Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/31/20230131

Drum dance accuracy detector: A CNN-based pose estimation framework for detecting inaccurate movements in Miao drum dance

Rouhan Qian[1,3], Weihong WU[2,4]

[1]The Madeira School 8328 Georgetown Pike, McLean, Virginia 22102
[2]University of California, Riverside, USA

[3]shmilyqian0929@outlook.com
[4]embarkwu@gmail.com

Abstract. As an intangible cultural heritage derived from the Miao culture, drum dance faces the crisis of losing its inheritance. The main cause behind the issue is the lack of teaching resources and distribution of the drum dance. This work is dedicated to resolving this issue with the usage of the recent advances in artificial intelligence. More specifically, the solution of Incorrect Movement Detection (IMD) is invented, where the core technique lies in the CNN-based pose estimation algorithm. IMD enables the real-time detection of the inaccurate dance movements, which realizes the possibility of self-learning of drum dance. The effectiveness of IMD is already validated by its application in a field research project at a Miao village in Xiangxi, Hunan, China. The proposed system may have a significant impact on the preservation and promotion of Miao culture.

Keywords: Drum Dance Accuracy Detector, CNN, Incorrect Movement Detection, Transmit the Culture.

1. Introduction

Figure 1. Two female drum dancers performing Miao drum dance with a group of men wearing traditional Miao costumes standing behind [1].

DOI: 10.54254/2755-2721/31/20230131

Drum dance is a type of folk dance in Miao culture, derived from and distributed in the regions like the western part of Hunan and the southeastern part of Guizhou. In a piece of drum dance, people play the drums to make rhythm and dance with the drum beats meanwhile. For example, drum dancers sometimes pretend the drum sticks as rice and make the movements similar to transplanting the rice in a field. Unlike the dance genres that are created for ornamental objectives, drum dance contains huge cultural significance. The dancers are not only doing simple movements for aesthetics. Instead, the dance movements are recording and narrating the daily life of Miao people. As an intangible cultural heritage, drum dance is a significant representation of the Miao culture and a tool to transmit the culture from generations to generations.

Nevertheless, drum dance currently faces a huge difficulty: it almost lost its inheritance. As the younger generations move out of the village, less people would be willing to put in the effort and spend a long time on learning this cultural performing art. Additionally, the aesthetic standard transfers, so less people in the young generation of dancers can appreciate the beauty of drum dance, which causes the Miao dance to not be included in the syllabus of the universities with dance majors. All the factors together lead to one potential terrible consequence, which is the loss of inheritance of an intangible cultural heritage. Therefore, it is emergent to solve the problem and promote the drum dance to be continually transmitted to the future generations.

One solution for the problem would be reducing the required consumption of effort and time for learning the drum dance. Specifically, when people know the basic movements for the drum dance, a system of automatic correction, almost functioning as a drum dance teacher, is the key to enable them to improve the movements' accuracy. To fulfill this goal, the main idea proposed in this paper is the usage of deep learning. Deep learning [2, 3] is developed with the progress in the field of artificial intelligence, and the advent of deep learning makes this target possible. With deep learning-based pose estimation algorithm [4, 5, 6, 7], the human poses can be accurately estimated directly using a camera, without the need of attaching a specific device , such as sensors, on the human body while dancing. This usage brings the learners the benefit of saving the costs while maintaining the learning efficiency. As shown in Figure 2, the pose estimation system contains the input of an image and the output of the representation of human pose using a coordinate system. Every human body joint is illustrated by a 3-dimensional coordinate, as well as the connections among the joints, which can clearly distinguish the individual from other objects. Given the accurate pose obtained from the pose estimation system, it is now possible to detect the nonstand movement with the Incorrect Movement Detection (IMD) system proposed in this paper. The IMD system transforms the criteria of the drum dance into a computer program written in Python. For instance, when the arm position of a drum dancer is at a wrong angle, the IMD system can detect and mark the deviation.

