A SYSTEM FOR AUTOMATIC AND FAST GYMNASTICS POSE ESTIMATION AND KEY FRAME IDENTIFICATION

Ilias Masmoudi[1], Johannes Link[1], Sebastian Möck[2], Petra Nissinen[2], Anne Koelewijn[1]

Machine Learning and Data Analytics (MaD) Lab, Faculty of Engineering, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany[1]
Department of Exercise Science, Olympic Training and Testing Center of Hessen, Otto-Fleck-Schneise 4, Frankfurt, Germany[2]

In artistic gymnastics, performance analysis tools that provide immediate feedback to athletes and coaches are key to optimizing performance. Therefore, we developed a system that provides direct video feedback and automatically identifies key frames relevant to achieve high performance. This system estimates the gymnast's two-dimensional (2D) and three-dimensional pose from a single-view video and uses a long short-term memory (LSTM) network to identify key frames, in which contact with the ground or an apparatus starts or ends. We compared three 2D pose estimation algorithms and found that MoveNet yielded the highest accuracy, being able to detect 84% of the joints accurately. The LSTM network detected the key frames with a 97% F1 score. In conclusion, our system provides direct video feedback to gymnasts and coaches relevant performance parameters.

KEYWORDS: markerless, Joint Angles, gymnastics, 3D pose, performance analysis.

INTRODUCTION: Artistic gymnastics is a sport characterized by its rigorous demands on precision, form, and timing. For good technique, and thus good scoring in competition, pose is highly important. The slightest deviation can significantly impact performance and the overall quality of movements. Therefore, feedback based on movement visualizations can optimize performance. Marker-based motion capture is impractical in a dynamic sports environment due to its long setup time and complexity (Wade et al., 2022). Instead, coaches and trainers have traditionally relied on observation and video playback to evaluate performances.
The current feedback process involves a time-consuming method where key frames, particularly those marking the start or end of ground or apparatus contact, are manually identified. Additionally, the joint angles of these key frames are labelled for feedback. This manual process leads to a significant delay, with feedback becoming available only hours after the practice session. These key frames are used instead of the full movement, since they are most relevant for the evaluation of movement quality and thus competition scores.
However, athletes can benefit from immediate visual feedback, such that errors in form, posture, and key contacts can be flagged and corrected almost immediately. The reduced time between error detection and correction can potentially accelerate learning (Rhodes et al., 2014). Recent technology advancements have led to an influx of sophisticated tools designed to offer real-time insights in sports. Particularly, deep learning, a subset of machine learning, has demonstrated promising results to enhance athletic performance (Zhang et al., 2022). One main application of deep learning is image analysis, such as two dimensional (2D) (Votel & Li, 2021) and three dimensional (3D) (Gong et al., 2021) pose estimation. This application is highly relevant for gymnastics, where feedback on joint angles, especially during the key frames, is most important. The body's position, particularly the centre of gravity during the contact phases, allows coaches and scientists to draw conclusions about the impact forces and angular impulses. However, since gymnastics movements are fast and different from normal human activity, conventional pose estimation algorithms, which are trained on normal human activity, might not perform well. Previously, MoveNet was found to outperform other pose estimation algorithms, including challenging dynamic actions like the ”Jump Rope” exercise and the ”Golf Swing” (Chung et al., 2022).

Therefore, we investigated if we can develop a system to provide direct video feedback to coaches and athletes in gymnastics, together with accurate pose estimation information, without disturbing the athletes. Specifically, this system should be able to perform 3D pose estimation and identify the key frames to provide precise feedback for performance improvement. Furthermore, for the system to be useable by coaches and athletes, it should be deployable on lightweight computer systems, like a tablet. Specifically, due to the challenges with gymnastics movements, we compared different pose estimation approaches to see which could output joint angles most accurately. We also investigated if we could identify key frames using a long short-term memory (LSTM) network.

