MoCaPose: Motion Capturing with Textile-integrated Capacitive Sensors in Design-centric Loose-fitting Smart Garments

ANONYMOUS AUTHOR(S)
SUBMISSION ID: 3574

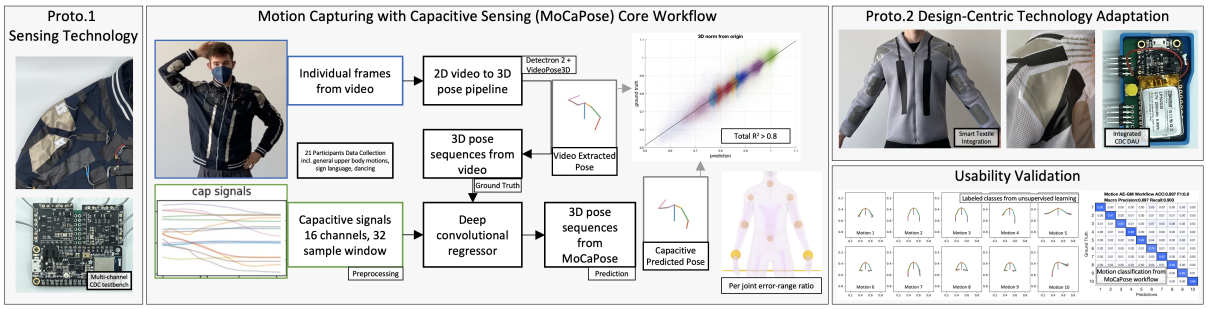


Fig. 1. Overall MoCaPose concept from textile-integrated capacitive sensors and deep regression for pose reconstruction towards design-centric smart garments.

We present MoCaPose, a novel wearable motion capturing (MoCap) approach to continuously track the wearer' upper body's dynamic poses through multi-channel capacitive sensing integrated in fashionable, loose-fitting jackets. Unlike conventional wearable IMU MoCap based on inverse dynamics, MoCaPose decouples the sensor position from the pose system. MoCaPose uses a deep regressor to continuously predict the 3D upper body joints coordinates from 16-channel textile capacitive sensors, unbound by specific applications. The concept is implemented through two iterations of prototypes to first solve the technical challenges, then establish the textile integration through fashion-technology co-design towards a design-centric smart garment. A 38-hour dataset of synchronized video and capacitive data from 21 participants was recorded for validation. The motion tracking result was validated on multiple levels from statistics ($R^2 \sim 0.91$) and motion tracking metrics ($MPJPE \sim 86mm$) to the usability in pose and motion recognition (0.9 F1 for 10-class classification with unsupervised class discovery). While the sensor placement within the jacket is not fully arbitrary, the sensing requirements impose few constraints on the actual fashion design. Overall, MoCaPose demonstrates that textile-based capacitive sensing with its unique advantages, can be a promising alternative for wearable motion tracking and other relevant wearable motion recognition applications.

CCS Concepts: • **Human-centered computing** → **Ubiquitous and mobile computing design and evaluation methods**; • **Computing methodologies** → **Neural networks**; *Knowledge representation and reasoning*.

Additional Key Words and Phrases: wearable sensing, capacitive sensing, deep learning, motion tracking, smart textile

ACM Reference Format:
Anonymous Author(s). 2022. MoCaPose: Motion Capturing with Textile-integrated Capacitive Sensors in Design-centric Loose-fitting Smart Garments. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 1 (August 2022), 40 pages. https://doi.org/10.1234/1234567.1234567

2474-9567/2022/8-ART $15.00
https://doi.org/10.1234/1234567.1234567

1 INTRODUCTION

Dynamic poses are fundamental sources of information in individual human activities and interactions. Fully capturing dynamic poses requires information from all degrees of freedom of all joints in the upper or lower body. Recent progresses from computer vision has led to systems that can robustly estimate full body dynamic poses from monocular camera videos without special visual markers [93, 98]. In terms of solutions with wearable sensors, most of the research focus on the intuition of inverse dynamics for pose reconstruction, for example, calculating the pose from the joint acceleration with inertial measurement unit sensors (IMUs) [51] or angle with proximity or stretch sensors [10]. Inverse dynamics requires reproducible, exact and mechanically stable sensor attachment [86], despite recent attempts on IMU-based motion capturing in loose clothing [61]. The need for a large number of sensing nodes firmly, exactly fixed to specific body locations makes wearable motion tracking systems unsuitable for many real-life applications. This restricts further fashion-technology co-design towards smart garments that can be welcomed by the broader public and professions, especially with soft textiles as IMU sensors are rigid bodies. Problems such as error drifting also requires special algorithms which further increases the computational complexity [3]. Consequentially, most established works were limited to sports or rehabilitation analytic domains where it is acceptable to wear specialized sensing suits [12, 19, 77].

As a result, wearable systems, such as those aiming to track user activities throughout daily life situations forego the dynamic pose information and instead depend on recognizing characteristic patterns in the signals from isolated sensors (e.g. in the smart watch on the wrist) to match certain activity contexts [68, 76, 94]. While such approaches have shown satisfying results in many areas in human activity recognition (HAR) and human computer interaction (HCI), e.g. daily activities [30] and interactions [79, 100], health [8, 75, 83], sign language [49], production and logistics [78], etc.; they fail to capture many relevant information and often struggle with complex or subtle activities, particularly when intra-class or inter-subject variability is involved. There are strong indication from the computer vision landscape that working with information on the motion capture level can be beneficial in such situations [20, 36, 74], as from the human perspective, we rely on more intuitive visual cues such as the natural movement of the body rather than the obscure sensor signals elusive to human intuition.

Capacitive sensing has received major attention in wearable sensing, mainly thanks to the sensor simplicity, that the only sensing element required is conductive traces or patches, which can be easily integrated with the fabrics of clothing [103, 104]. The sensing principle is also relatively easy to implement and has been widely studied. However, capacitive sensing has been mostly used in proximity detection and classification tasks, such as recognizing gestures or shape profiles [10, 13, 15, 34, 64, 95]. Our proposed method MoCaPose investigates using capacitive sensing as an alternative to wearable motion capturing, that does not require individual sensors to be fixed precisely according to a bio-mechanical model or rely on tight-fitting garments such as state-of-the-art IMU-based motion capturing suits. On the countary, MoCaPose facilitates smart garments that fuse technology and fashion, and can be worn in daily life settings.

1.1 Motivation

The intuition of solving motion tracking problems as mentioned above is usually placing the sensor directly correlated with the node or joint that is being tracked, and inversely solve a bio-mechanical problem. In this work, we introduce a paradigm shift that at the first glance, might be counterintuitive for motion tracking, but is persistent in human activity pattern recognition tasks which have been dealing with different sensor placements that might not be directly at the source of the motion [50]. We decouple the sensor placement from the points and joints in the bio-mechanical system (e.g. body skeleton). Specifically, we use the multi-channel capacitive sensing modality. The sensor signals might not be directly utilised to reverse-engineer the geometry information; however, the signals are definitively influenced by the change in the overall 3D space geometry. This geometry-to-sensor relationship indicate there might also be a definitive inverse sensor-to-geometry relationship. We utilize deep

regression to model the relationship between the sensor signals and the pose joints, bypassing the problems of complex 3D reconstruction or inverse dynamics with decoupled sensor placement and joints system.

This decoupling paradigm can largely benefit the wearable technology beyond the novel sensing modality and deep learning algorithms. For activity recognition, moving from specific sensor signals of varying channels and placements across different hardware prototypes, to the universal standard human pose system to describe activities provides a common ground among the heterogeneity nature of wearable devices. Such a common ground would help preserving knowledge from previous studies for the future studies. This also provides bridges to trans-modality knowledge transfer with computer vision technologies.

For smart garment designs, our approach allows more cohesion during fashion and technology co-design. For example, smart fashion designers can incorporate functional fabric patches into their designs that not only consider the sensing technology but also aesthetics, garment structures and fashion trends. Then the deep regression model can be trained or fine-tuned with data from the new design layout to realize motion tracking. This changes the traditional smart wearable process where design is dictated by the requirements of technology, to a design-centric approach that brings design and style forward with adaptive technology.

1.2 Hypothesis

Our approach can be summarized as the following hypothesis: there exists a complex yet definitive relationship between the wearer's body pose, and the multi-channel capacitive sensor's values. This relationship is a manifestation of the geometry of the body and tissues which influences the positioning and deformation of the capacitors' plates, and the dielectric's micro composition inside the capacitors. Such a relationship can be abstracted as a multi-input (pose joints) multi-output (capacitive signals) complex system. We currently lack the necessary understanding to precisely model such a physics system at reasonable costs to either predict the capacitive channels from poses or, inversely, reconstruct the poses from capacitive signals, especially when the output sensor points are not physically placed at the input joint locations. However, we can leverage deep learning, with sufficient observations data points of the system in sufficient states (poses and motions) to model the behavior of such a system.