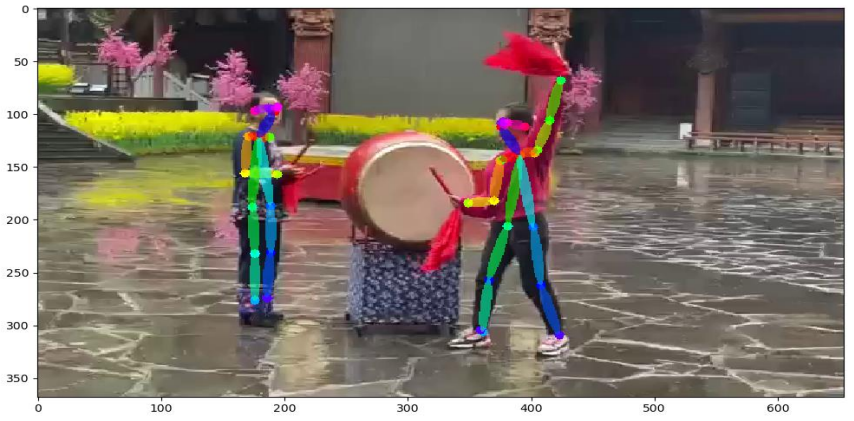

Figure 2. The picture is the product of the pose estimation system that shows the obtained pose. The 18 circles represent the major joints of the body parts, and the lines that connect the circles represent the human skeletons.

DOI: 10.54254/2755-2721/31/20230131

The pose of the drum dancers can be estimated through two deep learning algorithm categories, which are the top-down approaches [8, 9] and the bottom-up approaches [10]. Because top-down approaches contain the defects of error accumulation and resource consumption, this research is conducted based on the bottom-up approach, which can estimate human poses more accurately than the top-down approaches.

Moreover, the two currently existing approaches mentioned above are mainly designed for static image data with standard illumination, yet both face a special challenge when applied to video data: the lighting condition in a video can be unstable. Under this circumstance, the result of pose estimation is incomplete or inaccurate in certain frames with a bad lighting condition. Based on the observation of the video's continuity, which means that the neighboring frames are similar to each other, this paper proposes a solution to the challenge of lighting: Auto-Completion Algorithm. The missing estimation for poses in some specific frames can be completed by using the interpolation between the information in neighboring frames. Therefore, this solution enhances the consistency and robustness of consecutive predictions.

Figure 3. This picture visualizes the angles automatically calculated by the system. 1) Distance between two arms. 2) Angle between the upper arm and lower arm. 3) Angle between the upper leg and lower leg.

The criteria for the automatic correction would be split into three sections. The first one would be the timing for the light squad. Since a movement that continues along the entire choreography for drum dance is the rhythmic light squad, it is important to evaluate if the squad corresponds with the rhythm of the drum beat. For example, it can be determined with the angle between the thigh and the calf, and the angle needs to change in a fixed period and pattern. The second section would be the position of the arm. The arms are required to be fully extended with power in the drum dance, so it is essential to have the arms reach sufficient energy to make the dance look aesthetic. In order to have the arms located at the accurate position, the judgment can be done by measuring both the distance between the two hands and the angle between the upper arm and the lower arm. The optimal moment for the measurement would be at either the starting or ending position of a set of movements, in which the movements are repeated for several rounds before getting into the next set, because there is usually a brief pause between each set in a choreography, and it would be more easily identified. The third criterion can be the area of the drum being exposed in front of the camera. Different positions would result in the dancer blocking the different percent of the area of the drum, and measuring the percent of exposure of the drum can determine if the dancer is standing at the correct position at a specific moment of the dance. Because the position changes among each set of movement in one choreography, the criteria for drum exposure also

DOI: 10.54254/2755-2721/31/20230131

needs to change based on the switch of the set of movements. Basically, the criteria of the system are settled down based on the known knowledge about the drum dance.

According to the Drum Dance Accuracy Detector (DDAD) in this paper, we would achieve multiple contributions as follows:

1. Open new possibilities for promoting the drum dance and preserving cultural identity. The DDAD continues the inheritance of the drum dance by broadening the access of learning drum dance to a wider group with less limitation on teaching resources, which would preserve the heritage of Miao culture.
2. Propose a framework or a solution to robustly estimate human pose, based on which people can evaluate the accuracy of dance movement through the angle between thigh and calf, upper arm and lower arm, and two arms without human guidance.
3. Extend the existing OpenPose algorithm to video data, and solves the problem of inconsistent estimation among video frames via our proposed approach, termed pose estimation stabilization.