METHODS: Our system performs four different steps to provide video feedback. First, 2D pose estimation is performed on the collected video recordings. Second, this 2D pose estimation is converted to a 3D pose estimation. Third, the key frames are identified where contact with the ground or an apparatus starts or finishes. Fourth, a graphical user interface (GUI) provides the video feedback to the user, combined with the relevant joint angles. We chose an approach involving separate 2D and 3D pose estimation, since we found in preliminary work that direct 3D pose estimation could not identify joints accurately due to motion blur of fast movements. Instead, we first estimated the 2D spatial location of different keypoints (mainly joints) of the human body, and converted these to a 3D pose using PoseAug (Gong et al., 2021).
The data collection process for the 2D pose estimation algorithms and the key frame identification evaluation was conducted at the Hesse Olympic training centre, employing a single camera positioned at a 90-degree angle to the gymnastics vault runway to capture lateral movement. We used two recording settings: 1080p at 50 fps to provide high-definition clarity for detailed observation of the gymnast's movements, balancing motion capture with file size and processing needs, and 720p at 240 fps to capture rapid movements with high temporal resolution, essential for analysing fast manoeuvres in gymnastics. We then labelled the collected data in two ways. We created a ground truth for the pose using the highly accurate real-time multi-person one-stage (RTMO) model (Lu et al., 2023), which is not suitable for real-time applications on a CPU. We manually corrected this labelling to rectify inaccuracies in the model's pose detection, particularly in instances of motion blur or occlusion. We also manually labelled the frames where the gymnast touched the ground or apparatus as "contact frames" to train the key frame identification model. Out of the 58 clips collected, 44 were used for training, and 14 for testing. We evaluated three lightweight pose estimation models, namely PoseNet (Papandreou et al., 2018), MoveNet (Votel & Li, 2021), and BlazePose (Bazarevsky et al., 2020) by determining the percentage of detected joints (PDJ). The PDJ represents the percentage of joints for which the predicted location from each of the 3 models is within 5% of the torso diameter from the ground truth. The torso diameter is defined as the Euclidean distance between the left shoulder and right hip. We also trained an LSTM to identify key frames, specifically, the start and end of contact between the athlete and the ground or an apparatus, based on the manually labelled contact-frames. We chose LSTMs (Hochreiter et al., 1997), since we expected that image-based neural networks would not be effective due to variation in lighting, athlete physiques and attire, camera angles and positions, and numerous gymnastic apparatus. We designed the LSTM to be lightweight, meaning that it features a single-layer architecture with a hidden size of 256. This design enables efficient processing while still capturing the necessary temporal dynamics in the data. To enhance the LSTM model's ability to generalize across various gymnastic scenarios, we implemented specialized data augmentation techniques on the 3D pose data estimated by MoveNet. Firstly, we applied random rotations to the 3D poses to simulate different camera angles and ensure that the model does not become biased towards a specific viewpoint. Secondly, we introduced controlled pose corruption into the data to mimic real-world scenarios where data might be noisy or imperfect and ensure reliability in diverse settings. The model is trained over 40 epochs with a learning rate of 0.01, using the Adam optimizer. For evaluation, we estimate the 3D pose using MoveNet on the test set and use it as input to the LSTM. We focused on binary cross-entropy loss, accuracy, and precision-recall metrics as the primary evaluation criteria. We finally evaluated the required computational resources using the system's speed for

different image resolutions and model configurations. To do so, we compared the frame rate of the 2D pose estimation only to the frame rate combining the 2D pose estimation with 3D pose estimation, and with 3D pose estimation and contact frame identification.

RESULTS AND DISCUSSION: We found that MoveNet was able to identify the 2D keypoints (i.e., joints) most accurately, with a PDJ of 84%, while BlazePose's PDJ was 78%, and PoseNet's 76%. Figure 1 shows an example of the 2D pose estimation during a rapid movement. This image highlights that the 2D pose estimation by MoveNet is most accurate in this scenario, since BlazePose had trouble identifying the keypoints in the feet and the upper torso and PoseNet was not able to identify the keypoints in the legs at all. Therefore, we selected MoveNet for 2D pose estimation in our system. It is important to note that the system's accuracy hinges on the quality of the 2D pose estimation, and errors here will cascade into the 3D pose estimation, and contact event detection. Therefore, other lightweight pose estimation algorithms could be explored, or transfer learning could be used to improve 2D pose estimation accuracy for these gymnastics' movements.

Figure 1: Side-by-side comparison of 2D pose estimation by PoseNet, BlazePose, and MoveNet on a gymnastics movement frame.

We used an LSTM to identify key frames in the video and found that a relatively lightweight model was ample to attain high accuracy in detecting contact frames. As illustrated in Figure 2, representing the confusion matrix, the LSTM model demonstrated an F1 score of 97%, signifying its effectiveness in precisely classifying contact and no-contact frames. The true positive rate was high for both no-contact and contact predictions, with low false negative and false positive rates. This highlights the capability of the lightweight LSTM to effectively distinguish between contact and no-contact.

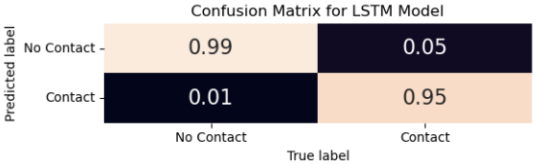

Figure 2: Confusion Matrix of the LSTM Model, highlighting the precision in distinguishing between no-contact and contact phases during gymnastic routines.

Table 1: Comparison of Frame Rates (frames per second) on i5 13600k for Different Resolutions and Model Configurations.

