Journal of Statistics & Management Systems
ISSN 0972-0510 (Print), ISSN 2169-0014 (Online)
DOI : 10.47974/JSMS-1565

Posture recognition in exercise frames via HOBV-based deep sparse learning

Shivani Singhai *
Department of Computer Application
Rabindranath Tagore University
Raisen 462026
Madhya Pradesh
India

Pratima Gautam †
Department of Computer Science and Application
Rabindranath Tagore University
Raisen 462026
Madhya Pradesh
India

Jitendra Singh Kushwah §
Department of Computer Science and Engineering
Institute of Technology and Management
Gwalior 474001
Madhya Pradesh
India

Abstract

In the age of machine learning, Human Exercise Recognition (HER) is an area that has been extensively researched. When it comes to computer vision, action recognition is the process of classifying a human exercise that is seen in a video as belonging to one of a selection of prepared actions. This paper proposes a new method for recognizing physical activity using a Hybrid Object Boundary Value Deep Sparse Auxiliary Network (HOBV-

* *ORCID-ID:* 0009-0000-5092-8067
† *ORCID-ID:* 0000-0003-0401-2335
§ *ORCID-ID:* 0000-0001-5102-2160

* *E-mail:* shivanisinghai5@gmail.com (Corresponding Author)
† *E-mail:* pratima_shkl@yahoo.com
§ *E-mail:* jitendra.singhkushwah@itmgoi.in

DSAN). The technique combines object boundary value characteristics with deep sparse learning frameworks to improve the depiction of complicated human motions. Our model captures the global structure as well as the fine-grained motion features of physical activities by integrating boundary-aware representations with a sparsity-driven deep auxiliary network. Experimental results show that the proposed technique outperforms traditional models such as YOLO, Alpha Pose, and Deep Pose on many performance criteria such as Accuracy (95.33%), Sensitivity (99.05%), Specificity (97.47%), and Precision (94.57%). The findings demonstrate the HOBV-DSAN model's durability and accuracy in detecting a broad range of workouts, making it a suitable option for intelligent physical activity monitoring systems.

Subject Classification: *68T07.*

Keywords: *Exercise recognition, Sparse coefficient, Gabor filter, Deep learning.*

I. Introduction

Movies, in contrast to photographs, which only give an appearance cue, comprise both the motion cue acquired from consecutive frames as well as the appearance cue that is generated from individual frames. Therefore, the motion cue is an essential component in the Exercise Recognition. Nevertheless, the accuracy of modern Exercise Recognition algorithms is still at a lower level than that of human beings. The process of Exercise detection is plagued by a number of challenges, including as the presence of dynamic background clutter, changes in camera viewpoint, and fluctuations in light.

More than fifteen years of study have been conducted on the subject of Physical Exercise Recognition (PExR), which is a key issue in the field of computer vision. It has been shown that several deep learning algorithms have been presented up to this point; however, none of them have demonstrated excellent performance in movies that include multiple activities or in situations where precisely recognizing the exact activity is difficult. For instance, human exercise posture estimation permits higher-level reasoning in the context of human-computer interaction and activity recognition, while PExR may be carried out at a variety of different levels of abstraction. "Crunches" and "pushups" are examples of PExR, which is a more complex physical movement that is composed of a large number of action primitives (gestures) that are structured in a chronologically chronological order. Although years of work have been put into it, posture estimation continues to be a challenging and mostly unsolved problem.

The development of wearable sensors that are able to objectively measure the motion and dynamics of the body has been accomplished [1]- [4]. Despite this, the invasive nature of these technologies makes them unsuitable for application in real-world situations and situations that occur often. To a large extent, they are analogous to conventional neural networks, which are made up of neurons that have weights and biases that have been preset. On the other hand, neural networks have a limited capacity to scale when dealing with larger

pictures. In recent times, Deep Learning (DL) algorithms have included the processes of feature extraction and classification.

