

Discover human poses similarity and action recognition based on machine learning

Mohammed Moath Abdulghani¹, Mohammed Talal Ghazal², Anmar Burhan M. Salih²

¹Department of Basic Sciences, College of Agriculture and Forestry, Mosul University, Mosul, Iraq

²Department of Computer Engineering Technology, Northern Technical University, Mosul, Iraq

Article Info

Article history:

Received Oct 3, 2022

Revised Nov 2, 2022

Accepted Nov 16, 2022

Keywords:

Action recognition

Computer vision

Cosine distance

OpenPose

Pose similarity

Support vector machine

ABSTRACT

In the computer vision field, human action recognition depending on pose estimation recently made considerable progress, especially by using deep learning, which improves recognition performance. Therefore, it has been employed in various applications, including sports and physical activity follow-up. This paper presents a technique for recognizing the human posture in different images and matching their pose similarity. This aims to evaluate the viability of employing computer vision techniques to verify a person's body pose during exercise and determine whether the pose is executed properly. Exercise is one strategy we use to maintain our health throughout life. Gymnastics and yoga are two examples of this type of exercise. The proposed algorithm identifies human action by recognizing the body's key points. The OpenPose library has been used to detect 18 key points of the human body. The action classification task is performed using the support vector machine (SVM) algorithm. Then, the algorithm computes the similarity of the human pose by comparing a model image to a test image to determine the matching score. Evaluations show that our method can perform at a competitive or state-of-the-art performance on a number of body pose datasets.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Mohammed Talal Ghazal

Department of Computer Engineering Technology, Northern Technical University

Mosul, Iraq

Email: mohammed.ghazal@ntu.edu.iq

1. INTRODUCTION

Recently, human pose estimation based on computer vision has been employed in understanding human activities from videos or images [1], [2]. Researchers in this field developed many efficient models in terms of speed and detection poses of multiple persons in the same scene. The most famous of these models are PoseNet [3], OpenPose [4], and MoveNet lightning [5]. Pose estimation uses key points and body joints, such as the elbows and wrists, to locate and track human posture automatically, in addition to determining the orientation of body limbs [6]. This task has a lot of potential in many applications, which include tracking human body movement or analyzing and detecting inappropriate human behavior [7]–[9]. Pose estimation is used in sports performance analysis to track and evaluate human movement accuracy, as well as in a variety of other fields like human-computer interaction (HCI) and augmented reality (AR). Several types of research on the two-dimensional (2D) and three-dimensional (3D) pose estimation tasks have been presented; some focus on recognizing a person's pose in video or image, while others work on locating every joint for multiple persons [10], [11]. In the present paper, the researchers have focused on the 2D pose estimation approach. According to the deep learning method, the 2D single-person pose estimation approaches have accomplished the best performance in comparison with the pose estimation of multiple-person, particularly in complicated

scenes with significant occlusions. While the 3D human pose estimation field has some limitations, represented by the challenge of recovering the 3D ground-truth information of human joint placement, especially in open-air space, and as a result, there aren't many large datasets that have the 3D human pose annotated. The majority of 3D datasets are collected via motion capture (MoCap) systems in a lab environment. Because there are few differences in background, viewpoint, and illumination in such environments, the existing models can not be used in general within unconstrained conditions.

Pose estimation techniques can be divided into single-person approaches and multi-person approaches. Images taken during sports activities may contain multiple persons in the same scene [12]. Many single-person techniques use a human detection model for cropping the region, which contains a single person, to estimate the pose. However, determining a human posture might be counted as either detection or a regression issue. Techniques based on regression attempt to map the acquired image directly to body key point positions. While the methods based on detection usually try to detect key points individually by generating key point heatmaps after that combining them in the next processing stages to produce the final pose prediction. There are various difficulties when attempting to infer each person's posture in an image, especially when the scene is crowded. Actually, an image could include a variable amount of people in it. Additionally, interactions between people lead to several prediction problems, mostly involving occlusions, which complicates the detection of joint and body part association. Moreover, the processing time duration increases significantly with the number of persons in the image, directly affecting the system's performance. As a result, all these elements must be considered to obtain a high accuracy rating during inference. Two different methods can be used to deal with multi-person pose estimation. The first technique is the top-down method, which uses a human detection model to identify the area which only includes one person after that, applies a pose estimation model to its output. While the bottom-up method is the second technique, this involves first identifying each key point, then connecting the resulting joints according to the person's instance.