1.3 Novelty and Contribution

Our novel approach provides similar pose estimation results as state-of-the-art (SOTA) systems, but without considering the complex inverse dynamics engineering. At the same time, it opens up more flexible possibilities for smart garment designs. Through the deep interdisciplinary convergence of wearable sensing, smart textile integration, deep learning, and computer vision, our contributions revolve around validating the above-mentioned hypothesis:

(1) We developed two iterations of prototypes: the first iteration focused on building a viable sensing hardware beyond the SOTA solutions, with the second towards a design-centric integrated smart fashion piece.
(2) To provide the validation data, we conducted one experiment with each prototype, in total containing 21 participants with 38 hours of recording, one of the larger datasets among the SOTA.
(3) The MoCaPose approach can predict poses from independent short time windows of capacitive signals with a straight-forward yet appropriately designed deep convolutional regressor. From various statistical and pose tracking metrics, we proved our hypothesis and placed our novel approach on par with SOTA motion tracking methods from other modalities.
(4) We further validated the usability aspects relevant to smart wearables including classification tasks, reducing channels, motion speed relevance, and interference.
(5) We open the access to our data, hardware design and the code to perform the core MoCaPose workflow for new designs, to the wearable community to inspire future works in this direction.

Table 1. Comparison of Related Work in Motion Tracking with Wearable Sensors and Capacitive-driven Activity Recognition

Study	**Sensing Modality and Form Factor**	**Sensor and Tracking Points**	**Algorithms**	**Recognition Performance**	**Tracking Performance**	**Dataset** (p:participants)
DIP [43]	IMUs	17 sensors for full body pose	bi-directional RNN	none	15.85° angular error	10 p + TotalCapture
Teufl, 2019 [87]	IMUs	7 sensors for 7 joints from the lower body	EKF, global translation and inverse dynamics	none	RMSE 2.28°– 2.58°	28 p
Boddy, 2019 [19]	IMU sensor (Motus-BASEBALL TM)	1 forearm sensor for the elbow baseball motion	proprietary	none	average R^2 = 0.724, best arm slot R^2=0.975	10 p, 10 to 14 throws
TIP [46]	IMUs	6 sensors for full body pose	transformer model	none	root error 0.129 meters	TotalCapture (public)
HybridCap [56]	IMUs and camera combined	4 IMUs with 1 camera	gated recurrent unit and inverse dynamics solver	none	MPJPE 43.3mm	AIST++ (public)
EM-Pose [48]	on-body EM field transceivers	12 nodes for full body pose	learnt gradient descent with public dataset prior	none	MPJPE 31.8mm, 13.3°	5 p, 36 minutes
Liu, 2020 [59, 60]	microflow sensor and IMU on a wrist band	1 sensor for a single joint angle	EKF, regression neural network	none	RMSE 0.4 °	4 p
Atalay, 2018 [10]	capacitive strain sensor on the knee	1 sensor for 1 knee joint	statistical analysis	none	R^2=0.997	1 p
PersiSense [95]	capacitive ring-shaped sensor	4 capacitive sensors on a ring for all finger joints	LSTM based regression model	none	MAE 13.02°	17 p
Frediani, 2021 [34]	piezo-capacitive stretch sensors	2 sensing strips on the back of the trunk	statistical analysis	none	RMSE 8° ∼ 15°	5 p
C-Stretch [7]	capacitive stretch-sensitive sensor	2 sensors for transversal neck rotation	statistical analysis	none	RMSE 5.86°	2 p
Zheng, 2018 [105]	capacitive	6 channels for forearm motions	quadratic discriminant analysis (QDA)	16 motions 92%, 5 motions 98.7%	none	7 p
Bian, 2019 [17]	passive capacitive sensing	3 sensors on wrist, calf and in pocket	residual deep convolutional neural network	63% for 7 exercises, 91% counting	none	11 p
MoCapaci [13, 14]	capacitive sensors from OpenTheramin	4 channels for 20 upper body poses and gestures	Conv2D	97.17%	none	14 p
Wong, 2021 [96]	capacitive textile sensors	5 capacitive electrodes, 1 on each finger	bayesian classification	99% for 26 hand gestures	none	10 p
TouchPose [5]	capacitive touch screen	capacitive touch screen for hand pose	multi-task neural network	finger detection 91.1%	20.7 mm depth error	10 p, 65,374 records
Ours	multi-channel capacitive patches in 2 iteratively designed jacket prototypes	16 channels, for 8 upper body joints	deep convolutional regressor, auto-encoder, gaussian mixture clustering	79.2% F1 for 10 self-clustered pseudo-classes of poses and 90% F1 for 10 pseudo-classes of motions	R^2=0.892 (3D), 0.932 (horizontal), 0.895 (vertical); MPJPE 0.032 (pose normalized), approx. 86mm ; RMSE(P) approx. 89mm; RMSE(A) 12.8°	21 p, 2 experiments, in total 154 recordings, around 38 hours

1.4 Paper Structure

Section 1 introduces the motivation and contribution. Section 2 reviews the related work from the aspects of smart textiles, motion sensing and capturing with wearables, and capacitive sensing. Section 3 describes our smart garment design process of Proto.1 and Proto.2 from the sensing hardware to the textile integration. Section 4 details the procedure of the data collection experiment for the two prototypes. Section 5 explains the algorithms for the core pose estimation approach of MoCaPose, including data processing, establishing ground truth and the deep convolutional regressor. In Section 6, we discuss the pose estimation results with the dataset of Proto.1, from the signal level to the motion tracking metrics and statistical correlation. Section 7 presents the pose estimation results of the improved Proto.2 with similar key metrics, together with some comparison with Proto.1 especially on faulty signals and predictions. In Section 8, we evaluate our approach from various usability and practical perspectives of human activity recognition and smart garment design. In Section 9, limitations and outlook prospects are discussed with the focus on the wearable community. Section 10 concludes the entire paper.

2 RELATED WORK

2.1 Smart Textile Integration Techniques

As ubiquitous items, sensors integrated into clothing are always worn, and users do not need additional (hand-held) devices [24]. To adapt the electronic components to the properties of the textile substrate, e.g. flexibility and drapability, the electronic components or the sensing elements should also have textile-like properties if possible. It therefore makes sense to use conductive textile materials as sensors. In addition to the properties that match the substrate, this material has the advantage of wide area coverage and freedom of tailoring into desired shapes. Techniques such as that proposed in [97] is easier and faster to implement than the conventional textile processing techniques of sewing.

Other published papers use the integration of conductive textile elements at fiber level [102, 103], at yarn level [4, 67, 84], or at surface level [10]. For the integration at fiber level, the conductive fibers are spun into the yarn during the spinning process. In case of the integration at yarn level, conductive yarns are introduced into the textile surface of the wearable using various techniques, e.g.: stitching. For the integration at surface level, conductive fabrics can be cut into the desired shape and used as a textile sensor. This technique can be used with or without a substrate during fabrication.

2.2 Motion Sensing and Capturing with Wearables

Pose tracking or motion capturing (MoCap) are usually implemented by computer vision methods including digital cameras or depth sensor arrays. Because many methods including visual landmark detection, markers, multi-camera, temporal dilation, optical flow, converge to provide high precision tracking results [31, 33, 72, 91, 93]. Computer vision solutions have several practical problems, however, including diverse clothes, arbitrary occlusion, occlusion due to viewing angles in monocular camera settings, background, need of robust camera calibration model in multi-camera settings, difficult to perform inference in the wild for multi-camera pose estimation, need of time to set up, lighting conditions, etc.[82, 93].

Currently, most of the mature wearable pose tracking methods are based on IMU sensors [37, 69], with many commercially available solutions for both general purposes and specialized applications [77]. Inverse dynamics is the classical technology of pose reconstruction from wearable IMUs which is still being continuously improved [21, 25, 61, 87, 106] with available open-source frameworks [6]. The pose is considered a bio-mechanical system joined together by rigid bodies representing major body parts. Inertia and dynamic properties such as acceleration and rotation from defined points on the bio-mechanical system can be used to derive the spatial positions of the connected rigid bodies [12]. Error drifting is a major problem in pure IMU-based systems as the acceleration error accumulates to the position errors with the progression of time. Both algorithmic solutions such as extended

Kalman filters (EKF) [87], and hardware solutions such as sensor fusion with other sensors [59] are needed. A method for resolving drift and instability problems of inertial measurement methods for limb motion detection was proposed by Liu, et al. An IMU and an additional microflow sensor data was fused to gather precise velocity, acceleration and attitude data by obviating mathematical integration before training an intra-limb coordination neural network model [59, 60].

The motion capturing field from the computer vision society has built an archive of many publicly available datasets, some of which contain multi-modal data including precision marker-based tracking systems, multiple cameras, and even IMU sensors, such as AIST++ [53, 90] and TotalCapture [89]. With the wide availability of millions of data points including synchronized IMU sensors and motion captures, wearable IMU-based pose reconstruction has seen a major shift in the research directions. Instead of solving the inverse bio-mechanical system approach, cutting edge deep learning methods such as recurrent or transformer models can leverage the prior knowledge existing in the datasets and provide pose estimation results beyond what is possible through inverse dynamics, especially in the error drifting problem [43, 46, 56]. In other words, the pose reconstruction algorithm can learn the certain ways human moves from those datasets, and provide more reliable pose estimation. The mean-per-joint-position-error (MPJPE) can be around 40mm or 15°[56, 63].

In the wearable sensing discipline, other sensing modalities were also investigated for recognizing or tracking motions. Kaufmann, et al. used up to 12 electromagnetic (EM) field transceivers around the wearer's body, to estimate the pose with the help of prior knowledge from the public dataset AMASS[63] by learned gradient descent, showing MPJPE of 31.8mm and 13.3°. However, the system is susceptible to EM distortion caused by metal or electronic objects within 1.5 meters to the user. HybridCap[56] improved the tracking accuracy by employing a hybrid solution fusing the data of a single camera and four wearable IMUs with 43.3mm MPJPE and 60 frames-per-second real-time performance.

