GAIT ANALYSIS WITH TRINOCULAR COMPUTER VISION USING DEEP LEARNING

Odysseas Stavrakakis[1], *Athanasios Mastrogeorgiou*[1], *Aikaterini Smyrli*[1], *Evangelos Papadopoulos*[1]

[1]Control Systems Lab, School of Mechanical Engineering,
National Technical University of Athens, Greece

ABSTRACT

The recent years' progress in deep learning (DL) technology has resulted in convolutional neural networks (CNNs) capable of producing fast and accurate results, with minimal data preprocessing. Currently, gait analysis (GA) is attracting the attention of the field of deep learning due to its seamless integration applicability. Our approach focuses on CNN-based GA application on healthcare and orthopaedics. Using CNNs and visual fiducial systems for recognition and 3D mesh reconstruction of the human form and the floor respectively, we can virtually recreate the human-floor interaction, which can be particularly useful in the study of gait dynamics, through the per-frame gait phase classification. However, the current state-of-the-art (SOTA) is that most CNN mesh reconstruction software produces a mesh from a monocular input. The use of photogrammetry could alternatively be implemented, but would require multiple cameras and expensive equipment. Our approach aims to create a refined mesh obtained from trinocular footage, along with an interactive and easy-to-use interface.

1. INTRODUCTION

Traditionally, gait cycle parameters, such as the step and stride duration, the ground reaction forces, etc., are captured using pressure sensitive walkways or force-plates. Following data collection, the data are imported and interpreted using dedicated computer software for the study of the subject's gait [1]. However, this process introduces some major disadvantages, including limited portability, restricted measurement area, high purchase and maintenance costs, invasive nature (barefoot use), limited temporal resolution and lack of contextual information (joint angles, body posture, muscle activation, etc.). The use of motion capture (Mo-Cap) systems is usually combined with pressure sensitive walkways, since it provides more contextual information (usually a full body model) and has generally better temporal resolution. But that comes at the cost of even higher costs of acquisition, increased setup times and increased patient discomfort, since the user has to be tailored with multiple Mo-Cap markers [2].

Recenty, significant progress has been made in the area of deep learning (DL) with constantly evolving and expanding databases [3]. This has provided new untapped opportunities for detection, reconstruction and classification, making them ideal for gait analysis application. However, the vast majority of their implementation has been in the fields of sports [4], and biometric identification [5] [6] [7] [8].

In this work, we focus on medical and orthopaedic gait analysis leveraging DL. Our goals are: (a) to create a method that is more versatile and descriptive than the SOTA techniques, (b) to attain an equivalent level of accuracy as the forementioned alternatives, and (c) to minimize cost and patient discomfort. Our method requires footage from 3 individual low-cost cameras, and performs detection and 3D reconstruction of the captured scene's gait-related elements, i.e., the human subject and the floor.

The foundational framework of the methodology relies on the integration and functionality provided by ECON [9], a research project focused on generating highly accurate representations of clothed individuals by combining detailed clothing models with human body scans. It integrates the best properties of implicit and explicit representations, to infer high-fidelity 3D humans from in-the-wild images, even with loose clothing or in challenging poses. However, even though ECON can produce results of high quality and adequacy from monocular footage, it is not uncommon for the object's opposite side to the camera to sometimes look uncanny. Other fundamental component of this project's framework is the utilisation of AprilTags [10] for floor detection, and the ICPA algorithm [11].

The golden standard of 3D reconstruction from 2D images is photogrammetry, for its ability to deliver high accuracy results. However, in the field of gait analysis, its strengths are out-weighted by its weaknesses like camera setup and calibration needs, sensitivity to subject movement and marker placement, high computation costs etc., which render photogrammetry, in its current state, an almost infeasible way to perform gait analysis.

On the contrary, neural network trinocular vision offers a simplified setup, with marker-less analysis and occlusion handling capabilities, while maintaining lower computational and monetary costs. Our methodology's objective is to accomplish all of the above, while achieving results directly comparable to those of photogrammetry.

Lastly, this method offers high portability making it optimal for patients with limited mobility. It operates in a nonintrusive and even contact-less manner, which is highly desirable and ensures greater convenience, comfort, and safety for the patients under analysis.

2. METHOD

The necessary equipment that comprises our setup is a set of 3 RGB camera devices for video footage capture, the tags needed to capture the floor information and a laptop computer, for post-processing. The tags are placed in an orthogonal configuration that represents the edges of the walking area-plane as shown on Fig.1, which presents our testing setup and equipment used.

