GymbrAIn: Artificial Intelligence at the Service of Your Muscles

Casali Cristian Flotta Aldo Lamberti Francesco
University of Modena and Reggio Emilia, Italy
Email: {284312, 282955, 272361}@studenti.unimore.it

Abstract

The increasing prevalence of incorrect exercise form in worldwide gyms poses a significant health risk to fitness enthusiasts. Many individuals who engage in resistance training perform exercises with improper technique, potentially leading to injuries and suboptimal results. To address this concern, we present GymbrAIn, an innovative artificial intelligence system designed to provide real-time form correction and exercise verification. By leveraging advanced pose estimation algorithms and self-attention mechanisms, our system offers users immediate feedback on their exercise execution. Our approach combines computer vision technology with sophisticated neural networks to analyze movement patterns and compare them against expert references. Through careful camera calibration and joint angle detection, the system achieves high accuracy in identifying form discrepancies and providing corrective guidance. Initial results demonstrate significant potential for improving exercise safety and effectiveness, with a relevant classification accuracy across multiple common exercises.

1. Introduction

The contemporary fitness landscape has been dramatically transformed by social media influences, with platforms like Instagram and TikTok driving an unprecedented surge in gym membership among young adults seeking to achieve idealized physiques. This phenomenon, while potentially positive for public health, has created a concerning trend where many individuals begin intensive resistance training without proper professional guidance. The decision to forego expert consultation, whether due to financial constraints or perceived inconvenience, is particularly prevalent among two vulnerable categories: young enthusiasts driven by social media trends and older adults attempting to maintain their health. Both groups often find themselves navigating complex exercise routines without consistent professional oversight, primarily due to the prohibitive costs of personal training services or the limited availability of qualified trainers.

The implications of improper exercise execution extend far beyond mere inefficiency, they pose real risks to physical well-being, potentially leading to both acute injuries and chronic musculoskeletal conditions. The lack of accessible professional guidance has created a significant gap in fitness education, particularly affecting those who cannot afford regular personal training sessions or those in areas with limited access to qualified fitness professionals. Additionally, even when trainers are available, the intermittent nature of supervised sessions means that form corrections are not continuously available during every workout.

To address this critical gap in fitness education and supervision, we have developed an artificial intelligence-based solution that leverages computer vision technology. Our system combines advanced pose estimation neural networks with sophisticated GRU (Gated Recurrent Unit) architecture and self-attention mechanisms, enabling real-time exercise classification and form analysis through a standard smartphone camera. The system's core functionality rotates around precise articulation angle analysis, comparing user execution against reference movements performed by certified fitness professionals.

Through careful camera calibration techniques detailed in [6], our system achieves remarkable accuracy in joint angle detection and movement pattern recognition. The result is a comprehensive feedback mechanism that provides users with immediate guidance for improving their exercise form. This technology has the potential to significantly reduce injury risk and improve training outcomes for a big range of users, from beginners to experienced athletes trying to refine their technique.

2. Related Work

The intersection of artificial intelligence and fitness instruction has emerged as a promising field of study, with several notable contributions advancing our understanding of automated exercise analysis. A particularly significant contribution comes from the International Conference on Information Technology (ICIT), where researchers conducted an extensive comparative analysis of three distinct neural network architectures applied to human pose estima-

tion in fitness contexts. Their investigation regarded Feed-Forward Neural Networks (FNN), Long Short-Term Memory (LSTM) networks, and Gated Recurrent Units (GRU), with each model evaluated for its efficacy in exercise pose classification.

The GRU model demonstrated superior performance, achieving an impressive 97.27% accuracy in pose classification tasks. This exceptional performance was closely followed by the LSTM model, which maintained strong reliability with an accuracy of approximately 95% [4]. These findings provided crucial validation for the potential of recurrent neural architectures in movement analysis applications.

Our project builds upon these foundational insights while introducing significant architectural innovations. While we maintain the GRU-based approach that proved so effective in previous studies, we've enhanced the architecture through the strategic integration of self-attention layers and optimized dropout parameters, as detailed in our Methods section 4. The incorporation of GRU architecture is particularly crucial for our application, for their inherent ability to learn and maintain long-term dependencies while mitigating the vanishing gradient problem makes it ideally suited for capturing the temporal dynamics of exercise execution [4].

3. Data

Our project utilizes a carefully curated dataset sourced from Kaggle, comprising high-quality video recordings of exercise executions [1]. We strategically selected a subset of videos to enable rapid prototyping and development of our system. While the original dataset contains a much larger collection of exercise recordings, we deliberately limited our initial scope to four fundamental resistance training movements. This focused sampling approach allowed us to speed up the training process and validate our methodology efficiently, while still maintaining sufficient data for meaningful analysis.

