14TH ANNUAL
MIT SLOAN
SPORTS ANALYTICS CONFERENCE
MARCH 6 - 7, 2020 | BOSTON, MA
presented by ESPN

Gan, Q., Nguyen, S. M., Fenaux, E., Clemencon, S., El-Yacoubi, M. A., and Jelassi, O. (2023). Reconstruction of Trajectories of Athletes Using Computer Vision Models and Kinetic Analysis. MIT Sloan Sports Analytics Conference.

Reconstruction of Trajectories of Athletes Using Computer Vision Models and Kinetic Analysis

Paper Track: Other sports
Paper ID: 443997

Qi GAN[1], Sao Mai Nguyen[2], Eric Fenaux[3], Stéphan Clémençon[1], Mounim El Yacoubi[4], Ons Jelassi[1]

[1]LTCI, Telecom Paris, IP Paris, France [2]Flowers Team, U2IS, ENSTA Paris, IP Paris & Inria, France [3]ef-e-sciences [4]Samovar, Télécom SudParis, IP Paris, France

Abstract

Athlete’s pose acquisition and analysis is promising to provide coaches with details of athletes performance and thus help to improve athletes’ performances with more detailed supervision from coaches. Compared with traditional ways of acquiring an athlete's gesture, such as using wearable sensors, computer vision technology has advantages of low-cost, high-efficient and non-intrusive. This paper aims to bridge these two fields, by reconstructing athletes’ trajectory using monocular (i.e. single-camera-shot) videos. Under a few assumptions that are applicable to most of the sports of athletics, we proposed a method combining computer vision techniques and physics laws to reconstruct athletes’ trajectories from monocular videos. The method first estimates 3D pose of athletes from video inputs, then performs kinematic analysis on estimated poses to reconstruct the trajectories of athletes. We tested this algorithm on videos from the triple jump finals of the 2016 Olympics in Rio de Janeiro. We achieved the best performance with 9.1% mean average error when using ground-truth foot-ground contact signal and 21.4% mean average error when using predicted foot-ground contact signal.

1. Introduction

In this paper, we focus on the reconstruction of an athlete's trajectory from dynamic monocular RGB videos of athletics sports games. The reconstruction of a player's pose or trajectory is helpful to provide detailed physics parameters (positions, velocities, accelerations, forces, etc.) to coaches to help correct poses and improve the performance of athletes. Thanks to advances in computer vision and artificial intelligence, estimating human poses from video recordings has become possible [1-6]. Videos for athletes are usually shot with a single moving camera or independent, unsynchronized moving cameras so that it can track the athlete and to ensure that the athlete appears large enough and remains in the center of the frame. Such videos are characteristic of TV broadcasting footage of sports games or videos shot by trainers during training sessions. Therefore, it is important to reconstruct athletes’ trajectory from dynamic monocular RGB videos.

The estimation of 3D pose of humans from monocular RGB videos is challenging because of the nature of recovering 3D coordinates from 2D images. To recover the trajectory from dynamic monocular videos is more cumbersome with the uncertainty of camera positions and a variable background. To recover the trajectory, one needs the global position of every pose across the whole video. The most common solution is to use triangularization with multiple-camera-shot videos [7-10], which is not applicable in monocular videos. There are papers that solve this problem by introducing the SLAM algorithm [11]. However, it is mentioned by Yuan *et al*. [12] that moving objects in the environment could cause SLAM to fail. The authors proposed an algorithm to estimate the translation and rotation matrices , but this algorithm requires a dataset of dynamic cameras for training. In our application, public sports-specific datasets of dynamic monocular footage are still rare. Therefore, we propose an algorithm to take advantage of state-of-the-art human pose estimation models, and introduce physics laws to help analyze and recover human body trajectories from videos.