2.3 Capacitive Sensing in Wearable, Ubiquitous and Mobile Computing

Capacitive sensing is well studied in wearable, ubiquitous and mobile computing, as the flexible nature of the sensing element and the versatility to couple with various physical properties (e.g. touch, proximity, deformation, etc.) enable an expansive landscape of possibilities of garment and gadget designs [9, 40, 62]. The sensing principle can be categorized to measuring the variation of charge [18, 28] and the frequency [16, 29] caused by body actions.

Charge-based capacitive sensing is relatively easier to implement, as the circuit only needs to measure the current or voltage variation signal, which can be realized by common operational amplifiers (Op-amps) and analog-to-digital converters (ADCs) with slow sampling rates. Choi, et al. integrated capacitive sensors inside a driver's seat, and showed statistical correlation between the voltage measurement and the drivers' cardio and body activities [27]. In [17], passive capacitive sensors were placed on three body positions (wrist, calf and in pocket) to detect gym workouts exercises with a residual deep convolutional neural network. Al-Nasri, et al. used two capacitive stretch sensors placed on each side of the neck and showed statistical correlation between the measured voltage and the neck rotation angle [7]. CapGlasses [65], a sensing glasses prototype with several passive metal capacitive electrodes placed around the frame and a transparent indium tin oxide (ITO) capacitive plate covering the lens, demonstrated reliable recognition performance among 12 facial and head gestures.

It has been pointed out that frequency-based capacitive sensing is more robust against EM interference than charge-based sensing in studies such as [64]. In [64] the circuit driver of the above-mentioned CapGlasses prototype was upgraded from instrumentation amplifiers that measures current to a resonant circuit based on an LC tank. This enabled the prototype to show robust performance while untethered from the ground, which is important for smart wearable devices. Cheng, et al. [23] developed a neckband with conductive textile patches driven by Colpitts oscillators to monitor nutrition intake. With 5 capacitive electrodes driven by 555 timers, one

tied on each finger, 26 hand gestures could be recognized with Bayesian classification in [96]. In [39], 555 timers were also used to drive capacitive proximity sensors to improve accelerometer-based daily activity recognition in the form of a smartwatch.

Although frequency-based sensing requires more sophisticated resonant excitation signals and frequency counting, modern integrated circuit (IC) technologies has already packaged most necessary electronic front-end in miniaturized packages such as the recent chip models FDC1004 and FDC2214 (Texas Instruments, USA) [88]. Wilhelm, et al. integrated 4 passive capacitive electrodes driven by FDC1004 in a ring-shaped device PeriSense placed on the middle finger to predict the finger angles with a regression model based on long short-term memory (LSTM) [95], demonstrating 13.02° mean absolute error (MAE). Bian, et al. demonstrated a smart wristband with 4 single-end electrodes driven by FDC2214, which performs runtime gesture recognition with a lightweight convolutional classifier in the micro-controller [15].

Atalay, et al. [10] used a capacitive strain sensor on the knee joint and showed strong correlation between the bend angles and the capacitive values. While the method was performed on a single participant and a single knee joint in lab settings, it reveals a promising idea that capacitive sensing may be scaled to estimate motions of more body parts. Bello, et al. connected a theremin circuit to 4 capacitive textile antennas integrated inside a MoCaBlazer prototype to recognize up to 20 different upper body movements with a deep convolutional classifier [13, 14]. While MoCaBlazer was performing only classification for defined movements, the classification with 97.17% accuracy might have been supported by reproducible relationships between capacitive signals and the body pose.

As further summarized in Table 1, the SOTA has shown a barrage of motion tracking with wearable sensors and capacitive-driven activity recognition applications. However, to the best of our knowledge, no previous work has shown a viable solution for continuous upper body pose estimation like our work, which can be brought on par with other modalities such as computer vision and IMUs. Moreover, our approach welcomes design-centric smart garment integration as the sensing pathes are not bound to precise locations, which is generally not a concern of the SOTA solutions in wearable pose estimation.

3 SENSING HARDWARE AND SMART GARMENT DESIGN

We adopted an iterative technology-fashion co-design approach during the implementation of the hardware to validate our hypothesis. Two prototypes, Proto.1 and Proto.2 were designed consequentially. Table 2 shows a comparison of key aspects that we considered while choosing the capacitive sensing modality as opposed to IMUs for designing our prototypes, that could provide reference for consideration for future garment designs. In general, IMU based pose estimation has a strong algorithmic background support from inverse dynamics, EKF to the recent deep inertial posers. But the IMU modality is still not as flexible for textile integration on the similar level as smart textile based capacitive sensing, despite the recent attempt in loose fitting garments [61].

An alternative approach is the use of textile sensors. The use of textile materials as the carrier material (garment) and as the sensor material offers several advantages. Due to the same or similar fiber-based structure, similar properties can be derived in terms of elasticity, drapeability and breathability. These properties provide the basis for a less noticeable adaptation of the textile sensor on the carrier material, compared to conventional sensors. The use of textile-based sensors increases user acceptance, because the wearing comfort is not impaired even with clothing that is close to the body. Furthermore, damage to the carrier material can be prevented. Due to the textile structure of the sensor, processing with textile-typical machines and processes is possible to a large extent. The size and shape of textile sensors can be individually adapted, so that large-area or spatially resolved detection is also possible. [58]

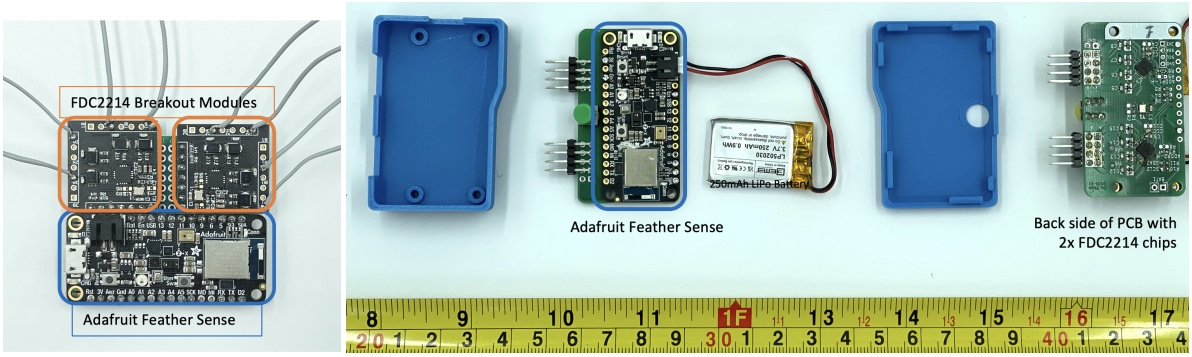
Table 2. Characteristic Differences Between IMU-based and Capacitive-based motion tracking wearables

Aspects	On-body IMU Network	Capacitive (MTCP)
Outputs	acceleration, angular velocity, orientation	capacitor value, or change in charge or frequency
Channel dependency	3 axis interrelated per sensor	independent
Calibration	factory calibration and runtime initial calibration (e.g. rotation transformations between nodes)	not needed (all channels are normalized during pre processing)
Influence factors	value depends on motion	proximity, dialectric, plate deformation
External interference	susceptible to physical impact on the sensor rigid body	physical contact from conductors (e.g. human body) on the capacitive patches
Error sources	transient, and error drifting with time progression	transient within independent short time window
Motion tracking algorithms	inverse dynamics, or deep learning without mechanics knowledge (e.g. transformer inertial poser)	deep convolutional regressor (this work)
Pose prediction time window	progressive sliding windows with context (temporal integration)	independent short time window
Textile Integration formats	rigid attachments or buttons	soft conductive fabric patches, traces or threads
Attachment	normally needs to be precisely fixed to the predefined locations, even glued to the skin in some cases	can be integrated with the fabric of different garment designs
Routing	power, ground and digital bus lines (at least 4 in total) from every node to the central controller	only a single unshielded line for connecting every conductive patch to the CDC

3.1 Multi-channel Capacitive Sensing Implementation

Among the existing works, high channel count capacitive sensors can be implemented in a matrix or honeycomb structure [35, 54], which increases the sensing nodes by multiplying the excitation and measurement channels; or through time-interleaving multiplexing which reuses the same channel circuit for different electrodes [70]. However, these implementations do not fulfill our requirement due to drawbacks such as slower sampling rate per sensing node, crosstalk between channels or the physical routing complexity between channels. We thus require a sensing electronics system capable of driving multiple individual channels that are highly robust against interference and matching the sampling rate of the video frames (usually 30 Hz). The sensing electronics should also have as little physical footprint as possible and can operate wirelessly as the requirements from smart garments.

