Pose estimation, tracking and comparison

Romeo Šajina
Faculty of Informatics
Pula, Croatia
romeo.sajina@unipu.hr

Marina Ivašić Kos
University of Rijeka
Rijeka, Croatia
marinai@uniri.hr

***Abstract*—Deep learning has become the number one research field with a lot of effort invested in computer vision. By providing possibilities as object detection, object tracking, and scene annotation, computer vision has found a lot of applications in the real world. In the field of sports, computer vision can be used to detect players, track players, detect and track the ball, detect player actions, detect objective score change, player pose estimation, etc. In this paper, we will describe player pose estimation, tracking, and comparison. This is particularly interesting as we can collect poses of a player executing an action (e.g. jump shot) and use it as a template for other players. By comparing other player's poses to the template poses we can provide them with the information of the needed corrections to their action execution. This information can help players to improve their overall action execution sequence where they can evaluate their pose in each video frame. The closes that the action sequence is to the template sequence, the better the score will they achieve. This application can be especially useful to beginners in the sport, as later on in the career a top player can develop their style of executing certain action sequences, thus trying to correct them might compromise their performance. In this paper, we will discuss player pose tracking while executing an action sequence with pose comparison and evaluation techniques.**

***Index Terms*—Pose estimation, pose tracking, pose comparison, pose sequence alignment**

I. Introduction

Machine learning (ML) has been widely applied in areas of human activities, such as sport, exercise, dance, etc. In these areas, ML is usually used for activity analysis, i.e. evaluating how was some action performed, determining was it performed correctly, and evaluating how similar was it to the action performed by a professional. Pose estimation is a very powerful technique to determine a persons' pose constructed from eighteen keypoints as shown in image 1. Collecting sequence of person poses enables further analysis of their actions and movement style. Meaning, we can collect two sequences from two different persons and compare them to calculate the difference between them. A variety of techniques can be applied to first align the two sequences and then calculate the difference between them. This type of analysis is very useful while learning or training certain actions in sport, dance moves, rehabilitation exercises, etc.

In this paper, we will explore all the necessary components needed to establish a system that will collect a sequence of person poses, align them with a sequence performed by a professional, and report back feedback containing the difference between the two sequences. The rest of the paper is organized as follows: section II describes pose estimation methodologies with a monocular camera along with different approaches that combat this problem, section III describes tracking algorithms that will enable joining detected poses in sequences in a multi-person environment, section IV finally describes methods for aligning and comparing sequences of poses.

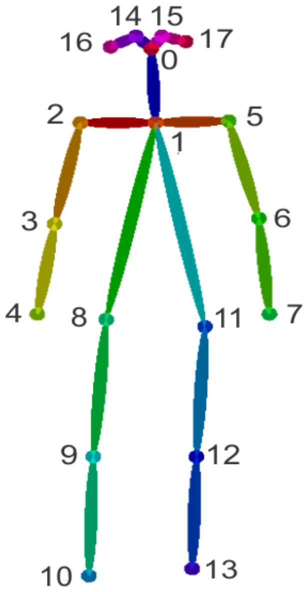

Fig. 1. Standard 18 persons keypoints in pose estimation.

II. Pose estimation

Pose estimation is a heavily researched field, finding purpose in action recognition, activity tracking, usage in augmented reality experience, usage in animation, gaming, etc. Different approaches have been introduced to achieve better results in pose estimation which can generally be divided into two categories: Single-person and Multi-person approaches as shown in 2. The single-person approach detects the pose of a person in an image given the position of the person and an implicit number of keypoints, making it essentially a regression problem. On the other hand, the objective of multi-person approach aims to solve an unconstrained problem, because the number and positions of persons within the image are unknown.

A. The single-person approach

The single-person approach is classified into two frameworks based on the keypoint prediction method: directly

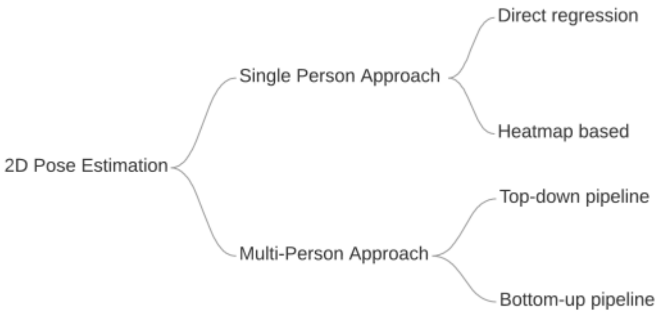

Fig. 2. Taxonomy of pose estimation approaches.

regressing keypoints from the features (i.e. direct regression based framework) or by generating heatmaps and inferencing keypoints via heatmap (i.e. heatmap based framework).

1) Direct regression based framework: Toshev and Szegedy presented DeepPose where they proposed a cascaded DNN regressor for keypoints prediction directly from feature maps. The model follows a simple architecture with convolutional layers, followed by dense layers that produce (x, y) values for keypoints. In [21] authors proposed the method that will iteratively refine the model output by feeding back error predictions, resulting in a significant increase of accuracy. Luvizon, Tabia, and Picard propose a Soft-argmax function to convert feature maps directly to joint coordinates by utilizing a *keypoint error distance based loss function* and a context-based structure to achieve competitive results compared to heatmap-based framework. Sun et al. proposed a structure-aware regression approach that adopts a reparameterized pose representation using bones instead of joints. Bones detection is an easier task because the bones are more primitive, more stable, cover a bigger area, and more robust to occlusion making it easier to learn than joints. Presented results show a performance improvement over previous direct regression based frameworks but are also very competitive with the heatmap based frameworks.

2) Heatmap based framework: Rather than directly predicting keypoints, an alternate approach can be used to create heatmaps of all keypoints within the image. Additional methods are then used to construct the final stick figure as shown in image 3. In [11] autors a graphical model with pairwise relations to make adaptive use of local image measurements. Using local image measurements can be used both to detect joints and also to predict the relationships between joints. Newell, Yang, and Deng designed a "stacked hourglass" network, closely related to encoder-decoder architecture, which is based on the successive steps of pooling and upsampling before producing the final set of predictions. They showed that repeated bottom-up, top-down processing with intermediate supervision is critical for improving the performance of human pose detection. A stacked hourglass network was commonly used in later research [39, 32, 33]. Adversarial PoseNet [32] uses a discriminator to distinguish between real poses and fake ones, which are usually the result of a complex scene or occlusion. Discriminator learns the structure of the stick figure, where he can decide if a pose is real (reasonable as body shape) or fake. Discriminator output is then used to further train the pose estimation model. Chu et al. employed a multi-context attention mechanism which will focus on the global consistency of the full human body and description for different body parts. Additionally, they introduce a novel Hourglass Residual Unit to increase the receptive field of the network. Martinez et al. introduce a baseline for 3D human pose estimation that uses an hourglass network to predict 2D keypoints which are then fed into a simple feed-forward network, outputting a 3d keypoints prediction.