To enhance the robustness and generalization capabilities of our model, we implemented strategic data augmentation techniques. This included horizontal translations of original videos within a range of +100 to -100 pixels, as well as mirror transformations of the original footage. This augmentation approach significantly expanded our effective dataset while maintaining the essential biomechanical authenticity of the movements.

The augmented dataset distribution in our working subset is as follows:

- 112 videos for bench press
- 104 videos for barbell biceps curl
- 90 videos for triceps push down
- 94 videos for lat pull down

for a total of 400 videos across four exercises.

The preprocessing pipeline incorporated sophisticated image processing techniques to optimize video quality and enhance feature detection. Initially, we applied a Gaussian filter to eliminate random noise artifacts that could potentially interfere with pose estimation accuracy. This was followed by a sharpening filter application, which enhanced the definition of human figures and movement patterns within the footage. This dual-filter approach ensured optimal input quality for our neural network while maintaining the integrity of the original movement patterns.

4. Methods

The pipeline of our system consists of several distinct but interconnected components, each optimized for its specific role in the overall system.

4.1. Preprocessing

The initial stage of our system involves rigorous video preprocessing to ensure optimal input quality. Raw video input undergoes a two-phase filtering process.

4.1.1 Gaussian filter

We implement a carefully tuned Gaussian filter to suppress random noise while preserving essential movement information. The filter parameters were empirically determined to maintain an optimal balance between noise reduction and detail preservation.

4.1.2 Sharpening Filter

Following noise reduction, we apply a sharpening filter to accentuate edge definition and improve the clarity of human figures within the frame. This step proves crucial for subsequent pose estimation accuracy.

4.2. Pose estimation

The preprocessed video frames are then analyzed using YOLO POSE [3], a state-of-the-art neural network architecture specifically optimized for human pose estimation. This component:

- Generates real-time skeletal tracking data
- Identifies key anatomical landmarks with high precision
- Maintains temporal consistency across video frames
- Operates efficiently within mobile device computational constraints

4.3. Exercise classification

Our neural network is designed to perform video-based exercise classification based on the poses extracted by YOLO POSE [3], identifying which of the four fundamental exercises (bench press, barbell biceps curl, triceps push down, or lat pull down) the user is performing. This classification task began by implementing the foundational architecture presented in [4], which consists of GRU-based sequential processing units with 256 neurons and a dropout rate of 0.5, trained on a training dataset of 320 videos from our initial dataset.

From this baseline, we conducted systematic modifications to enhance classification performance:

4.3.1 Core Architecture

The initial implementation included:

- GRU-based sequential processing units
- 256 neurons in the hidden layers
- Dropout rate of 0.5
- Standard temporal feature processing
- Output layer with 4 neurons (one for each exercise class) with softmax activation

We systematically explored several modifications to improve upon the base architecture.

4.3.2 Neuron Expansion

We increased the hidden layer size from 256 to 512 neurons, maintaining the original dropout rate of 0.5. The aim was capturing more complex temporal patterns.

4.3.3 Self-Attention

We also added self-attention layers implemented as multi-head attention for parallel feature processing. This enabled infinite receptive field for global context awareness and added a dynamic weighting of temporal segments. This enhancement helped the model focus on the most discriminative parts of each exercise movement.

4.3.4 Final optimized architecture

The final optimized architecture incorporates:

- Self-attention mechanisms
- Reduced neurons to 256
- Reduced dropout rate to 0.25

The key insight in our final architecture was recognizing that the combination of data augmentation techniques (translation and mirroring) already provided substantial regularization against overfitting. This allowed us to reduce then number of neurons from 512 to 256 and the dropout rate from 0.5 to 0.25, striking an optimal balance between model regularization and learning capacity. A higher dropout rate, when combined with our augmentation strategy, would have been redundant and potentially harmful to the model's ability to learn effective representations.

Once the network classifies the exercise being performed, this information is used by subsequent components of our system to apply the appropriate form-checking criteria and provide relevant feedback to the user. This modular approach allows us to maintain high accuracy in both exercise identification and form analysis.

This architectural decision was validated through experimental results (detailed in Section 5), which demonstrated that the combination of robust data augmentation and reduced dropout rate achieved superior performance compared to architectures with higher dropout rates or higher number of neurons.

4.4. Video retrieval and geometrical analysis

The final component implements a geometric analysis system that leverages video retrieval techniques for form comparison. The process follows these key steps.

4.4.1 Movement Analysis

Movement analysis identifies salient movement moments through velocity minima detection. These moments typically correspond to the starting and ending positions of each repetition. Velocity analysis helps isolate the most critical phases of each exercise.