To achieve the goals of this paper, three main tasks must be accomplished: human pose estimation, recognizing human action, and finding the similarity between poses. The proposed method involves recovering the orientations and locations of a person's body parts from images and videos. The OpenPose library has been used to obtain high-accuracy pose estimation; then, a supervised machine learning model is applied to classify activity using the pose information. Finally, we matched the key point sets of different persons by applying the cosine formula to find the similarity of human poses. Section 2 of this paper introduces a survey of related works, while section 3 explains the strategy of the proposed algorithm to discover human pose similarity and action recognition. The experimental work and results are in section 4, and finally, in section 5, the conclusion and future works are presented.

2. RELATED WORKS

The purpose of 2D pose estimation is to predict the spatial location or 2D position of human body joints. Deep learning techniques have significantly improved this field. The previous works are categorized into single-person and multi-person-based pose estimation; detecting the pose for one person is usually easier than detecting the pose for multiple people. Both categories experience interesting development. In this section, we go over some current research field approaches. The first study to suggest using deep learning to capture the full-body context completely was deep pose [13]. The pose estimation problem in this study was considered a regression issue toward the joints of the body. According to Sun *et al.* [14] researchers proposed compositional pose regression as a structure-aware method. The suggested method uses a new representation of posture that employs bones rather than joints because they are more stable. To set a composition loss function that represents the long-term interactions between body bones, the network thus uses a joint connection structure. To enhance the pose estimator's capabilities for key points that are obscured in complex or crowded environments, [15] made a good multi-scale structure-aware method that combines information scheme (I.S.), multi-scale supervision (M.S.), multi-scale feature combination network (MFC-Net), structure-aware loss (S.L.), and a key point masking training technique to address the shortcomings of the latest hourglass model's version. It was suggested in [16] to use ResNet instead of hourglass blocks in a novel cascade feature aggregation (CFA) method. The combination of features extracted from various stages yields information about both local and global contexts, making CFA more resistant to changes like lighting and partial occlusions. The methods mentioned in [13]–[16] are related to single-person pose estimation. To infer a multi-person pose, [17] introduced a novel framework based on region to guarantee an accurate pose prediction, even with the inaccurate human bounding boxes prediction. The system primarily consists of three parts: The first part is a symmetric spatial transformer network (SSTN), followed by a parametric pose non-maximum suppression (NMS) approach, and finally, a novel posture distance measure to assess pose matching and remove those that are redundant. A bottom-up multi-person pose estimator called OpenPose has been proposed in [18]. This algorithm uses part affinity fields (PAFs) for key points association, in which

the position and orientation of human limbs are encoded. During the experiments, the proposed method demonstrated its efficacy in identifying poses accurately, even with crowded people in the image or video. Several researchers have begun to recognize human activity in images or videos. Guerra *et al.* [19] detect activities of daily living by preprocessing collected data from the Microsoft Kinect motion-sensing device to minimize systematic error. Reserach by Reily *et al.* [20] proposed a new method to recognize human activity by simultaneous feature extraction from human posture and the activity's objects.