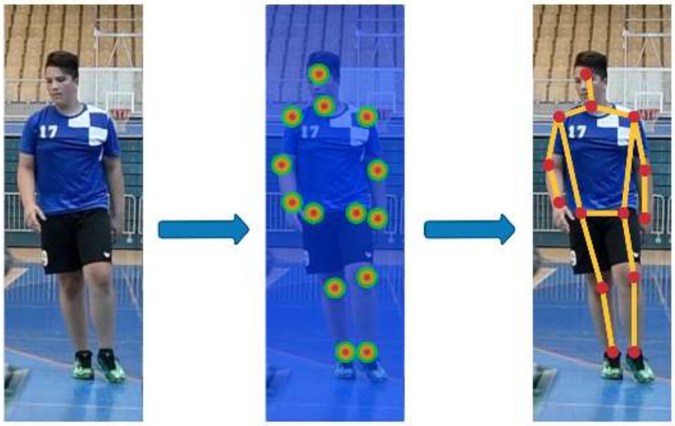

Fig. 3. Heatmap pose estimation.

B. *The multi-person approach*

The multi-person approach is a more complex task because the number and positions of persons within the image are not given, thus the framework has to detect keypoints and assemble an unknown number of persons. To combat this task, two pipelines have been proposed: top-down pipeline and bottom-up pipeline.

1) Top-down pipeline: The top-down pipeline starts by detecting all persons within an image, producing bounding boxes. The next step uses the detected bounding boxes and performs a single-person approach on each of them. The single-person approach will produce keypoints for each person that is detected, after which the pipeline may involve additional steps of post processing and improving final results as shown in image 4. The first top-down method was proposed by Toshev and Szegedy where authors used a face detector based model to determine the human body bounding box. The next stage involved a multi-stage cascade DNN based join coordinate regressor to estimate joint coordinates. He et al. build a segmentation model as an extension of Faster R-CNN [19] model by adding a branch for predicting object mask. The robustness of the proposed model made it easy to use the model for the human pose estimation where it achieved state-of-the-art results. Mask R-CNN simultaneously predicts

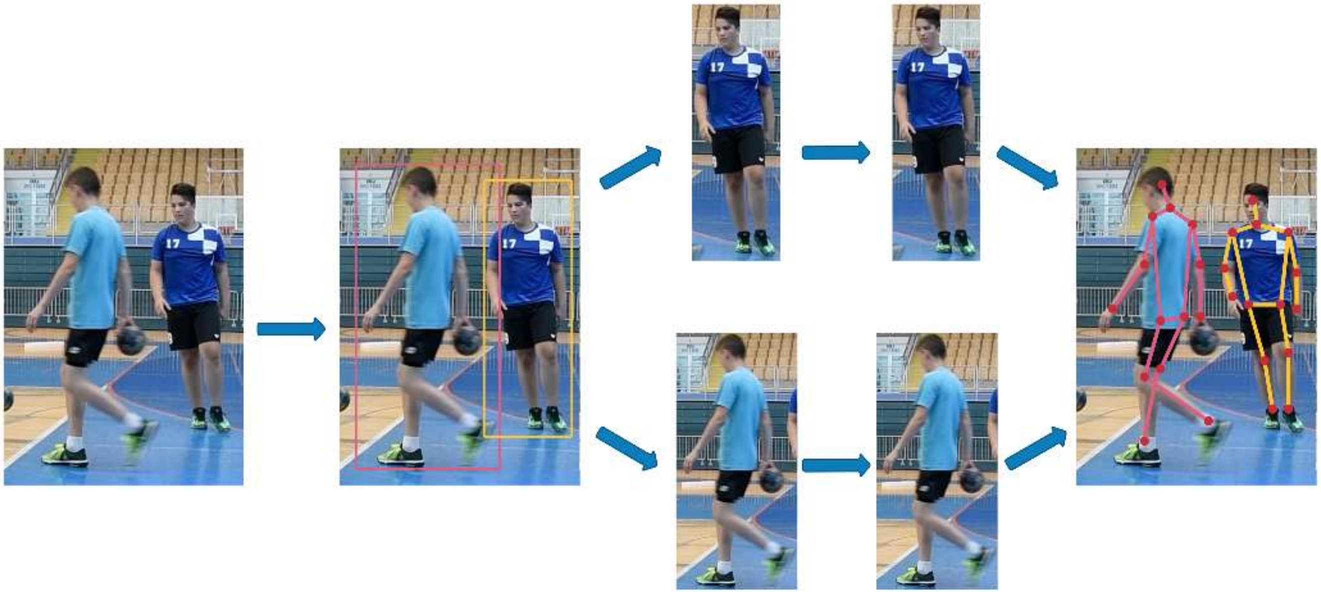

Fig. 4. The top-down pipeline in multi-person approach for pose estimation.

human bounding box and human keypoints, thus making the detection faster by sharing the features of the underlying model. Radosavovic et al. exploited omni-supervised learning that involves data augmentation prediction on unlabeled data, later used as additional data to train the model. Experiments were performed on the Mask R-CNN model and produced a robust detector able to apply self-training techniques to challenging real-world data. Fang et al. exploited the sensitivity of single-person pose estimation to bounding box detection. By employing a Symmetric Spatial Transformer Network (SSTN) and Pose-Guided Proposals Generator (PGPG) authors created a method that is able to handle inaccurate bounding boxes and redundant detections.

2) Bottom-up multi-person pipeline: The bottom-up pipeline works like a reversed top-down pipeline. The bottom-up pipeline starts by detecting all the keypoints, which are then associated with human instances as shown in image 5. Compared to the top-down pipeline, the bottom-up pipeline is likely to be faster because it does not detect human bounding boxes and run pose estimation for each human detection separately.

The bottom-up multi-person pipeline for pose estimation was first proposed by Pishchulin et al. They formulated it as a joint subset partitioning and labeling problem. The model jointly infers the number of people, their poses, spatial proximity, and part level occlusions. Their formulation implicitly performs non-maximum suppression on the set of part candidates and groups them to form configurations of body parts respecting geometric and appearance constraints. Insafutdinov et al. improved the method proposed in [27] by using a deeper neural network to achieve better body part detections, introducing a novel image-conditioned pairwise terms between body parts that improve the performance in complex scenes, and achieving faster pose estimation by combining the previous components. Insafutdinov et al. further improved the method from [22] by simplifying and sparsifying the body-part relationship graph and leveraging recent methods for faster inference, and by offloading a large share of the reasoning about body-part association onto a feed-forward convolutional architecture. Cao et al. proposed a non-parametric representation, referred to as Part Affinity Fields (PAFs), to learn to associate body parts with individuals in the image. Their model generates a set of confidence maps for body part locations, and a set of vector fields of part affinities, which are finally parsed by greedy inference to output the keypoints. Zhu, Jiang, and Luo presented an improved approach based on PAFs by including a deeper pre-trained model on COCO [12] dataset and introducing redundant PAFs which increases the robustness of joint connections. Newell, Huang, and Deng proposed associative embeddings, a method that simultaneously outputs detection and group assignments. The embeddings serve as tags that encode grouping: detection with similar tags should be grouped, i.e. body joints with similar tags should be grouped to form a single person. Findings show the method outperformed methods of likes [30, 37, 27], but also a top-down method proposed in [34].

