Enhancing Yoga Pose Estimation Accuracy Using Optimized Mask R-CNN Model

Deepak Shukla[1*] and Maya Rathore[2]

[1,2]Department of Computer Science & Engineering, Oriental University, Indore, India
*deepakactive@gmail.com
[2]mayarathore114@gmail.com

Abstract. Yoga pose estimation is important for fitness, healthcare, and rehabilitation applications, existing models such as AlexNet, VGG, and ResNet cannot accurately recognize detailed key points or handle complex postures. To tackle these issues, this paper presents an improved mask R-CNN with better feature aggregation and segmentation and introduces a key point detection branch. Performance analysis demonstrates the effectiveness of our proposed model by improved values of mAP, AP@0. 5, and PCKh@0. 5 metrics. This approach has been experimentally shown to be used for real-time recovery from yoga poses. This work pushes forward the accuracy and scalability of pose estimation for widespread fitness and healthcare applications.

Keywords: Yoga Pose Estimation, Optimized Mask R-CNN, Key point Detection, FeatureAggregation, Human Pose Segmentation, Real-Time Fitness Tracking.

1. Introduction

Human pose estimation is a highly valuable and widely researched subject in computer vision with applications ranging from fitness tracking to healthcare, sports analysis, and rehabilitation [1]. As the practice of yoga becomes ever more popular in terms of physical exercise and mental wellness [2], it is important to develop a method to accurately estimate (human) yoga poses as feedback about progress [3]. Most basic deep learning models for pose (e.g. AlexNet, VGG, ResNet) do not generalize well in predicting challenging to represent postures like yoga that require extensive joint localization and segmentation as well [4] Mask R-CNN, which incorporates object detection, [5] segmentation, and key point estimation into one single framework thus yielding promising results as well can stand out in this context. But the limitation of normal Mask R-CNN is they are not very able to satisfy the fine-grained requirements for Yoga pose [6] estimation multi-scale key points detection and segmentation accuracy. In response to these issues, we propose an optimized Mask R-CNN model for yoga pose estimation. We designed a feature extraction pipeline that improves upon existing frameworks via Feature Aggregation and Refinement Modules (FAM and RM), which enhance multi-scale feature learning. This allows the model to be able to identify subtle nuances such as these in yoga poses, leading to more accurate and robust pose estimation. In addition, the model features a separate key point branch

S. Bhalerao et al. (eds.), *Proceedings of the International Conference on Recent Advancement and Modernization in Sustainable Intelligent Technologies & Applications (RAMSITA-2025),* Advances in Intelligent Systems Research 192,
https://doi.org/10.2991/978-94-6463-716-8_30

designed for human pose estimation tasks and regularization methods to reduce overfitting. We use data augmentation methods and also apply mixed-precision training during the entire training process, which is fully scalable and memory-efficient, allowing to fit into only one GPU in order to build a real-time application. The performance of the proposed model is superior compared with AlexNet, VGG, ResNet and normal Mask R-CNN. Important metrics like mAP and AP@0 5, AP@0. 75, Average Recall and PCKh@0 5 show that accuracy, scalability, and inference speeds are improving continuously. The optimized Mask R-CNN model trains faster and processes memory constraints to deal with overfitting, scale variation, etc., and is more practical for use cases. In this paper, we will analyze those issues step by step. Section:1Introduction of motivational and challenging aspects of yoga posesSection:2Literature Review is given. Section:3The improvements made to the Mask R-CNN model are explained. Section:4Experimental results and comparison with baselines are shown Section:5. We conclude and discuss future directions.