2. Related work

2.1. Applications of movement detection

People have developed various types of approaches for movement monitoring in the field of sports. One popular method that people adopt is wearable performance devices [11, 12] The sensors in the wearable devices enable the athletes and their physicians to monitor real-time physiologic and movement parameters, which is very helpful for the detection of position-specific patterns, invention of efficient sports-specific training programs, and analysis of potential cause of injury. The monitor of functional movements, workloads, and biometric markers are beneficial for the optimization of athletes' performance and reduction in the injuries to the largest degree [13]. Another method that people utilize for movement detection for sport is machine and deep learning [14], which is more similar to the method used in this work. Through the input of an inertial measurement unit (IMU) [15] and computer vision data, the deep learning system can process the information to function with automated detection and sport-specific movement recognition. The method is beneficial because of its competence over manual performance analysis methods.

In the medical field, automatic pose and body part detections are also significant, and the detections are conducted through the technique of thermal image [16]. The detection process can be divided into multiple steps. First, only the pixels located inside the body with the usage of Otsu's thresholding approach were subjected to the histogram equalization (HE) method. Then, feature extraction adopts DarkNet-19 architecture, while feature selection uses the approaches of principal component analysis (PCA) [17] and t-distributed stochastic neighbor embedding (t-SNE) [18]. The final step is the examination of the performance of the classification methods that contain various machine learning algorithms, with a result of high accuracy in upper vs. lower body parts, back vs. front of upper body, and back vs. front classification of lower body parts. The approach would bring enhancement to the automatization process of thermal images [16].

2.2. Pose estimation methods

To address the issue of pose estimation, there are methods of two categories of leading deep learning algorithms to choose from. The first category is the top-down approach [8, 9], which breaks down the solution process into two stages. In the first stage, the system relies on an object detection algorithm to detect and mark the position of each human individual with bounding boxes. Then, within each bounding box, the system labels the exact position of human body joints. Despite its effectiveness, some disadvantages exist in this top-down approach. First, it is hard to recover from incorrect or inaccurate object detection, which means the potential mistake in bounding an incorrect object during the first stage. That is to say the error will be transferred into the second stage and influence the accuracy of the pose estimation. Secondly, the process of objection detection is complicated and resource consuming. In comparison, the bottom-up approach [10] is more efficient by solving the problem in an end-to-end manner. The approach uses one stage to directly estimate the position of each human body joint and the

DOI: 10.54254/2755-2721/31/20230131

joint connection of each distinct individual. This approach avoids the extra step of objection detection and its corresponding error. Especially in the cases where blockage among human bodies and objects exists, the accuracy of bounding boxes in the top-down approach can be severely affected, whereas it barely affects the accuracy of the bottom-up approach. Therefore, for more accurate estimation, the proposed framework adopts the bottom-up approach.

2.3. Convolutional neural networks

Convolutional neural networks (CNN) [19] is mainly utilized for feature extraction from the data with convolution structures. The unique characteristic of CNN is that the network does not require manual feature extraction, and its architecture is inspired by visual perception [20]. Moreover, CNN contains multiple kernels, and each kernel functions as a receptor that responds to a specific feature. There are also two functions of CNN [21]: activation function and loss function. The activation function sets up a threshold for the neural electric signals to exceed in order to reach the next neuron. The loss function helps the users to train the entire CNN system to learn about and reach a result of their expectations.

There are multiple advantages [21, 22] of CNN compared to other networks. First, CNN has local connections that improve its effectiveness. Unlike other networks that have each neuron connected to all the neurons in the previous layer, the neurons in CNN are connected to a smaller group of neurons in the previous layer, which would be more effective by decreasing parameters and accelerating convergence. Second, further reduction in parameters can be realized through the weight sharing feature of CNN, which means that the same weight is shared between a group of connections. Third, CNN is advantageous because of its feature of downsampling dimension reduction. Basically, according to the principle of image local correlation, a pooling layer can downsample an image, which brings the benefits of the reduction in data without loss of useful information and the reduction in parameters via trivial feature removal.