As we discussed in Section 2.3, the most promising option for our design is the FDC2214 4-channel 28-bit capacitance-digital converter (CDC). It has many industry-leading features such as adjustable low power consumption and high sensitivity [88]. The features that interested us the most are that it has strong robustness against interference since it operates with the frequency as we mentioned in Section 2.3 and supports both single-end or differential sensing modes. It integrates the entire analog front-end of 4 channels to a miniaturized package, requiring only one pair of external capacitor and inductor for every channel. We use the single-end mode as it is easier from the garment geometry design perspective; differential mode requires pairs of conductive patches which might complicate and restrict the design process. We configured it to sample at 30Hz, and with an external inductor of 18uH and capacitor of 33pF. The inductor and capacitor forms an LC oscillation circuit, which the CDC compares with the main 40MHz clock input to set the excitation frequency. These two passive components are needed only to set the excitation signal's frequency, thus small and low-cost selections (e.g. 0603 footprint and 20% tolerance) are sufficient, and other values are also possible. Our configuration operates with an average frequency of 13.7Mhz, with variations of approximately ± 1MHz from channel to channel caused by the passive components' manufacturing tolerance.



(a) Version 1 in Proto.1.

(b) Version 2 in Proto.2.

Fig. 2. Capacitive sensing modules and enclosure.

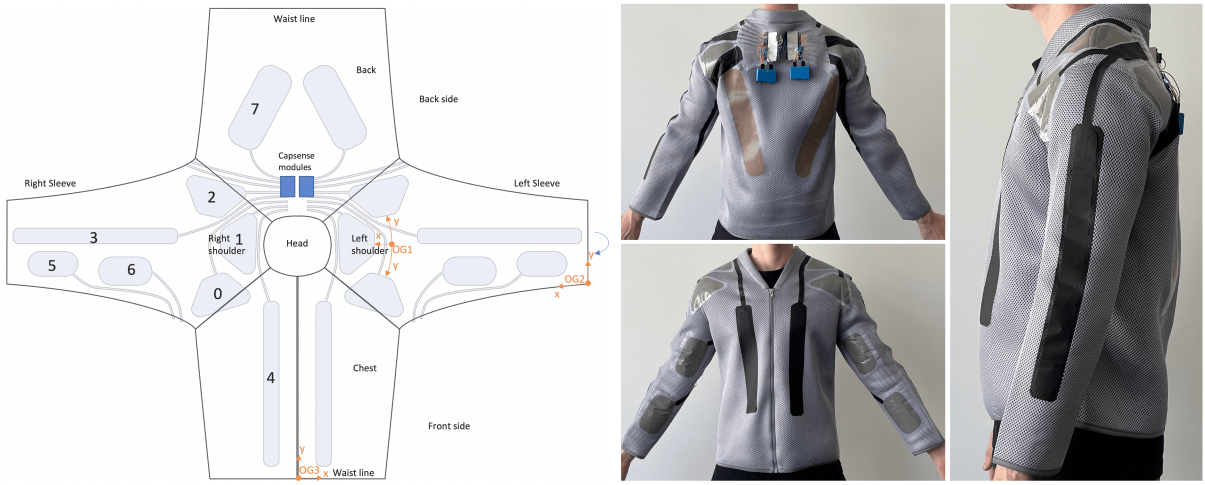
We implemented our custom data acquisition unit (DAU) as shown in Fig. 2. We used the Feather Sense board (Adafruit Industries) with the nRF52840 (Nordic® Semiconductors) as the microcontroller unit (MCU). Two FDC2214 CDCs were connected to the same MCU through the I2C bus with two addresses. Every DAU thus supports 8 individual single-end capacitive channels, or 16 differential channels. The realtime sampling data is sent to receiving devices via Bluetooth Low Engergy (BLE), which is compatable with a wide range of devices from smartphones and computers to other custom embedded systems. In our study, we use two DAUs to drive 16 single-end conductive textile patches as capacitive plates in a prototype. Both DAUs can be connected to the same computer via BLE for synchronized data logging.

Fig. 2a shows the initial testing module of the DAU, which we used in Proto.1. It included two custom-designed FDC2214 breakout modules which were connected to the MCU via jump wiring on a prototyping breadboard. For Proto.2, we improved the design shown in Fig. 2b by integrating two FDC2214 CDCs into the backside of the host printed circuit board (PCB) designed in KiCAD. We tested the power consumption of the DAUs with a HM7042-5 power supply (Rhode & Schwarz®) set to 3.70 Volt output with 0.100 Ampere fuse protection. During testing, the DAUs were fully operational, connected to all 8 sensing patches in the prototype while streaming data with BLE. The power consumption of a single DAU from Proto.1 is between 0.018 Ampere; and 0.016 Ampere in Proto.2. With a LiPo battery of 250mAh capacity, the low power consuming DAU could operate for more than 12 hours continuously. A custom-designed 3D-printed enclosure with screw-less clipping mechanisms protects the DAU. To ensure better fit, the PCB with components were modelled together with the 3D enclosure in Audodesk® Fusion 360™. We added usability features such as improved battery charging behavior, standard pin-headers for sensor connections, and a latch-action power switch that sits flush with the enclosure when turned on and extrudes when off.

3.2 Proto.1: Proof-of-concept

The first prototype was designed to test the hypothesis that multi-channel capacitive sensors in a smart clothing can be used to reconstruct body poses. We thus referred to the pose joints system from the Human 3.6M dataset [45], which considers hip, spine, thorax, shoulders, elbows and wrists of the upper body. We empirically placed 8 sensor patches on either side of the jacket as shown in Fig. 3a, based on observations that upper body motions should cause as much displacement among the patches as possible. For either side, at the shoulder joint, we placed three trapezoid shaped patches to cover the upper, front and back side (CH0, CH1, CH2). One long stripe covered most of the length of the arm through the outer elbow (CH3). Another long stripe was placed from the

(a) Proto.1.

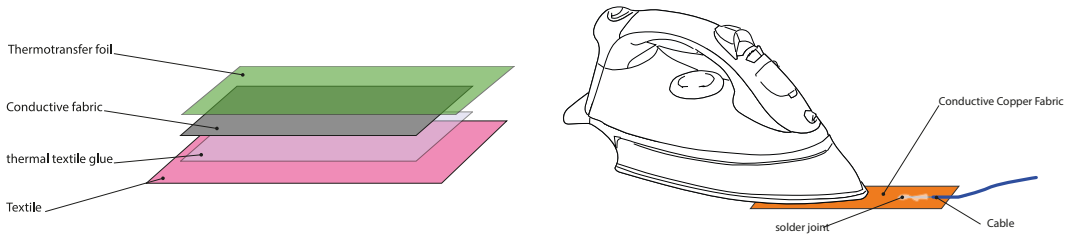


(b) Proto.2.

Fig. 3. The smart textile schematic and photos of both prototypes.

chest to the waistline (CH4). Two shorter and wider patches were placed on the inner side of the triceps (CH6) and lower arm (CH5). And one wide, long stripe was placed at the back (CH7).

For the functional sensor patches, we selected off-the-shelf conductive fabrics with nickel or copper threads and coating, which were typically sold for blocking EM signals, such as those on the linings of some wallets or car-key bags. The patches were fixed to a bomber-style jacket of men's L size by simple stitching. We connect the sensor patches to the DAU's single-end inputs by standard unshielded electrical cables of 28 American Wire Gauge (AWG). Since the conductive fabrics do not attach to solder, we used double-sided conductive tape Tesa® 60372 to establish stable electrical connection between the fabric and the cable. The capacitive patches were



(a) the layer stack of the protected textile capacitive sensor plate in the second prototype

(b) the layering technique of textile integration in the second prototype

Fig. 4. Illustrations of textile integration techniques in Proto.2.

exposed and all other parts from the sensing functions, including the cables, DAUs and LiPo batteries were attached to the jacket with Tesa® Eco Repair fabric tape.

3.3 Proto.2: Smart Textile Integration

The second prototype was a design-centric garment piece that fuses design and functionality as shown in Fig. 3b. It was built on a slicker jacket template with breathable mesh materials. With Proto.2, we focus on a more applicable jacket design ready for everyday usage. Therefore, DAUs are positioned on the center back to ensure optimal routing paths from each sensor patch, causing the least disturbance in the wearability of the jacket.

A suitable layer system was developed to place the textile sensors on the surface of the garment as shown in Fig. 4a. The conductive elements of the capacitive sensors and conductor tracks are made of conductive silver-plated fabric [2]. To create the capacitive sensors, the fabric is tailored into patches and 5mm wide connection traces to the DAUs(Fig. 3b). The garment fabric schematic followed the similar design for the first prototype, but also included aesthetic aspects. The exact dimensions and placement of both prototypes are listed in Table 5. The conductive panels and traces are prepared with double-sided adhesive material [1]. The adhesive material used is highly flexible and transparent. Due to its high flexibility, the material matches the typical textile properties of the underlying textile substrate.

The process of integrating the layer stack onto the garment is illustrated in Fig. 4b. First, the positions of the conductive patches were determined, and then gently fixed with an iron. Then the conductive traces were ironed onto the garment to route from the patches to the designated DAUs. There was sufficient overlapping area between conductive traces and conductive panels. Only then were the insulation fixed with a thermal transfer press, thus creating a strong bond between the textile carrier material and the layers of the textile sensor patch construct, as well as stable electrical connections with overlapping conductive textiles. The layers of the sensor were covered with thermotransfer foil as protective insulation. After this process, the sensor patch is firmly adhered to the textile and can only be separated again with strong heat of approx. 90 degrees Celsius.