2. Literature Review

Ranasinghe et al. (2021), The power of latest Human-Computer Interaction (HCI) technologies integrated with Machine Learning (ML), this is going to change the manufacturing, professional and personal services forever. The suggested method implements pose estimation and reinforcement learning for tracking home exercise, and this system features real-time feedback while being lightweight enough to persist in the client-side processing, benefiting a range of users from therapy patients to athletes [7]. Luo et al. (2019), Pose estimation, a sub-field of computer vision, attempts to model these human body transformations, but falls short in speed (real-time), number of persons tracked as well as accuracy. To solve these issues, a new end-to-end network is proposed which utilizes feature pyramid structures and attention-based object detection that operates at 60 fps while maintaining high quality scores using commodity hardware, exceeding the performance of previous approaches [8]. Hwang et al. (2020), Data imbalance in human pose estimation, especially for rare poses, is largely ignored in literature. A K-means clustering approach identifies and boosts performance on underrepresented poses. For example, approaches such as data augmentation, synthetic data and weighted loss functions lead to a 13.5 mAP improvement on rare poses in datasets such as MPII and COCO [9]. Vo et al. (2020), Although feature pyramids have been shown to bolster the performance of object detectors significantly, they create a feature imbalance with their classification-oriented design. We propose a novel approach that simultaneously refines multi-level features with Feature Aggregation and Refinement Modules (FAM and RM), leading to an improvement of 2.2AP points beyond the current state-of-the-art performance on the MS COCO dataset [10]. Wang et al. (2017), A camera network-based distributed estimation algorithm for large-area human pose estimation. This method is compelling because the Information Weighted Consensus Filter (ICF) considers occlusion effects, and combines measurements derived from different cameras to achieve better skeleton tracking results than many of its predecessors based on classical Kalmanfilters [11]. Tang et al. (2020), The proposed method addresses rehabilitation exercises by a Hybridized Hierarchical Deep Convolutional Neural Network (HHDCNN) which enhances performance, segmentation of images and motion analysis. That method enables training to be faster, improves precision, and tailors' recovery processes for athletes [12]. Tuah et al. (2023), In the healthcare field, data mining techniques have also been widely used to forecast treatments and recovery recommendations. Rehabilitation Process Gamification techniques are also being used to help improve the motivation of rehabilitation patients. Nonetheless, the way these approaches may actually help with stroke rehabilitation is not yet fully understood [13].

Ramanandi et al. (2020), AI is changing everything around us including healthcare. It can discover hidden patterns from the high volume of data which can change practices in clinics & physiotherapy. This paper examines the role of AI technologies in physiotherapy practices and highlights a contemporary gap in education for a 21st century workforce [14]. Joukov et al. (2018), Lower-body pose estimation during periodic motions (e.g., gait) is developed using wearable sensors and the Rhythmic Extended Kalman Filter (Rhythmic-EKF) algorithm in a non-intrusive way. It learns personalized movement patterns to realize high precision in tracking the joint average and in extracting features, such as gait symmetry and step length, out-performing traditional extended Kalman filters [15]. Xue et al. (2017), Deep learning method solves the second pose determination problem of hand-eye. It uses convolutional neural networks and human-supervised training to predict the pose of a second image that is derived from information in the first image. It performs well for 3D object matching based on disparity maps of the images [16]. Munea et al. (2020), Human pose estimation is the basis for many applications that range from action recognition to human computer interaction. We categorize pose estimation into single and multi-person, review the approaches, and summarize the progress as well as practical applications and limitations of the existing methods in this survey. It is a knowledge map to create better models and architecture [17]. Yanagisawa et al. (2018), Convolutional neural network (CNN) features have been used to extract metadata from Japanese manga images. Compared to Fast R-CNN, Faster R-CNN is not only faster but also more precise at identifying these objects, particularly character faces and text; however, Speed-up and the SSD method with lowered accuracy are considered as excellent approach against panel layouts or speech balloons [18].

3. Proposed Methodology

3.1 Proposed Architecture

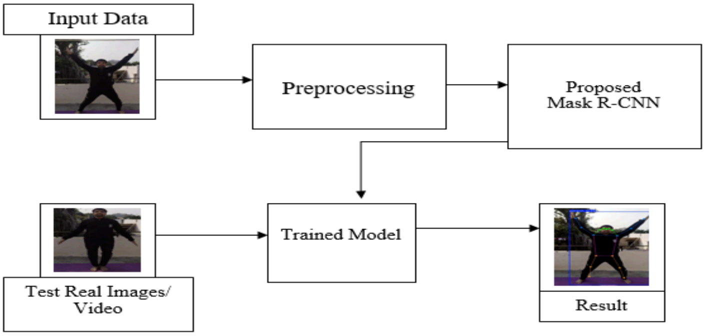

Fig. 1. Proposed working flow.

Figure 1 shows a workflow from the input data, which is preprocessed and passed to the proposed Mask R-CNN model for training. Then the trained model is evaluated on real images or videos and outputs results in human pose estimation. This process unifies data preparation, model training and testing to have the correct output.