Model Configuration	720p	1080p
MoveNet	41.93	35.21
MoveNet + PoseAug	39.04	34.07
MoveNet + PoseAug + Contact LSTM	37.10	32.29

The analysis of the system's performance shows that it achieves high frame rates across various configurations of MoveNet, including with PoseAug and the contact LSTM, at both 720- and 1080-pixel resolutions (Table 1). Given that standard gymnastics recordings are typically captured at 60 frames per second, this processing rate implies that the system processes slightly over half a frame for every frame of recording time. Therefore, a gymnastics movement that typically takes 1-10 seconds to record would take approximately 2-20 seconds to process. This performance, achieved on a CPU, suggests that the system does not require excessive computational resources, enhancing its practicality and accessibility for various gymnastics training environments.

The final GUI, as illustrated in Figure 3, provides athletes with comprehensive feedback. At the top, it presents real-time biomechanical analysis with the athlete's velocity, joint angles, and contact information. The bottom left section features a synchronized video with 2D and 3D pose visualizations, joint angles, and velocities for an in-depth movement analysis. The bottom right focuses on key frames, especially moments before flight and landing, aiding.

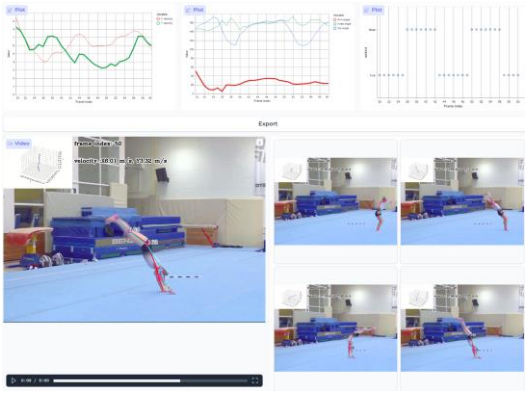

Figure 3: GUI Overview: Real-time Gymnastics Performance Analysis with Velocity, Joint Angles, and Key Frames Visualization.

CONCLUSION: We have shown that our system can provide direct video feedback of gymnastics movements to athletes and coaches while being unobtrusive. This feedback consists of the recorded motion together with the relevant joint angels and velocities. By offering near-instantaneous feedback that is recorded unobtrusively, we have created many new opportunities in sports training. The system is now being employed by coaches at the Hesse Olympic training centre, showcasing its real-world application and value in sports training. In the future, we aim to analyse and quantify the benefit of such direct video feedback to athletes and to extend to other sports where form is highly important, such as figure skating or diving. We also aim to integrate a system to not only identify errors but also suggest corrective measures in real-time could be a significant advancement.

REFERENCES

Bazarevsky, V., Grishchenko, I., Raveendran, K., Zhu, T., Zhang, F., & Grundmann, M. (2020). Blazepose: On-device real-time body pose tracking. arXiv preprint arXiv:2006.10204.

Chung, J.-L., Ong, L.-Y., and Leow, M.-C. (2022). Comparative analysis of skeleton-based human pose estimation. Future Internet, 14(12):380.

Dìaz-Pereira, M. P., Gomez-Conde, I., Escalona, M., and Olivieri, D. N. (2014). Automatic ´ recognition and scoring of olympic rhythmic gymnastic movements. Human Movement Science, 34:63–80.

Gong, K., Zhang, J., & Feng, J. (2021). Poseaug: A differentiable pose augmentation framework for 3d human pose estimation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 8575-8584).

Lu, P., Jiang, T., Li, Y., Li, X., Chen, K., & Yang, W. (2023). RTMO: Towards High-Performance One-Stage Real-Time Multi-Person Pose Estimation. arXiv preprint arXiv:2312.07526.

Papandreou, G., Zhu, T., Kanazawa, N., Toshev, A., Tompson, J., Bregler, C., & Murphy, K. (2018). Towards accurate multi-person pose estimation in the wild.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

Votel, R., & Li, N. (2021). MoveNet: Lightning fast pose detection. Google AI Blog. https://blog.tensorflow.org/2021/05/next-generation-pose-detection-with-movenet-and-tensorflowjs.html

Rhoads, M. C., Da Matta, G. B., Larson, N., & Pulos, S. (2014). A meta-analysis of visual feedback for motor learning. Athletic insight, 6(1), 17.

Wade, L., Needham, L., McGuigan, P., and Bilzon, J. (2022). Applications and limitations of current markerless motion capture methods for clinical gait biomechanics. PeerJ, 10:e12995.

Zhang, Y., Tang, H., Zereg, F., and Xu, D. (2022). Application of deep convolution network algorithm in sports video hot spot detection. Frontiers in Neurorobotics, 16.