Compared to the hand stitching in Proto.1, the method of thermal transfer pressing used in Proto.2 is more efficient and durable [97], and all the layers were tightly bond in one textile stack together with the garment's fabric. In addition, compared to Proto.1, replacing cables with conductive textile traces with insulation creates connections that are more robust against tearing and pulling of the garment piece. To connect the conductive

[1] https://www.lotustransfers.com/WebRoot/SageSMB/Shops/Lotus-YourOneStopShopForTextilePrinting/MediaGallery/Info_PDF/Products/DatasheetFolien/SISER/PS-Adhesive-EasyWeed-Adhesive/PS-ADHESIVE-EASYWEED-ADHESIVE-DE.pdf

traces to the DAUs, small connecting cables of 30 AWG were soldered onto the terminal of the conductive traces. A conductive copper textile stripe was used as a bridge between the conductive traces and the cables, as the textile material used for the conductive traces and patches were not suitable for soldering.

4 EXPERIMENT DESIGN

To validate our approach and hypothesis, we designed an upper body motion capturing experiment which would provide synchronized video and capacitive data from our MoCaPose jacket prototypes. SOTA marker-less computer vision models were used to extract pose sequences from the video of a smartphone's monocular camera, that are also synchronized with the rest of the data sources, as attaching additional markers is not easily compatible with custom smart garment designs. Being able to gather the ground truth via commercial smartphone cameras also simplify the scalability and reproduction requirements of our approach. The purpose of the experiment was to introduce as many variations of poses and motions as possible, to provide observations of different system states of the multi-input multi-output system described in Section 1.2.

The experiment was set in a spacious room, where the participant would stand in front of a white wall wearing the jacket prototype and follow along videos that showed various movements. The participants were instructed that they did not have to follow the motions from the videos exactly, but rather perform them in the most natural way for themselves. That is because the primary goal of our approach is to continuously track different poses rather than to classify exact gestures. The instructive videos themselves were composed after discussion of typical activities related with the upper body, which were segmented into two parts as listed in Table 4. The whole protocol was followed for every recording session and participant. Additionally, we gathered body measurements and their feeling about the fitting of the jacket to ensure a balanced set of participants representing different body sizes. The participants gave informed consent in accordance with the policies of the Ethics Team of the [Redacted], which approved the experimental protocol. More details for reproducible purposes are introduced in Appendix A.

In total, our experiments were divided into two phases:

- The first phase was to gather enough data to verify our hypothesis with the first proof-of-concept prototype. We have recruited 21 participants for this phase, consisting of eight women and 13 men. We recorded on average six sessions per person, with some participants recorded less or more sessions depending on their availability. Overall, we conducted 118 recording sessions with a total of 29 hours. The set of participants has the following body dimensions: the height in a range of 170cm to 189cm with a median of 180cm; the shoulder girth in a range of 95cm to 128cm with a median of 115cm; the bust girth in a range of 80cm to 105cm with a median of 93cm; the waist girth in a range of 59cm to 112cm with a median of 82cm. 14 participants reported comfortable fit and seven reported loose fit wearing the jacket.
- The second phase was to validate the second prototype with better textile and electronics integration. Six participants from the participants pool of the first phase were invited back for this phase, consisting of one woman and five men. Each of them recorded six sessions, in total 36 recording sessions and cumulatively 9 hours of data.

In either phase, both instruction videos were played for every participant.

5 POSE TRACKING WITH DEEP REGRESSION

The core algorithm that drives the MoCaPose jacket to perform upper body pose estimation from multi-channel capacitive signals is a deep convolutional regressor. Fig. 1 summarizes the core workflow of training the deep regressor.

5.1 3D Pose Ground Truth

The ground truth of the dynamic poses were extracted from the experiment videos by a computer vision pipeline with public access state-of-the-art models. First, every video frame was processed by the Detectron2 library [98] with its pose estimation model Keypoint R-CNN, which generates 2D poses according to the 17-keypoint pose system based on the COCO dataset [57]. Then we used VideoPose3D [71] to interpolate the temporal sequences of 2D poses into 3D poses with another 17-joint kinematics system based on the Human 3.6M dataset [45]. VideoPose3D was selected because the model leverages the time domain to interpolate the depth information from 2D poses, which has been shown to provide superior 3D estimation than other models [93]. Our ground truth for the 3D pose was thus the output of VideoPose3D.

The poses generated from this pipeline were scaled with the video frame. Thus, the sizes of the skeleton and bones were influenced by how tall the person is, and how close they were to the camera. To reduce such variation and have all the poses from the recording correspond to a relatively stable scale across different persons and recordings, we normalize the joint coordinates with the following process. First, within each recording, we calculate the average length $avg(bonelengths)$ of the bones of hip-spine, spine-thorax, left-right-shoulder, left shoulder-elbow, right shoulder-elbow, left wrist-elbow and right wrist-elbow. Then the pose joint coordinates are normalized according to the following formula:

$$Joint_{new}(x, y, z, t) = (Joint_{old}(x, y, z, t) - HipJoint_{old}(x, y, z, t)) / (4 \times avg(bonelengths)) \quad (1)$$

So that the normalized poses are centered at the hips, and scaled according to the average value of the above mentioned six bones.

Since our prototype only covers the upper body, we remove the joints that are below the hip and above the neck and use hip as the center point (0.5, 0.5, 0.5) in the range between [0, 1] of the 3D space. In the end, 8 joints were considered: {Spine, Thorax, Left Shoulder, Left Elbow, Left Wrist, Right Shoulder, Right Elbow, Right Wrist}, which we use as the ground truth for the pose estimation.

The pose normalization also provides the approximated scaling from the normalized relative unit to the absolute meter measurements, even though the pipeline does not provide absolute distances. We select several participants and compare their shoulder width measurements with the normalized pose joints and decided the global scaling factor from the pose coordinates to meters to be 1 unit = 1.28 meters. However, since this pipeline was not calibrated to provide precise absolute distance measurements, and our approach is aimed towards wearable motion pattern sensing in HAR where relative units typically suffice; we refer primarily to the normalized units in presenting the results, unless when explicitly mentioned otherwise. The approximated meter unit is referred to in further discussions on the usability and comparison with the state-of-the-art. Furthermore, since the VideoPose3D has a Mean Per Joint Position Error (MPJPE) of 0.046 m, we will add this error to our results in approximated distance error results. Thus errors (such as RMSE) in our normalized scale can be approximated to distance scale of meters with the following conversion:

$$Error_{distance} = Error_{normalized} \times 1.28 + 0.046 \quad (2)$$

5.2 Capacitive Data Pre-processing

The capacitive sensing chip FDC2214 in our prototype provides very high signal quality which does not require extra signal conditioning. In rare conditions, we observed gradual drifting during a recording session, of which the trend-line approximates a straight slope. Thus, we remove such drifting by subtracting the trend-line determined by 1-degree polynomial fitting, which also centers the signal around zero. The capacitive values are essentially the frequency. And in our case, we have fixed excitation frequencies below 2 MHz for the channels, we normalize all sensor values by dividing with 2000000 (2 MHz) then plus 0.5, to place the values in the range of [0, 1]. In the first prototype, there were rare faulty peaks in the sensor values due to unstable connected and exposed conductive

patches. We removed those anomalies with outlier removal by manually setting the percentile threshold from observing the data. In the second prototype, however, with improved textile integration and isolation, neither drifting and outlier removals were observed.

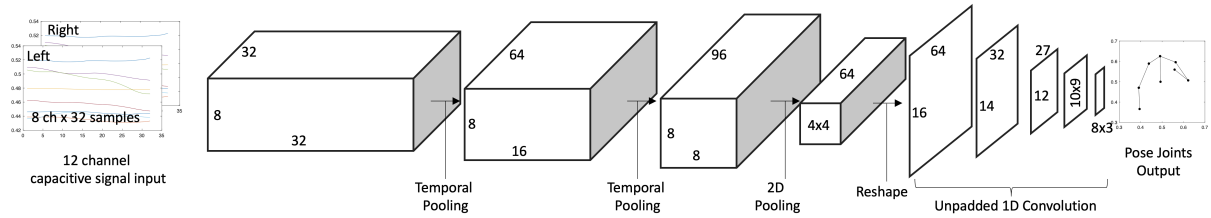


Fig. 5. The model architecture of the MoCaPose deep convolutional regressor.

5.3 Deep Convolutional Regressor

The appropriately designed yet simple core regressor shown in Fig. 5 takes the task of modelling the complex relationship between the dynamic capacitive sensor signals and the wearer's body poses which we introduced in Section 1.2. We consider a transformation from temporal sequences of multiple capacitive channels from a short time-window to the pose at the end of the window. The time window was selected as 32 samples for the ease of temporal pooling, equivalent to 1.07 seconds. The step between adjacent time windows is 2 samples. However, different from IMU-based deep inertial poser solutions such as [43, 46], where the model requires sequential input of the sliding windows; the windows in our approach are independent from each other. The model generates prediction on individual windows alone and does not require recurrent operations with previous windows. Since the capacitive channels were placed symmetrically between the left and right side of the jacket, we also incorporate this placement into the neural network design by separating the capacitive channels into two dimensions. The first dimension of 8 represents the placement (e.g. front shoulder, etc.) and the second dimension of 2 represents the left or right side. Thus, the input to the model is shaped as (32, 8, 2) for every data point. For the output, we consider pose estimation of the 8 joints in the upper body in the 3D space (8, 3).

Our design policy for the deep regressor is that it shall first combine the time, sensor channel and side information into a single layer, then compress the channels of information to the desired output. The regressor should hold enough information capacity to model the multi-input multi-output complex system as we described in Section 1.2. And with enough observations of the system, the training progress would optimize the learnable parameters to best approximate such system.