One of the most important and challenging areas of study is known as Physical Exercise Recognition (PExR). When it comes to action recognition, one of the most significant issues we have is the fact that the same action may be performed in a number of different ways, even by the same person. Therefore, the most significant challenge is to develop an adequate representation of activities that is both discriminative (that is, it enables us to discern between distinct actions) and universal (that is, it enables us to group videos that include the same actions performed by a variety of humans who appear and move). The paper is organized into several sections for clarity and coherence.

The paper is systematized into several sections for clarity and coherence, Section second discussed previous work related in this area. Section third describes the datasets used for in our experiments. Section fourth gives a detail description of proposed method. The experimental results along with a discussions are presented in Section fifth. Finally, Section sixth concludes the paper.

II. Literature Survey

The purpose of this section is to provide a concise summary of the historical work that has been done by a large number of researchers in the field of human exercies identification, specifically eye-based human exercies recognition.

Meng et al. [5] introduced AdaFuse, an Adaptive Temporal Fusion Network, to enable vigorous temporal modelling. The AdaFuse network merges information from earlier feature maps with the current refined ones to create a richer set of temporal features, which helps enhance the accuracy of activity detection. Li et al. [6], MediaPipe was employed to identify basic fitness exercise. The fitness action recognition process involves three main steps, with the first step separating upward and downward movement states into individual image samples. In the next step, the K-Nearest neighbors (KN) algorithm is used to classify the different fitness based on the prepared training dataset. Kumar and Sinha [7] present an Android application for yoga posture estimation that employs the OpenPose model for key landmark identification. The suggested pipeline employs a collection of yoga poses that are preprocessed using OpenPose and saved on the local computer, and then the posture is determined by comparing the real-time recorded event to the local machine data. Taware et al. [8] employed occlusion-simulating augmentation to evaluate the posture detector under strong occlusions, which differs from standard testing. Before implementing user alignment, the estimator identifies the user's 33 important spots. It employs regression and heatmaps. Kanase et al. [9] describe a posture estimation approach based on the pre-trained OpenPose model, which is a multistage CNN model. This approach identifies crucial spots in videos. Anilkumar et al. [10] employed Mediapipe's BlazePose model to determine body joint coordinates. Coordinates are analyzed and compared to yoga position

data. Chen et al. [11] used deep transfer learning to develop a fitness database. A neural network was taught to recognise fitness motions. Yolov4 and Mediapipe were employed by researchers to identify brief fitness movements. They used the fitness signal from each movement to create the database. MLP was utilized to categories the fitness signal waveforms. Pauzi, et al. [12] track body movement and overlay annotated skeletal joints onto a video. Using the PoseNet dataset and the Mediapipe Blazepose algorithm, this method applies deep learning to sports and physically demanding vocations. Fang et al. [13] proposed AlphaPose, a technique that improves posture prediction even with imperfect human bounding boxes using a Regional Multi-person posture prediction (RMPE) framework.

However, many current models are primarily concerned with recognizing key spots or bounding boxes, leaving out important information about the body's border forms and finer structural details. Furthermore, deep learning models sometimes suffer from overfitting or need enormous volumes of annotated data, limiting their usefulness in real-world, varied exercise conditions. To address these issues, this study presents a unique technique called the Hybrid Object Boundary Value Deep Sparse Auxiliary Network (HOBV-DSAN). By combining object boundary value extraction with a deep sparse auxiliary learning framework, the model can detect both high-level posture patterns and minor motion changes. The introduction of boundary values improves contour and form representation, whilst sparsity-based auxiliary networks increase learning efficiency and concentrate on critical characteristics. This hybrid architecture enables more robust and accurate identification of a broad range of physical workouts. Extensive trials show that the suggested technique surpasses current methods such as YOLO, Alpha Pose, and Deep Pose, with higher accuracy, sensitivity, Specificity, and precision.