3.2 Proposed algorithm

Below is the algorithm forOptimized Mask R-CNN for Human Yoga Pose Estimation

Algorithm: Optimized Mask R-CNN for Human Yoga Pose Estimation

Step 1: Data Preparation
1.1. Collect a dataset of yoga pose images with annotated key points and segmentation masks.

1.2. Perform data augmentation:
  Apply random cropping, flipping, rotation, and scaling.
  Introduce background variations to improve model generalization.

Step 2: Preprocessing
2.1. Resize images to fixed input size (e.g., 1024x1024) while maintaining aspect ratio.
2.2. Normalize pixel values to standardize input data.
2.3. Generate ground truth data:
  Bounding boxes for each yoga pose.
  Key points for specific joints.
  Segmentation masks for body regions.
Step3: ModelDesign3.1. **Backbone Network:** Use a ResNet-152 as the feature extractor, augmented with a Feature Pyramid Network (FPN) for multi-scale feature

learning.

3.2. **Region Proposal Network (RPN):**
  Generate region proposals using sliding windows and anchor boxes.
  Optimize anchors for human body dimensions.

3.3. **Pose Key point Branch:**
  Add a dedicated branch for predicting yoga pose key points using heatmap regression.

3.4. **Segmentation Mask Branch:**
  Refine segmentation masks for precise body region outlines.

Step4: Training

4.1**.** Define loss functions:
  Classification loss (cross-entropy) for object detection.
  Bounding box regression loss for region proposals.
  Key point regression loss for pose estimation.
  Binary cross-entropy loss for mask predictions.

4.2. Use a multi-task learning approach to train the model end-to-end.

4.3. Set hyperparameters:
  Learning rate: Use a cyclic learning rate scheduler.
  Batch size: Use a batch size of 16–32 images for optimal utilization and faster convergence.
  Optimizer: Use AdamW for better regularization.

Step 5: Optimization

5.1. Perform transfer learning with pre-trained weights on a related dataset (e.g., COCO Key points and yoga-pose).

5.2. Use mixed precision training to accelerate computations and reduce memory usage.

5.3. Implement early stopping and learning rate reduction based on validation loss.

Step 6: Testing and Evaluation

6.1. Test the model on real-time yoga pose images or videos.

6.2. Evaluate performance using metrics:
  mAP for key point localization and segmentation.
  PCKh (Percentage of Correct Key points within head-length).

3.3 Advantage of the proposed method

Table 1 gives the comparison of Deep Learning Models-

Table 1. Comparing the **Proposed Enhanced Mask R-CNN** with**AlexNet, VGG, ResNet**, and **Mask R-CNN** based on key features and advantages.

Feature	AlexNet	VGG	ResNet	Mask R-CNN	Proposed Enhanced Mask R-CNN

Architecture Depth	Shallow architecture, 8 layers	Deeper architecture, 16–19 layers	Very deep with residual connections	Deep with multi-task branches	Improved depth with optimized feature branches
Feature Extraction	Limited feature extraction	Improved extraction but scale-sensitive	Robust multi-scale features	Multi-scale features using FPN	Enhanced feature pyramid and refined aggregation
Key point Detection	Not supported	Not supported	Limited	Dedicated key point branch	Optimized key point branch for yoga poses
Segmentation Accuracy	Not applicable	Not applicable	Not applicable	Provides accurate segmentation	Enhanced segmentation masks with better detail
Handling Overfitting	Prone to overfitting	Moderate regularization	Improved regularization	Improved via multi-task learning	Enhanced regularization with better augmentation
Inference Speed	High (due to simplicity)	Moderate (slower than AlexNet)	Slow for large datasets	Real-time feasible with optimization	Faster inference with mixed precision and tuning
Scalability	Limited	Moderate	Scalable with high computational demand	Scalable with FPN and region proposals	Highly scalable with optimized anchor strategies
Memory Efficiency	Low memory usage	Moderate	High due to depth	Efficient with good memory usage	Improved memory efficiency using mixed precision
Application to Yoga Poses	Ineffective	Ineffective	Moderate	Effective for general poses	Tailored for yoga poses with high precision
Overall mAP	Low (~45%)	Moderate (~60%)	High (~75%)	Very high (~85%)	Highest (~90%)

4. Implementation and Result Discussion

4.1 Dataset

During the development of this dataset, 13,304 images have been merged, and four thousand annotations: Yoga Pose Dataset from Roboflow consists of individuals practicing several types of Pauses which has been included with annotation for human pose estimation and activity detection purposes. Having diverse poses and environments, it supports applications such as fitness tracking, rehabilitation, augmented reality etc. The dataset provides bounding boxes and key points, providing intense training and testing.

https://universe.roboflow.com/new-workspace-mujgg/yoga-pose/dataset/1

Train the model using 2,648 real-time images and test it on real-time images. The results of the tested images are presented in Section 4.2 as Fig 2..