In this work, we rely on pre-trained existing models to predict 3D skeletons of athletes from dynamic monocular videos. We estimate the foot-ground contact signal to help split athletes' movements into foot-on-ground phase and foot-off-ground phase. Then using kinematic analysis, we recover athletes' trajectories of each phase. The advantage of this algorithm is that it does not require an extra dataset for training and could be applied to most TV footage of sports of athletics.

2. Related work

Analyzing an athlete's trajectory thus relies on human pose estimation, physics models and action analysis.

Human pose estimation. In recent years, there are numerous models proposed to estimate 3D human pose by key points, but most of which estimate key points in camera coordinates or root-relative coordinates [1-6]. To obtain global/absolute positions, one could use multiple synchronized cameras with triangularization [7-10] or extra information like Inertial Mounted Unit (IMU) sensor measurements [13, 14]. Some works consider recovering occlusion by estimating translation and rotation matrices [12], which are learned on specific datasets*.*

In the computer vision field, there are two popular ways to represent a human body, i.e. skeleton representation [24, 25] and body mesh representation [26]. While the body mesh representation focuses more on the surface of the body in a 2D planar or a 3D volumetric manner, the skeleton representation represents a human body by a few key points and the connections between the key points. It thus directly provides us with synthetic features such as the body part positions and orientations, and can be easily processed by physical law models. To integrate dynamic analysis in our approach, we opt for the key points representation.

The methods estimating key points of human pose from monocular cameras have made tremendous improvements in the last years. The most successful algorithms such as Openpose, HRNet [29], Alphapose, estimate 2D coordinates of each key point, in the plane of the camera image. However, as the plane of the camera image is not always parallel to the athlete's trajectory, these results can not be easily exploited for our problem. More recently, 3D estimation of key point coordinates have also developed, such as RepNet[33], SRNet [32] or MHFormer [4]. MHFormer utilizes Yolov3 [28] for detecting bounding boxes of humans in frames and HRNet [29] for predicting 2D key points of humans. With the detected 2D key points as inputs, it estimates 3D key points coordinates: this process is called “lifting”.

In this work, we need a 3D human pose estimation to analyze the athlete’s trajectory, thus we choose MHFormer for its highest scores on benchmarks.

Sports action analysis. Computer vision techniques have been applied to athletes’ action recognition and analysis [19-21]. A review [22] on the application of camera-based HPE lists recent research in the field of sports and physical exercise.For example, Wang *et al.* [19] applied computer vision to estimate the pose of skaters and their boards, in order to evaluate the movements of skaters. Although the videos used in this work are shot by dynamic cameras, they only need the information about the poses of athletes, instead of their trajectories. Li *et al.* [20] applied a human posture estimation algorithm to an auxiliary training system for golf, which is shot by fixed cameras. Fastovets *et al.* [21], focused also on monocular TV sports footage, but their work is to estimate 2D poses, instead of 3D poses or trajectories. This work tackles 3D poses and trajectories, despite using a 2D video from a moving monocular camera.

Physics-law based models. Works considering physics laws by introducing physics engines [15-18] first estimate human pose (2D or 3D) from videos, then optimize human models so that the pose fits the human pose predictions. Although these models could provide poses constrained by physics law, to our knowledge, there is not yet one that could apply to dynamic cameras. Indeed, this would imply that the physics law can model the dynamics of the camera trajectories. In this work, we get inspired by this idea that physics law is important, and can compensate for the uncertainty introduced by a mobile camera.

In the proposed approach, we tackle sports action analysis for 3D trajectories and use physics law based models to recover the trajectory from a moving camera.

3. Methods

Our proposed method for reconstructing human trajectory from video inputs combines two analyses: human pose estimation and kinematic analysis. An overview of the approach is presented in Fig. 1. Based on these outputs, the signals of body-ground contact information and body center of mass of each pose are calculated. Finally, kinematic analysis is performed on previous results to reconstruct the trajectory of the human in the input video.