III. Methodology

Over the past few years, many new techniques have emerged to make holistic feature extraction more effective. Among these methods, the use of Gabor representations has proven to be particularly effective for image processing. HD Features, an advanced form of holistic descriptors, leverage this approach. A Gabor filter (or wavelet) functions as a band-pass filter whose response is shaped by combining a sinusoidal function with a Gaussian envelope. This combination results in a two-dimensional filter that captures specific spatial frequencies and orientations, making it highly effective for analyzing image textures and structures.

Gabor features are widely recognized for their strong performance in feature representation. However, most methods focus on the magnitude of the Gabor coefficients rather than the phase, as magnitude-based features tend to yield better results in practical applications. As such, in feature extraction tasks, the magnitude is usually considered more valuable. This technique is effective in capturing information in both spatial and frequency domains, making it a powerful tool for visual pattern recognition.

We developed a 2D odd-symmetric Gabor filter I our method, described by the following equation:

$$HD(F)_{\theta_k, f_i, \sigma_x, \sigma_y}(x, y) = exp\left(-\left[\frac{x_{\theta_k}^2}{\sigma_x^2} + \frac{y_{\theta_k}^2}{\sigma_x^2}\right]\right).\cos(2\pi f_i x_{\theta_k} + \varphi)$$

This 2D Gabor filter is applied on all dataset to extract Holistic Features.

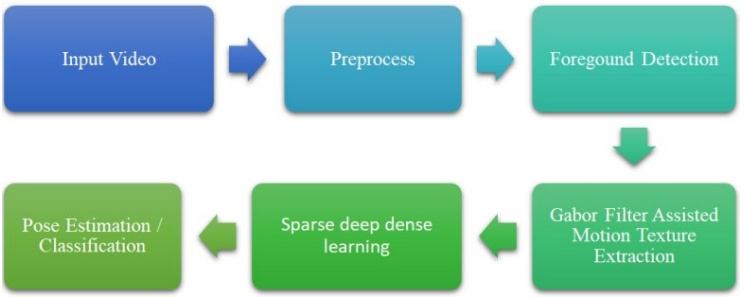

Figure 1
Proposed Approach

Figure 1 is shows proposed approach of my research. Due to their strength in identifying spatial ad frequency features across various directions and scales, Gabor filters are extensively used for analyzing textures and detecting edges. Gabor filters, in their standard form, can be costly in terms of computational and effectiveness in some applications. The Sparse Gabor Filter addresses these issues by applying a sparse demonstration of the filter's response, effectively reducing the computational burden while maintaining high-quality feature extraction. SGFs are particularly effective in detecting fine-grained designs and ends in hand positions and movements, even under challenging conditions like varying lighting, occlusions, or background clutter. By capturing the multi-scale spatial and frequency components of hand positions, the Sparse Gabor Filters boosts the system's capability to distinguish amongst subtle gesture differences and improve the robustness of recognition algorithms. The Sparse Gabor filter efficiently extracts discriminative spatial and frequency features from hand gestures, capturing important details while reducing computational complexity. These features are then fed into the Random Forest classifier, which leverages its ensemble of decision trees to classify the gestures. The combination allows for highly accurate recognition, as the Random Forest model effectively handles the high-dimensional feature space produced by the SGF, while also being robust to variations such as noise, lighting changes, and background clutter. This approach benefits from both the rich texture information captured by the Sparse Gabor features and the ensemble strength of the Random Forest, making it well-suited for real-time gesture recognition tasks.

Mini-batch stochastic gradient descent is used to minimize the error function below,

$$E_{sgd} = \frac{1}{M}\sum_{n=1}^{M}\left\|\widehat{S_n}(T_{n-\tau}^{n+\tau}, W, b) - S_n\right\|_2^2$$

E_{sgd} is the mean squared error.

$\widehat{S_n}(T_{n-\tau}^{n+\tau}, W, b)$ is estimated signal frame

S_n is reference normalized frame at index n

M is Mini batch size

(W, b) indicating the trainable parameters of weight and bias.