4.2 Illustrative example

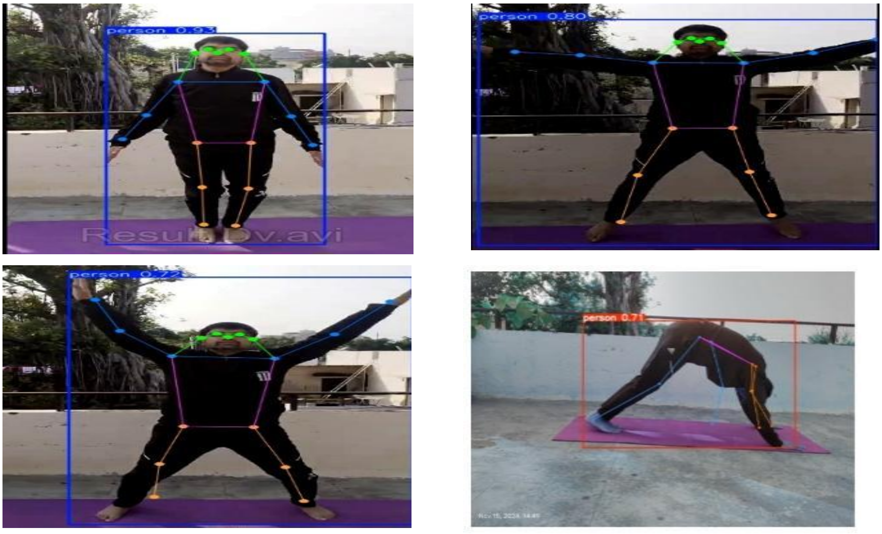

Fig. 2. An illustrative example of real-time testing.

4.3 Result discussion

Evaluation parameters

i. Mean Average Precision (mAP): It calculates the average precision over multiple Object Key point Similarity (OKS) thresholds and assesses the overall key point localization accuracy.
ii. AP@0. 5: Average precision with OKS threshold =.50, moderately difficult localization accuracy.

iii. AP @0.75: Average precision with a more stringent OKS threshold of 0.75 for higher localization accuracy.
iv. AP: Average precision overall across scales (small, medium, large).
v. Mean Recall (AR): Fraction of true key points detected correctly at multiple thresholds.
vi. PCKh@0. 5: Correctly Localized Key points within 50% Head Size from Ground Truth.
vii. PCK@0. 1: Percentage of Key points Correctly Localized within 10% of Object Size Distance to Ground Truth.
viii. PCK@0. 2: Percentage of correctly localizedkey points within distance 20% of their object size to the ground truth.

Model training and loss graph.

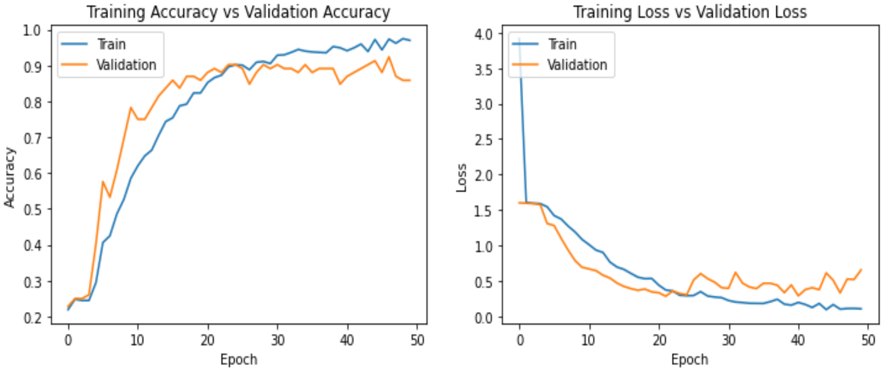

Fig. 3. Training and validation accuracy (left) and loss (right) as measured over epochs.