Generally, because of the advantages that convolutional neural networks contain, CNN achieves remarkable results in a variety of areas, such as object detection [23], image classification [24], medical diagnosis [25], action recognition [26], face recognition [27], and more other fields. In this work, we successfully apply it to the task of action recognition, under the context of the action of drum dance movements.

3. Background

3.1. Convolutional neural network

Convolutional neural networks (CNNs) consist of three types of layers: convolutional layers, pooling layers and fully-connected layers [24]. A CNN architecture would be established when the three layers are piled up together.

As mentioned above, the core component of the CNNs is convolutional layers. It equation is defined as follows [19]:

$$conv(I,K)_{x,y} = \sum_{i=1}^{n_H} \sum_{j=1}^{n_W} \sum_{k=1}^{n_C} K_{i,j,k} I_{x+i-1,y+j-1,k} \quad (1)$$

in which I symbolizes image and K is the kernel or convolutional filter. The variables x and y represent the horizontal and vertical position of the output activation in the feature map, while the variables i, j, and k are the weight index of the convolutional filter.

The convolution operation relies on applying learnable kernels. Even though these kernels usually exist as the sliding windows that are spatially small, they fully cover the depth of the input. The weighted sum of the kernel weight and the input value in the current window can be calculated as the sum of the scalar product. The calculation process is shown in Figure 4. Each kernel captures a specific type of feature, and the kernel will react (or the output will be large) only when the specific feature appears in the current window. This is called as acticications.

DOI: 10.54254/2755-2721/31/20230131

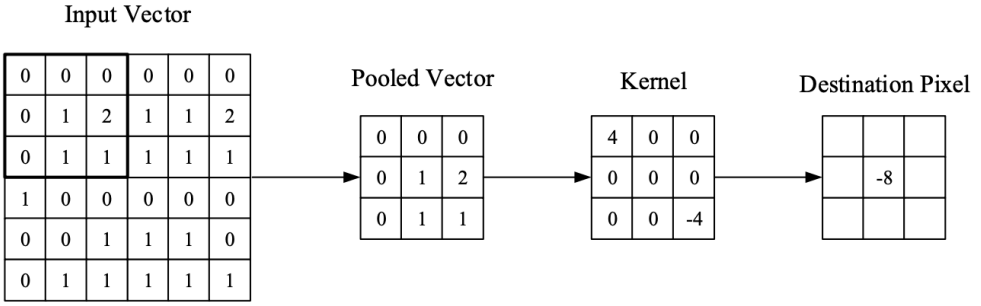

Figure 4. The graph visually illustrates a convolutional layer. The boden frame in the input vector is a representation of the location of the sliding window. The destination pixel displays the weighted sum of the scalar product between pooled vector and kernel [28].

3.2. Pose estimation

The common procedure of the bottom-up approach contains two steps. First, an image (I) is fed into a pre-trained convolutional network for feature extraction, such as VGG-19. Then, the extracted feature maps (F) functions as an input to the second convolutional network to generate the output of an estimation of the pose (S):

$$F = CNN_{pretrainedfeatureextractor}(I)$$
$$S = CNN_{poseesti,ator}(F) \quad (2)$$

4. Method

4.1. Single frame pose estimation

RDP code (Row-Diagonal Parity) is a binary error correction code specially designed for RAID 6. Its main application is to correct the error in RAID 6 system. It has the advantage of its low computational complexity, which makes it advantageous in scenarios with high performance requirements. However, the RDP code can only correct two errors, which limits its fault tolerance [9].

4.1.1. Transformation of binary keypoint annotations into heatmap labels. To evaluate the predicted pose S, we transform the annotated two dimensional keypoints into the groundtruth confidence map S*. Each confidence map represents the probability of each specific body part being located in any given pixel in the two dimensional coordinates. In an ideal situation, for a single individual in the image, each body part should correspond with a confidence map with the existence of a single peak. Conversely, for an image with multiple people, each visible part j for each person k should have a peak in the confidence map.