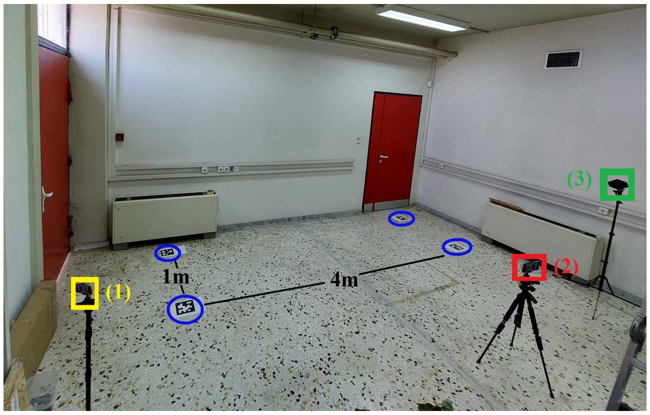


Fig. 1. Setup. The tag positions are annotated in blue colour. Cameras (1), (2) and (3) are annotated in yellow, red, and green respectively.

By splitting the video footage captured from each camera into individual frames and importing them into the AprilTag detection software, we can perform floor detection and 3D reconstruction for the floor element of the scene, based on the tag placement. We then export this information with relation to a specific camera as a point of reference. The same frames are imported to ECON, where they are processed to generate 3D reconstruction of the human in the scene and to obtain information about the SMPL-X body model [12] that is used to express its spatial and temporal information.

The next step is to align the three 3D human meshes that are produced from the frames and correspond to the same time instant. The orientation process, preferably to the same point of reference as the floor from the former task, is performed using the ICPA. For better alignment results, we initially implement the transformation matrices of all the incorporated elements of the setup and scene, performing a first rough alignment and subsequently using the ICPA, to fine-tune and refine the alignment. We then perform a Poisson reconstruction, fusing them into one solid element that represents the collective mesh of all camera views for the corresponding time frame.

By using the Blender 3D computer graphics software, we combine the human and floor elements, recreating the captured scene in 3D space. Utilizing the Blender Python API to automate the whole process for x iterations, we can reconstruct the entire video sequence in a high fidelity 3D animation. Tasks like masking the foot's regions, performing collision detection with colour or/and text activation, making an educated guess on which phase of the gait cycle the subject is in, extracting information about the joint angles, the ground force, etc. and many other clinical oriented features can then be performed. Our framework is summarized in Fig.2.

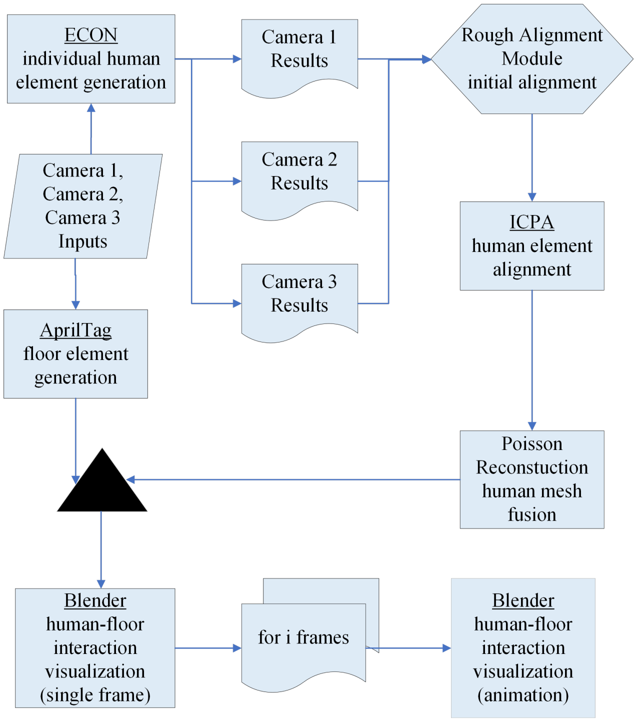

ECON
individual human element generation
Camera 1 Results
Rough Alignment Module initial alignment
Camera 2 Results
Camera 1, Camera 2, Camera 3 Inputs
ICPA
human element alignment
Camera 3 Results
AprilTag
floor element generation
Poisson Reconstuction human mesh fusion
Blender
human-floor interaction visualization (single frame)
for i frames
Blender
human-floor interaction visualization (animation)

Fig. 2. Flowchart. The camera footage is processed separately in order to obtain the human and floor meshes. The combined product of these for multiple spatial iterations results in a 3D animation of the elements

3. DISCUSSION AND RESULTS

3.1. Relation to prior work

The framework presented here expands on the work of ECON [9], by introducing a trinocular application. While the ECON project focuses exclusively on monocular footage results, our implementation employs 3 separate monocular footages' results into a single one, with refined qualities, using the ICPA [10].