4.4.2 Reference Video Retrieval

Once the exercise is classified, the system retrieves a corresponding reference video using an L2 norm between reference and user poses at minimum velocity frames. The reference video shows the same exercise performed by an expert with proper form.

4.4.3 Comparative Analysis

Then it extract critical joint angles at the identified keypoints for both the user's video and the retrieved reference. It aligns the user's movement phases with the reference video's salient moments and performs direct comparison of joint angles between user and reference at corresponding temporal points.

4.4.4 Feedback Generation

Finally it calculates angular differences between user execution and reference at key points and generates corrective feedback when differences exceed predetermined thresholds. This video retrieval-based approach offers several advantages over static template matching:

- Accounts for the natural variability in human movement
- Captures the dynamic nature of exercise execution
- Provides realistic reference points for form correction
- Enables more nuanced comparison of movement patterns

You can see some real photo example in Appendix A.

5. Experiments

To validate our approach and assess the effectiveness of various architectural decisions, we conducted a comprehensive series of experiments focused on optimizing the exercise classification component of GymbrAIn. Our experimental journey began with a baseline GRU implementation inspired by previous work in the field [4], and systematically evolved through several architectural variations. Each iteration was designed to test specific hypotheses about the relationship between model complexity, regularization, and performance. Our experimental framework was designed to not only measure raw classification accuracy but also to understand the model's learning dynamics through careful tracking of training progression and loss patterns. We paid particular attention to the impact of self-attention mechanisms, which we hypothesized would enhance the model's ability to capture the temporal dependencies inherent in exercise movements. The results of these experiments, detailed below, provide valuable insights into the effectiveness of our architectural choices and their implications for real-world deployment. These results were obtained splitting randomly our dataset into 320 videos as a training set, as seen before, 40 as a validation set and 40 for test set.

Architecture	Neurons	Dropout	Self-Attention	Accuracy (%)
GRU	256	0.5	No	77
GRU	512	0.5	No	67
GRU	512	0.5	Yes	35
GRU	256	0.5	Yes	77
GRU	256	0.25	Yes	80

Table 1. Test performances comparison of the architectures

The accuracy values are lower than the accuracy of the original architecture [4] because of the smaller training set that we used. Training accuracy and loss plots are instead in the Appendix B as well as the confusion matrix C.

6. Conclusion

This project presents GymbrAIn as a promising advancement in automated exercise form analysis and correction. While our integration of GRU-based architecture with self-attention mechanisms and optimized dropout rates demonstrates potential, there remains significant margin for improvement in classification accuracy. Although the current accuracy of 80% is promising, it indicates that further architectural refinements and expansion of the dataset could lead to more robust results. The system successfully addresses the need for accessible guidance for exercise form, but with notable limitations that future work can address.

6.1. Key results

We developed a robust exercise analysis system. We achieved a competitive accuracy in movement classification. We successfully implemented real-time form correction capabilities. We design a flexible architecture that has potential for mobile deployment.

6.2. Limitations and future work

Several promising directions for future research emerge from our current findings.

6.2.1 Dataset Expansion

Incorporation of additional exercise variations. Integration of different training modalities. Collection of more diverse subject populations

6.2.2 Technical Enhancements

Implementation of automatic camera calibration. Development of more sophisticated form correction algorithms. Integration of user-specific adaptation mechanisms.

6.2.3 Mobile Implementation

Optimization for various mobile hardware configurations. Development of user-friendly interface designs. Integration of real-time performance optimization techniques.

6.2.4 Clinical Applications

Validation for physiotherapy applications. Integration with rehabilitation protocols. Development of specialized movement analysis modules.

A significant current limitation is the manual camera calibration requirement, which impacts user experience and system accessibility. The integration of deep learning-based

calibration methods like DeepFocal [5] and DeepCalib [2] presents a promising solution to automate this process in future iterations.

6.3. Impact and implications

While GymbrAIn demonstrates the potential for AI-driven solutions in personal fitness and rehabilitation contexts, its current implementation should be viewed as a proof of concept rather than a final product. The system's limitations in accuracy and camera calibration requirements indicate the need for continued development. However, these challenges also present clear paths for improvement. As the system evolves and incorporates automated calibration and improved accuracy, its applications could extend beyond traditional gym environments into physical therapy, sports training, and general wellness monitoring. The foundation laid by this project suggests that with continued refinement, AI-driven exercise form analysis can become a practical tool for improving exercise safety and effectiveness.