The revised prediction of W^{ι} and b^{ι} in the i^{th} layer, at a certain learning rate, may be repeatedly calculated as follows:

$$\Delta(W_{n+1}^{\iota}, b_{n+1}^{\iota}) = -\lambda\frac{\partial E_{sgd}}{\partial(W_n^{\iota}, b_n^{\iota})} - \kappa\lambda(W_n^{\iota}, b_n^{\iota}) + \omega\Delta(W_n^{\iota}, b_n^{\iota}), 1 \leq \iota \leq L+1$$

Where L shows how many layers are present under the surfaces and $L+1$ show the last output layer. κ is the weight decay coefficient? And ω is the momentum. If given enough training samples, network may automatically learn the complex connection required to isolate speech from the noisy and reverberant sounds.

This sparse approach leads to more efficient processing in real-time video analysis, making it well-suited for applications such as gesture-based control in interactive systems, sign language translation, and human-robot interaction.

IV. Expeiment & Results

Proposed model is implemented using Python on Laptop with Intel core i3 2.40GHz CPUs, having 16GB. To minimize computational cost, the input image dimension is reduced to 512*512 by bilinear interpolation.

Performance Analysis:

In order to assess and analyses the categorization effects of each model more precisely, this article makes use of a confusion matrix, accuracy, sensitivity, specificity, and positive prediction rate.

Accuracy represents the rate at which the true condition of the pose is detected; specificity represents the rate at which different pose are distinguished; sensitivity represents the rate at which the absence of a particular pose is detected; and the positive predication rate represents the rate at which accurate positive identifications are made. An overview of the method's performance evaluation is provided below.

Proposed model (HOBV-DSAN) achieve F1 score (97%), Recall (98%), precision (94%), and accuracy (95%). The outcomes are broken down by technique for all three groups below.

$$\text{ACCURACY} = \frac{Correct\ Prediction[\text{TP} + \text{TN}]}{Total\ Cases\ [P + N]}$$

$$\text{SENSITIVITY} = True\frac{Positives[\text{TP}]}{Total\ Actual\ Positives\ [P]}$$

$$\text{SPECIFICITY } = [\text{TN}]\ /\ [\ \text{N}\ ]$$

$$PRECISION = [\text{TP}]\ /\ [\ \text{TP} + \text{FP}]$$

From figure we can find out that the proposed method performs better in recognition of sign than others. This can be reflected by the improvement of all the criteria. The reason is probably that the frequency resolution of proposed method is more adaptive to classify action from other action. That is to say this method is a better feature representation than traditional method in recognition of gesture.

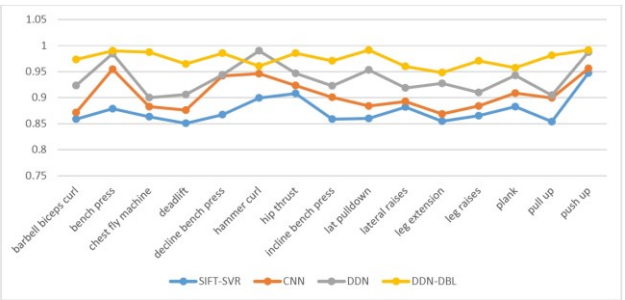

Figure 2
Class- wise Accuracy Comparison

Proposed model gives superior results compared to other classifiers, which indicates that this will provide a very good recognition. Figure 2 compares the performance of four different methods — Proposed, YOLO [14], Alpha Pose [13], and Deep Pose [15]— based on four key evaluation metrics: Accuracy, Sensitivity, Specificity, and Precision. It can be observed that Among the three classifiers, the proposed approach demonstrates more than 96% accuracy in recognition, as depicted in the figure.

