2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

MarkerLess Motion Capture: ML-MoCap, a low-cost modular multi-camera setup

Jinne E. Geelen[1], Mariana P. Branco[2], Nick F. Ramsey[2],
Frans C. T. van der Helm[1], Winfred Mugge[1], Alfred C. Schouten[1]

Abstract— **Motion capture systems are extensively used to track human movement to study healthy and pathological movements, allowing for objective diagnosis and effective therapy of conditions that affect our motor system. Current motion capture systems typically require marker placements which is cumbersome and can lead to contrived movements.**

Here, we describe and evaluate our developed markerless and modular multi-camera motion capture system to record human movements in 3D. The system consists of several interconnected single-board microcomputers, each coupled to a camera (i.e., the camera modules), and one additional microcomputer, which acts as the controller. The system allows for integration with upcoming machine-learning techniques, such as DeepLabCut and AniPose. These tools convert the video frames into virtual marker trajectories and provide input for further biomechanical analysis.

The system obtains a frame rate of 40 Hz with a sub-millisecond synchronization between the camera modules. We evaluated the system by recording index finger movement using six camera modules. The recordings were converted via trajectories of the bony segments into finger joint angles. The retrieved finger joint angles were compared to a marker-based system resulting in a root-mean-square error of 7.5 degrees difference for a full range metacarpophalangeal joint motion.

Our system allows for out-of-the-lab motion capture studies while eliminating the need for reflective markers. The setup is modular by design, enabling various configurations for both coarse and fine movement studies, allowing for machine learning integration to automatically label the data. Although we compared our system for a small movement, this method can also be extended to full-body experiments in larger volumes.

I. Introduction

Capturing natural movement is a key technique in various disciplines, such as biomechanics, sports engineering, neuroscience, rehabilitation and robotics [1]. For instance, motion capture is crucial to evaluate healthy and pathological movements in order to enable objective diagnosis, and effective therapy for conditions that affect our motor system [2]. Further developments enabling objective and contactless measures will benefit the diagnostics of neurological disorders, such as the finger-tapping test to evaluate bradykinesia in Parkinson's Disease [3] or the follow up gait identification in the rehabilitation of children with cerebral palsy [4].

This work is part of the STW perspective programme NeuroCIMT with project number 14906, which is financed by the Dutch Research Council (NWO).

[1]Department of BioMechanical Engineering, Faculty of Mechanical, Maritime and Materials Engineering, Delft University of Technology, The Netherlands

[2]Department of Neurology and Neurosurgery, University Medical Center Utrecht, The Netherlands

corresponding author: j.e.geelen@tudelft.nl

Generally, motion capture studies are performed in a laboratory using expensive and cumbersome equipment to study movement [1]. Current passive motion capture systems use cameras to record reflective markers, requiring marker attachment prior to the recording and marker labelling afterwards. Active motion capture systems resolve the inconvenient and labour-intensive labelling process yet still require accurate marker placement. Markers can introduce soft tissue artefacts as the markers move with and relative to the skin. For example, during finger movements, markers typically move 0.55 mm for each 10º of flexion around either the proximal interphalangeal (PIP) or the distal interphalangeal (DIP) joints [5]. Furthermore, participants move less naturally with markers attached [6]. Ideally, motion capture systems allow for measurements of natural movements under minimal environmental constraints and without labour-intensive post-processing.

Triangulation from multiple cameras provides the most reliable motion recordings in 3D, especially for delicate movements sensitive to occlusion. Synchronization of the cameras is essential to capture the recorded movement from the videos accurately. Standard computer workstations do not allow for simultaneous acquisition of multiple video streams due to the required high data transfer rates [7]. External triggers allow for simple direct synchronization. Nevertheless, cameras allowing for external triggers are expensive and impede out-of-the-lab experiments due to a lack of flexibility. Most low-cost cameras do not allow for direct synchronization, but post-processing can be performed based on audio triggers (clapboard), audio fingerprints, or blinking LEDs [8] [9]. These methods are effective as post-processing techniques but prevent the use of real-time applications. A computer network, where each unit operates one or more cameras, allows for a real-time approach to camera synchronization [7].

