Evaluating the Limits of Real-Time Complex Pose Estimation and Reference Matching

Maaz Shamim - mshamim2 | Max-Peter Schrøder - mschrod2 | Lukas Geer - lgeer3 | Valerie Liang - vliang5

Problem Area:

For our project, we want to develop a model capable of real-time comparison of live webcam body poses to a reference pose, estimating and scoring the posing accuracy between input and reference, and ensuring robustness to temporal and physical disturbances. One practical use case is for video games that require mimicking poses with fast movements (e.g, JustDance), but could also potentially be extended to sports analysis and beyond. While body tracking has become increasingly common in gaming and motion-based applications, common challenges such as **occlusion**, **temporal jitter**, and **latency** may lead to unstable detections, causing the models to be unstable in continuous interactions like dynamic activities. Furthermore, mapping the current input pose to a reference presents challenges such as **body proportion scaling** and **posing accuracy**.

Our project will evaluate and attempt to improve these methods by directly comparing state-of-the-art body tracking models with our algorithms in scenarios that demand precise and efficient body alignment to estimate the position of obscured or jittered body parts frame-by-frame. We will focus the training, testing, and evaluation of our models to a restricted subspace of poses, targeting motions that require precise body mapping and timing.

Context:

In class, we discussed SIFT (Scale-Invariant Feature Transform), a traditional computer vision approach that has been used to detect and match objects across images despite changes in translation and scale. While effective for static objects, SIFT and similar algorithms struggle with real-time detection and tracking, particularly for dynamic, articulated objects like humans.

More recent methods address these challenges using pose estimation models and skeleton-based tracking, which represent the body as a set of keypoints or joints. While these approaches improve tracking speed and handle dynamic motion more effectively, they continue to struggle with challenges like occlusion and jitter, which are worsened by motion blur. Benchmarks have demonstrated that even SOTA models remain susceptible to real-world disruptions and are “unsolved” problems in modern pose-estimation systems **[1]**. More recent deep learning approaches have sought to address occlusion using cost volumes, which encode similarities of pixels across frames and allows networks to learn and recover motion features via optical flow **[2]**.

Data:

We will use the AIST++ dataset for training and evaluation, which offers over 10 million annotated frames of 3D dance motion across ten genres. This provides the complex, dynamic poses necessary to rigorously test pose estimation.We will use a single-camera view to focus the challenge on accurately predicting occluded joints without multi-view cues. The dataset's video structure allows us to test both static frame accuracy (for handling self-occlusion) and temporal performance (for consistency and latency). With a total dataset size of over 10 million images (frames), AIST++ also provides a good foundation for testing how well our model generalizes across diverse poses, movements, and subjects. Our experimental approach involves selecting a diverse subset of poses for training, emphasizing body contortions that lead to self-occlusion. We will then test the model on held-out poses with similar limb configurations to evaluate its ability to generalize and accurately infer the position of occluded joints.

Proposed Solution:

The proposed system captures real-time video of a user performing gestures and extracts 2D joint keypoints using a pose estimator. Each keypoint is evaluated for reliability and adjusted over time to reduce jitter and handle brief occlusions or misdetections. The processed keypoints are then compared against reference poses to assess how accurately the user mimics the intended action. We will explore multiple approaches for tracking the subject’s movements. The idea is to implement both a reliable baseline and a more experimental method, then compare their performance.

Baseline Method: We will use pre-trained pose estimation models such as MediaPipe Pose and MoveNet to track body joints in real time. These models will allow us to evaluate the strengths and limitations of existing approaches, particularly in scenarios involving self-occlusion, limb overlap, and rapid motion. Building on these baselines, we will develop algorithms that combine pose-based tracking with motion segmentation to improve temporal stability and accuracy.

Experimental Method: Our primary approach will be to develop a robust model capable of maintaining pose and estimation coherence with low-latency to real time video, directly addressing the challenges of occlusion and jitter.