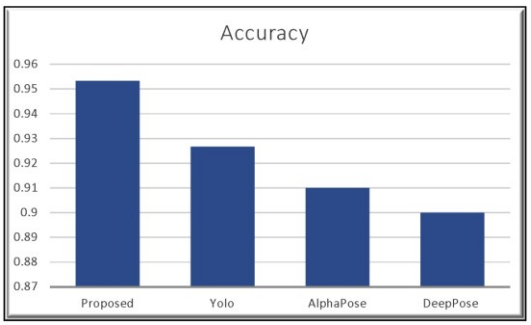

Figure 3
Comparison of Accuracy

Figure 3 illustrates the comparative accuracy performance of various models used for exercise posture recognition, including YOLO, Alpha Pose, Deep Pose, and the proposed HOBV-DSAN model. The bar graph clearly demonstrates that the HOBV-DSAN model outperforms the existing state-of-the-art methods in terms of classification accuracy. While YOLO, Alpha Pose, and Deep Pose show competitive results, their accuracy levels fall short when compared to the 95.33% achieved by HOBV-DSAN.

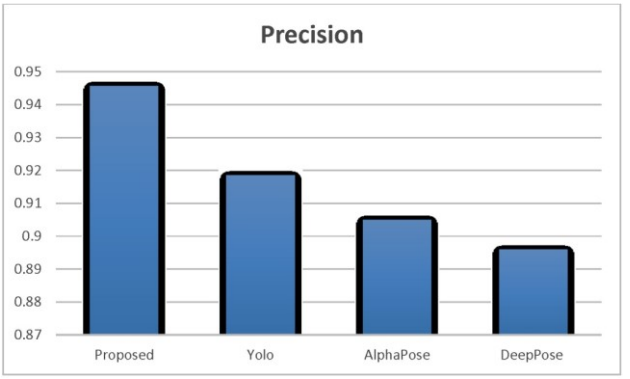

Figure 4
Comparison of Precision

Figure 4 presents a comparative analysis of the precision achieved by various models used for posture recognition in exercise videos, including YOLO, Alpha Pose, Deep Pose, and the proposed HOBV-DSAN model. Precision focuses on that proportion which are correctly identified positive instances in all the positive instances classification, is a critical metric in assessing the reliability of recognition systems. As shown in the figure, the HOBV-DSAN model achieves the highest precision score of 94.57%, significantly outperforming the other models.

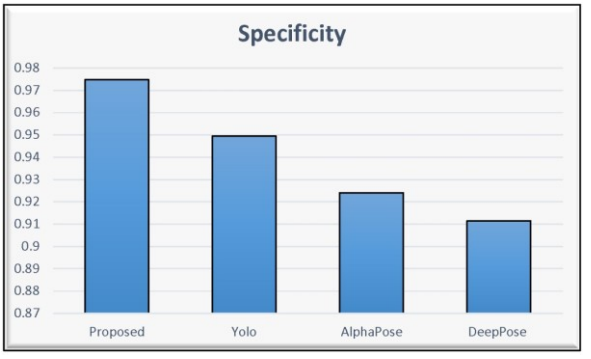

Figure 5
Comparison of Specificity

Figure 5 illustrates the specificity performance of different models—YOLO, Alpha Pose, Deep Pose, and the proposed HOBV-DSAN—used for recognizing exercise postures from video frames. Specificity refers to the model's ability to correctly identify negative instances, ensuring that non-target postures are not mistakenly classified as relevant. As depicted in the figure, the HOBV-DSAN model achieves a specificity of 97.47%, outperforming the other models by a significant margin. This high specificity indicates the model's strong capacity to minimize false positives, making it particularly effective in distinguishing between correct and incorrect or irrelevant movements. The superior performance can be attributed to the integration of object boundary value analysis with deep sparse auxiliary learning, which enhances the model's discriminative power. Overall, Figure 5 confirms the robustness of HOBV-DSAN in delivering accurate posture classification with minimal misidentification.