Recent developments in computer vision algorithms, computational processing power, and electronic hardware provide opportunities for optimizations of the motion capture workflow [10]. Time-consuming and error-prone marker labelling can benefit from integration with upcoming machine learning techniques. Recent results indicate that OpenPose-based markerless motion capture achieves positional errors smaller than 30 mm for full-body movements [11]. Elaborate post-processing can reconstruct occluded markers or correct the joint centres [6]. Nevertheless, marker-based systems remain the trusted tool in clinical and scientific settings due to their superior accuracy. Further development of computer vision

algorithms can improve the accuracy of the captured three-dimensional joint positions in markerless motion capture. In combination with low-cost hardware systems, the novel software-based solutions will improve the performance of markerless solutions while allowing for out-of-the-lab systems. Thereby, markerless motion capture is becoming more and more a competitive alternative for marker-based systems.

Here we present a cost-effective markerless system that allows for motion capture outside the laboratory environment. Our system is modular by design, enabling various configurations for both coarse and fine movement studies. This study validated our system with finger movement, a full range metacarpophalangeal (MCP) joint motion. Capturing these fine movements with old-style marker-based recordings is affected by skin movement, occlusion by the fingers, restricted movements (in case of more oversized markers), or dropping of markers (in case of tiny 3mm markers). We validated our system by comparing it to a conventional motion capture setup using passive reflective markers as a gold standard.

II. Methods

A. Experimental Setup

The MarkerLess Motion Capture (ML-MoCap) system contains multiple interconnected single-board microcomputers (Raspberry Pi 4 Model B) (Figure 1). Every microcomputer is coupled to a camera (HQ Camera, Raspberry Pi), forming a camera module, and provides the encoding of the video stream. The camera modules are connected to the network via a network switch (NETGEAR), which also supplies power to the connected microcomputers via the Power over Ethernet (PoE) board extensions. One other microcomputer acts as the control module of the system. Network time protocol (NTP) synchronization provides sub-millisecond synchronization between the microcomputers in the system. All hardware is commercially available. The total cost for the setup is relatively low: starting around 100 euros (for a controller and PoE switch), adding about 100 euros for each camera module (depending on the cameras chosen, exclusively cameras compatible with Raspberry Pi).

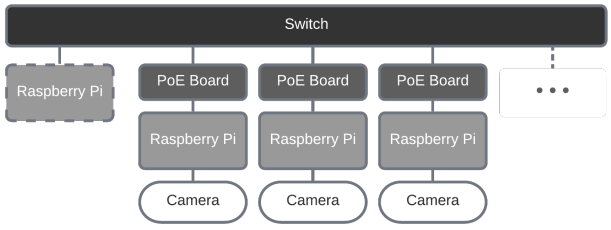

Fig. 1. System overview. Camera modules are interconnected via a Power over Ethernet (PoE) network switch (top) and controlled by a control module (dashed block on the left). Each camera module has a PoE board for power supply, a microcomputer to encode the video stream, and a camera board including an image sensor and lens. The number of camera modules is flexible (modularity indicated by the dots on the right).

The system is operated using a web application, which communicates with the control module. After recording a trial, the videos were transferred from the camera modules to a temporary network folder. At the end of the experiment, all videos were uploaded to an encrypted network folder to ensure data safety and participant privacy. The software to operate and connect the controller and camera modules is available with an open-source license (BSD 3-Clause)[1].

Figure 2 provides an overview of the pipeline for studies equipped with an ML-MoCap system. When initializing a new project, two manual steps should be performed once: calibration and training data. The calibration determines all camera positions relative to each other and a 3D coordinate system. We performed a ChArUco board calibration (OpenCV). Once a sufficient number of labels are applied manually (usually between 50-200 frames), the remaining frames will be labelled automatically by the model trained by DeepLabCut [10]. After the automatic labelling of all video frames from the multiple camera modules, the triangulation process was performed with AniPose [14], a 3D extension to the 2D DeepLabCut toolbox. Finally, the 3D marker trajectories are converted into joint angles.