- **Image Segmentation:** We will use basic computer vision techniques to segment the subject and their body parts from the background using object labeling and binarization. This approach allows us to account for variations in body shape, clothing, and skin tone by testing thresholds across different data images. Movements will be then detected by analyzing frame-to-frame changes and spatial moments of the silhouettes to estimate overall body orientation and limb position via a heuristic of general bone shape/structure.

- **Optical Flow [3]:** We will use optical flow to capture the pixel-level motion of the subject across frames. This provides a continuous, dense motion field that is inherently robust to the occasional keypoint detection failures. By fusing these precise motion vectors with the structural keypoint data from our primary pose models, we can create a more resilient tracking system. This synergy allows us to infer the likely position of occluded or lost joints by propagating their last known location through the observed motion field, directly addressing a key weakness of standalone pose estimation.

- **Temporal Smoothing via Kalman Filter [4, 5]:** Kalman filters are typically applied as a post processing technique in SOTA models to achieve temporal smoothing. The filter incorporates multiple frames in the process of correcting joint predictions which may be subject to jitter/occlusion. It consists of two stages, the prediction stage and the correction stage. We would be training the A, H, Q, and R matrices. The A matrix maps how states evolve, the H matrix encodes how observations relate to the latent state, the Q matrix processes noise covariance, and the R matrix measures noise covariance. To specifically account for the nature of our data, which contains quick, high speed movements, we can additionally model velocity and acceleration while keeping these variables in the A matrix.

Evaluation:

We will benchmark our pose estimation models against the current state of the art (MoveNet). Overall, this comprehensive evaluation will determine if our models can offer improved reliability, precession, and speed especially in challenging scenarios with dynamic movements and complex poses.

- Performance will be measured using two metrics:
    1. Binary Accuracy: A score of 1 for a correct pose prediction within a set threshold, otherwise 0.
    2. Quantitative Precision: The average pixel distance between the predicted joint location and the true location.
- Inference latency will be computed by considering the average processing time per frame to ensure the model meets sufficient requirements for interactive applications. We will consider whether our model's latency falls within an acceptable margin of SOTA models, balancing accuracy and responsiveness.

Limitations:

We anticipate several challenges in developing and improving the performance of pre-existing models. Aside from the limitations we seek to address, some other key problems may include:

- **Generalization across subjects and poses:** Models trained on a limited set of dance sequences may struggle to generalize to unseen body types, movements, or styles outside the training data. Furthermore, the reference pose needs to be able to map to similar input poses despite differences in body proportions.
- **Error Propagation in Temporal Models:** If we employ a recurrent or tracking-based architecture to handle video data, errors in a single frame may propagate forward, corrupting the pose estimation in subsequent frames. We may need to consider how we can correct for these errors without being too conservative
- **Balancing temporal smoothing with lag:** Applying temporal smoothing techniques for joint prediction to hook up to game inputs requires a balance between lag (predictions fall too far behind to make instantaneous detections) and smoothing (minimizing jitter which may lead to multiple quick false detections).

References:

1. Ma, S., Zhang, J., Cao, Q., & Tao, D. (2024). PoseBench: Benchmarking the Robustness of Pose Estimation Models under Corruptions. arXiv. https://doi.org/10.48550/arXiv.2406.14367
2. Shafie, A. A., Kamaru Zaman, F. H., & Ali, M. H. (2009). Motion detection techniques using optical flow. ResearchGate. https://www.researchgate.net/publication/265538405_Motion_Detection_Techniques_Using_Optical_Flow
3. Jiao, Y., Shi, G., & Tran, T. D. (2021). Optical flow estimation via motion feature recovery. arXiv. https://arxiv.org/abs/2101.06333
4. G. Revach, N. Shlezinger, X. Ni, A. L. Escoriza, R. J. G. van Sloun and Y. C. Eldar, "KalmanNet: Neural Network Aided Kalman Filtering for Partially Known Dynamics," in IEEE Transactions on Signal Processing, https://ieeexplore.ieee.org/document/9733186
5. Welch, G., & Bishop, G. (2006). An introduction to the Kalman filter. University of North Carolina at Chapel Hill, Department of Computer Science. https://www.cs.unc.edu/~welch/media/pdf/kalman_intro.pdf