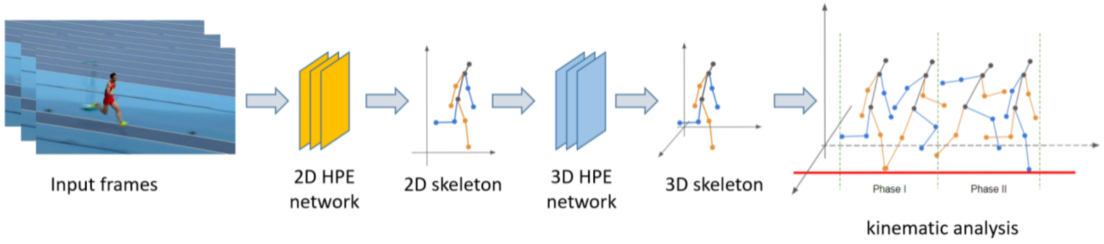

Figure 1. Overview of proposed solution

This algorithm introduces a few assumptions in order to perform kinematic analysis. Firstly, it assumes that the ground is flat, i.e. the gravity is perpendicular to the ground floor, which is true in most of the athletics. The second assumption is that the feet of the athlete do not slide when running, which is also valid for running and jumping movements that are considered in this study.. Beside these assumptions, we limit our applications to single-athlete videos, as usually in videos during training sessions.

Human Pose Estimation

2D & 3D HPE networks: In this work, we use the lifting network MHFormer [4] with the skeleton representation shown in Fig. 2, which follows the definition of dataset Human3.6M [27]. The output of the 3D key point coordinates are relative, with key point 'Hip' as root, with coordinates fixed at zero. The networks in this algorithm utilize pre-trained models. Therefore, no training process is required, though it would be helpful to train the model on an athletics-specific dataset.

First, we could observe for the 2D HPE network which estimates the 2D skeleton of input frames of a video, joint swaps between left and right body parts [23] and noise. We corrected these obvious errors of 2D skeletons as detailed in the next subsection and the results are sent to the 3D HPE lifting network to generate 3D skeletons.

Then, as the 3D key point coordinates of the predicted poses by MHFormer are relative ones, and the bone length is not explicitly set to be fixed throughout the video by MHFormer, we normalized the bone lengths by ensuring the sum of every estimated 3D pose to be consistent. Then using the official height of each athlete, we could estimate the absolute lengths of bones by approximating the height with the sum of lengths of leg, spine, neck and head.

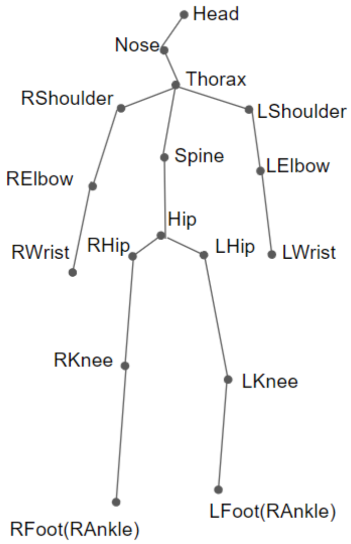

Figure 2. Skeleton representation of human body and key point definitions

2D Prediction corrections: In order to improve the performance of the 3D key point estimation, we propose an algorithm to correct 2D predicted key points. In practice, 2D predictions are prone to swap errors, i.e. wrongly assigning left and right leg key points. Under the assumption that swap errors occur rarely, we claim that the existence of one swap causes the trajectory of both key points to be longer (see an example in Fig. 3). Based on this principle, we detect swap errors point by point. In addition to swap errors, there are also some missing detections that are corrected by interpolation.