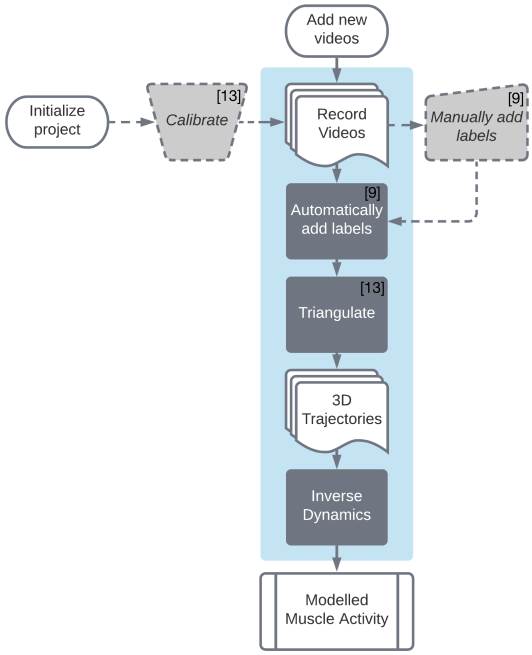

Fig. 2. Overview of the motion capture process. In blue are all automated recurring steps. The two dashed blocks are manual steps performed at the start of a new project to calibrate the cameras and initialize the automatic labelling.

B. Camera module synchronization

Synchronization of the camera modules is essential for accurate triangulation and extracting movement. The clocks of all camera modules are synchronized. However, the time needed for camera initialization at the start of a recording varies slightly between cameras and between recordings. Ten

[1]https://github.com/JinneGeelen/ML-MoCap.git

recordings in which all cameras simultaneously recorded a laptop screen showing a digital clock with milliseconds were initiated to check the synchronization. The laptop screen had a refresh rate of 60 Hz. The videos from the six cameras were stored and converted into frames. The displayed time from the first ten frames was used to calculate the synchronization between the cameras.

C. Validation of 3D finger motion capture

To validate the system simultaneous recordings with our ML-MoCap system (six regular cameras with a resolution of 1024 by 768 pixels at 40 Hz) and with an established reflective marker-based system with twelve infrared cameras at 100 Hz (Oqus 300 series & Qualisys Track Manager) were performed on one healthy participant. The participant completed a sequence of three MCP full range of motion flexion of the index finger.

Following the Covid-19 regulations, solely movement of the first author was captured. The participant sat in front of the system, see Figure 3. The experiment was approved by the Human Research Ethics Committee of the Delft University of Technology.

Twelve 3 mm reflective markers were attached to the index finger; three markers on each bony segment to reconstruct each segment movement and the corresponding joint angles. Reflective markers were positioned on the distal phalanx (DP), intermediate phalanx (IP), proximal phalanx (PP), and on the metacarpal (MC) of the index finger. The similarity between the two systems was more relevant than the biomechanics in this application. Therefore, the DIP, PIP, and MCP joint angles were assessed as rotations between the two consecutive bony segments over a single rotation axis. Although no markers are required for a typical use of the proposed pipeline, in this case, we decided to use them for once to assure the best comparison between the two systems.

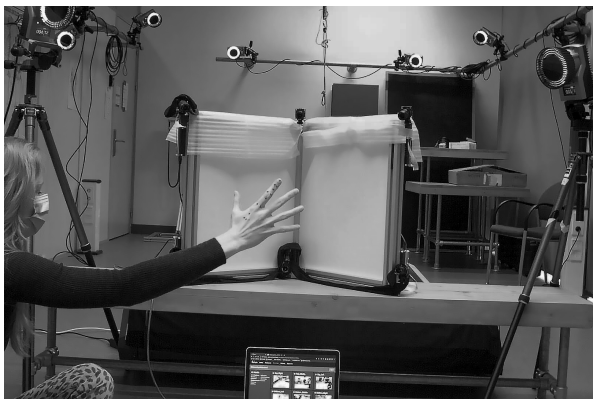

Fig. 3. Experimental setup for validation of the ML-MoCap, including simultaneous Qualisys recordings. The participant is seated in front of the ML-MoCap frame, which is placed in the middle of the Delft BioMechaMotion Lab (including twelve Qualisys cameras). The Pi camera modules are attached to each corner of the frame. Other parts as described in Figure 1 are attached to the back of the frame. The laptop on the floor initiated the recordings from a web application.

DeepLabCut uses transfer learning and therefore requires minimal training data. We manually labelled five frames of 4 repetitions from all six camera angles, resulting in 120 training frames. Subsequently, the model (ResNet-50, with a 90% training fraction, p-cutoff of 0.6, and the DeepLabCut defaults settings) was trained for 500,000 iterations, which took less than 10 hours on a computer with a discrete GPU (NVIDIA GeForce RTX 3090). The 3D virtual marker coordinates were converted to joint angles similar to the procedure with the marker-based system.