3. METHOD

Our strategy for discovering human pose similarity and action recognition consists of three sequential tasks: starting with pose estimation from images based on the OpenPose framework, then classifying the actions using obtained pose key points as input and the support vector machine (SVM) classification algorithm [21], [22] and finally matching the poses between the input image and reference image to find the similarity between them based on cosine distance metrics. The OpenPose library is a human body recognition project built with the Caffe framework and based on convolutional neural networks (CNN). It has the capability of detecting different human 2D poses for single or multiple people. The OpenPose architecture is shown in Figure 1. The framework extracts the image's original feature map with the VGG-19 deep neural network as a backbone and splits it into two branch inputs. The first part employs CNN for human body joints heat map prediction, while the second employs CNN to obtain the partial affinity fields (PAF) for every connected joint. The PAF is represented by a 2D vector that stores the limb's orientation and position. At each stage, the PAF mapping with the input feature layer and the key point heat map is taken into consideration and is known as S^t and L^t . For the input layer, excluding the first layer, represents the feature layer generated by the model of the VGG-19 network [23]. The acquired heat map and PAFs from the prediction stages give the position information and the direction vector of all connected joint points. The Hungarian greedy optimization algorithm has been used for optimizing and matching the limbs.

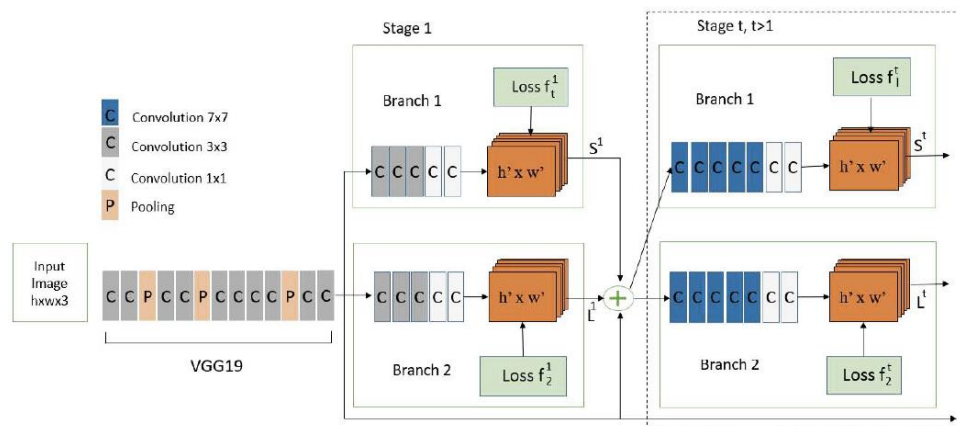


Figure 1. OpenPose architecture [2]

Figure 2 illustrates the OpenPose algorithm pipeline: firstly, the features map is extracted from the input image by the VGG-19 model, as shown in Figure 2(a). These features are then split into two sections and fed to the CNN for confidence (S) and affinity vector (L) prediction of the key points for all points, as shown in Figures 2(b) and 2(c) respectively. For each key point, both S and L are analyzed to see how the key points are grouped together by the greedy reasoning algorithm, as shown in Figure 2(d). Finally, the skeleton's assembly in Figure 2(e).

The analysis of human body movement is primarily accomplished by annotating key points in the human body and changing those key points. The COCO key points dataset has been used in OpenPose firmware to estimate the human pose; the dataset consists of about 200 k images, each labeled with key points. The annotation of 18 human body points is used in this work. The reference diagram of human pose key points is shown in Figure 3. The activity categories proposed in this paper are divided into four groups: 'sitting', 'standing', 'jumping', and 'running'. The change of key point position leads to a change in the human action categories. To determine the human action class, we addressed the issue of activity classification as a multi-class classification approach that can be modeled using several classification

algorithms. The 18 key points of the human body are mapped to the classification algorithm for training and testing to predict the action class.

