Physics aware Human Action Recognition from Monocular Videos (PhARMov)

Introduction

While current methods have shown promising progress in estimating 3D human motion from monocular videos (video captured using a single camera from a single viewpoint), their motion estimates are often physically unrealistic because they mainly consider kinematics. Human body kinematics refers to studying and analyzing human movement without considering the forces or torques that cause it. Kinematics focuses on describing motion in terms of measurable parameters such as position, velocity, acceleration, angles, and trajectories of body parts (joint angles based on observed or measured motion data; e.g. tracking how the knee joint moves during running, i.e. trajectories of joints and body postures). Furthermore, the estimation of kinematics highly depends on the correct detection of the human skeleton from the image where the detection of landmarks plays a pivotal role in the construction of the skeleton. **Firstly**, we aim to enhance the accuracy of landmark detection by utilizing human body kinematics, such as positions, velocities, accelerations, and joint angles, guided by established rigid body equations from physics. **Secondly**, We want to improve the performance of estimating 3D human motion by combining kinematics with dynamics [1]. Since kinematics itself doesn't involve force analysis (which belongs to dynamics), we would like to propose techniques that can estimate forces based on the observed human motions (kinematics). For example, if we know the trajectory and speed of a leg during a kick (kinematics) then we can infer the muscular forces needed to produce that motion using the existing mathematical model of biomechanics and physics. The underlying idea is to incorporate physics principles governing human motions where we would like to build a physics-based body representation and contact force model (to capture the physical properties of the human body and the forces it experiences). **Thirdly**, our focus is on leveraging "Active Learning" techniques to minimize the annotation costs for new real-time video frames. Specifically, we aim to develop innovative methods for selecting the optimal set of frames that can most effectively improve the model's training performance. **Fourthly**, we intend to deploy our model or algorithm on edge devices (e.g., NVIDIA Jetson Nano, Google Coral, RaspberryPi, etc.) equipped with camera systems. To achieve this, we aim to develop compact, efficient, and task-specific algorithms (e.g. detecting abnormal falls of isolated elderly persons, specific actions like ball-pass, off-side, shooting goal in soccer, etc.) capable of performing real-time inference.

Keywords: Human Action Recognition, 2D/3D Skeleton of Human Body, Land-mark Detection, Feature Extraction, Spatio-temporality, Biomechanics, Kinematics, Dynamics, Rigid body analysis, Active Learning, Edge Device.

Background

Recent advances in deep learning, along with the progress in 3D human modeling [2] have substantially improved the reconstruction of 3D humans from a single image [3, 4]. With video inputs, current research aims to enhance model performance by exploiting temporal information. Some other authors devise temporal models that extract meaningful features from video to improve performance [5, 6, 7, 8]. Other works learn motion priors that capture natural 3D body movement patterns. Integrating the learned priors into model training can promote smooth motion estimates [9]. Although these approaches improve reconstruction to some degree, they often result in unrealistic outputs, marked by prominent physical artifacts like motion jittering and foot sliding.

That's why to address this limitation, a promising strategy is to leverage physical principles, governing body movements. In such an approach, the human body is treated as an articulated rigid body, and the human body dynamics can be represented through existing mathematical models of biomechanics (e.g. Euler-Lagrange equations). Such mathematical models can link the body mass, inertia, and physical forces (including joint actuation and contact forces) to body motions through ordinary differential equations. Some other research works [10, 11] in the literature formulate optimization frameworks that jointly estimate unknown physical parameters and refine kinematics-based estimates by aligning them with physics-based equations. Alternatively, other direction of research works [12, 13] employ the learning-based frameworks by keeping aside the cumbersome manual parameter tuning which is an inherent portion of optimization-based techniques that trains neural networks to predict the parameters.

The key challenge in these kinds of approaches is that the physics information and physical properties of human bodies and motion forces are absent in the current 3D motion capture datasets [14]. Existing methods generally rely on physics engines to incorporate physics [15]. This includes creating proxy bodies with geometric primitives to capture body properties, importing these proxies into the physics engine, and then leveraging the physics engine to compute the necessary physical parameters and simulate body motions. These approaches face challenges due to the inefficiency in computing gradients from physics engine outputs [16], hindering smooth integration with deep learning models. Additionally, most physics-based models are trained using 3D annotated videos, which are difficult to obtain in practice. As a result, these models struggle to generalize effectively to new, unseen scenarios.