To begin with, we create individual confidence maps $S^*_{j,k}$ for each individual k. First, $x_{j,k} \in R^2$ can be set as the groundtruth position of body part j for person k in the image. The value of the confidence map $S^*_{j,k}$ at any given point $p \in R^2$ is defined as follows:

$$S_{j.k}(p) = exp((-\|p - x_{j.k}\|_2)^2/\sigma^2)$$
$$C*_j(p) = \max_k S*_{jk}(p) \quad (3)$$

in which σ determines the concentration of the distribution. Through a max operator, accumulating the individual confidence maps can then allow the network to predict the groundtruth confidence map.

DOI: 10.54254/2755-2721/31/20230131

4.1.2. Training the pose estimator. Based on the transformed heatmap labels, the loss between the estimated predictions and the groundtruth maps and fields can be denoted as L. There is an issue that some datasets are not capable of labeling every person in the image, and weighing the loss functions spatially can solve the problem. The equation form of the loss function of the confidence map is written as:

$$L = \sum_{j=1}^{J} \sum_{p} W(p) \cdot ||S_j(p) - S_j^*(p)||_2^2 \quad (4)$$

L stands for the groundtruth part confidence map, and W represents a binary mask. If the annotation is missing at the pixel p, then it can be illustrated as W(p) = 0. One main function of the mask is to prevent the penalization for the true positive predictions (the correct predictions) during training.

4.2. Neighboring frame completion

Because of the varying illumination and other disturbances during the filming process, some moments would not be able to be completely recorded, and certain frames would be incomplete or lost. One method to address the issue is using the neighboring frame as a reference and substitute to complete the lost frame.

For instance, at the t^{th} frame, if $C_{jt} = 0$, and C_{jt+1} or C_{jt-1} do not equal to 0, this means the particular joint on the t^{th} frame is missing. Therefore, we should interplate the missed frame using the average coordinates of the previous frame and the subsequent frame.

$$C^t{}_j = 1/2(C_J{}^{t-1} + C_J{}^{t+1}) \quad (5)$$

We apply this method iteratively until every missing point is replaced by the approximation.

5. Experiment and analysis

5.1. COCO people challenge dataset

The COCO [30] training set is composed of the person samples of over the quantity one hundred thousand. In the training set, the instances overall are labeled with more than one million keypoints, which can symbolize the mark for body parts. The testing set is divided into three subsets, which are "test-challenge," "test-dev," and "test-standard," with each sub-section containing approximately twenty thousands images as samples. Besides, the object key-point similarity (OKS) is defined through the COCO evaluation that has the main competition metric of using the mean average precision (AP) larger than the thresholds of 10 OKS. In the object detection process, the functions and roles of OKS and the IoU are identical. Calculation of the two values, the scale of the person and the distance between predicted points and GT points, can thus obtain the result of OKS.

5.2. Comparison of computation efficiency

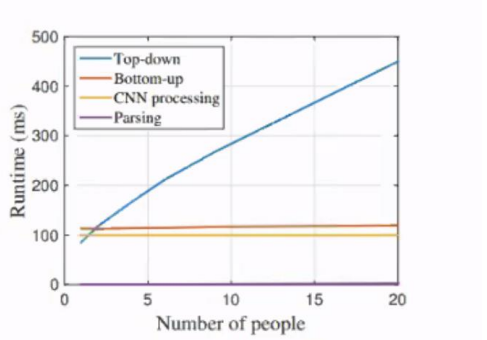

Figure 5. According to the graph, for the bottom-up model, with the increased number of people, the runtime is steady. However, for the top-down method, as the number of people increases, the runtime also increases [10].