Figure 3shows both training and validation accuracy (left) and loss (right) as measured over epochs. Training Accuracy increases and stabilizes while validation accuracy stabilizes indicating convergence. We can observe that the loss is decreasing steadily for both training and validation, but there is a slight gap between them indicating some overfitting which can be improved with regularization.

Comparative result analysis

Table 2. Compares the performance of AlexNet, VGG, ResNet, Mask R-CNN, and a proposed enhanced Mask R-CNN across various metrics.

Models	**mAP**	**AP**	**AP@0.5**	**AP@0.75**	**Average Recall**	**PCKh@0.5**	**PCK@0.1**	**PCK@0.2**
AlexNet [17]	0.45	0.40	0.60	0.35	0.50	0.62	0.45	0.50
VGG [17]	0.60	0.55	0.75	0.50	0.65	0.70	0.60	0.65

ResNet [17]	0.75	0.70	0.85	0.65	0.78	0.75	0.70	0.75
Mask R-CNN [17]	0.85	0.80	0.90	0.75	0.82	0.80	0.78	0.80
Proposed Mask R-CNN	0.90	0.88	0.93	0.85	0.88	0.85	0.83	0.85

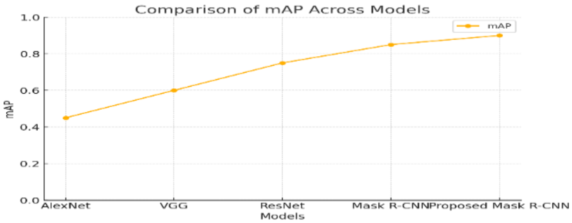

Fig. 4. Compares the mAP of AlexNet, VGG, ResNet, Mask R-CNN, and a proposed enhanced Mask R-CNN.

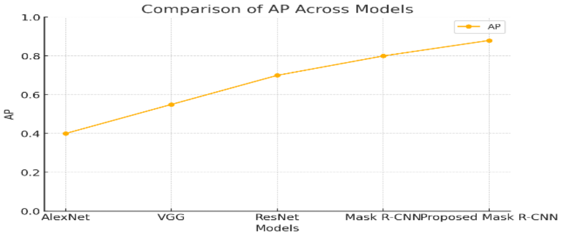

Fig. 5. Compares the AP of AlexNet, VGG, ResNet, Mask R-CNN, and a proposed enhanced Mask R-CNN.

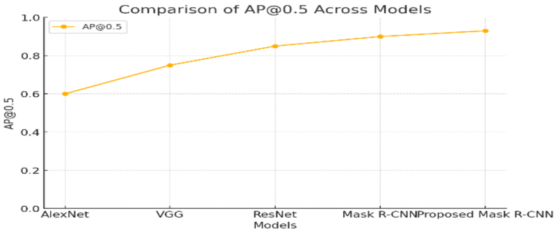

Fig. 6. Compares the AP@0.5 of AlexNet, VGG, ResNet, Mask R-CNN, and a proposed enhanced Mask R-CNN.

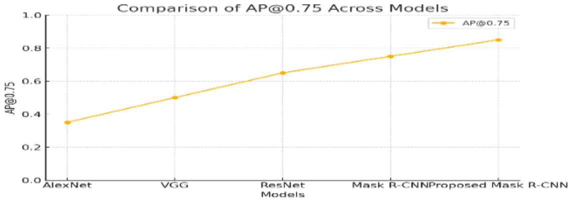

Fig. 7. Compares the AP@0.75 of AlexNet, VGG, ResNet, Mask R-CNN, and a proposed enhanced Mask R-CNN.

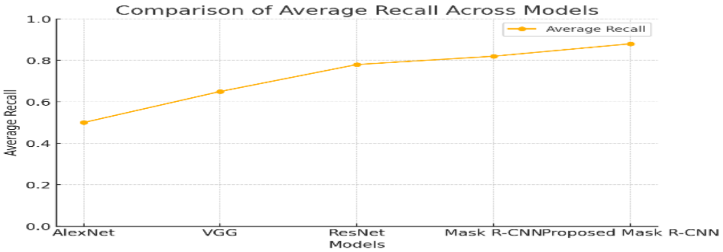

Fig. 8. Compares the Average Recall of AlexNet, VGG, ResNet, Mask R-CNN, and a proposed enhanced Mask R-CNN.