III. Results

A. Camera module synchronization

The synchronization test (Figure 4) revealed that the average time difference between cameras is below 10 ms, which is well below the required 25 ms to prevent a frame shift. All cameras provide a recording in which the event in front of the cameras will be captured by consecutive frames, not more than one frame number apart.

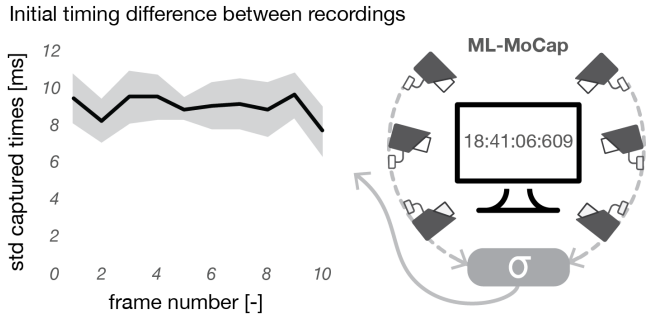

Fig. 4. Synchronization test setup and results. All cameras face one monitor displaying a digital millisecond clock. The time difference for the first ten frames between the six cameras (std of displayed time). The solid line shows the average std of the displayed time over the ten repetitions. The shaded area shows the std over the repetitions.

B. Validation of 3D finger motion capture

The 2D model trained by DeepLabCut labelled frames from the six cameras with a training error of 2.13 pixels mean average Euclidean error (MEA) between the manual labels and the ones predicted by DeepLabCut, and a test error of 2.39 pixels MEA. After calibration and triangulation with Anipose, the 3D coordinates retrieved were transformed to joint angles. As the coordinate systems of the two systems were not aligned, the relative joint angles were compared, see Figure 5. The difference between both systems for a full range metacarpophalangeal (MCP) joint motion was 7.5 degrees (expressed root-mean-square-error, RMSE). The DIP and PIP joints moved across a smaller range and thus resulted in smaller errors of 2.3 degrees and 3.2 degrees, respectively.

IV. Discussion

We described and evaluated a new markerless system for motion caption named ML-MoCap. The system is modular, i.e. it can contain a variable number of cameras. The first results presented here indicate that the ML-MoCap can record finger movements.

We performed a synchronization test with a setup with a limited screen resolution of 60 Hz. This limit resulted in an

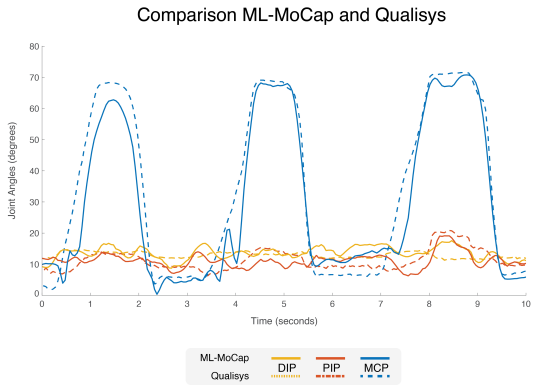

Fig. 5. Comparison of the joint angles retrieved with a marker-based system (Qualisys, dashed lines) and our markerless system (ML-MoCap, full lines). Three angles are compared: MCP (blue), DIP (yellow) and PIP (red) joints.

overestimation of the error between cameras, adding at most 16 milliseconds of error. A test conducted with a screen with a higher frame rate or blinking LEDs could show the actual delays between the recordings more accurately.

Our comparison based on simultaneous recording with two capture systems resulted in suboptimal circumstances for either system. ML-MoCap was blocking the 'sight' of the Opus cameras resulting in reduced visibility and suboptimal positioning of the hand. Currently, an accuracy of 7.5 mm RMSE was achieved over a range of 0 to 70 degrees flexion, a substantial difference between the two systems, the accuracy should be improved to provide biomechanically relevant data. Improvements can be made in the hardware, creating better lighting conditions and angles towards the cameras. The post-processing could be improved by extended training and increased iterations to specify the model for use in hands.

The model used to compare our system to a marker-based system was trained to recognise the reflective markers on one hand. While this provided the best comparison to the marker-based system, it does not reflect the optimal use of the ML-MoCap system. Markers reduce the advantage of the minimal environmental constraints of the system. Ideally, a hand model should be trained on multiple bare hands and without requiring attachments to the skin.