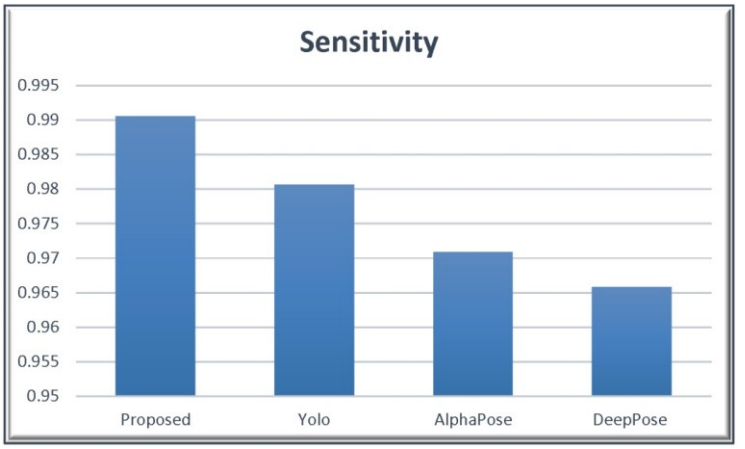

Figure 6
Comparison of Sensitivity

Figure 6 displays the sensitivity (or recall) performance of various models—YOLO, Alpha Pose, Deep Pose, and the proposed HOBV-DSAN—for exercise posture recognition. Sensitivity measures the model's ability to correctly identify true positive cases, reflecting how well it detects actual exercise postures without missing relevant instances. As shown in the figure, the HOBV-DSAN model achieves the highest sensitivity score of 99.05%, significantly outperforming the other compared methods. This exceptional result indicates that HOBV-DSAN is highly effective at capturing nearly all correct postures in a video sequence, which is crucial in applications such as fitness monitoring and rehabilitation where missing a correct posture could lead to inaccurate feedback. The model's advanced hybrid architecture contributes to this performance by enhancing feature extraction and motion understanding.

Overall, the figure emphasizes HOBV-DSAN's strength in maintaining high recall accuracy under varied conditions and complex movements.

It's clear that the proposed framework offers several advantages compared to traditional methods of the conventional methods. Therefore, our proposed framework is presented for Exercie classification using a PExR-net to improve the accuracy of current machine learning models in pose estimation. The **Proposed method** (HOBV-DSAN) demonstrates superior performance across all metrics, achieving the highest Accuracy (0.9533), Sensitivity (0.9905), Specificity (0.9747), and Precision (0.9457). **YOLO [14]** comes second, with strong performance but slightly lower values, particularly in Accuracy (0.9267) and Precision (0.9186). **Alpha Pose [13]** shows moderately good results, with an Accuracy of 0.91 and Precision of 0.9050, indicating reliable but less optimal detection compared to the Proposed method and YOLO. Finally, **Deep Pose [15]** records the lowest values among the four, with an Accuracy of 0.9 and a Precision of 0.8959, suggesting that while it still performs reasonably well, it lags behind the other approaches. Overall, the Proposed method clearly outperforms the others, making it the most effective technique according to the presented metrics.

V. Conclusion

The most important addition that this thesis makes is the ability to identify human Exercies in videos by using deep learning algorithms while taking into consideration a wide variety of challenges. This paper proposes a new method for recognizing physical activity using a Hybrid Object Boundary Value Deep Sparse Auxiliary Network (HOBV-DSAN). Experimental results show that the proposed technique outperforms traditional models such as YOLO, Alpha Pose, and Deep Pose on many performance criteria such as Accuracy (95.33%), Sensitivity (99.05%), Specificity (97.47%), and Precision (94.57%). These results show that HOBV-DSAN is both effective and reliable when handling the challenges and unpredictability of physical exercises.

The improved recognition capability can be attributed to the hybrid architecture that combines object boundary analysis with deep sparse auxiliary networks, allowing the model to extract and interpret fine-grained features from video sequences. This ensures precise posture recognition even in complex and dynamic exercise scenarios. The model's robustness to variations in movement and background further establishes its ability to support in real-life applications like fitness monitoring, rehabilitation, and sports analytics.

Used in daily life for things like tracking workouts, helping people recover from injuries, and analyzing sports performance.