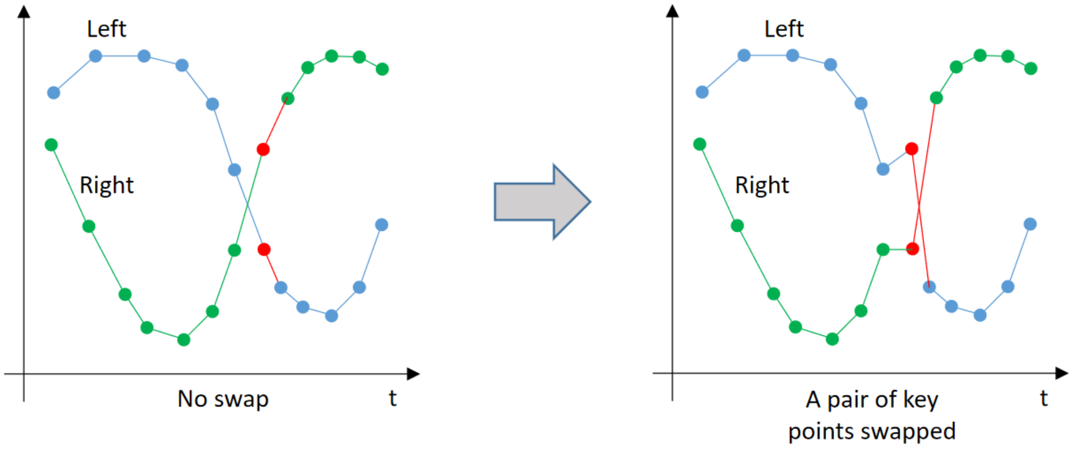

Figure 3. An example showing that swap error causes key point trajectory to be longer

Body mass center estimation: In order to use kinematic analysis by applying Newton's laws, we need to estimate the body mass center. It can be obtained from the position of each body part and its mass. Some literature studied the distribution of body mass [30, 31]. We use Table. 1 from [30] to approximately calculate the center of mass for any 3D pose. We first estimate the centers of mass of different parts of the body, which are represented by connections of the skeleton representation. For connections below the head (key point 'Thorax'), the center of body mass is estimated by the midpoint of the connections. The center of gravity of the combination of head and neck is estimated by the midpoint of keypoints 'Head' and 'Thorax'. Finally, as there is no key point representing hands and feet, but the weights of hands and feet are not negligible, we roughly approximate the centers of gravity of hands with key points 'LWrist' and 'RWrist' and similarly the centers of gravity of feet are approximated with key points 'LAnkle' and 'RAnkle'.

Segment	Mass percentage (%)	
	female	male
Head	6.68	6.94
Trunk	42.57	43.46
Upper Trunk	15.45	15.96

Segment	Mass percentage (%)	
	female	male
Forearm	1.38	1.62
Hand	0.56	0.61
Thigh	14.78	14.16

Mid Trunk	14.65	16.33	Shank	4.81	4.33
Lower Trunk	12.47	11.17	Foot	1.29	1.37
Upper Arm	2.55	2.71	-	-	-

Table 1. Mass distribution of different body segments [30]

Kinematic analysis: Phase splitting, Phase I, Phase II.

Our kinematic analysis will differentiate the phases the athlete is touching the ground and the phases s/he is on air.

Phase splitting: The athlete's running process could be split into two different phases (see Fig. 4). The first phase (Phase I) is when the whole body of an athlete is in the air. The second phase (Phase II) is defined as when there is some part of the body touching the ground, especially one foot or both of the feet touch ground. We split the phases by detecting whether the athlete's body contacts the ground.

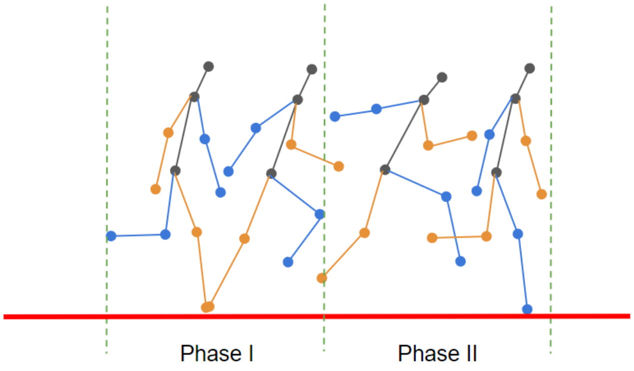

Figure 4. Splitting of two phases during running

There are multiple solutions to split the phases. One possible solution is to train a machine learning model with a supervised learning scheme to detect whether the foot or a part of the body is contacting the ground floor or not. However, we do not have datasets labeled with foot/body contacting ground information to train such models. Therefore, we chose a simpler way, which is based on signal processing algorithms on 2D key point estimation results.