C. *Oclusion*

Occlusion is the prevailing problem in human pose estimation problem and a couple of works tried to tackle the problem. Iqbal and Gall considered multi-person pose estimation as a joint-to-person association problem and used linear programming to resolve the association problem for each person. Chen et al. proposed a novel network structure called Cascaded Pyramid Network (CPN) that includes GlobalNet and RefineNet. GlobalNet is used to localize simple visible keypoints, and RefineNet is used to handle hard invisible or

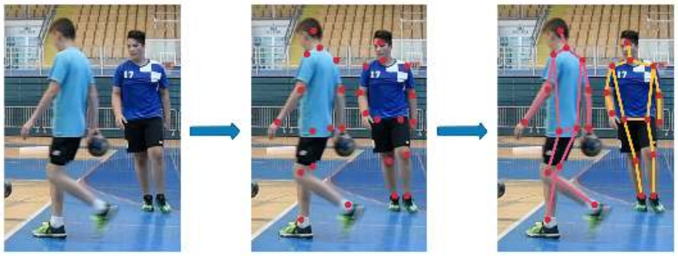

Fig. 5. The bottom-up pipeline in multi-person approach for pose estimation.

occluded keypoints. Fang et al. used Non-Maximum Suppression to solve the occlusion problem and eliminates redundant poses, the problem raised from redundant detections. A similar approach was implemented in [41] to eliminate redundant detections.

D. Metrics

In the early works, the metric that was widely used was the Percentage of Coorreclty estimated body Parts (PCP) [3], where a limb is considered detected and a correct part if the distance between the two, predicted joint locations, and the true limb joint location is at most half of the length (PCP at 0.5). Another widely used metric is PCK (probability of correct keypoint) [10] and its variant PCKh. In both metrics, a joint is considered detected and correct part if it falls within a certain amount of pixels from the ground truth joint, determined by the height and the width of the person bounding box (or person's head in the case of PCKh). More recent metrics include Percentage of Detected Joints (PDJ) [13] and Object Keypoint Similarity (OKS) [42]. PDJ considered a joint correctly detected if the distance between the predicted and the true joint is within a certain fraction of the bounding box diagonal. OKS is calculated from the distance between predicted points and ground truth points normalized by the scale of the person. The OKS metrics only show how close the predicted keypoint is to the ground truth, with a value from 0 to 1. Calculating the final performance usually involves thresholding the OKS metrics and calculating the Average Precision (AP) and Average Recall (AR) scores.

III. Tracking

Multiple object tracking (MOT) is well researched problem and in this paper, we will present the most important methods used to achieve state-of-the-art performance. Generally, a multiple object tracking problem involves using detected bounding boxes of an object and a method that associates them between sequences of frames, thus producing object trajectories. The taxonomy of tracking methods used in this paper is shown on image 6.

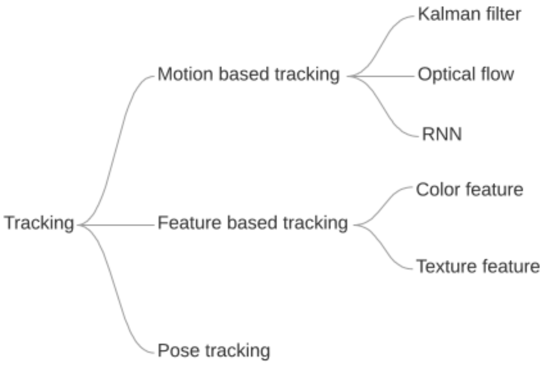

Fig. 6. Taxonomy of tracking methods.

A. Motion based tracking

Motion based tracking or tracking where objects are tracked by their motion or trajectories. Well known example is by using the Kalman filter to estimate the position of a linear system, assuming that the errors are Gaussian. Kalman filter is usually incorporated with different kinds of techniques for object feature representation or improvement in target position estimation [18, 1, 51, 79]. One of the most popular tracking systems that use the Kalman filter is SORT [20] which additionally uses the Hungarian algorithm to perform data association (connecting bounding boxes across frames). In some works [60, 15, 52, 85] optical flow was used to perform object tracking by separating the moving foreground objects from the background and generating an optical flow field vector for the moving object between subsequent frames. Other works, such as [40, 57, 53, 80], employed Recurrent Neural Network (RNN) to learn object movement behavior and use it for object tracking, usually applied on the bounding box coordinates.

B. Feature based tracking

Feature based tracking or tracking where objects are tracked by feature representation of their appearance, ie. object color, texture, shape, etc. Wojke, Bewley, and Paulus improved the method proposed in [20] by introducing deep association metric. It is accomplished by capturing object feature representation within the bounding box to enable object tracking through longer periods of occlusions, thus reducing the number of identity switches. Following works, such as [71, 81, 61], focused on improving object associations between frames using different methods or constructing a single model to perform object tracking and association. Further improvements were made by segmenting objects within the detected bounding box to eliminate unnecessary information (background, other objects, etc.) as proposed in [73], and subsequent improvements of the new approach [83, 82, 86].

C. Pose Tracking

Iqbal, Milan, and Gall first formulated the problem of multi-person pose estimation and tracking and introduced a challenging ”Multi-Person PoseTrack” dataset. The authors proposed a baseline method for solving this problem by representing body joint detection with a spatio-temporal graph and solving an integer linear program to partition the graph into sub-graphs

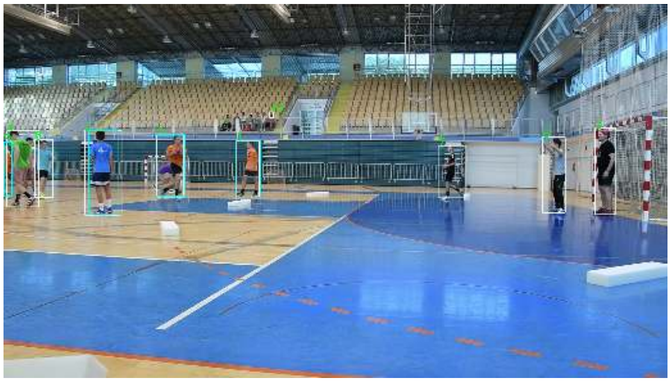

Fig. 7. Example of a scene where a player detection and tracking is executed.