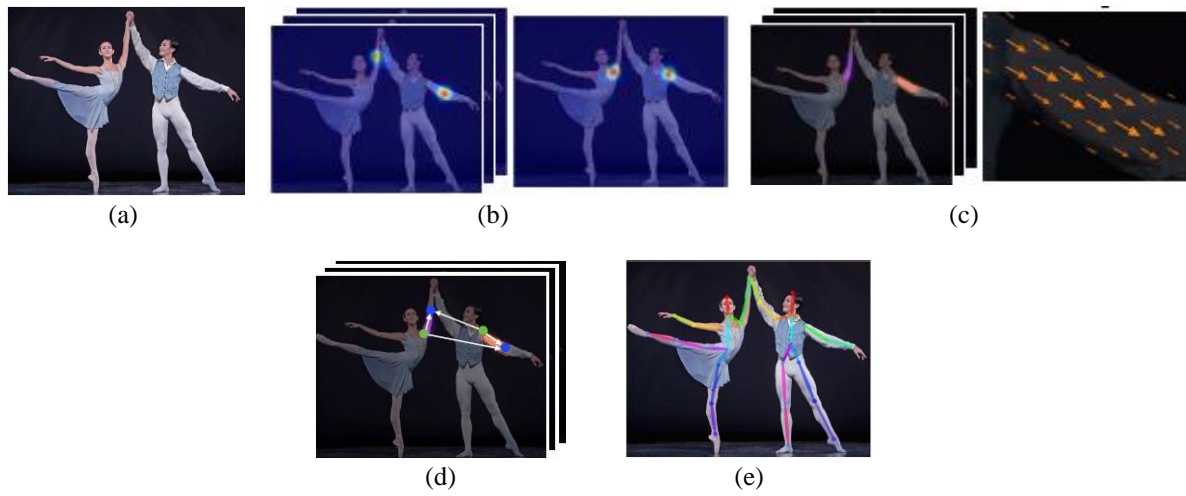


Figure 2. The OpenPose algorithm pipeline (a) input image, (b) part confidence maps, (c) part affinity fields, (d) bipartite matching, and (e) parsing results [2]

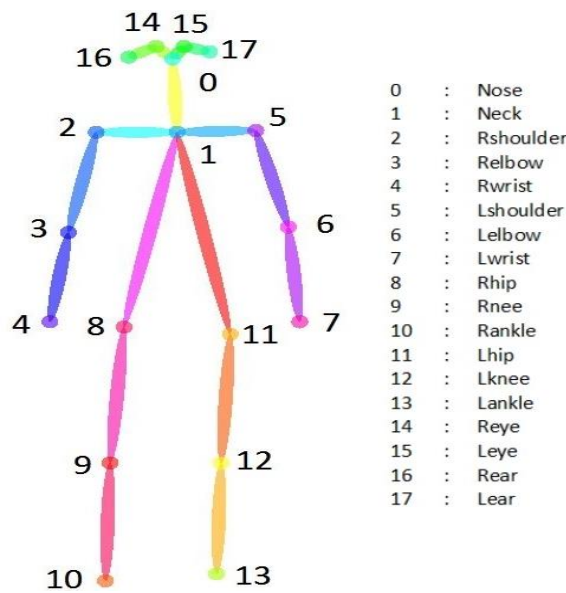


Figure 3. The human key points reference diagram

Many challenges must be considered in order to find similarities between multiple people; these challenges are related to differences in people's physiques and the pose estimator's error margins [24], [25]. The people should have the same poses in tall, skinny, and heavy to match each other. As a result, we cannot describe a pose using absolute distances between points. This problem can be fixed by cropping and scaling the pose vectors to align them on top of one another. To find the similarity between the poses in the two images, we must detect the human body's key points and then compare them. The reference image would be placed on the test image to determine the degree of similarity between the poses depicted in both images. The pose vectors for both images are normalized before calculating the distance [26]. Then, the cosine distance formula, shown in (1), is applied to calculate the angle between the two vectors and subtracts one from it.

$$\text{Cosine Distance}(\text{pose1}, \text{pose2}) = 1 - \frac{\text{pose1} \cdot \text{pose2}}{\|\text{pose1}\| \|\text{pose2}\|} \tag{1}$$