To overcome these issues, we would like to work on deep learning models which will be trained on the data set of human motion along with corresponding kinematic and physical parameters for the calculations and predictions of motion properties such as positions, velocities, accelerations, and joint angles which will inherently improve the computation of human body skeleton which is constructed based on detected landmarks. Then, these refined motion data i.e. joint trajectories or movement patterns align better with the laws of physics and real-world bio-mechanical constraints. For instance, joint movements must respect joint limits and avoid unnatural postures. The forces and accelerations must be consistent with the observed motion. In this manner, the machine learning model can ensure that the motion data looks realistic and adheres to the principles of biomechanics and physics. The motion forces (muscular forces, external forces like gravity, or ground reaction forces) are responsible for producing the observed motion. The model can infer forces that are realistic and "favorable" (in biomechanics, it means forces within the range of human muscular capabilities) for replicating the desired motion.

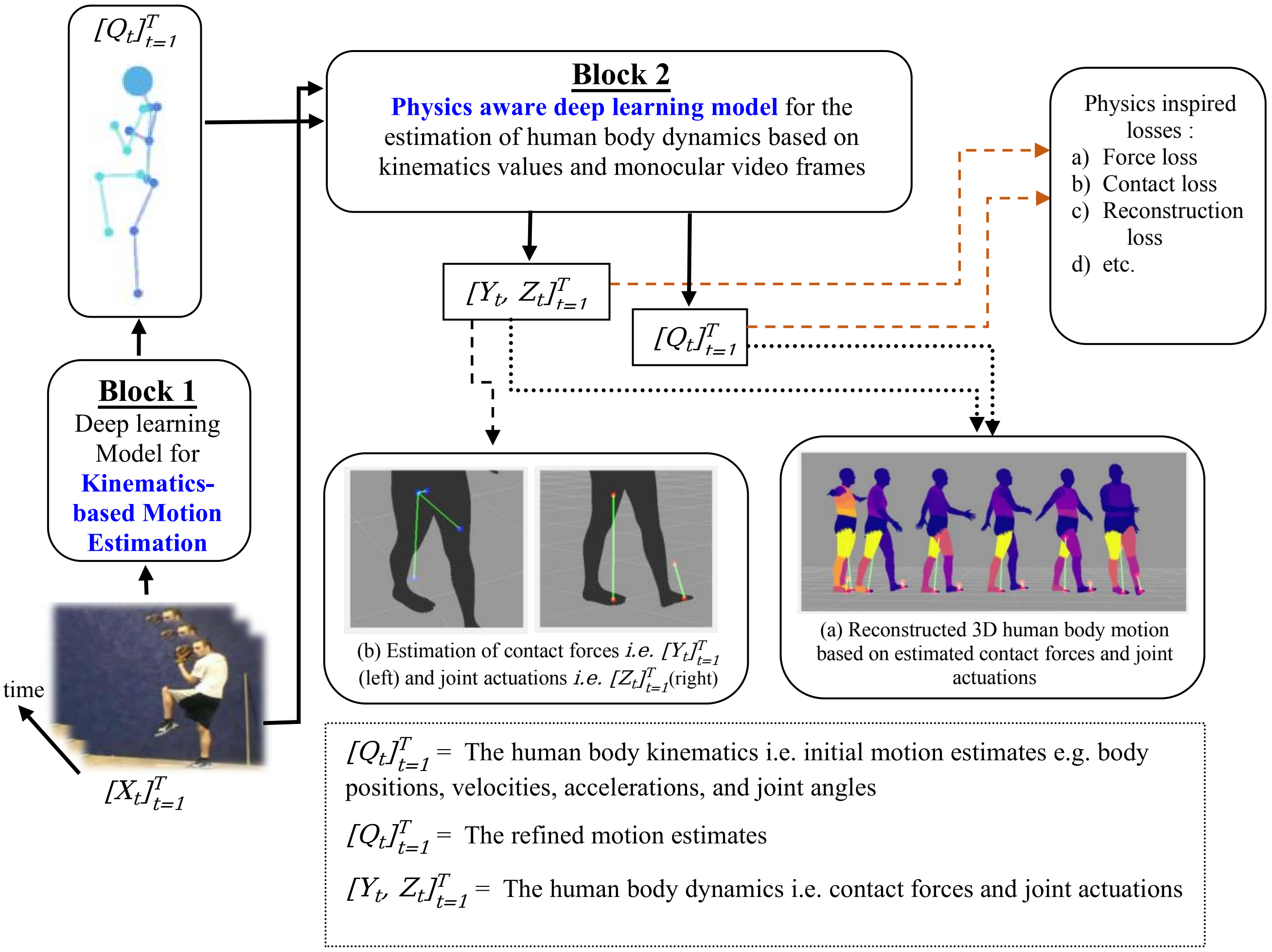
$[Q_t]^T_{t=1}$
Block 2
Physics aware deep learning model for the estimation of human body dynamics based on kinematics values and monocular video frames
Physics inspired losses :
a) Force loss
b) Contact loss
c) Reconstruction loss
d) etc.
$[Y_t, Z_t]^T_{t=1}$
$[Q_t]^T_{t=1}$
Block 1
Deep learning Model for Kinematics-based Motion Estimation
time
$[X_t]^T_{t=1}$
(b) Estimation of contact forces i.e. $[Y_t]^T_{t=1}$ (left) and joint actuations i.e. $[Z_t]^T_{t=1}$(right)
(a) Reconstructed 3D human body motion based on estimated contact forces and joint actuations
$[Q_t]^T_{t=1}$ = The human body kinematics i.e. initial motion estimates e.g. body positions, velocities, accelerations, and joint angles
$[Q_t]^T_{t=1}$ = The refined motion estimates
$[Y_t, Z_t]^T_{t=1}$ = The human body dynamics i.e. contact forces and joint actuations