We consider two 2D predicted key points: left and right foot ('LFoot' and 'RFoot'). The predicted 2D key points are in image coordinates, which are influenced by the dynamics of the camera. However, when the camera moves slowly and smoothly, the key points in a frame sequence appear to be periodic, which corresponds to the steps of the athlete. The algorithm considers the change of y coordinate of each key point versus time. For either key point ('LFoot' or 'RFoot'), if it has a local minima at time t and the y coordinate value is lower than its counterpart, it will be detected as contacting ground at time t. For t's neighboring time if the y coordinate value of the same key point does not exceed the local minima by a threshold, that time will be detected as the foot still on the ground.

Phase I: During Phase I the whole body of the athlete is in the air, which means the body of the athlete can be approximated as only affected by gravity, if other effects like wind are neglected. Using methods introduced in the previous section, the center of body mass could be calculated frame by frame across the whole video. Suppose that the speed of the mass center at the beginning of Phase I is known, using the acceleration of gravity, the trajectory of the mass center across the whole phase could be easily calculated and recovered.

Phase II: During Phase II, there is some part of the body (represented by a key point) touching ground. Because of the assumption of no sliding, the absolute position of the key point is fixed. With this fixed key point during Phase II, the positions of the other parts of the body can be recovered. Since the key points of the skeleton across the whole video is calculated, the positions and speeds of the center of body mass could also be recovered, which will provide the initial speed for the Phase I following the current phase.

4. Results

Dataset and evaluation metrics

We experimented with the proposed algorithm on the TV broadcasting videos of men's and women's triple jump finals of the Rio Olympics games in 2016 to test the performance. These videos are monocular and dynamic, and track a single athlete, which makes them suitable for our requirements. The results of each jump attempt also provide us an easily accessible metric for evaluation. Moreover, the heights of the athletes are available on the Internet.

The videos we selected include part of jumps of medal-winners in triple jump finals of the Rio Olympics. The following table summarizes the jumps included in our dataset. There are in total 19 jump attempts in the videos, some of the attempts have two jump videos of normal speed and slow motion, respectively. We discarded some low-quality videos, and cut the videos so that each one contains only one jump. Finally, we have 27 jump videos of 18 attempts.

		Results (m)					
Attempts		1	2	3	4	5	6
Women	Caterine Ibarguen	14.65	15.03	14.38	15.17	14.76	14.8
	Yulimar Rojas	14.32	X	14.87	14.98	14.66	14.95
	Olga Rypakova	14.73	14.49	14.52	14.2	14.74	14.58
Men	Christian Taylor	17.86	17.77	X	17.77	X	X
	Will Claye	17.76	X	X	17.61	X	17.55
	Dong Bin	17.58	X	X	–	–	–

Table 2. Results of 6 medal-winners in the triple jump finals of the Rio Olympic games. 'X' represents an attempt is foul. '-' represents missed attempt because of injury. The yellow highlighted results are included in

our dataset, among which those attempts highlighted with underline have two videos. There are in total 27 videos in our dataset.

Based on the prediction of the algorithm, we accumulate the distance of the last three steps as the predicted result of the jump attempt. We take the average accuracy of all predictions as the major metric to evaluate the performance of the algorithm.

We tried two methods to calculate total jump length, which is illustrated in Fig. 5. Method 1 directly calculates the distance between the set-off position and the final landing position. Method 2 calculates the sum of each step length.