In the evaluation experiment, we tracked relatively slow finger motions. For these motions, a frame rate of 40 Hz was sufficient. However, for faster motions such as natural grasping, higher frame rates are required. Current commercially available cameras compatible with Raspberry Pi systems can record up to 60-90 Hz depending on the chosen image resolution. Next-generation products will allow for higher frame rates and thus serve a wider range of experiments, including natural and fast behaviour.

V. Conclusion

We present a new markerless motion capture setup and proposed integrating it with an automated analysis pipeline based on recently developed machine learning toolboxes. The ML-MoCap system is a compact, markerless, and modular motion capture solution. The setup eliminates the need for special-purpose hardware or labour-intensive post-processing and allows for out-of-the-lab experiments.

We evaluated the system by recording joint flexion sequences of the index finger in 3D using six camera modules and compared the trajectories to the results from a commercial system based on passive reflective markers. We obtained sub-frame synchronization of the videos from all camera modules and achieved a difference of 7.5 degrees RMSE between both approaches for index finger movements.

The ML-MoCap system can potentially be extended to capture motion of the full-body, specific body parts or other animals. This framework promotes integration with upcoming machine learning techniques such as DeepLabCut and AniPose, providing an alternative for commercial camera-based motion capture while eliminating the need for reflective markers and labour-intensive labelling. Further validation has to be performed for more configurations of the system.

References

[1] N. Seethapathi, S. Wang, R. Saluja, G. Blohm, and K. P. Kording, "Movement science needs different pose tracking algorithms," arXiv:1907.10226 [cs, q-bio], Jul. 2019.

[2] E. Knippenberg, J. Verbrugghe, I. Lamers, S. Palmaers, A. Timmermans, and A. Spooren, "Markerless motion capture systems as training device in neurological rehabilitation: a systematic review of their use, application, target population and efficacy," J NeuroEngineering Rehabil, vol. 14, no. 1, p. 61, Dec. 2017.

[3] S. Williams et al., "The discerning eye of computer vision: Can it measure Parkinson's finger tap bradykinesia?," Journal of the Neurological Sciences, vol. 416, p. 117003, Sep. 2020.

[4] E. Pantzar-Castilla et al., "Knee joint sagittal plane movement in cerebral palsy: a comparative study of 2-dimensional markerless video and 3-dimensional gait analysis," Acta Orthop., vol. 89, no. 6, pp. 656–661, 2018.

[5] C. D. Metcalf et al., "Quantifying Soft Tissue Artefacts and Imaging Variability in Motion Capture of the Fingers," Ann Biomed Eng, vol. 48, no. 5, pp. 1551–1561, May 2020.

[6] A. Perez-Carrillo, "Finger-String Interaction Analysis in Guitar Playing With Optical Motion Capture," Front. Comput. Sci., vol. 1, 2019.

[7] G. Litos, X. Zabulis, and G. Triantafyllidis, "Synchronous Image Acquisition based on Network Synchronization," in 2006 Conference on Computer Vision and Pattern Recognition Workshop, Jun. 2006.

[8] P. Shrestha, M. Barbieri, and H. Weda, "Synchronization of multi-camera video recordings based on audio," in Proceedings of the 15th international conference on Multimedia - MULTIMEDIA '07, Augsburg, Germany, 2007, p. 545.

[9] H. Aghajan and A. Cavallaro, Multi-Camera Networks: Principles and Applications. Academic Press, 2009.

[10] A. Mathis et al., "DeepLabCut: markerless pose estimation of user-defined body parts with deep learning," Nature Neuroscience, vol. 21, no. 9, Art. no. 9, Sep. 2018.

[11] N. Nakano et al., "Evaluation of 3D Markerless Motion Capture Accuracy Using OpenPose With Multiple Video Cameras," Front. Sports Act. Living, vol. 2, 2020.

[12] L. G. Wiedemann, M. Kampel, R. Planinc, and I. Nemec, "Performance evaluation of joint angles obtained by the Kinect v2," in IET International Conference on Technologies for Active and Assisted Living (TechAAL), London, UK, 2015, p. 6 .-6 .

[13] N. Vignais, D. M. Cocchiarella, A. M. Kociolek, and P. J. Keir, "Dynamic Assessment of Finger Joint Loads Using Kinetic and Kinematic Measurements," p. 6.

[14] P. Karashchuk et al., "Anipose: a toolkit for robust markerless 3D pose estimation," Neuroscience, preprint, May 2020.