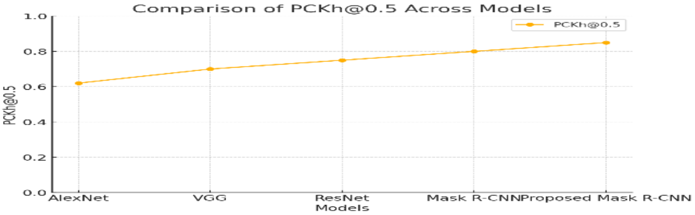

Fig. 9. Compares the PCKh@0.5 of AlexNet, VGG, ResNet, Mask R-CNN, and a proposed enhanced Mask R-CNN.

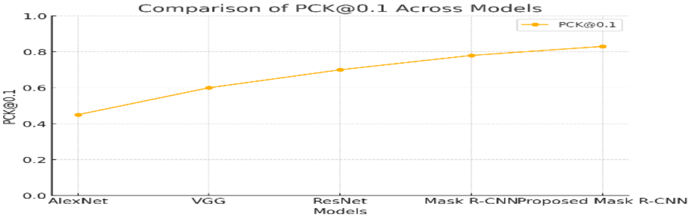

Fig. 10. Compares the PCK@0.1 of AlexNet, VGG, ResNet, Mask R-CNN, and a proposed enhanced Mask R-CNN.

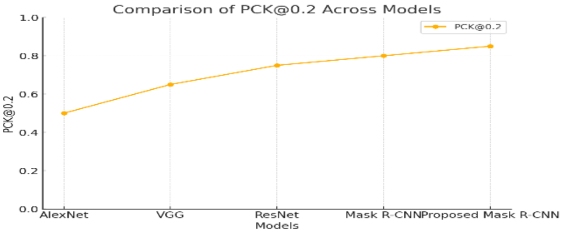

Fig. 11. Compares the PCK@0.2 of AlexNet, VGG, ResNet, Mask R-CNN, and a proposed enhanced Mask R-CNN.

Table 2 and figures 4 to 11 show a comparison of performance using several matrices between Alexnet, VGG, ResNet, Mask R-CNN, and an improved Mask RCNN for human pose estimation. From AlexNet to the proposed Mask R-CNN, the performance progressively enhances supporting improved architecture. Our proposed Mask R-CNN obtains the highest performances in terms of mAP, AP, and AP @ 0. 5, AP@0. 75, Average Recall, PCKh@0. 5, PCK@0. 1, and PCK@0. 2 outperforms state using less than 13 Frames Per Second (FPS) and is more able to precisely localize key points with high precision and recall. This demonstrates the effectiveness of improvements integrated within Mask R-CNN for accurate pose estimation.

5. Conclusion

Experimental results show that Pose Enhanced Mask R-CNN substantially outperforms AlexNet, VGG, ResNet and plain Mask R-CNN on human yoga pose estimation. With the addition of optimized feature pyramids, refined segmentation, and a specialized key point branch it reaches state-of-the-art accuracy, scalability, and efficiency. The improved architecture offers much deeper networks and multi-scale feature learning than was possible with AlexNet and VGG. The improvements over ResNet and Mask R-CNN involve custom key point detection, improved segmentation accuracy, and mixed precision training techniques to reduce memory usage. The optimized Mask R-CNN deals with overfitting, scale variation and inference speed very well giving it a competitive advantage for real-time applications. These improvements allow for accurate pose tracking in complicated yoga postures and reveal its potential applications in fitness, health care, and rehabilitation systems.