4. EXPERIMENTAL WORK AND RESULTS

Several tools and libraries have been utilized in the experimental work for the proposed work, which includes: The OpenPose (multi-person key point detector) generates 18 key points of the human body to be used as a reference point for human action recognition. The key points obtained are the (X) and (Y) coordinates and their confidence value of: 0-nose, 1-neck, 2-RShoulder, 3-RElbow, 4-RWrist, 5-LShoulder, 6-LElbow, 7-LWrist, 8-RHip, 9-RKnee, 10-Rankle, 11-LHip, 12-LKnee, 13-LAnkle, 14-Reye, 15-LEye, 16-Rear, and 17-Lear, as mentioned in Figure 3 above. Tensorflow (Google's deep learning library) was used to create the action recognition classifier model. Other libraries integrated into Tensorflow were also used during development, including Keras, Numpy, and Panda. The model has been developed under Jupyter Notebook (an interactive Python development environment) with Python version 3 (scripting language). The experiments were executed on a computer with Ubuntu 22.04 64-bits, Intel (R) Core (TM) i7-1165G7 (11th Gen) CPU, 32 GB RAM, and one NVIDIA GeForce MX450 GPU.

For the action classification task, 1,000 images have been collected for each class, using COCO human pose dataset with Google image search. To achieve the best results for training and testing, the selected image could identify all of the 18 body key points using OpenPose firmware. For each image, the key points normalized to 128*128 pixels. The key points information must be saved into a comma separated value (CSV) file for classification purposes, which include: key points and action labels. The SVM classifier model, which is a supervised machine learning that is used for solving classification problems, has been employed to classify the four categories of human body actions. Experimentally, the radial basis function (RBF) kernel is used to train the SVM classifier, which is built in the scikit-learn python library; the model results in considerable accuracy in predicting the four classes of human action. Figure 4, shows the confusion matrix of the experimental results, where the classification accuracy reached to 87%.

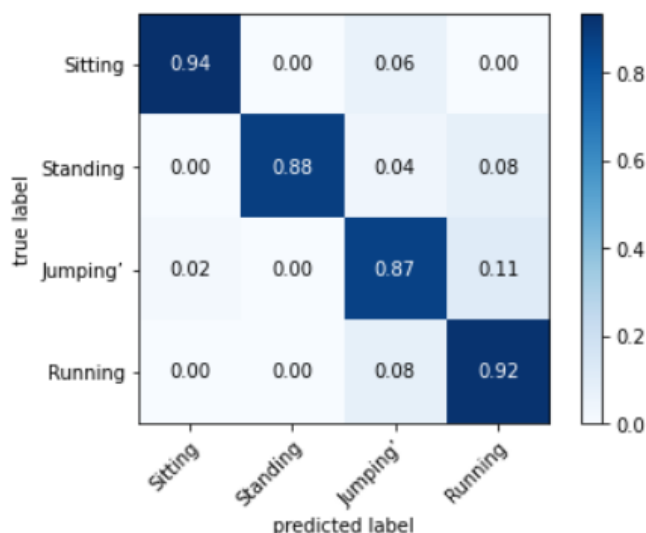


Figure 4. Classification results of each pose

To make sure of pose similarity between the reference image and the target image for different body poses, a similarity score is calculated. The person's pose is considered sufficiently correct if the similarity score reaches a specific threshold based on the cosine similarity comparison method. The coordinates from two images are centered, scaled, flattened to a single vector, and normalized before the comparison. This is performed to ensure that the coordinates are compared using the same criteria. The sklearn library in python is used to calculate cosine similarity. Figure 5, shows some examples of pose similarity between different images. The confidence score for each key point in the two pose images has been calculated, which represents the likeliness that the key point exists in the calculated spot. In the first example, the returned results of yoga exercise is very similar in the two images which is shown in Figure 5(a). In the second example, the returned results of gymnastic exercise look very similar too, while the final example return a less similar degree because the exercise in the two images is not similar as shown in Figures 5(b) and (c) respectively.