3.2. Initial results

We demonstrate the results of the steps from our approach on trinocular detection and 3D reconstruction in Fig.4 through

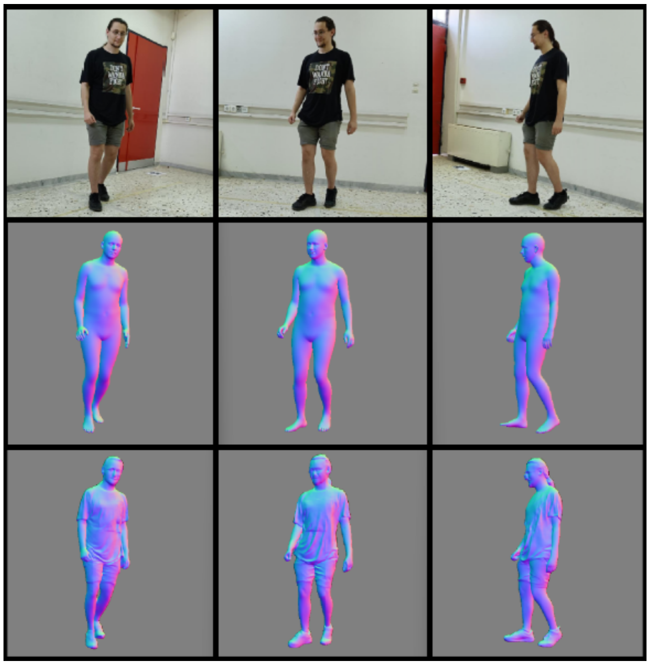


Fig. 3. ECON results. The first row is comprised of the input images. The second row is comprised of the SMPL-X fitted models. The third row is comprised of the output objects

Fig.7. The input images are used to determine the shape parameters of the SMPL-X models used to capture the human actor's physique. These models are enhanced with the actor's features and turned into point clouds (PCs). Subsequently, they are aligned with each other, through designating one as fixed and transforming the others relative to it. One can see the results of the rotation in Fig.5, first relative to their initial position individually and also with relation to each other. These PCs are then fused into one single mesh that better encapsulates the 3D information of the actual human actor. By separating the video footage into its individual frames and performing this pipeline for each set of frames that refer to the same time period we can obtain the corresponding mesh for each one of them.

3.3. Limitations

The majority of our results' limitations migrate from ECON [9]. Recovering body pose information using ECON is still an open problem. This could lead to ECON failures, like bent legs, artifact stitching or wrong thickness depictions. As the generated data is getting sufficiently realistic, their domain gap with real data will be significantly narrowed, resulting in the elimination of such limitations.

The background of the setup should be as uncluttered and uniform as possible. On the other hand, the human actor should avoid clothing that is long, dark and/or uniform, since it disrupts joint and on-body-shadow visibility. Background or clothing patterns or colours that obscure the human actor's outline should also be avoided. It's also recommended to avoid using as input footage with motion blur, occlusions and hard shadows, since it can severely affect the final result. Furthermore, there is a limit to the possible camera resolution that can be used as input. However, this comes with a beneficial trade-off in terms of lower costs and computational sources, as well as improved portability.

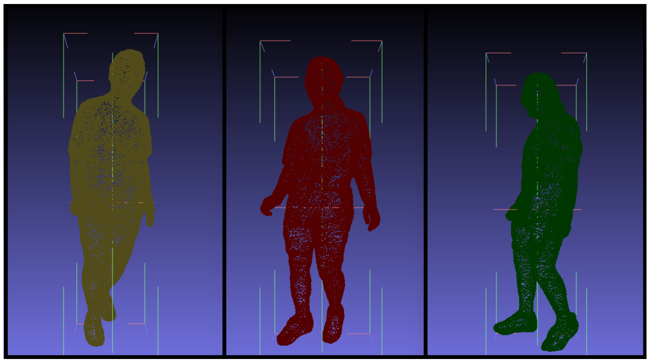


Fig. 4. Object to PC Transformation. The generated PCs of each camera (1,2,3) perspective annotated with the respective colours (yellow, red, green).