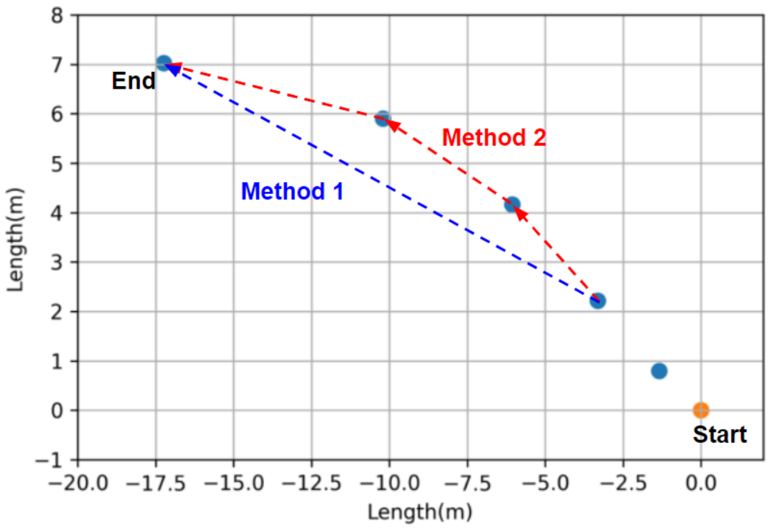

Figure 5. Recovered trajectory of Dong Bin's jump, with illustration of two methods to calculate the jump result

Quantitative evaluation

Calculation method	Foot-ground contact signal	With 2D corrections	Mean absolute error (%)
1	Ground-truth	No	10.6
1	Ground-truth	Yes	10.6
1	Predicted	Yes	22.0
2	Ground-truth	No	**9.1**
2	Ground-truth	Yes	9.2
2	Predicted	Yes	21.4

Table 3. Summary of results. Best performance is highlighted in **bold**.

The major results are reported in Table 3. We tested the performances of different calculation methods, ground truth versus prediction of foot-ground contact signal and with and without 2D corrections. In general, the best MAE (Mean Absolute Error) we achieved is 9.1% using method 2 with ground-truth foot-ground contact signal and without 2D corrections.

Different calculation methods: From above results, it can be seen that under the same experimental settings, method 2 outperforms method 1. Fig. 5 is a trajectory recovered from Dong Bin's jump, which also illustrates the definition of the two methods. From this figure, we can see that the recovered trajectory is not a curved line, rather than the original straight one. The distortion explains why method 2 performed better than method 1.

The distortion of the trajectory is resulted from the rotation of the camera. Although we assumed that there is no foot-sliding, our model only takes one key point for each foot, which means the rotation around the foot is neglected. In fact, there are models able to predict more details of foot (e.g. key points including toes and heels in [25] or body surface model [26]), which can help solve this problem.

Influence of 2D corrections: Even though we performed corrections to 2D key points predictions, we surprisingly found that the 2D corrections do not help to improve the performance. An explanation is that joint swap error occurs not very frequently, thus does not influence the result very much. To better examine the result, we made a boxplot of the results. We found that there are some outliers. The outliers are caused by some uncorrected swaps in 2D poses. Fox example, in Fig. 6 is 3 sequential frames of the same phase. As we rely on foot-ground contact information to perform kinematics analysis, in this case the foot key points are exchanged, thus causing significant errors in prediction.

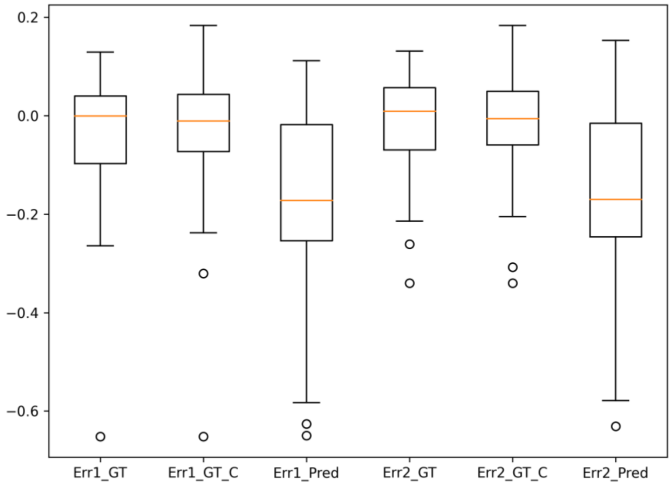

Figure 6. Boxplot of results.