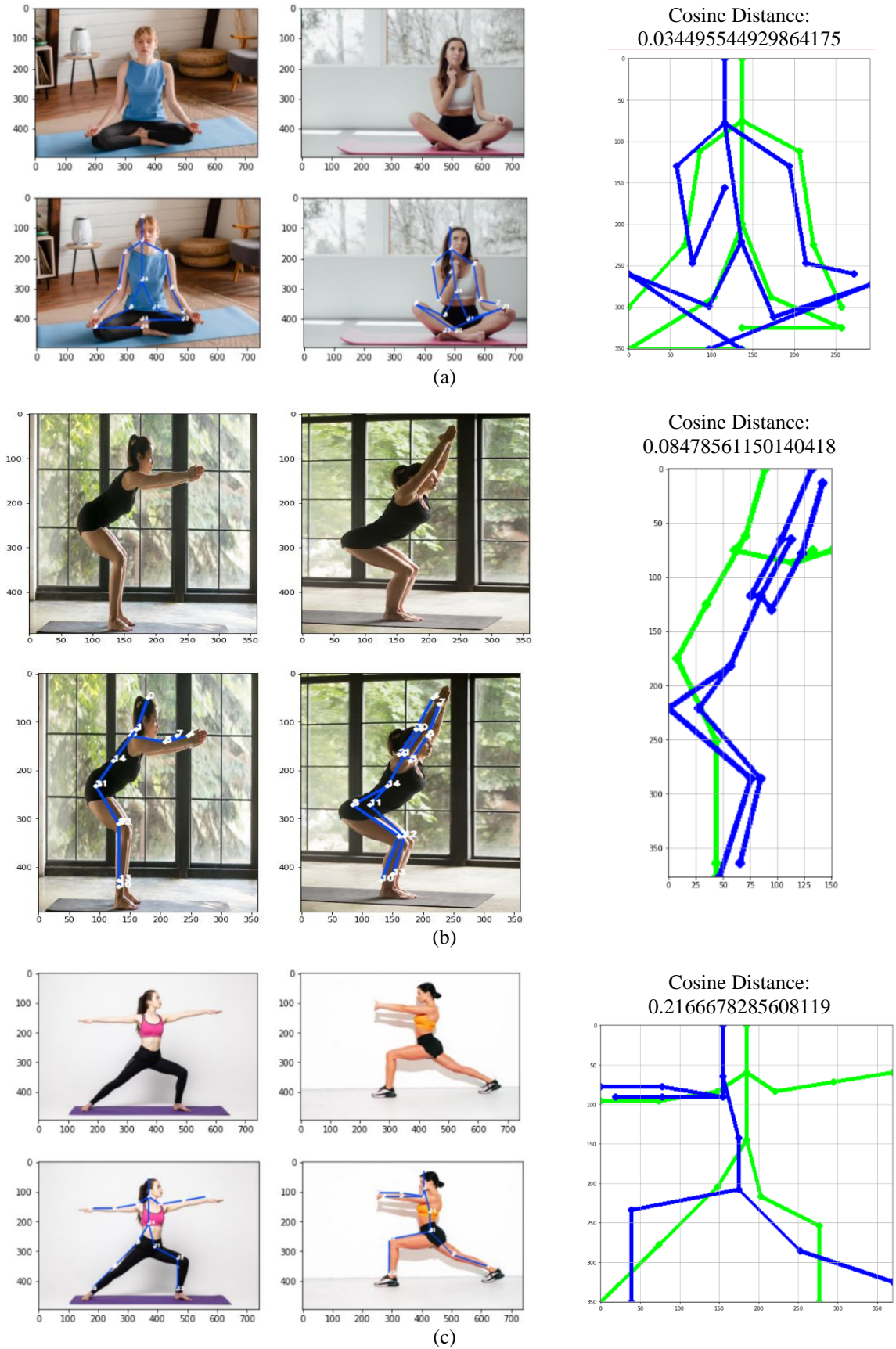


Figure 5. Examples of pose similarity calculation based on cosine distance (a) yoga exercise similarity check, (b) gymnastic exercise similarity check, and (c) gymnastic exercise similarity check

5. CONCLUSION

In this study, we presented our method for extracting pose from images using the OpenPose application programming interface (API), after that, using the extracted pose information for activities classification with the aid of a supervised machine learning algorithm. We created a dataset for this study that includes four activities: sitting, standing, jumping, and running. To enhance our model's performance, we used SVM algorithms. The results of our experiment revealed that SVM has an accuracy of 87%. The detected poses are then mapped to the pose similarity checker algorithm to find the matching between input and reference images, which is useful in checking whether the sports exercises are performed correctly or not. For future work, we will work on classifying more human actions and try different pose-matching algorithms to enhance detection accuracy.