DOI: 10.54254/2755-2721/31/20230131

The graph illustrates the runtime performance of the OpenPose method. The experimental data is obtained from the videos that have different numbers of people in each of them. The blue lines show the performance of the top-down method with the usage of person detection and single-person CPM, in which the runtime increases rapidly with the number of people in the image. Conversely, as the number of people increases, the runtime of the bottom-up approach increases substantially slower than that of the former approach. Nevertheless, since the CNN processing time is two orders of magnitude more than the parsing time, the latter does not affect the overall runtime substantially.

5.3. Comparison to existing approaches

Table 1. The table shows the comparison of the results of the test-challenge subset between the open-pose method and other approaches on the COCO 2016 keypoint challenge. AP stands for the average precision of the methods. AP50 corresponds to the condition of OKS equals to 0.5, and APL represents the value for large scale persons [10].

Team	AP	AP50	AP75	APM	APL
OpenPose	60.5	83.4	66.4	55.1	68.1
G-RMI	59.8	81.0	65.1	56.7	66.7
DL-61	53.3	75.1	48.5	55.5	54.8
R4D	49.7	74.3	54.5	45.6	55.6

As shown in Table 1, the OpenPose method over-competes the other currently existing methods in all cases except for APM. Even though the runtime of the OpenPose method is relatively low, it presents a great performance in the AP values. This advantage over the other methods can be attributed to the high efficiency of the CNN backbone, as well as the end-to-end bottom-up framework.

6. Body angles calculation and comparison

For real-life application, the CNN technique is applied to the DDAD system for establishing and identifying the body skeleton of the drum dancer. With the body skeleton obtained and visualized, DDAD also needs to realize the function of detecting inaccurate poses, in which a standard to evaluate accurate poses and differentiate them from inaccurate ones is required. The standard is chosen to be the angles among different fields of the body skeleton, such as the angle between the upper and lower arm (both left and right arms), the angle between the upper and lower leg (both left and right legs), and the angle between left and right arms. A number of standard videos with diverse tempos and movements are recorded in the system, so the users can choose and imitate one of the standard videos to film their test video and upload it to DDAD. DDAD enables the calculation of the body angles in every frame of the video, which would form a set of data. With the settled data set of a standard video, in which the dancer performs standard drum dance movements, DDAD can compare it with the data set of a test video, in which the dancer may perform inaccurate dance movements. In the same frame of the two videos' data set, if the difference between the angle of the same part of the body skeleton exceeds a certain threshold, the frame of the test video would be detected as inaccurate. Just as Figure 7 illustrates, DDAD identifies and shows the frames with inaccurate poses via a visualized form.

DOI: 10.54254/2755-2721/31/20230131

Figure 6. This picture shows one frame of the test video. Since the angle between upper and lower left leg exceeds the threshold, DDAD identifies it as an inaccurate pose, which is shown in the lower left corner.

7. Conclusion

In this work, the Drum Dance Accuracy Detector is developed. I adopt the pose estimation approach on the foundation of convolutional neural networks, which successfully completes the task of identifying inaccurate poses. The inaccuracy can be informed through the calculation of body angles and the comparison between the standard video and a test video. However, to further improve and perfect the system, future improvements can be implemented in creating a more comprehensive standard of movement accuracy. If the drumbeats can be integrated into the measurement of accuracy, the fixed timing for an angle to reach a specific degree can enhance the comprehensiveness and quality of the requirements for accuracy, resulting in more precise detection of inaccurate movements.

References

[1] "The national intangible cultural heritage of Xiangxi Miao drum dance," Sina, http://k.sina.com.cn/article_7064782218_p1a518058a0010117d2.html#p=1 (accessed Jul. 12, 2023).

[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," nature, vol. 521, no. 7553, pp. 436–444, 2015.

[3] I. Goodfellow, Y. Bengio, and A. Courville, Deep learning. MIT press, 2016.

[4] X. Qian et al., "Pose-normalized image generation for person re-identification," presented at the Proceedings of the European conference on computer vision (ECCV), 2018, pp. 650–667.

[5] H. Joo, T. Simon, and Y. Sheikh, "Total capture: A 3d deformation model for tracking faces, hands, and bodies," presented at the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8320–8329.