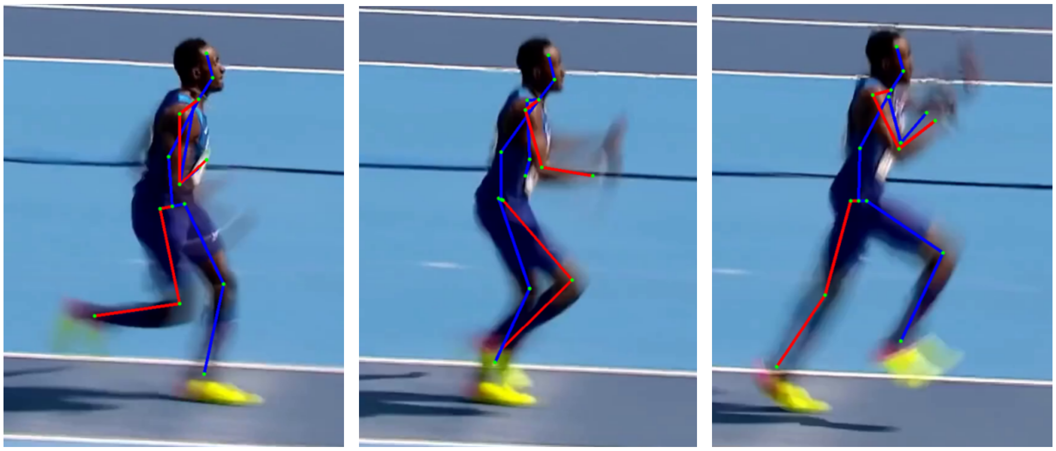

Figure 7. Example of swap error.

Influence of foot-ground contact signal accuracy: The performance using foot-contact signal predicted by our model is much worse than that using ground truth. Because the foot-contact information is important in splitting two phases for kinematic analysis, it has a large influence on the final performance.

5. Conclusion

In conclusion, this work proposed an algorithm to combine existing human pose estimation models with kinematic analysis, to enable athletes' trajectory recovery from monocular dynamic videos. This algorithm is applicable to most sports of athletics. We take triple jump as an example and achieved the lowest mean absolute error of 9.1%. The result validated the feasibility of the proposed algorithm. Moreover, this method could be improved by applying better human pose estimation models or better foot-ground contact prediction models. Finally, this method provides an option to analyze athletes' movement with low cost and high efficiency, and thus help with the improvement of athletes' performances.

References

[1] Martinez, Julieta, et al. "A simple yet effective baseline for 3d human pose estimation." Proceedings of the IEEE international conference on computer vision. 2017.

[2] Pavllo, Dario, et al. "3d human pose estimation in video with temporal convolutions and semi-supervised training." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

[3] Cheng, Yu, et al. "Monocular 3D multi-person pose estimation by integrating top-down and bottom-up networks." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

[4] Li, Wenhao, et al. "Mhformer: Multi-hypothesis transformer for 3d human pose estimation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[5] Li, Wenhao, et al. "Exploiting temporal contexts with strided transformer for 3d human pose estimation." IEEE Transactions on Multimedia (2022).

[6] Zhang, Jinlu, et al. "MixSTE: Seq2seq Mixed Spatio-Temporal Encoder for 3D Human Pose Estimation in Video." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[7] Dong, Junting, et al. "Fast and robust multi-person 3d pose estimation from multiple views." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

[8] Dong, Zijian, et al. "Shape-aware multi-person pose estimation from multi-view images." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[9] Usman, Ben, et al. "MetaPose: Fast 3D Pose from Multiple Views without 3D Supervision." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[10] Kadkhodamohammadi, Abdolrahim, and Nicolas Padoy. "A generalizable approach for multi-view 3d human pose regression." Machine Vision and Applications 32.1 (2021): 1-14.