References

[1] S. S. Intille and L. Bao, "Physical activity recognition from acceleration data under semi-naturalistic conditions," *Technical Report* (2003).

[2] H. Zhang, L. Li, W. Jia, J. D. Fernstrom, R. J. Sclabassi, and M. Sun, "Recognizing physical activity from ego-motion of a camera," in Proc. IEEE Int. Conf. Eng. Med. Biol. Soc. (EMBS), pp. 5569–5572 (2010), doi: 10.1109/IEMBS.2010.5626794.

[3] E. M. Tapia, S. S. Intille, W. Haskell, K. Larson1, J. Wright, A. King, R. Friedman, "Real-time recognition of physical activities and their intensities using wireless accelerometers and a heart monitor," in Proc. Int. Symp. Wearable Computers, pp. 37–40 (2007).

[4] L. Li, H. Zhang, W. Jia, J. Nie, W. Zhang, and M. Sun, "Automatic video analysis and motion estimation for physical activity classification," in Proc. IEEE 36th Annu. Northeast Bioeng. Conf., pp. 1–2 (2010).

[5] Y. Meng, R. Panda, C. C. Lin, P. Sattigeri, L. Karlinsky, K. Saenko, and R. Feris, "AdaFuse: Adaptive temporal fusion network for efficient action recognition," arXiv preprint, arXiv:2102.05775 (2021).

[6] X. Li, M. Zhang, J. Gu, and Z. Zhang, "Fitness action counting based on MediaPipe," in Proc. 15th Int. Congr. Image Signal Process., Biomed. Eng. Informatics (CISP-BMEI), Beijing, China, pp. 1–7 (2022), doi: 10.1109/CISPBMEI56279.2022.9980337.

[7] D. Kumar and A. Sinha, Yoga Pose Detection and Classification Using Deep Learning. Saarbrücken, Germany: LAP Lambert Academic Publishing (2020).

[8] G. Taware, R. Kharat, P. Dhende, P. Jondhalekar, and R. Agrawal, "AI-based workout assistant and fitness guide," in Proc. 6th Int. Conf. Comput., Commun., Control Autom. (ICCUBEA), Pune, India, pp. 1–4 (2022), doi: 10.1109/ICCUBEA54992.2022.10010733.

[9] R. R. Kanase, A. N. Kumavat, R. D. Sinalkar, and S. Somani, "Pose estimation and correcting exercise posture," in ITM Web Conf., vol. 40, pp. 03031, EDP Sciences (2021).

[10] A. Anilkumar, K. T. Ardra, S. Athulya, K. A. Sarath, and S. Sreeja, "Pose estimated yoga monitoring system," *SSRN Electron. J.* (2021), doi: 10.2139/ssrn.3882498.

[11] K. Y. Chen, J. Shin, M. A. M. Hasan, J. J. Liaw, Y. Ouchi, and Y. Tomioka, "Fitness movement types and completeness detection using a transfer-learning-based deep neural network," *Sensors*, vol. 22, no. 15, p. 5700 (Jul. 2022), doi: 10.3390/s22155700.

[12] A. S. B. Pauzi, F. B. M. Nazri, S. Sai, A. M. Bataineh, M. N. Hisyam, M. H. Zafar, M. . A. Wahab, A. S. A. Mohamed, "Movement estimation using MediaPipe BlazePose," in Advances in Visual Informatics (IVIC 2021), H. B. Zaman et al., Eds. Cham, Switzerland: Springer, Lecture Notes in Computer Science, vol. 13051, pp. 574–583 (2021), doi: 10.1007/978-3-030-90235-3_49.

[13] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), pp. 2334–2343 (2017).

[14] J. Ding, S. Niu, Z. Nie, and W. Zhu, "Research on human posture estimation algorithm based on YOLO-Pose," Sensors, vol. 24, no. 10, p. 3036 (2024), doi: 10.3390/s24103036.

[15] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 1653–1660 (2014), doi: 10.1109/CVPR.2014.214.

Received June, 2025