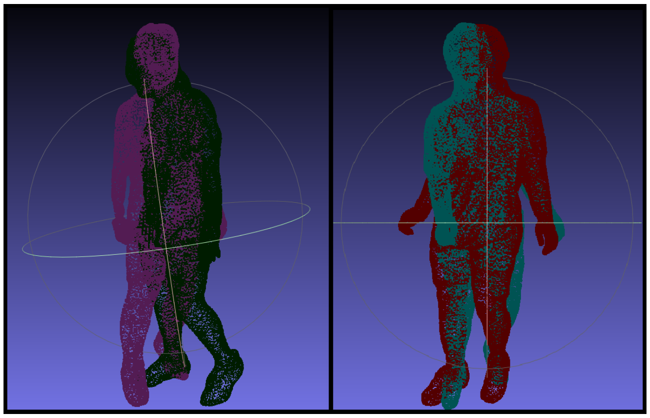


Fig. 5. ICPA results. On the left we see the initial (green) and final (pink) position of the PCs from Camera (3). On the right we see the initial (red) and final (cyan) positions of the PCs from Camera (2).

Finally, footage captured from very tilted camera positions, often results in high artifact and joint angle variance amongst the meshes, due to depth ambiguity.

3.4. Future work

Due to the innovative character of the proposed pipeline, its apparent that there is still significant room for refinement

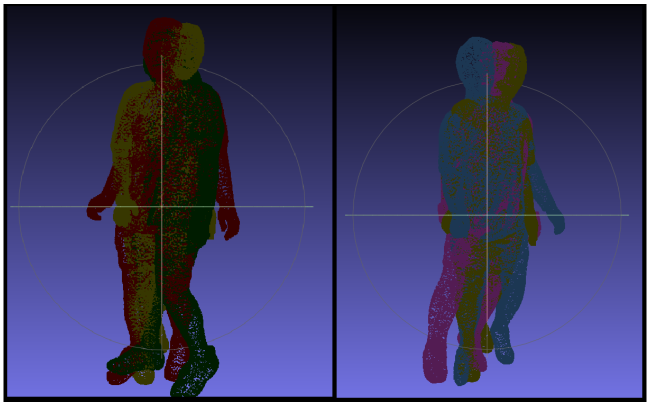


Fig. 6. ICPA results. Left: the alignment of the PCs from all cameras BEFORE THE ICPA. Right: the alignment of the generated PCs from all cameras after the ICPA.

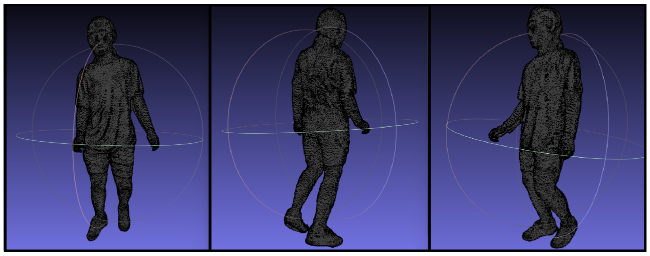
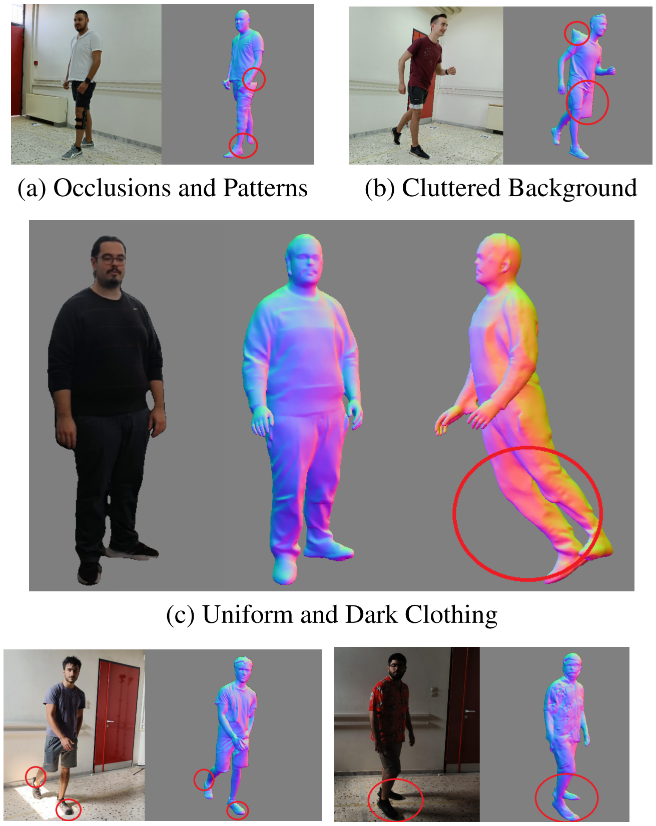


Fig. 7. Poisson Reconstruction. Different views of the fused resulting mesh produced from the Poisson reconstruction of the aligned meshes.

and optimization. Fig. 8 (a-d) presents an overview of the method's current shortcomings. A major aspect that could be improved is the optimization of the trinocular vision setup, and of the way in which the reconstructed 3D meshes from each of the three cameras are integrated into one. Future work could include the development of a novel CNN that builds a correspondence relationship across different models. By incorporating constraints like shape and pose consistency into an energy function we could perform geometry inference across our trinocular input, resulting in a much higher quality and accuracy final mesh. Note that programs like PIFu [13] and PaMiR [14] demonstrate results from a multi-image input. However, given ECON's monocular input reconstruction superiority to these programs, we expect improved results from the integration of trinocular input to the ECON framework.