ACKNOWLEDGEMENTS

The current research is supported by a collaboration between Northern Technical University NTU and Mosul University in Iraq. The website of NTU is <https://www.ntu.edu.iq> and Mosul University is <https://www.uomosul.edu.iq>.





REFERENCES

- [1] M. Ghazal, R. Albasrawi, N. Waisi, and M. Al Hammoshi, "Smart meeting attendance checking based on a multi-biometric recognition system," *Przeglad Elektrotechniczny*, vol. 98, no. 3, pp. 93–96, 2022, doi: 10.15199/48.2022.03.21.
- [2] N. Y. Abdullah, M. T. Ghazal, and N. Waisi, "Pedestrian age estimation based on deep learning," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 22, no. 3, pp. 1548–1555, Jun. 2021, doi: 10.11591/ijeecs.v22.i3.pp1548-1555.
- [3] A. Kendall, M. Grimes, and R. Cipolla, "PoseNet: A convolutional network for real-time 6-dof camera relocalization," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015, pp. 2938–2946, doi: 10.1109/ICCV.2015.336.
- [4] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using Part Affinity Fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 172–186, Jan. 2021, doi: 10.1109/TPAMI.2019.2929257.
- [5] B. Jo and S. Kim, "Comparative analysis of OpenPose, PoseNet, and MoveNet models for pose estimation in mobile devices," *Traitement du Signal*, vol. 39, no. 1, pp. 119–124, Feb. 2022, doi: 10.18280/ts.390111.
- [6] A. W. R. Emanuel, P. Mudjihartono, and J. A. M. Nugraha, "Snapshot-based human action recognition using OpenPose and deep learning," *Snapshot-Based Human Action Recognition using OpenPose and Deep Learning*, vol. 48, no. 4, pp. 2–8, 2021.
- [7] P. N. Huu, N. N. Thi, and T. P. Ngoc, "Proposing posture recognition system combining MobilenetV2 and LSTM for medical surveillance," *IEEE Access*, vol. 10, pp. 1839–1849, 2022, doi: 10.1109/ACCESS.2021.3138778.
- [8] L. Song, G. Yu, J. Yuan, and Z. Liu, "Human pose estimation and its application to action recognition: A survey," *Journal of Visual Communication and Image Representation*, vol. 76, pp. 1–12, Apr. 2021, doi: 10.1016/j.jvcir.2021.103055.
- [9] V. K. Kambala and H. Jonnadula, "A multi-task learning based hybrid prediction algorithm for privacy preserving human activity recognition framework," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 6, pp. 3191–3201, Dec. 2021, doi: 10.11591/eei.v10i6.3204.
- [10] Z. Shu, P. Wang, and W. Zhan, "The research and implementation of human posture recognition algorithm via OpenPose," in *2020 2nd International Conference on Artificial Intelligence and Advanced Manufacture (AIAM)*, Oct. 2020, pp. 90–94, doi: 10.1109/AIAM50918.2020.00023.
- [11] M. B. Gamra and M. A. Akhloufi, "A review of deep learning techniques for 2D and 3D human pose estimation," *Image and Vision Computing*, vol. 114, pp. 1–23, Oct. 2021, doi: 10.1016/j.imavis.2021.104282.
- [12] J. G. da S. Neto, J. M. X. N. Teixeira, and V. Teichrieb, "Analyzing embedded pose estimation solutions for human behaviour understanding," in *Anais Estendidos do XXII Simpósio de Realidade Virtual e Aumentada*, Nov. 2020, pp. 30–34, doi: 10.5753/svr_estendido.2020.12951.
- [13] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2014, pp. 1653–1660, doi: 10.1109/CVPR.2014.214.
- [14] X. Sun, J. Shang, S. Liang, and Y. Wei, "Compositional human pose regression," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 2621–2630, doi: 10.1109/ICCV.2017.284.
- [15] L. Ke, M.-C. Chang, H. Qi, and S. Lyu, "Multi-scale structure-aware networ for human pose estimation," in *Proceedings of the european conference on computer vision (ECCV)*, 2018, pp. 731–728, doi: 10.1007/978-3-030-01216-8_44.
- [16] Z. Su, M. Ye, G. Zhang, L. Dai, and J. Sheng, "Cascade feature aggregation for human pose estimation," *arXiv preprint*, Feb. 2019, [Online]. Available: <http://arxiv.org/abs/1902.07837>
- [17] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 2353–2362, doi: 10.1109/ICCV.2017.256.
- [18] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 7291–7299, doi: 10.1109/CVPR.2017.143.
- [19] B. M. V. Guerra, S. Ramat, R. Gandolfi, G. Beltrami, and M. Schmid, "Skeleton data pre-processing for human pose recognition using neural network," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Jul. 2020, pp. 4265–4268, doi: 10.1109/EMBC44109.2020.9175588.
- [20] B. Reily, Q. Zhu, C. Reardon, and H. Zhang, "Simultaneous learning from human pose and object cues for real-time activity recognition," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, May 2020, pp. 8006–8012, doi: 10.1109/ICRA40945.2020.9196632.
- [21] E. A. Mahareek, A. S. Desuky, and H. A. El-Zhni, "Simulated annealing for SVM parameters optimization in student's performance prediction," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 3, pp. 1211–1219, Jun. 2021, doi: 10.11591/eei.v10i3.2855.