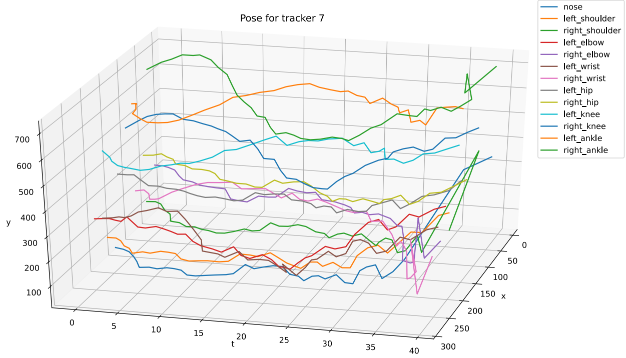

Fig. 8. A 3-dimensional plot visualization of joints in space and time when executing a jump shot.

that correspond to plausible body pose trajectories for each person. Xiu et al. proposed a PoseFlow method that consists of two techniques, namely, Pose Flow Builder (PF-Builder) and Pose Flow non-maximum suppression (PF-NMS). PF-Builder is used to associate the cross-frame poses that indicate the same person by iteratively constructing pose flow using a sliding window, where PF-NMS takes pose flow as a unit in NMS processing thus stabilizing the tracking. Doering, Iqbal, and Gall proposed a temporal model that predicts Temporal Flow Fields, i.e. vectors fields which indicate the direction where each body joint is going to move between two subsequent frames. Raaj et al. build upon Part Affinity Fields (PAF) [30] representation and propose an architecture that can encode and predict Spatio-Temporal Affinity Fields (STAF). Their model encodes change in position and orientation of keypoints across time in a recurrent fashion, i.e. the network ingests STAF heatmaps from previous frames and estimates those for the current frame. Bao et al. proposed a pose-guided tracking-by-detection framework that fuses pose information into both video human detection and human association procedures. The framework adopts the pose-guided person location prediction in the detection stage, thus exploiting the temporal information to make up missing detections. Furthermore, the authors propose PoseGCN for person association, a model task that exploits the human structural relations in addition to a person's global features. Bazarevsky et al. focused on creating a lightweight single-person pose estimation and tracking method. They followed the top-down pipeline and used a face detector along with certain calculations to determine a person's bounding box's width and height, making the detection fast. In the pose estimation step, the authors adopted a combined heatmap, offset, and regression approach where heatmaps and offset loss are only used in training. Kong et al. proposed a framework that consists of the Pose-based Triple Stream Networks (PTSN) and a multi-state online matching algorithm. PTSN is responsible for calculating the similarity scores between the history tracklets and the candidate detection in the current frame, where the scores come from three network streams that model three pose clues, i.e., pose-based appearance, motions, and athletes' interactions.

D. *Metrics*

Evaluation of tracking algorithms usually involves a couple of metrics. The most basic metric is the number of ID switches (IDs) [7] that count how many times an algorithm switched (or lost) object ID. The most widely used metric is Multiple Object Tracking Accuracy (MOTA) [2] that combines three error sources: false positives missed targets and identity switches into a single number. Another popular metric is Multiple Object Tracking Precision (MOTP) [2] that calculates misalignment between the annotated and the predicted bounding boxes. The metric Mostly Tracked targets (MT) [5] measure the completeness of tracking by calculating the ratio of ground-truth trajectories that are covered by a track hypothesis for at least 80% of their respective life span. Complementary to MT is the ML metric, or Mostly Lost targets [5], which calculates the ratio of track hypothesis that are covered for at most 20% of their respective life span.

IV. Pose comparison

Pose comparison is a complex problem that aims to provide insight into the difference between two poses. This could be useful for finding images with similar poses as the target pose, comparing the target pose with a template pose to determine the accuracy of certain action execution (eg. executing a yoga pose), etc. A vast combination of pose data (i.e. person size, person position within the image, keypoint misdetection, etc.) only complicates the task of pose comparison.

A. *Spatial alignment and normalization*

Given that the images can be different sizes, a person can appear in a different part of the image, a preprocessing step is needed to enable consistent pose comparison. The first step is to resize and scale a bounding box cropped person detection to a consistent size. The second step is to normalize the resulting keypoints coordinates by treating them as an L2 normalized vector array. Additionally, the poses can be alignment by a chosen pose point (e.g. point between the hips [70]) or by procrustes analysis as in [8, 72, 31, 58].

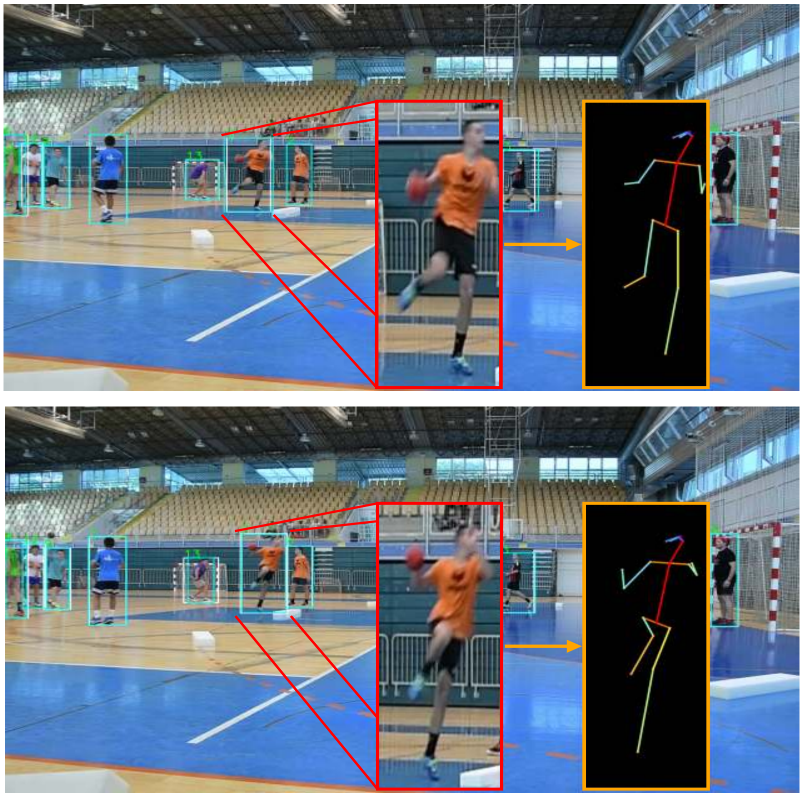

Fig. 9. Two frames of tracked player executing a jump shop where poses are estimated after performing necessary transformation.

B. Pose comparison

A simple approach to compare poses is using cosine similarity [44] that computes the similarity between two vectors, wherein these case vectors contain pose coordinates. The output of the calculation is -1 if they are exactly opposite and 1 if they are exactly the same. With a few calculations we arrive at scores: pose cosine distance that indicates pose similarity, and euclidean distance that indicates how different are the poses. The previous approach can be further improved by using keypoint confidence score, where a higher confidence score will have a bigger effect on the distance, and a lower confidence score will have less effect respectively, as expressed in [54] and used in [49]. Borkar, Pulinthitha, and Pansare proposed Matchpose, an approach that calculates the difference between angles of joints for pose comparison. For example, Matchpose can distinguish if arm position is inwards on outwards and use it to calculate similarity score.