Another area of improvement, could be the refinement of the capturing technique introduced in our pipeline. Hard shadow artifacts could be eliminated directly with better lit setups with direct but ambient lighting, or computationally, with foot shape evaluation modules. Self-occlusions and object outline obscurities could be addressed with an implementation of robust human pose and shape estimation methods. PARE [15] has already demonstrated such results, but only for SMPL models.



(a) Occlusions and Patterns

(b) Cluttered Background

(c) Uniform and Dark Clothing

(c) Inconsistent Lighting

(b) Hard Shadows

Fig. 8. Limitation Examples

Although ECON provides very accurate capturing of the subject's clothing, sometimes this can work counter-productively. Introducing a dataset of more simply dressed individuals, and training the neural network to predict the shape of a human regardless of their clothing could substantially improve the robustness to variations, and potentially even background elements.

4. CONCLUSION

The field of medical gait analysis holds substantial promise for advancements through the integration of 3D computer vision. Our methodology provides a trinocular application, utilizing the potential of deep neural networks. It achieves an effective detection and three-dimensional reconstruction of the human-floor interaction for each time frame, and incorporates all the time frames into a 3D animation sequence. Using this sequence and dedicated modules, it aims to facilitate the extraction of information about gait for medical professionals.

5. REFERENCES

[1] Nathaniel Goldfarb, Alek Lewis, Alex Tacescu, and Gregory S. Fischer, “Open source vicon toolkit for motion capture and gait analysis,” *Computer Methods and Programs in Biomedicine*, vol. 212, pp. 106414, 2021.

[2] McGuirk TE, Perry ES, Sihanath WB, Riazati S, and Patten C, “Feasibility of Markerless Motion Capture for Three-Dimensional Gait Assessment in Community Settings ,” *Front. Hum. Neurosci.*, June 2022.

[3] Abdullah S. Alharthi, Syed U. Yunas, and Krikor B. Ozanyan, “Deep learning for monitoring of human gait: A review,” *IEEE Sensors Journal*, vol. 19, no. 21, pp. 9575–9591, 2019.

[4] Darwin Gouwanda and SMNA Senanayake, “Emerging trends of body-mounted sensors in sports and human gait analysis,” in *4th Kuala Lumpur International Conference on Biomedical Engineering 2008: BIOMED 2008 25–28 June 2008 Kuala Lumpur, Malaysia*. Springer Berlin Heidelberg, 2008, pp. 715–718.

[5] Munif Alotaibi and Ausif Mahmood, “Improved gait recognition based on specialized deep convolutional neural network,” *Computer Vision and Image Understanding*, vol. 164, pp. 103–110, 2017.

[6] Maryam Babaee, Linwei Li, and Gerhard Rigoll, “Gait recognition from incomplete gait cycle,” in *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 768–772.

[7] Maryam Babaee, Linwei Li, and Gerhard Rigoll, “Person identification from partial gait cycle using fully convolutional neural networks,” *Neurocomputing*, vol. 338, pp. 116–125, 2019.

[8] Chao Fan, Junhao Liang, Chuanfu Shen, Saihui Hou, Yongzhen Huang, and Shiqi Yu, “Opengait: Revisiting gait recognition towards better practicality,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 9707–9716.

[9] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black, “ECON: Explicit Clothed humans Optimized via Normal integration,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023.

[10] John Wang and Edwin Olson, “AprilTag 2: Efficient and robust fiducial detection,” in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, October 2016.

[11] Philipp Glira, Norbert Pfeifer, Christian Briese, and Camillo Ressl, “A correspondence framework for als strip adjustments based on variants of the icp algorithm,” *Photogrammetrie-Fernerkundung-Geoinformation*, vol. 2015, no. 4, pp. 275–289, 2015.

[12] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black, “Expressive body capture: 3D hands, face, and body from a single image,” in *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 10975–10985.

[13] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo, “Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2020.

[14] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai, “Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021.

[15] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black, “PARE: Part attention regressor for 3D human body estimation,” in *Proc. International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 11127–11137.