training of a convolutional network and a graphical model for human pose estimation. NIPS (2014)

15. L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In Proc. CVPR, 2009.
16. G. Gkioxari, B. Hariharan, R. Girshick, and J. Malik. Using k-pose lets for detecting people and localizing their key points. In Proc. CVPR, 2014.
17. K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. Proc. NIPS, 2014.
18. Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A.: Synthetic data and artificial neural networks for natural scene text recognition. arXiv preprintarXiv:1406.2227 (2014)
19. T. Pfister, K. Simonyan, J. Charles, and A. Zisserman. Deep convolutional neural networks for efficient pose estimation in gesture videos. Proc. ACCV, 2014.
20. S. Johnson and M. Everingham. Learning effective human pose estimation from inaccurate annotation. In CVPR, 2011.
21. Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures of parts. In CVPR, 2011.
22. Yale Song, David Demirdjian, and Randall Davis. Continuous body and hand gesture recognition for natural human-computer interaction. ACM Transactions on Interactive Intelligent Systems (TiiS), 2(1):5, 2012.
23. Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. ICLR, 2016.
24. Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, editor=" Fleet David Zitnick, C. Lawrence", Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars. Microsoft coco: Common objects in context. In Computer Vision – ECCV 2014, pages 740–755. Springer International Publishing, 2014.
25. Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(7):1325–1339, jul 2014.
26. Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2014.
27. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.
28. Leonid Pishchulin, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele. Poselet conditioned pictorial structures. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 588–595, 2013.
29. Varun Ramakrishna, Daniel Munoz, Martial Hebert, James Andrew Bagnell, and Yaser Sheikh. Pose machines: Articulated pose estimation via inference machines. In European Conference on Computer Vision, pages 33–47.Springer, 2014.
30. Benjamin Sapp, Chris Jordan, and Ben Taskar. Adaptive pose priors for pictorial structures. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 422–429. IEEE, 2010.
31. Yi Yang and Deva Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In CVPR 2011, pages 1385-1392. IEEE, 2011.
32. Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1653–1660, 2014.
33. 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.
34. Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In European Conference on Computer Vision, pages 34–50. Springer, 2016.
35. K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5693–5703.
36. I. Kligvasser, T. Rott Shaham, and T. Michaeli, "xunit: Learning a spatial activation function for efficient image restoration," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2433–2442.