The deep regressor first use 3 iterations of 2D convolutional and 2D average pooling layers, considering the last dimension of left and right sides as channels of the layer. We consider anisotropic kernels as the time and sensor placement channels are different physical concepts. For the time dimension, we use a convolutional kernel size of 5, in combination with the following 2D pooling size of 2. For the sensor placement dimension, we use kernel size of 8 so that all sensor channels are considered in the input of the following layer, and without pooling on this dimension (pooling size 1). The first temporal dimension of 32 is gradually reduced to 8, same as the capacitive channels. Then we use isotropic pooling of (2, 2) and another 2D convolutional layer with a kernel of (4, 4) and 64 filters. Till this point in the network, all the temporal and spatial information from the input are compressed into an interrelated tensor of (4, 4, 64), where time and space are no longer isolated from each other. We thus reshape this tensor to (16, 64), essentially flatten the first two dimensions. Then we use several layers of 1D convolution to narrow down the information channels. Specifically, the dimension of 16 is reduced to 8 (representing 8 joints in the upper body) by 4 convolutions with a kernel size of (3) without padding; and the

filter channels of 64 is gradually reduced by less filters in each consecutive layer. Ultimately the 1D convolutions reach the output of (8, 3) for the pose coordinates.

6 POSE PREDICTION QUALITY WITH PROTO.1

6.1 Signal Examples

Fig. 6 shows some examples of the time sequences of the capacitive signal, video-extracted poses, MoCaPose predicted poses and the errors between prediction and ground truth. The '3D pose ground truth' row of subplots indicates the participant's movement as extracted from the video recording. We can observe that the normalized multichannel 'capacitive sensor data' change in accordance with the wearer's motion. This is further reflected in the 'Predicted 3D pose from the capacitive sensors'. The subplots at the bottom row are created from single time steps in the '3D pose' subplots (the 3D coordinates are projected onto the front-facing 2D plane). Overall, Fig. 6 shows the first evidence that our approach can successfully extract 3D poses from multi-channel capacitive signals from Proto.1 (Fig. 3a). In the remainder of this section, we further investigate the results representative of the entire dataset.

6.2 Motion Capturing Specific Results

We converted our results to multiple motion tracking specific metrics used from state-of-the-art as listed in Table 1 thus to investigate the precision of the joint tracking results. First of all, the standard Mean-Per-Joint-Position-Error ($MPJPE$) was used, defined as:

$$MPJPE = \frac{\sum_{p=1}^{N} \left(\sum_{j=1}^{8} \left\| (x_j, y_j, z_j) - (\dot{x}_j, \dot{y}_j, \dot{z}_j) \right\|_2 \right)}{8N} \tag{3}$$

where j denotes one of the 8 joints in the upper body; p represents a single pose from the dataset N; (x, y, z) are the predicted pose coordinates in the horizontal, vertical and depth directions; and $(\dot{x}, \dot{y}, \dot{z})$ are the corresponding ground truth.

To compare with some other related work, we also converted our results to root-mean-squared-error of the joint position coordinates $RMSE(P)$ and angles $RMSE(A)$. We also converted the $MPJPE$ and $RMSE(P)$ values from the normalized pose scale to the meters according to Eq. (2) which considers the inherent ground truth error from the 3D video pose extraction pipeline.

During the training process of the deep convolutional regressor, we used mean absolute error (MAE) as the loss function and early stopping metric, as it is provided by the Tensorflow framework and is common in regression training tasks. Optimization with MAE while inspecting the results with MPJPE further avoids overfitting in our case.

We first break down the results to separate joints. Fig. 7 shows the per joint results with the regressor learning 3D coordinates. Table 3 lists all of the above-mentioned metrics for every joint.

We can observe that the errors of the corresponding joints from the left and the right sides of all cases are symmetrical. The prediction errors of the joints at the torso, including the spine, thorax and shoulders, are significantly smaller than the wrist. The elbow joints also have higher prediction error in the 3D space. A cause might be that the 3D pose is interpolated from temporal sequences of 2D poses by the VideoPose3D model, and thus the depth information itself might not be accurate. This may be exacerbated by the fact that the joints on the arms have more range of motion on the depth direction compared to the torso in our experiment, which can also be observed from $RMSE(P)$ and $RMSE(A)$. However, if we consider the error proportionally with the range of motion of the specific joint, we can see in Fig. 7b that the joints on the arm are comparable with the other joints in terms of the error-range ratio.

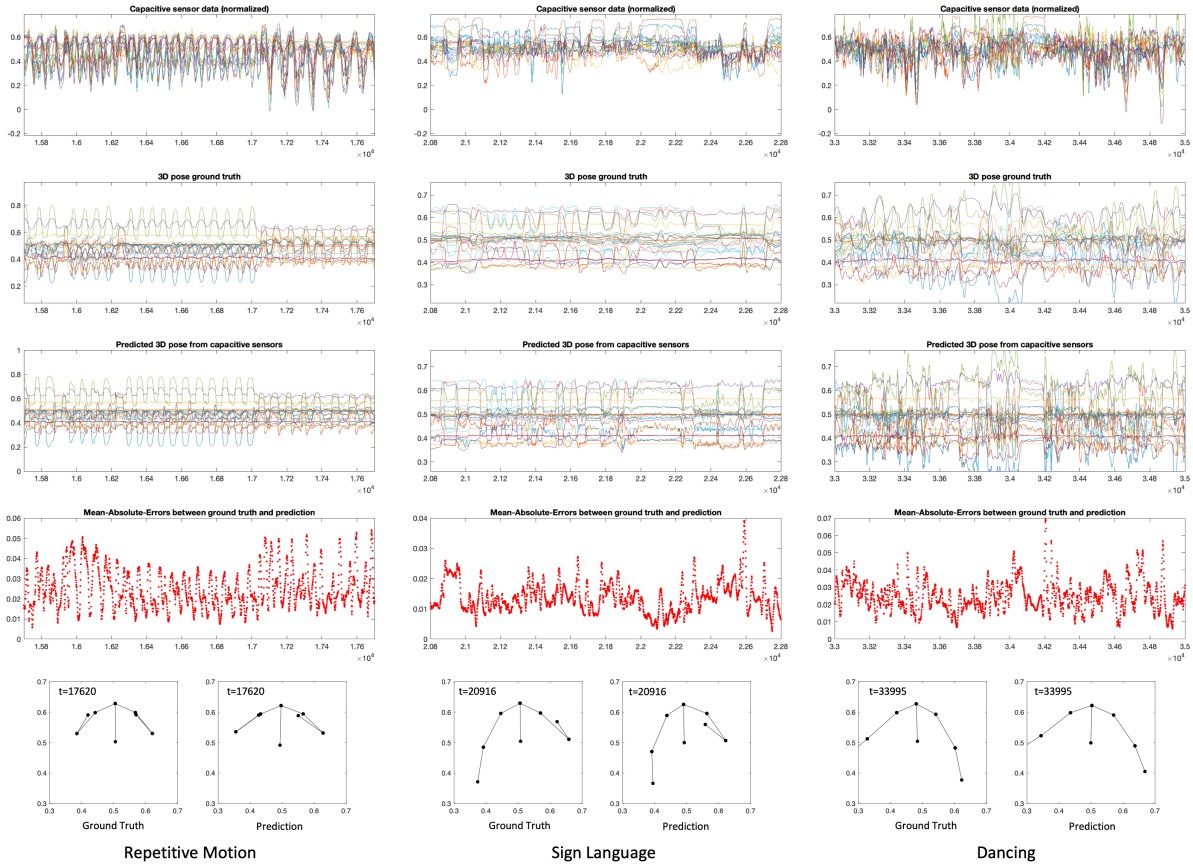
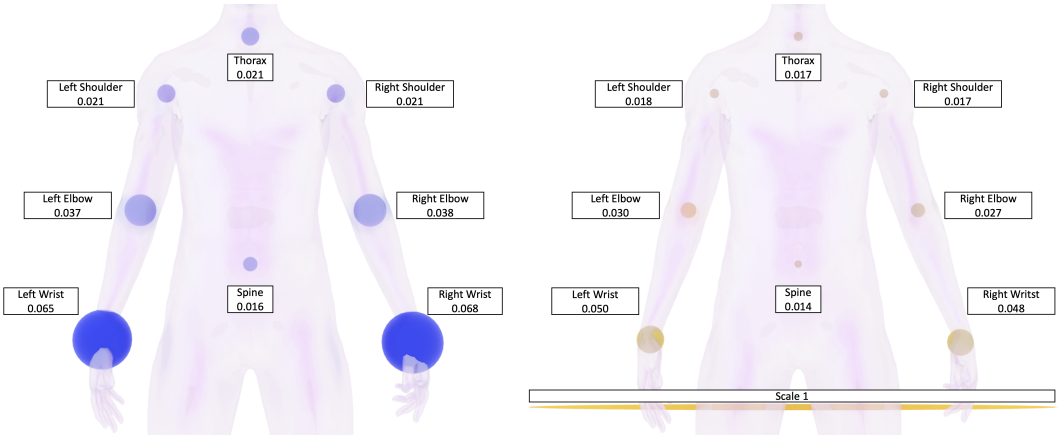


Fig. 6. The data examples from participant 1 including the input to the deep regressor - capacitive signals; ground truth and prediction of the regressor output - 3D poses; and the mean-absolute-error (MAE) between the prediction and ground truth. The Y axes are relative scales after normalization.