Fig 1: Overview of the proposed method: The framework integrates a kinematics-driven motion estimation model with a physics-aware model to estimate human dynamics from a monocular video. **Inset (a)** depicts the actuation of the right pelvis joint and the contact forces applied at each foot. **Inset (b)** shows the reconstructed body motion and the inferred forces, where lighter color shades indicate higher joint actuation magnitudes, such as upper body joints during standing and leg joints during walking. The above diagram is inspired from [1].

Research plan

Inspired by the approach mentioned in [1], as the **first part** of this research work, we would like to propose a framework for learning human body dynamics along with human body kinematics which will bypass the need for 3D annotated videos and effectively integrate physics based bio-mechanical models within the deep learning framework. To address these challenges, we aim to bypass the reliance on unrealistic body proxies and physics-based engines [15] by directly deriving the physical properties of the human body using widely adopted 3D body models, such as SMPL [17]. Our approach will integrate contact forces into the deep learning network, leveraging training sequences to derive motion forces. The framework will incorporate a total loss function that includes “force loss”, “contact loss”, and losses related to the bio-mechanical properties of the human body. Once trained, this model can be applied on top of any kinematics-based reconstruction model to generate improved motion and force estimates from monocular videos. The planned improvement of enhancement in motion and force estimates will have the potential to improve the accuracy in tasks like human action recognition. In Fig. 1, it is depicted that from the input video frames (i.e. $[X_t]_{t=1}^{T}$), we would like to initially extract the human body skeleton (see Block 1 in Fig. 1), based on the detected landmarks. The computed skeleton will provide us with the initial motion estimates i.e. kinematics ($[Q_t]_{t=1}^{T}$) e.g. body positions, velocities, accelerations, and joint angles. We plan to further improve this section of human body skeleton computation by incorporating existing physics-based mathematical models for rigid bodies (i.e. human body in our case). Then, these kinematics-based parameters and video frames will be given as input into physics aware deep learning model (see Block 2 in Fig. 1) to generate improved/refined kinematics estimates ($[Q_t]_{t=1}^{T}$) and dynamics or force estimates (i.e. contact forces $[Y_t]_{t=1}^{T}$ and joint actuation $[Z_t]_{t=1}^{T}$) from monocular videos. Based on this, we will perform the reconstruction of 3D human body motions. In this respect, we plan to employ several existing physics-based mathematical equations as the loss function (i.e. force loss, contact loss, 3D human body reconstruction loss, etc.) of this end-end deep learning model. Furthermore, **our ambitious objective** is to incorporate other existing physics-based equations of kinematics and dynamics of the human body to further improve the estimation of 3D human motion which inherently improves the accuracy of human action recognition, such as :