[11] Liu, Miao, et al. "4d human body capture from egocentric video via 3d scene grounding." 2021 International Conference on 3D Vision (3DV). IEEE, 2021.

[12] Yuan, Ye, et al. "GLAMR: Global occlusion-aware human mesh recovery with dynamic cameras." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[13] Guzov, Vladimir, et al. "Human poseitioning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted sensors." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

[14] Von Marcard, Timo, et al. "Recovering accurate 3d human pose in the wild using imus and a moving camera." Proceedings of the European Conference on Computer Vision (ECCV). 2018.

[15] Shimada, Soshi, et al. "Physcap: Physically plausible monocular 3d motion capture in real time." ACM Transactions on Graphics (ToG) 39.6 (2020): 1-16.

[16] Shimada, Soshi, et al. "Neural monocular 3d human motion capture with physical awareness." ACM Transactions on Graphics (ToG) 40.4 (2021): 1-15.

[17] Gärtner, Erik, et al. "Trajectory optimization for physics-based reconstruction of 3d human pose from monocular video." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[18] Gärtner, Erik, et al. "Differentiable Dynamics for Articulated 3d Human Motion Reconstruction." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[19] Wang, Jianbo, et al. "Ai coach: Deep human pose estimation and analysis for personalized athletic training assistance." Proceedings of the 27th ACM International Conference on Multimedia. 2019.

[20] Li, Chunguang, and Jianbiao Cui. "Intelligent sports training system based on artificial intelligence and big data." Mobile Information Systems 2021 (2021).

[21] Fastovets, Mykyta, Jean-Yves Guillemaut, and Adrian Hilton. "Athlete pose estimation from monocular tv sports footage." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2013.

[22] Badiola-Bengoa, Aritz, and Amaia Mendez-Zorrilla. "A Systematic Review of the Application of Camera-Based Human Pose Estimation in the Field of Sport and Physical Exercise." Sensors 21.18 (2021): 5996.
[23] Zecha, Dan, Moritz Einfalt, and Rainer Lienhart. "Refining joint locations for human pose tracking in sports videos." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2019.
[24] Andriluka, Mykhaylo, et al. "Posetrack: A benchmark for human pose estimation and tracking." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.
[25] Jin, Sheng, et al. "Whole-body human pose estimation in the wild." European Conference on Computer Vision. Springer, Cham, 2020.
[26] Loper, Matthew, et al. "SMPL: A skinned multi-person linear model." ACM transactions on graphics (TOG) 34.6 (2015): 1-16.
[27] Ionescu, Catalin, et al. "Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments." IEEE transactions on pattern analysis and machine intelligence 36.7 (2013): 1325-1339.
[28] Redmon, Joseph, and Ali Farhadi. "Yolov3: An incremental improvement." arXiv preprint arXiv:1804.02767 (2018).
[29] Sun, Ke, et al. "Deep high-resolution representation learning for human pose estimation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.
[30] Adolphe, Melvin, et al. "Center of mass of human's body segments." Mechanics and Mechanical Engineering 21.3 (2017): 485-497.
[31] Dumas, Raphaël, Laurence Cheze, and J-P. Verriest. "Adjustments to McConville et al. and Young et al. body segment inertial parameters." Journal of biomechanics 40.3 (2007): 543-553.
[32] Ailing Zeng (2020) SRNet: Improving Generalization in 3D Human Pose Estimation with a Split-and-Recombine Approach. European Conference on Computer Vision pp. 507-523.
[33] Bastian Wandt (2019) RepNet: Weakly Supervised Training of an Adversarial Reprojection Network for 3D Human Pose Estimation. IEEE conference on computer vision and pattern recognition pp. 7782-7791.