Fig. 8 shows the MPJPE per participant in both leave-session-out (LSO) and leave-person-out (LPO) cross-validations. First of all, the average MPJPE of all participants is 0.031 for the LSO and 0.033 for the LPO. Consider the scaling factor of meters and the VideoPose3D inherent error in Eq. (2), this is approximately 86mm and 88mm. For most participants, there is a slight performance degradation from LSO to LPO, which is expected as the model is predicting on the data from a stranger. However, this degradation is not significant considering both the magnitude of MPJPE values and the actual range of the joints in the pose. There are also no observable correlation between the fit of the jacket and the prediction performance. Thus, the per person results demonstrate our MoCaPose method remains robust when new users are being tested.

6.3 Statistical Correlation

Fig. 9 presents the results of the entire dataset with Proto.1 in the form of scatter plots between the predicted joints and the corresponding ground truth with the LPO condition. We can observe the clusters are mostly on the line of the diagonal, which represents a match between the prediction and the ground truth. In the horizontal



(a) The MPJPE represented by the diameter of the blue spheres, scaled with the pose system.

(b) The MPJPE-range ratio represented by the diameter of the golden spheres, scaled of 1 is indicated at the bottom.

Fig. 7. MoCaPose pose estimation results per joint in the 3D space.

Table 3. MoCaPose Results from Proto.1 According to Various Motion Capturing Specific Metrics

Metric	Spine	Thorax	L-shoulder	L-elbow	L-wrist	R-shoulder	R-elbow	R-wrist	All
MPJPE-LSO	0.016	0.021	0.021	0.037	0.065	0.021	0.038	0.068	0.031
MPJPE-LPO	0.017	0.022	0.022	0.039	0.069	0.022	0.040	0.071	0.033
MPJPE-LSO (m) *	0.067	0.072	0.073	0.093	0.130	0.073	0.094	0.133	0.086
MPJPE-LPO (m) *	0.068	0.074	0.075	0.096	0.135	0.075	0.097	0.136	0.088
RMSE(P)-LSO	0.017	0.018	0.018	0.024	0.048	0.019	0.026	0.050	0.033
RMSE(P)-LPO	0.018	0.019	0.019	0.026	0.050	0.020	0.028	0.052	0.036
RMSE(P)-LSO (m) *	0.067	0.070	0.069	0.077	0.108	0.070	0.079	0.110	0.089
RMSE(P)-LPO (m) *	0.068	0.071	0.071	0.079	0.110	0.072	0.081	0.112	0.091
RMSE(A)-LSO (°)	-	+4.446	9.362	19.074	-	9.470	20.068	-	12.801
RMSE(A)-LPO (°)	-	+4.828	9.713	19.938	-	9.835	21.327	-	13.479

+ Averaged between the Spine ~ Thorax ~ L-shoulder and Spine ~ Thorax ~ R-shoulder

* Approximated to absolute meters including the ground truth error according to Eq. (2)

subplot of Fig. 9 specifically, we can also observe that the prediction performance is evenly distributed between the left and right side of the jacket prototype.

Several clusters also exhibit the behavior of concentrations on a vertical direction centered at the cluster, for example, in the 3D subplot of Fig. 9 of Spine, Thorax, Left and Right Shoulders. This means for various actual positions of the joint, the regressor predicted them over-conservatively at the usual positions. These usual positions of the above-mentioned joints compose the normal relaxed up-right torso frame, which occupy a larger majority of the dataset. A lack of sufficient variations of these joints to train the regressor model might have caused the error in the predictions. Such over-conservative predictions are not obvious for the other joints, which might be contributed by that these joints do not have a typical position in our dataset. In other words, the

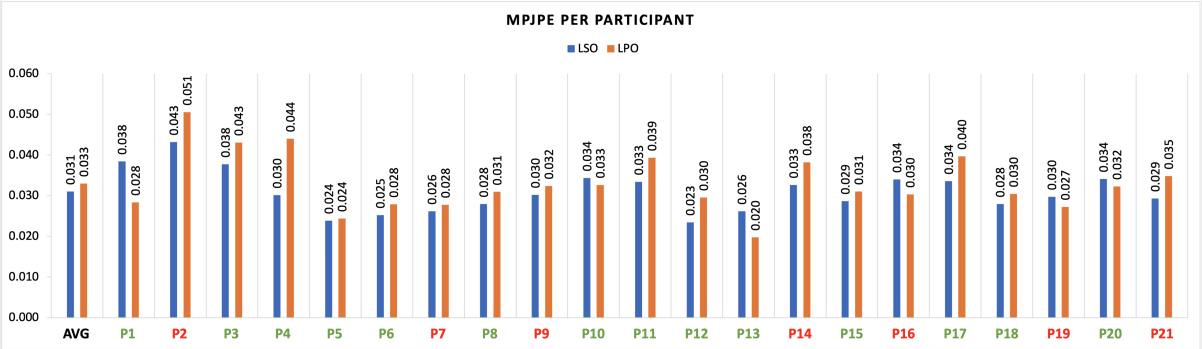


Fig. 8. The MPJPE values of individual participants in LSO and LPO cross-validations, scaled to the normalized pose. Lower is better. Green on the participants' ID indicates a comfortable fit and red indicates the jacket was too loose.

participants were standing in an up-right position where their torso and shoulders were mostly limited in small ranges of motion compared to the arms.

To quantify the performance of pose estimation, we use the standard statistical method, coefficient of determination or R^2 value, which is common in the relevant literature [11, 19, 73]. According to the literature, the R^2 values are usually defined as three levels: weak ($R^2 \in [0, 0.5)$), moderate ($R^2 \in [0.5, 0.7)$) and strong ($R^2 \in [0.7, 1]$) correlation. We fit a linear regression model for each of the scatter plots in Fig. 9, the resulting R^2 values are: 0.8230, 0.9145, 0.8344, 0.5133. This shows our approach predicts the poses from the 3D space, x and y direction with very strong correlation from the ground truth, especially the horizontal direction with the R^2 over 0.9. On the depth direction, however, the correlation is relatively moderate. This might be the result of several factors. The depth information is beyond the video frame and is interpolated by the VideoPose3D model from the temporal sequence of poses, which might introduce more error. The motion in the depth direction of a person is limited compared with the other two dimensions, which can be observed in the depth distribution as the major center cluster is the joints from the torso, and the smaller cluster closer to the origin is joints from the arm, as in most cases only the arms might move closer to the camera during the experiment. But we argue that in many application scenarios of human activities and interactions, the depth information is the least relevant compared to the two frontal dimensions, as one can distinguish activities using 2D poses or through a monocular camera.

7 SCALABILITY VALIDATION WITH PROTOTYPE 2

With the outcome of the deep regressor from the concept validation dataset with Prototype 1 demonstrated in Section 6, we further validate if the hypothesis of Section 1.2 would generalize to a second prototype following appropriate smart textile integration techniques described in Section 3.3. As mentioned in Section 4, 6 out of the 21 participants from the Proto.1 experiment were invited back to record 6 sessions each with Proto.2 following the same experiment prototol. We primarily focus on the MPJPE of the joint coordinates which is a common metric in pose estimation as we listed in Table 1.

To establish baselines for comparison, we considered the following aspects:

(1) The leave-session-out (LSO) and leave-person-out (LPO) results of the dataset from Proto.1 from Section 6.
(2) The LSO result of training and validating the deep regressor with only the same six participants from Proto.1 dataset.
(3) The best regressor model trained from the dataset of Proto.1 directly performing inference prediction with the dataset of Proto.2 without further training (Inference).