a) Gait Analysis (to perform whole body motion analysis)
b) Force-length relationships (incorporating the constraint that force varies with muscle length)
c) Force-Velocity relationships (force depends on the speed of contraction)
d) Finite Element Analysis (simulates stress, strain, and deformation under external forces)
e) Mass-spring models (to simplify the soft tissue dynamics for real-time applications)

As the **second part** of the work, we will focus on applying our technique to the new real-time datasets for real-time applications. Human motion estimation and its associated task like human action recognition requires spatio-temporal annotation on each frame of the video in addition to the video level annotations. Cost of such annotation is much higher compared to the classification tasks where only video level annotations is sufficient. In this work, we would like also to study how this high annotation cost for spatio-temporal detection can be reduced with minimal performance trade-off. The traditional active learning (AL) techniques [18] typically focuses on classification tasks and the selection is performed at the sample level. But in the case of video level action recognition, a frame-level spatio-temporal localization is required in addition to the video level class prediction. Therefore the AL techniques should also consider detection on every frame within a video apart from video-level decisions. The informativeness and diversity of samples, both are important criteria for the selection of samples in AL based techniques. Incorporating the informativeness and diversity of samples/frames in the paradigm of spatio-temporal context for video is challenging and not much explored in the literature. To handle the spatio-temporal nature of video, we plan to introduce “spatio-temporal loss” which will take into account temporal continuity of a human action in the video.

Two promising directions of AL are “task-agnostics” approach, in which we focus on selecting the samples which are highly informative (can enhance the information learning of the task model) and the second direction is “task aware approach” that relies on the perspective of task model in which we select the diverse & difficult samples which can introduce diversity in the training of the task model. Unfortunately, the former does not exploit the structures from the tasks and the later does not seem to well-utilize the overall data distribution. In the first class of methods, the goal is to identify influential samples, *e.g.* which are lying in the high density regions such that once labeled, a large numbers of neighboring samples can benefit from propagating these labels. The major drawback of these approaches are that they do not take into account how the output of the model will depend on the input. For example, for classification task, it would be more effective to label data instances that lie in the vicinity of decisions boundaries than the samples, lying in high-density regions where most data points belong to the same class. The task aware approaches explicitly address this limitation by effectively identifying difficult data samples (*e.g.* the ones which are close to the decision boundaries). But these family of approaches does not directly consider about how the labeled samples make influence on the entire data set. Recently, the TAVAAL [18] technique propose an AL scheme that combines the benefits of these two groups of approaches by offering the capabilities of identifying difficult and influential data samples.

Another issue with classical AL based techniques is that most of the techniques are generally studied on balanced datasets where an equal amount of images per class is available. However, real world datasets suffer from severe imbalanced classes, the so called long-tail distribution. Hence, we argue that this further complicates the active learning process, since the imbalanced data pool (which is case for real life applications) can result in sub-optimal classifiers. Hence the objective is to corrects the class-imbalances presence in the unlabeled data pool. In [19], the authors proposed a AL method for class imbalance unlabeled data-set that encourages the selection of class-balanced samples. This method can independently/separately work on “task-agnostics” approaches and also on “task aware” approaches. But it is not compatible to work on the techniques like TAVAAL which combines both the family of approaches.

Hence, in this direction, firstly, we would like to work on the AL techniques which can firstly incorporate the spatio-temporal nature of video (which is not much explored yet in the literature because most of AL techniques are designed for images and not for videos). Secondly, inspired by TAVAAL techniques for images, our plan would be to adapt and incorporate the underlying method for videos. Thirdly, we will also remedy the problem of class imbalances in the real-life data-set and mould our AL techniques in the paradigm of class balance AL technique.

External Collaboration

In this aspect, we would like to have some external collaborations which can help us to propoerly realize this project according to the plan.