[6] L.-Y. Gui, K. Zhang, Y.-X. Wang, X. Liang, J. M. Moura, and M. Veloso, "Teaching robots to predict human motion," presented at the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2018, pp. 562–567.

[7] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros, "Everybody dance now," presented at the Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 5933–5942.

[8] L. Pishchulin, A. Jain, M. Andriluka, T. Thormählen, and B. Schiele, "Articulated people detection and pose estimation: Reshaping the future," presented at the 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 3178–3185.

DOI: 10.54254/2755-2721/31/20230131

[9] U. Iqbal and J. Gall, "Multi-person pose estimation with local joint-to-person associations," presented at the Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II 14, Springer, 2016, pp. 627–642.
[10] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," presented at the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 7291–7299.
[11] Z. Zhang, "Microsoft kinect sensor and its effect," IEEE multimedia, vol. 19, no. 2, pp. 4–10, 2012.
[12] L. Yang, L. Zhang, H. Dong, A. Alelaiwi, and A. El Saddik, "Evaluating and improving the depth accuracy of Kinect for Windows v2," IEEE Sensors Journal, vol. 15, no. 8, pp. 4275–4285, 2015.
[13] R. T. Li, S. R. Kling, M. J. Salata, S. A. Cupp, J. Sheehan, and J. E. Voos, "Wearable performance devices in sports medicine," Sports health, vol. 8, no. 1, pp. 74–78, 2016.
[14] E. E. Cust, A. J. Sweeting, K. Ball, and S. Robertson, "Machine and deep learning for sport-specific movement recognition: A systematic review of model development and performance," Journal of sports sciences, vol. 37, no. 5, pp. 568–600, 2019.
[15] F. Höflinger, J. Müller, R. Zhang, L. M. Reindl, and W. Burgard, "A wireless micro inertial measurement unit (IMU)," IEEE Transactions on instrumentation and measurement, vol. 62, no. 9, pp. 2583–2595, 2013.
[16] A. Özdil and B. Yılmaz, "Automatic body part and pose detection in medical infrared thermal images," Quantitative InfraRed Thermography Journal, vol. 19, no. 4, pp. 223–238, 2022.
[17] S. Karamizadeh, S. M. Abdullah, A. A. Manaf, M. Zamani, and A. Hooman, "An overview of principal component analysis," Journal of Signal and Information Processing, vol. 4, no. 3B, p. 173, 2013.
[18] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE.," Journal of machine learning research, vol. 9, no. 11, 2008.
[19] Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series," The handbook of brain theory and neural networks, vol. 3361, no. 10, p. 1995, 1995.
[20] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," The Journal of physiology, vol. 160, no. 1, p. 106, 1962.
[21] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: analysis, applications, and prospects," IEEE transactions on neural networks and learning systems, 2021.
[22] M. Jogin, M. Madhulika, G. Divya, R. Meghana, and S. Apoorva, "Feature extraction using convolution neural networks (CNN) and deep learning," presented at the 2018 3rd IEEE international conference on recent trends in electronics, information & communication technology (RTEICT), IEEE, 2018, pp. 2319–2323.
[23] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," Advances in neural information processing systems, vol. 28, 2015.
[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Advances in neural information processing systems, vol. 25, 2012.
[25] S. Dabeer, M. M. Khan, and S. Islam, "Cancer diagnosis in histopathological image: CNN based approach," Informatics in Medicine Unlocked, vol. 16, p. 100231, 2019.
[26] G. Yao, T. Lei, and J. Zhong, "A review of convolutional-neural-network-based action recognition," Pattern Recognition Letters, vol. 118, pp. 14–22, 2019.
[27] M. Coşkun, A. Uçar, Ö. Yildirim, and Y. Demir, "Face recognition based on convolutional neural network," presented at the 2017 International Conference on Modern Electrical and Energy Systems (MEES), IEEE, 2017, pp. 376–379.
[28] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," arXiv preprint arXiv:1511.08458, 2015.
[29] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," arXiv 1409.1556, Sep. 2014.

DOI: 10.54254/2755-2721/31/20230131

[30] T.-Y. Lin et al., “Microsoft COCO: Common Objects in Context,” May 2014.