C. Aligning sequence of poses

Pose comparison is performed on only two poses, but when we want to compare differences while two persons are performing the same action, we need to compare sequences of poses to determine the similarity of action execution. Sequences usually have different lengths and we only want to compare certain periods where an action is performed, so the main problem in this area is aligning starting poses for both sequences. A simple approach is using a sliding window

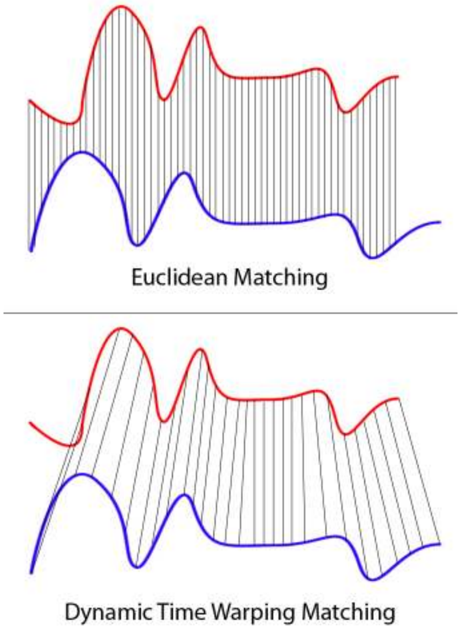

Fig. 10. Euclidean distance vs DTW comparison on two series that follow the same pattern. By calculating Euclidean distance to match the series we apply the one-to-one match, and the series are not correctly aligned. DTW overcomes the issue by applying one-to-many match so that the troughs and peaks with the same pattern are perfectly matched [Wiki Commons: File:Euclidean_vs_DTW.jpg].

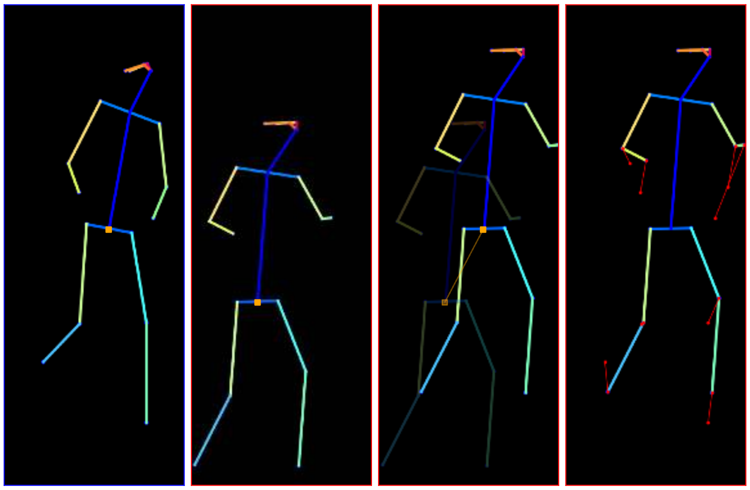

Fig. 11. Aligning poses based on the points which are calculated as midpoint between the hips of the two poses. After aligning the poses we can visualize the difference in the second pose compare to the first one. The first image shows the template pose, while the second image shows the pose that we are trying to align. The third image shows the aligned pose from the second image, and the last image shows the needed corrections to the second pose when comparing to the template pose.

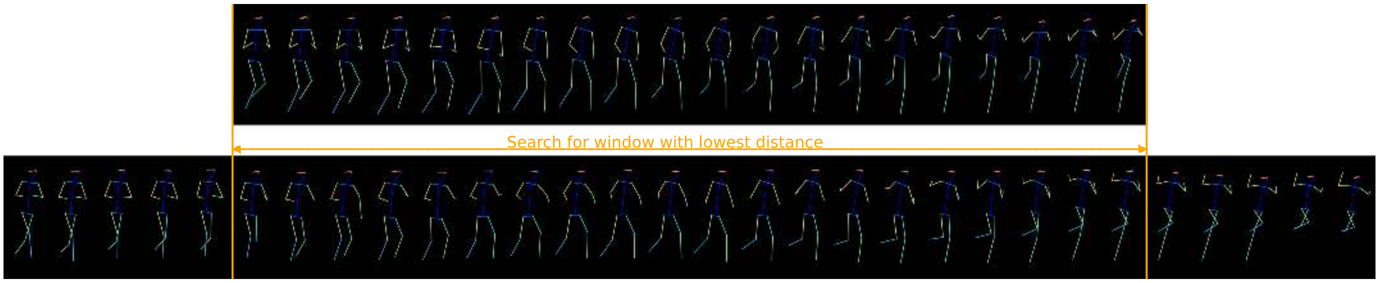

Fig. 12. Finding an appropriate sequence of poses by finding a window with a minimal distance of the two sequences.

approach where a window will capture action execution frames from the first sequence and find the window on the second sequence. The window will be slid across the second sequence to find the window of frames where the total distance between poses is minimal, as shown in image 12. As the poses can be viewed as multi-variate time series, a method called Dynamic Time Warping (DTW) [4] can be used to align sequences. DTW allows temporal sequences to be locally shifted, contracted, and stretched, and under some boundary and monotonicity constraints, it searches for a global optimal alignment path. DTW is a popular method because of its robustness against variation in speed or style in performing an action and is frequently used in the context of pose alignment [6, 16, 48, 35, 88]. An example of aligning two simple series by Euclidean Matching and DTW Matching is shown on image 10. Vemulapalli, Arrate, and Chellappa proposed a combination of DTW method and a Fourier Temporal Pyramid (FTP) [9] to construct a classifier for action recognition. Another method used for alining poses is Discrete Fourier Transform (DFT) which reveals periodicities in input data as well as the relative strengths of any periodic components and is commonly used in action recognition [59] or style transfer [29].

D. *Action style*

Capturing the style of action execution across a sequence of frames is a challenging task. This task aims to describe how a person is performing a certain action (eg. punch, kick, running, jump, etc.) and recognize which are the most notable features of action execution. This can be used to help improve a person's action execution by pointing out differences compared to a target action execution but also be used for person identification. Kviatkovsky, Shimshoni, and Rivlin used Principal Component Analysis (PCA) to decrease the dimensionality of an action sequence and create a representation that is compact and less noisy, and additionally, they applied Linear Discriminant Analysis (LDA) on the PCA-transformed representations to improve person identification, which is the ultimate goal of their paper. Recent works [63, 76] focused on creating an embedding for representing persons' style, which offers a number of interesting possibilities as is representing certain style in high dimensionality space as a vector that enables measuring ”style distance” between them.