Collaboration	Expertise
Expertise from Dept. of Physics or Mechanics (bio-mechanics)	In kinematics & dynamics of a rigid body, human bio-mechanics
PRETIL, CRIStAL	In incorporating our algorithms in the edge devices (NVIDIA Jetson Nano, Google Coral, RaspberryPi etc.). For this portion of work, we plan to seperately hire interns, engineers.

References

[1] Bengar, J. Z., Van De Weijer, J., Fuentes, L. L., & Raducanu, B. (2022). Class-Balanced Active Learning for Image Classification. Proceedings - 2022 IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, 3707–3716.

[2] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Smin- chisescu. Ghum & ghuml: Generative 3d human shape and articulated pose models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6184–6193, 2020.

[3] Yufei Zhang, Hanjing Wang, Jeffrey O Kephart, and Qiang Ji. Body knowledge and uncertainty modeling for monocular 3d human body reconstruction. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 9020–9032, 2023.

[4] Shashank Tripathi, Lea Muller, Chun-Hao P Huang, Omid Taheri, Michael J Black, and Dimitrios Tzionas. 3d human pose estimation via intuitive physics. In Proceedings ofthe IEEE/CVF Conference on Computer Vision and Pat- tern Recognition, pages 4713–4725, 2023.

[5] Boyang Zhang, Kehua Ma, SupingWu, and Zhixiang Yuan. Two-stage co-segmentation network based on discriminative representation for recovering human mesh from videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5662–5670, 2023.

[6] Zhenhua Tang, Zhaofan Qiu, Yanbin Hao, Richang Hong, and Ting Yao. 3d human pose estimation with spatio-temporal criss-cross attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4790–4799, 2023.

[7] Xiaolong Shen, Zongxin Yang, Xiaohan Wang, Jianxin Ma, Chang Zhou, and Yi Yang. Global-to-local modeling for video-based 3d human pose and shape estimation. In Pro- ceedings ofthe IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8887–8896, 2023.

[8] Gun-Hee Lee and Seong-Whan Lee. Uncertainty-aware human mesh recovery from video by learning part-based 3d dynamics. In Proceedings of the IEEE/CVF Inter- national Conference on Computer Vision, pages 12375– 12384, 2021

[9] Mingyi Shi, Sebastian Starke, Yuting Ye, Taku Komura, and Jungdam Won. Phasemp: Robust 3d pose estimation via phase-conditioned human motion prior. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 14725–14737, 2023.

[10] Gengshan Yang, Shuo Yang, John Z Zhang, Zachary Manchester, and Deva Ramanan. Ppr: Physically plausible reconstruction from monocular videos. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 3914–3924, 2023.

[11] Erik Gartner, Mykhaylo Andriluka, Hongyi Xu, and Cristian Sminchisescu. Trajectory optimization for physics based reconstruction of 3d human pose from monocular video. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13106– 13115, 2022.

[12] Buzhen Huang, Liang Pan, Yuan Yang, Jingyi Ju, and Yan- gang Wang. Neural mocon: Neural motion control for physically plausible human motion capture. In Proceedings ofthe IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6417–6426, 2022.

[13] Jiefeng Li, Siyuan Bian, Chao Xu, Gang Liu, Gang Yu, and Cewu Lu. D &d: Learning human dynamics from dynamic camera. In European Conference on Computer Vision, pages 479–496. Springer, 2022.

[14] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In Proceedings ofthe IEEE/CVF International Conference on Computer Vision, pages 5442–5451, 2019.

[15] Erwin Coumans and Yunfei Bai. Pybullet, a python module for physics simulation for games, robotics and machine learning. 2016.

[16] Erik Gärtner, Mykhaylo Andriluka, Erwin Coumans, and Cristian Sminchisescu. Differentiable dynamics for articulated 3d human motion reconstruction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13190–13200, 2022.

[17] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. ACM transactions on graphics (TOG), 34(6):1–16, 2015.

[18] Kim, K., Park, D., Kim, K. I., & Chun, S. Y. (2021). Task-Aware Variational Adversarial Active Learning. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 8162–8171.

[19] Bengar, J. Z., Van De Weijer, J., Fuentes, L. L., & Raducanu, B. (2022). Class-Balanced Active Learning for Image Classification. Proceedings - 2022 IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, 3707–3716.