References

[1] Hasyim Abdillah. Kaggle. https://www.kaggle.com/datasets/hasyimabdillah/workoutfitness-video. 2

[2] Oleksandr Bogdan, Viktor Eckstein, Francois Rameau, and Jean-Charles Bazin. Deepcalib: a deep learning approach for automatic intrinsic calibration of wide field-of-view cameras. In *Proceedings of the 15th ACM SIGGRAPH European Conference on Visual Media Production*, CVMP '18, New York, NY, USA, 2018. Association for Computing Machinery. 5

[3] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection, 2016. 2, 3

[4] Sijie Shang, Rong Jin, and Kevin Desai. A study of human fitness pose classification using artificial neural networks. In *2023 International Conference on Information Technology (ICIT)*, pages 250–255, 2023. 2, 3, 4

[5] Scott Workman, Connor Greenwell, Menghua Zhai, Ryan Baltenberger, and Nathan Jacobs. Deepfocal: A method for direct focal length estimation. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 1369–1373, 2015. 5

[6] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, 2000. 1

Appendices

A. Real example

(a) Starting position (b) Ending position

Figure 1. Frames at minimum velocity of the reference for bench press

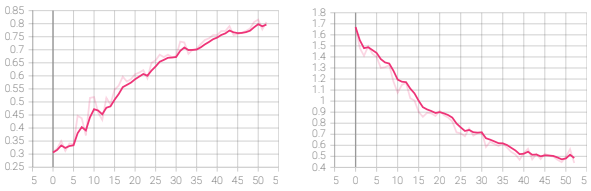
(a) Starting position (b) Ending position

Figure 2. Frames at minimum velocity of a user for bench press (wrong execution, the system found out a difference of 27° from the reference while 25° was the threshold)

(a) Starting position (b) Ending position

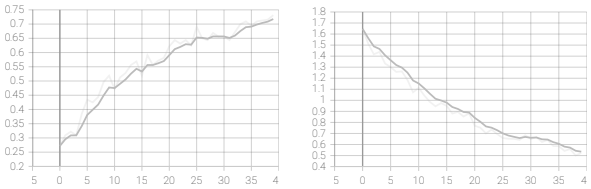
Figure 3. Frames at minimum velocity of a user for bench press (right execution)

B. Accuracy and loss plots



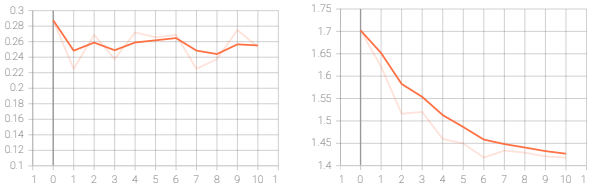
(a) Epoch Accuracy (b) Epoch Loss

Figure 4. GRU with 256 neurons and dropout 0.5



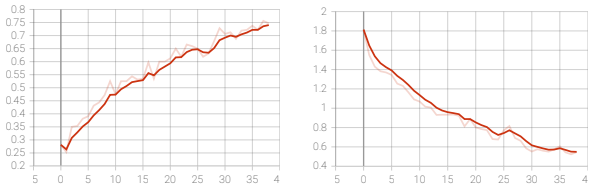
(a) Epoch Accuracy (b) Epoch Loss

Figure 5. GRU with 512 neurons and dropout 0.5



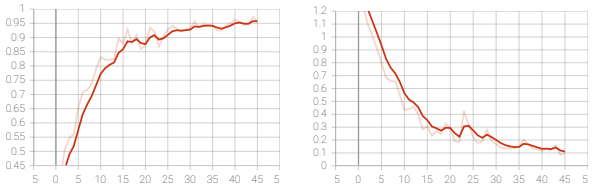
(a) Epoch Accuracy (b) Epoch Loss

Figure 6. GRU + Self-Attention with 512 neurons and dropout = 0.5



(a) Epoch Accuracy (b) Epoch Loss

Figure 7. GRU + Self-Attention with 256 neurons and dropout = 0.5



(a) Epoch Accuracy (b) Epoch Loss

Figure 8. GRU + Self-Attention with 256 neurons and dropout = 0.25

C. Confusion matrix

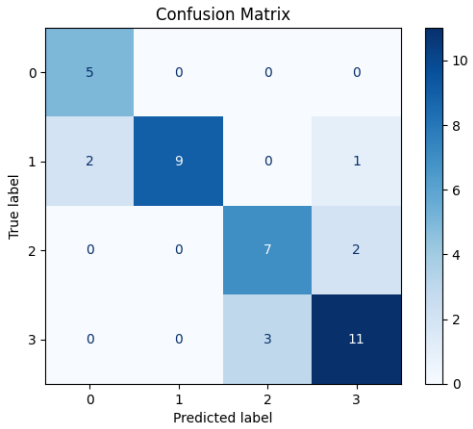


Figure 9. Confusion matrix