V. Action comparison in sport

Hachaj, Piekarczyk, and Ogiela used motion sensors to capture karate action poses. They focused on improving the

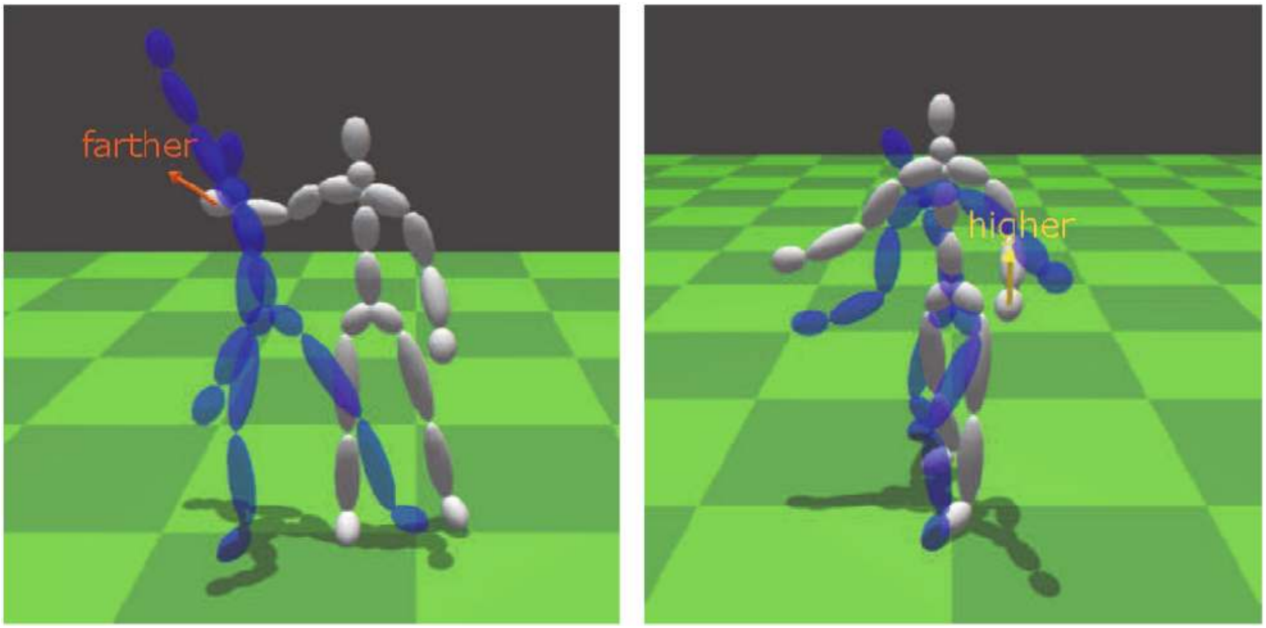

Fig. 13. Tenis shot evaluation - Visualizations of spatial, rotational, and temporal features with arrows. a The distance between the right and left feet, and the height of the left hand [67]

DTW method by normalizing distances between joints of two poses. The overall quality of action execution is measured as the median value of dynamic time warping normalized distances (DTWND). Oshita et al. used motion sensors to capture persons' keypoints while executing an action sequence. They collected a couple of sequences performing tennis forehand shot by professionals and compared them to action sequences performed by a novice. They extracted three crucial pose positions from the sequence to show the comparison, and the comparison was calculated as the difference between one-dimensional feature vector containing spatial, rotational, and temporal features based on the key sporting poses. Starting and ending pose for a sequence was calculated based on the middle pose, or the timing of hitting the ball, thus both sequences were aligned and ready for comparison. Example of their work is shown on image 13. Subsequent work [66] focused on improving motion visualization introducing new methods to visualize trajectories to create a more understandable comparison. Voulodimos, Rallis, and Doulamis used motion capture technologies as well as physically based modeling principles to introduce two methods for selecting key poses from the sequence: a clustering-based method for the selection of the basic primitives of choreography, and a kinematics-based approach that generates meaningful summaries at hierarchical levels of granularity. Rallis et al. build upon an idea similar to [84] and proposed a Bidirectional LSTM neural network to classify a short sequence of poses to determine to which basic primitive of choreography the poses belong. Deb et al. employed OpenPose pose estimation network to detect the pose of dancers performing a specific dance move. Then the pose is compared to a gallery of pre-recorded poses to determine similarity score using a weighted Cosine distance. Kamel et al. used a depth camera and custom CNN neural network for pose estimation in Tai Chi training. The system evaluates participants' action against a template motion by counting out 15 seconds and starting template motion at 0. Scores are computed for all body parts and presented to the participant. They proved the significant benefits of the system by evaluating the system against two groups: one learning by watching videos, and the other by using the training system. Several works [24, 28] explored using hidden Markov Models for gesture recognition and evaluation, where each hidden state is associated with a collection of similar body poses and a transition model encapsulates sequences of body-part configurations. Yasser et al. detected the misplaced joints of the athletes while lifting. It compared different models to see which one was more accurate in classifying correct postures in fundamental lifts such as deadlifts, shoulder presses, and squats. Wu and Koike developed a martial art training system based on real-time human pose forecasting while using VR to show the predicted movements of the trainees. It used recurrent networks (LSTM) to learn the temporal features of human motion and passes them to the 3D recovery network.

VI. Limitations and future work

In this paper, we presented the research on the task of pose comparison, pose alignment, and sequence comparison. Most of the presented research was conducted using motion sensors to collect persons' poses because it offers high accuracy and smooth movement of the keypoints across time. Managing the same level of accuracy and consistency using only one camera to estimate a person's pose is a very challenging task, but ultimately a cheaper and easier solution to apply in real-world situations. Future research should focus on employing pose estimation models and applying certain techniques to improve the consistency of detections and creating a smooth movement sequence. Techniques for pose comparison provide a good baseline for calculating the differences between keypoints. However, those methods assume that all poses have the same body proportions. Even though some papers tried to combat this issue, further progress can be made by creating and introducing *body proportions invariant method* for pose comparison. Another problem worth addressing is the speed of executing an action or part of the action. If two persons perform tennis forehand shot at different speeds but both with the same hand trajectory, the techniques described in this paper will not recognize this behavior, especially if it happens in only a certain part of the action. This is really important because the action was performed correctly trajectory-wise, but needs a different kind of correction, i.e. speed of action execution.

References

[1] Anders Erik Nordsjo. “A constrained extended Kalman filter for target tracking”. In: *Proceedings of the 2004 IEEE Radar Conference (IEEE Cat. No. 04CH37509)*. IEEE. 2004, pp. 123–127.

[2] Keni Bernardin and Rainer Stiefelhagen. “Evaluating multiple object tracking performance: the clear mot metrics”. In: *EURASIP Journal on Image and Video Processing* 2008 (2008), pp. 1–10.

[3] Vittorio Ferrari, Manuel Marin-Jimenez, and Andrew Zisserman. “Progressive search space reduction for human pose estimation”. In: *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2008, pp. 1–8.

[4] Pavel Senin. “Dynamic time warping algorithm review”. In: *Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA* 855.1-23 (2008), p. 40.

[5] Yuan Li, Chang Huang, and Ram Nevatia. “Learning to associate: Hybridboosted multi-target tracker for crowded scene”. In: *2009 IEEE conference on computer vision and pattern recognition*. IEEE. 2009, pp. 2953–2960.

[6] Samsu Sempena, Nur Ulfa Maulidevi, and Peb Ruswono Aryan. “Human action recognition using dynamic time warping”. In: *Proceedings of the 2011 International Conference on Electrical Engineering and Informatics*. IEEE. 2011, pp. 1–5.

[7] Kota Yamaguchi et al. “Who are you with and where are you going?” In: *CVPR 2011*. IEEE. 2011, pp. 1345–1352.