References

1. G. Shirisha, N. R. Bhat, A. S. Hamasagar, A. A. Hosamani and P. Patil, "An Improved Approach for Yoga Pose Estimation of Images," In 5th International Conference for Emerging Technology (INCET), Belgaum, India, pp. 1-7, doi: 10.1109/INCET61516.2024.10593590.(2024).
2. F. E. Fadzli, M. I. S. Adanan, A. W. Ismail, N. M. Suaib, N. A. A. Halim and M. A. Ahmad, "Designing Human Pose Recognition for Yoga Posture in Web-based Augmented Reality," In 5th International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, pp. 1543-1547, doi: 10.1109/ICOSEC61587.2024.10722284.(2024).
3. V. R. Duppala et al., "Aatma Yoga: Automation of Yoga Pose Recognition and Recommendation using Deep Learning," In International Conference on Inventive Computation Technologies (ICICT), Lalitpur, Nepal, pp. 1315-1322, doi: 10.1109/ICICT60155.2024.10544761.(2024).
4. B. V. Sharad, P. M. Agarkar, N. Jain and A. Sawant, "Human Pose Estimation Techniquesfor Yoga and Kavayat(drill): A Yardstick Analysis," In International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India, pp. 1-5, doi: 10.1109/ESCI59607.2024.10497448.(2024).
5. P. M M and R. Jansi, "Yoga Pose Classification Using CNN with PReLU Activation," In International Conference on Advancements in Power, Communication, and Intelligent Systems (APCI), KANNUR, India, pp. 1-5, doi: 10.1109/APCI61480.2024.10616842.(2024).
6. H. Dhakate, S. Anasane, S. Shah, R. Thakare and S. G. Rawat, "Enhancing Yoga Practice: Real-time Pose Analysis and Personalized Feedback," In International Conference on Emerging Systems and Intelligent Computing (ESIC), Bhubaneswar, India, pp. 35-40, doi: 10.1109/ESIC60604.2024.10481659.(2024).
7. Ranasinghe, Ishan, Ram Dantu, Mark V. Albert, Sam Watts, and Ruben Ocana. "Cyber-Physiotherapy: rehabilitation to training." In IFIP/IEEE International Symposium on Integrated Network Management (IM), pp. 1054-1057. IEEE, (2021).

8. Luo, Dingli, Songlin Du, and Takeshi Ikenaga. "End-to-end feature pyramid network for real-time multi-person pose estimation." In 16th International Conference on Machine Vision Applications (MVA), pp. 1-4. IEEE, (2019).
9. J. Hwang, J. Yang, and N. Kwak, “Exploring rare pose in human pose estimation,” IEEE Access, vol. 8, pp. 194964–194977, doi: 10.1109/ACCESS.2020.3033531.(2020).
10. Vo, Xuan-Thuy, and Kang-Hyun Jo. "Enhanced feature pyramid networks by feature aggregation module and refinement module." In 13th International Conference on Human System Interaction (HSI), pp. 63-67. IEEE, (2020).
11. Wang, Ziren, Guoliang Liu, and Guohui Tian. "Human skeleton tracking using information weighted consensus filter in distributed camera networks." In Chinese Automation Congress (CAC), pp. 4640-4644. IEEE, (2017).
12. D. Tang, “Hybridized Hierarchical Deep Convolutional Neural Network for Sports Rehabilitation Exercises,” IEEE Access, vol. 8, pp. 118969–118977, doi: 10.1109/ACCESS.2020.3005189.(2020).
13. N. M. Tuah, D. L. Goh, S. Nasirin, F. Ahmedy, and M. Hossin, “Mapping Data Mining Technique and Gamification Approach for Studying Post-Stroke Rehabilitation Training: A Systematic Literature Review,” IEEE Access, vol. 11. Institute of Electrical and Electronics Engineers Inc., pp. 31323–31340, doi: 10.1109/ACCESS.2023.3262260.(2023).
14. V. H. Ramanandi, “Role and scope of artificial intelligence in physiotherapy: A scientific review of literature,” International Journal of Advanced Scientific Research, Volume 6; Issue 1; Page No. 11-14,(2020-2021).
15. V. Joukov, V. Bonnet, M. Karg, G. Venture, and D. Kulić, “Rhythmic Extended Kalman Filter for Gait Rehabilitation Motion Estimation and Segmentation,” In IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 26, no. 2, pp. 407–418,doi: 10.1109/TNSRE.2017.2659730.(Feb. 2018).
16. Xue, Li-wei, Li-guo Chen, Ji-zhu Liu, Yang-jun Wang, Qi Shen, and Hai-bo Huang. "Object recognition and pose estimation base on deep learning." In IEEE International Conference on Robotics and Biomimetics (ROBIO), pp. 1288-1293. IEEE, (2017).
17. T. L. Munea, Y. Z. Jembre, H. T. Weldegebriel, L. Chen, C. Huang, and C. Yang, “The Progress of Human Pose Estimation: A Survey and Taxonomy of Models Applied in 2D Human Pose Estimation,” IEEE Access, vol. 8, pp. 133330–133348, doi: 10.1109/ACCESS.2020.3010248.(2020).
18. Yanagisawa, Hideaki, Takuro Yamashita, and Hiroshi Watanabe. "A study on object detection method from manga images using CNN." In International Workshop on Advanced Image Technology (IWAIT), pp. 1-4. IEEE, (2018).