- [22] N. S. A. Yasmin, N. A. Wahab, A. N. Anuar, and M. Bob, "Performance comparison of SVM and ANN for aerobic granular sludge," *Bulletin of Electrical Engineering and Informatics*, vol. 8, no. 4, pp. 1392–1401, Dec. 2019, doi: 10.11591/eei.v8i4.1605.
- [23] W. Setiawan, M. I. Utoyo, and R. Rulaningtyas, "Reconfiguration layers of convolutional neural network for fundus patches classification," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 1, pp. 383–389, Feb. 2021, doi: 10.11591/eei.v10i1.1974.
- [24] G. Mori *et al.*, "Pose embeddings: A deep architecture for learning to match human poses," Jul. 2015, [Online]. Available: <http://arxiv.org/abs/1507.00302>
- [25] J. Wilms, G. Beckers, T. Callemin, L. Geurts, and T. Goedem', "Human pose matching," in *Proceedings of the Dortmund International Research Conference*, 2018, pp. 72–76.
- [26] A. Gupta, K. Gupta, K. Gupta, and K. Gupta, "Human activity recognition using pose estimation and machine learning algorithm," in *ISIC'21: International Semantic Intelligence Conference*, 2021, pp. 323–330.

BIOGRAPHIES OF AUTHORS







Mohammed Moath Abdulghani     Obtained his M.Sc. degree from Computer Information Technology, Universiti Kebangsaan Malaysia UKM, Malaysia in 2016. His M.Sc. thesis entitled: "An Optimized Feature Set Based on Genetic Algorithm for Business Web Pages Named Entity Recognition" and his current research focuses on the development of human pose estimation algorithm. He can be contacted at email: albakri2@uomosul.edu.iq.



Mohammed Talal Ghazal     Obtained his M.Sc. degree from Computer Engineering Technology, Northern Technical University, Mosul, Iraq in 2016. His M.Sc. thesis entitled: "Wheelchair Robot Control Using EOG signals" and his current research focuses on the development of human pose estimation algorithm. He can be contacted at email: mohammed.ghazal@ntu.edu.iq.



Anmar Burhan M. Salih     Obtained his Ph.D. degree from Department of Electrical and Computer Engineering, Florida Institute of Technology, Florida, USA in 2020. His Ph.D thesis entitled "Cloud Computing Service Interoperability and Architectural Concepts". His current research focuses on development of human pose estimation algorithm. He can be contacted at email: anmar.salih@ntu.edu.iq.