[8] Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. “Reconstructing 3d human pose from 2d image landmarks”. In: *European conference on computer vision*. Springer. 2012, pp. 573–586.

[9] Jiang Wang et al. “Mining actionlet ensemble for action recognition with depth cameras”. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2012, pp. 1290–1297.

[10] Yi Yang and Deva Ramanan. “Articulated human detection with flexible mixtures of parts”. In: *IEEE transactions on pattern analysis and machine intelligence* 35.12 (2012), pp. 2878–2890.

[11] Xianjie Chen and Alan Yuille. “Articulated pose estimation by a graphical model with image dependent pairwise relations”. In: *arXiv preprint arXiv:1407.3399* (2014).

[12] Tsung-Yi Lin et al. “Microsoft coco: Common objects in context”. In: *European conference on computer vision*. Springer. 2014, pp. 740–755.

[13] Alexander Toshev and Christian Szegedy. “Deeppose: Human pose estimation via deep neural networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 1653–1660.

[14] Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa. “Human action recognition by representing 3d skeletons as points in a lie group”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 588–595.

[15] Kiran Kale, Sushant Pawar, and Pravin Dhulekar. “Moving object tracking using optical flow and motion vector estimation”. In: *2015 4th international conference on reliability, infocom technologies and optimization (ICRITO)(trends and future directions)*. IEEE. 2015, pp. 1–6.

[16] Kaustubh Kulkarni et al. “Continuous action recognition based on sequence alignment”. In: *International Journal of Computer Vision* 112.1 (2015), pp. 90–114.

[17] Igor Kviatkovsky, Ilan Shimshoni, and Ehud Rivlin. “Person identification from action styles”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2015, pp. 84–92.

[18] Nima Najafzadeh, Mehran Fotouhi, and Shohreh Kasaei. “Object tracking using Kalman filter with adaptive sampled histogram”. In: *2015 23rd Iranian Conference on Electrical Engineering*. IEEE. 2015, pp. 781–786.

[19] Shaoqing Ren et al. “Faster r-cnn: Towards real-time object detection with region proposal networks”. In: *arXiv preprint arXiv:1506.01497* (2015).

[20] Alex Bewley et al. “Simple online and realtime tracking”. In: *2016 IEEE international conference on image processing (ICIP)*. IEEE. 2016, pp. 3464–3468.

[21] Joao Carreira et al. “Human pose estimation with iterative error feedback”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 4733–4742.

[22] Eldar Insafutdinov et al. “Deepercut: A deeper, stronger, and faster multi-person pose estimation model”. In: *European Conference on Computer Vision*. Springer. 2016, pp. 34–50.

[23] Umar Iqbal and Juergen Gall. “Multi-person pose estimation with local joint-to-person associations”. In: *European Conference on Computer Vision*. Springer. 2016, pp. 627–642.

[24] Sohaib Laraba and Joëlle Tilmanne. "Dance performance evaluation using hidden Markov models". In: *Computer Animation and Virtual Worlds* 27.3-4 (2016), pp. 321–329.

[25] Alejandro Newell, Zhiao Huang, and Jia Deng. "Associative embedding: End-to-end learning for joint detection and grouping". In: *arXiv preprint arXiv:1611.05424* (2016).

[26] Alejandro Newell, Kaiyu Yang, and Jia Deng. "Stacked hourglass networks for human pose estimation". In: *European conference on computer vision*. Springer. 2016, pp. 483–499.

[27] Leonid Pishchulin et al. "Deepcut: Joint subset partition and labeling for multi person pose estimation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 4929–4937.

[28] Lili Tao et al. "A comparative study of pose representation and dynamics modelling for online motion quality assessment". In: *Computer vision and image understanding* 148 (2016), pp. 136–152.

[29] M Ersin Yumer and Niloy J Mitra. "Spectral style transfer for human motion between independent actions". In: *ACM Transactions on Graphics (TOG)* 35.4 (2016), pp. 1–8.

[30] Zhe Cao et al. "Realtime multi-person 2d pose estimation using part affinity fields". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 7291–7299.

[31] Ching-Hang Chen and Deva Ramanan. "3d human pose estimation= 2d pose estimation+ matching". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 7035–7043.

[32] Yu Chen et al. "Adversarial posenet: A structure-aware convolutional network for human pose estimation". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 1212–1221.

[33] Xiao Chu et al. "Multi-context attention for human pose estimation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 1831–1840.

[34] Hao-Shu Fang et al. "Rmpe: Regional multi-person pose estimation". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 2334–2343.

[35] Tomasz Hachaj, Marcin Piekarczyk, and Marek R Ogiela. "Human actions analysis: templates generation, matching and visualization applied to motion capture of highly-skilled karate athletes". In: *Sensors* 17.11 (2017), p. 2590.

[36] Kaiming He et al. "Mask r-cnn". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2961–2969.

[37] Eldar Insafutdinov et al. "Arttrack: Articulated multi-person tracking in the wild". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 6457–6465.

[38] Umar Iqbal, Anton Milan, and Juergen Gall. "Posetrack: Joint multi-person pose estimation and tracking". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 2011–2020.

[39] Julieta Martinez et al. "A simple yet effective baseline for 3d human pose estimation". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 2640–2649.

[40] Guanghan Ning et al. "Spatially supervised recurrent convolutional neural networks for visual object tracking". In: *2017 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE. 2017, pp. 1–4.

[41] George Papandreou et al. "Towards accurate multi-person pose estimation in the wild". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 4903–4911.

[42] Matteo Ruggero Ronchi and Pietro Perona. "Benchmarking and error diagnosis in multi-instance pose estimation". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 369–378.

[43] Xiao Sun et al. "Compositional human pose regression". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 2602–2611.

[44] Leilei Wang, Zhigang Chen, and Jia Wu. "An opportunistic routing for data forwarding based on vehicle mobility association in vehicular ad hoc networks". In: *Information* 8.4 (2017), p. 140.

[45] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. "Simple online and realtime tracking with a deep association metric". In: *2017 IEEE international conference on image processing (ICIP)*. IEEE. 2017, pp. 3645–3649.

[46] Xiangyu Zhu, Yingying Jiang, and Zhenbo Luo. "Multi-person pose estimation for posetrack with enhanced part affinity fields". In: *ICCV PoseTrack Workshop*. Vol. 7. 2017.

[47] Yilun Chen et al. "Cascaded pyramid network for multi-person pose estimation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 7103–7112.

[48] Hyo-Rim Choi and TaeYong Kim. "Modified dynamic time warping based on direction similarity for fast gesture recognition". In: *Mathematical Problems in Engineering* 2018 (2018).

[49] Suman Deb et al. "Interactive dance lessons through human body pose estimation and skeletal topographies matching". In: *International Journal of Computational Intelligence & IoT* 2.4 (2018).

[50] Andreas Doering, Umar Iqbal, and Juergen Gall. "Joint flow: Temporal flow fields for multi person tracking". In: *arXiv preprint arXiv:1805.04596* (2018).

[51] Pramod R Gunjal et al. "Moving object tracking using kalman filter". In: *2018 International Conference On Advances in Communication and Computing Technology (ICACCT)*. IEEE. 2018, pp. 544–547.

[52] Junjie Huang et al. "Optical flow based real-time moving object detection in unconstrained scenes". In: *arXiv preprint arXiv:1807.04890* (2018).

[53] F Lotfi, V Ajallooeian, and HD Taghirad. “Robust object tracking based on recurrent neural networks”. In: *2018 6th RSI International Conference on Robotics and Mechatronics (IcRoM)*. IEEE. 2018, pp. 507–511.

[54] George Papandreou et al. “Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 269–286.

[55] Ilija Radosavovic et al. “Data distillation: Towards omni-supervised learning”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 4119–4128.

[56] Yuliang Xiu et al. “Pose flow: Efficient online pose tracking”. In: *arXiv preprint arXiv:1802.00977* (2018).

[57] Yashu Zhang, Yue Ming, and Runqing Zhang. “Object detection and tracking based on recurrent neural networks”. In: *2018 14th IEEE International Conference on Signal Processing (ICSP)*. IEEE. 2018, pp. 338–343.

[58] Xiaowei Zhou et al. “Monocap: Monocular human motion capture using a cnn coupled with a geometric prior”. In: *IEEE transactions on pattern analysis and machine intelligence* 41.4 (2018), pp. 901–914.

[59] Zeeshan Ahmad and Naimul Mefraz Khan. “Multidomain multimodal fusion for human action recognition using inertial sensors”. In: *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*. IEEE. 2019, pp. 429–434.

[60] A Balasundaram, S Ashok Kumar, and S Magesh Kumar. “Optical flow based object movement tracking”. In: *Int. J. Eng. Adv. Technol.(IJERT)* 9 (2019), pp. 3913–3916.

[61] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. “Tracking without bells and whistles”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 941–951.

[62] Pradnya Krishnanath Borkar, Marilyn Mathew Pulinthitha, and A Pansare. “Match Pose-A System for Comparing Poses”. In: *International Journal of Engineering Research and Technology (IJERT)* 8.10 (2019).

[63] Han Du et al. “Stylistic Locomotion Modeling with Conditional Variational Autoencoder.” In: *Eurographics (Short Papers)*. 2019, pp. 9–12.

[64] Aouaidjia Kamel et al. “An investigation of 3D human pose estimation for learning Tai Chi: A human factor perspective”. In: *International Journal of Human–Computer Interaction* 35.4-5 (2019), pp. 427–439.

[65] Diogo C Luvizon, Hedi Tabia, and David Picard. “Human pose regression by combining indirect part detection and contextual information”. In: *Computers & Graphics* 85 (2019), pp. 15–22.

[66] Masaki Oshita. “Motion Volume: Visualization of Human Motion Manifolds”. In: *The 17th International Conference on Virtual-Reality Continuum and its Applications in Industry*. 2019, pp. 1–7.

[67] Masaki Oshita et al. “Development and evaluation of a self-training system for tennis shots with motion feature assessment and visualization”. In: *The Visual Computer* 35.11 (2019), pp. 1517–1529.

[68] Yaadhav Raaj et al. “Efficient online multi-person 2d pose tracking with recurrent spatio-temporal affinity fields”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 4620–4628.

[69] Ioannis Rallis et al. “Bidirectional long short-term memory networks and sparse hierarchical modeling for scalable educational learning of dance choreographies”. In: *The Visual Computer* (2019), pp. 1–16.

[70] Faegheh Sardari, Adeline Paiement, and Majid Mirmehdi. “View-invariant pose analysis for human movement assessment from rgb data”. In: *International Conference on Image Analysis and Processing*. Springer. 2019, pp. 237–248.

[71] ShiJie Sun et al. “Deep affinity network for multiple object tracking”. In: *IEEE transactions on pattern analysis and machine intelligence* 43.1 (2019), pp. 104–119.

[72] Hüseyin Temiz, Berk Gökberk, and Lale Akarun. “Multi-view reconstruction of 3D human pose with procrustes analysis”. In: *2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA)*. IEEE. 2019, pp. 1–5.

[73] Paul Voigtlaender et al. “Mots: Multi-object tracking and segmentation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 7942–7951.

[74] Erwin Wu and Hideki Koike. “Futurepose-mixed reality martial arts training using real-time 3d human pose forecasting with a rgb camera”. In: *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2019, pp. 1384–1392.

[75] Ammar Yasser et al. “Smart coaching: Enhancing weightlifting and preventing injuries”. In: *International Journal of Advanced Computer Science and Applications* 10 (2019), p. 01.

[76] Kfir Aberman et al. “Unpaired motion style transfer from video to animation”. In: *ACM Transactions on Graphics (TOG)* 39.4 (2020), pp. 64–1.

[77] Qian Bao et al. “Pose-guided tracking-by-detection: Robust multi-person pose tracking”. In: *IEEE Transactions on Multimedia* 23 (2020), pp. 161–175.

[78] Valentin Bazarevsky et al. “BlazePose: On-device Real-time Body Pose tracking”. In: *arXiv preprint arXiv:2006.10204* (2020).

[79] Fahime Farahi and Hadi Sadoghi Yazdi. “Probabilistic Kalman filter for moving object tracking”. In: *Signal Processing: Image Communication* 82 (2020), p. 115751.

[80] Wenjing Kang et al. "Online Multiple Object Tracking with Recurrent Neural Networks and Appearance Model". In: *2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. IEEE. 2020, pp. 34–38.

[81] Shyamgopal Karthik, Ameya Prabhu, and Vineet Gandhi. "Simple unsupervised multi-object tracking". In: *arXiv preprint arXiv:2006.02609* (2020).

[82] Lorenzo Porzi et al. "Learning multi-object tracking and segmentation from automatic annotations". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 6846–6855.

[83] Young-min Song and Moongu Jeon. "Online Multi-Object Tracking and Segmentation with GMPHD Filter and Simple Affinity Fusion". In: *arXiv preprint arXiv:2009.00100* (2020).

[84] Athanasios Voulodimos, Ioannis Rallis, and Nikolaos Doulamis. "Physics-based keyframe selection for human motion summarization". In: *Multimedia Tools and Applications* 79.5 (2020), pp. 3243–3259.

[85] Qingwen Xu et al. "An Optical Flow Based Multi-Object Tracking Approach Using Sequential Convex Programming". In: *2020 16th International Conference on Control, Automation, Robotics and Vision (ICARCV)*. IEEE. 2020, pp. 1216–1221.

[86] Yifu Zhang et al. "FairMOT: On the fairness of detection and re-identification in multiple object tracking". In: *arXiv e-prints* (2020), arXiv–2004.

[87] Longteng Kong et al. "Online Multiple Athlete Tracking with Pose-Based Long-Term Temporal Dependencies". In: *Sensors* 21.1 (2021), p. 197.

[88] Zhelong Wang et al. "Motion Analysis of Deadlift for Trainers with Different Levels based on Body Sensor Network". In: *IEEE Transactions on Instrumentation and Measurement* 70 (2021), pp. 1–12.