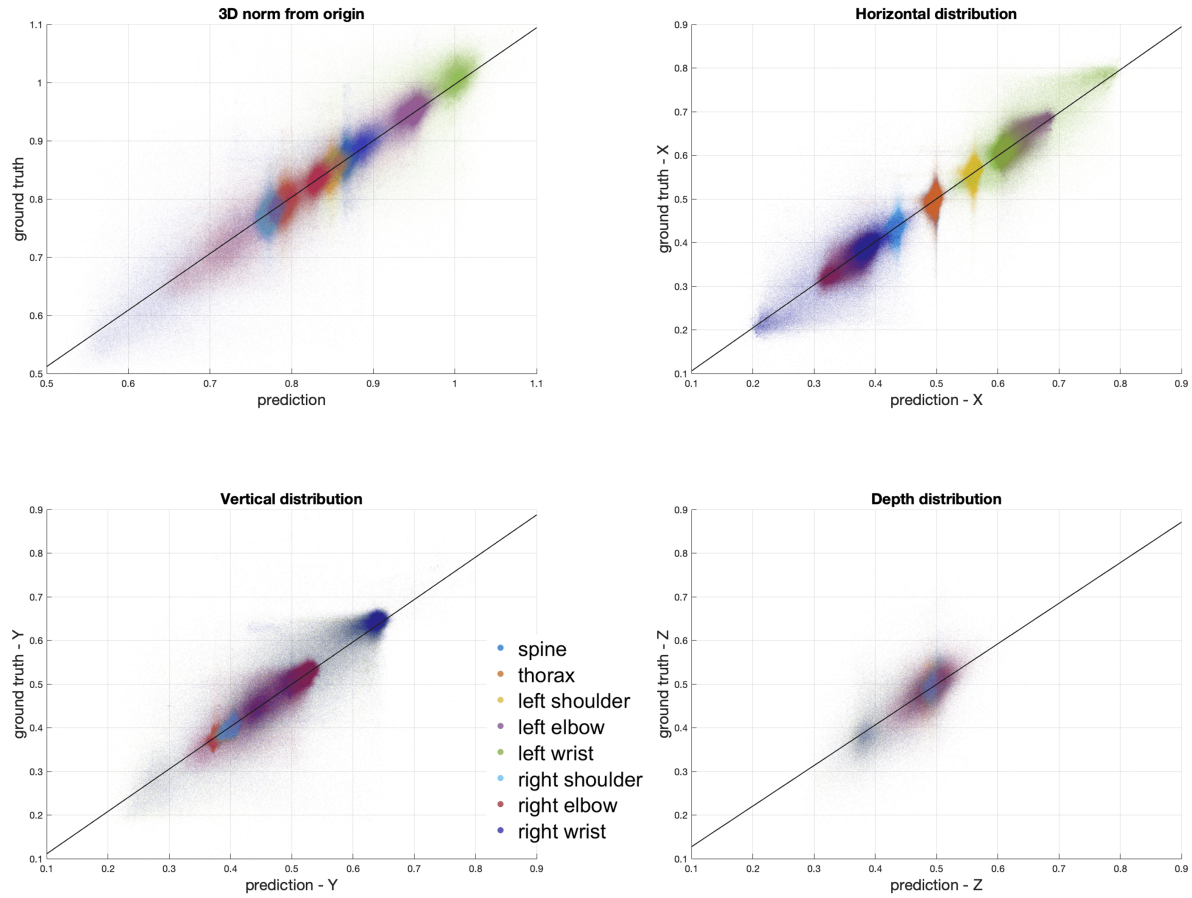


Fig. 9. The scatter plot of prediction and ground truth pose joint coordinates in the cases of vector norm to the origin in the 3D space, and the horizontal, vertical and depth directions separately. The density and transparency were adjusted to differentiate overlapping clusters from different joints.

(4) The same model architecture trained from scratch with the Proto.2 dataset following LSO and LPO validations (Vanilla).

The experiment of 21 participants of Proto.1 provides two forms of knowledge: the model weights of the trained deep convolutional regressor and the training data itself. We thus consider several approaches for passing this knowledge from Proto.1 to Proto.2 through transfer learning which has shown effective in sensor-based activity recognition [52, 85]:

(1) Transfer 1: model weight. The best model weight was loaded first to initialize the weight, and then the model is trained again with the dataset from Proto.2.
(2) Transfer 2: data of the 6 participants in the Proto.1 dataset. The last four recordings from the Proto.1 dataset were selected since two of them recorded 4 and 5 sessions only. And we do not intend to overwhelm the new data with the old data.
(3) Transfer 3: both model weight and the data of 6 participants from the Proto.1 dataset.

(4) Transfer 4: both model weight and the entire dataset of 21 participants from the Proto.1 dataset.

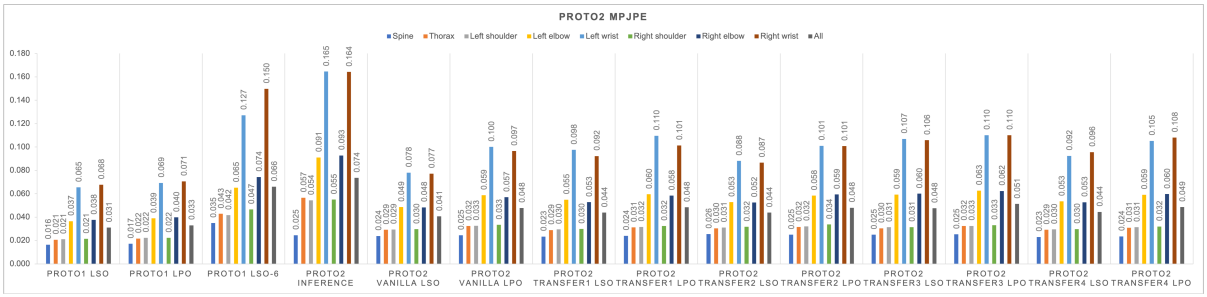


Fig. 10. The MPJPE values of the Proto.2 validation results, scaled to the normalized pose. Lower values are better.

Fig. 10 shows the comparison among all the validation approaches of Proto.2. First of all, the baseline without any knowledge of Proto.2 has the worst performance, with an average MPJPE of 0.074, more than twice that of the Proto.1 results, equivalent to 0.14 m. This is still on a similar level of magnitude with the work such as [46]. As there is no training with the new data involved, the result can still be considered significant. Across all the results, the trend that the joints on the elbow and wrist has higher median RMSE continues. However, as we mentioned in Section 6, this is also related to the fact that the hands had significantly higher ranges of motion compared with the torso in the dataset. Then comparing all the knowledge transfer results, only Transfer 2 has similar results as the Vanilla method in both the LSO and LPO validations, with the remaining offering worse results. At last, the best result from Proto.2 is on par with the Proto.1, which has significantly more data samples (more than 3 times the participants and the total recording duration). While we compare with the LSO-6 result, generated by training the deep regressor with the same 6 participants in the Proto.1 dataset, Proto.2 shows significant improvement. This may be the result of better textile integration techniques which offers much better stability. As the Vanilla method does not require any prior knowledge from Proto.1, this hints that for a new design, only the new data is sufficient to train a MoCaPose deep regressor, making it easier for continued adaptation.

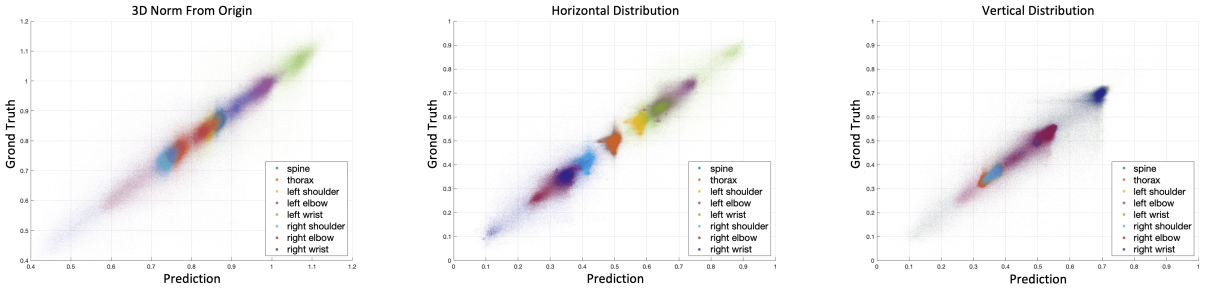
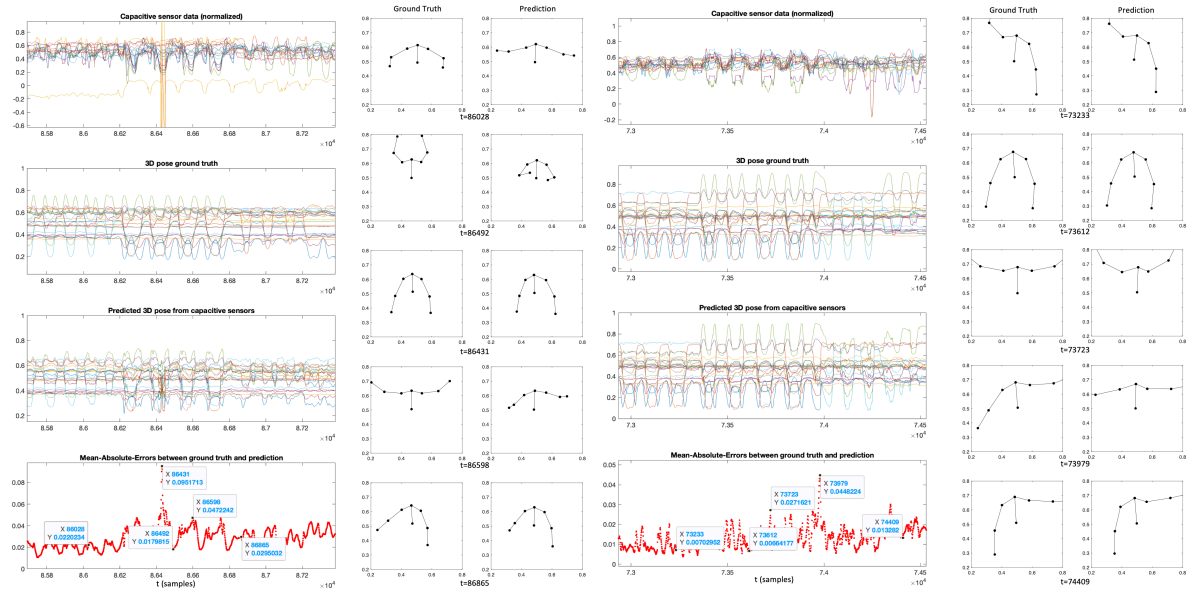


Fig. 11. The scatter plot of prediction and ground truth pose joint coordinates in the cases of vector norm to the origin in the 3D space, horizontal and vertical directions separately.

With the best performing method, Vanilla LSO, the R^2 values of the 3D vector length, and in the horizontal, vertical and depth directions are 0.8916, 0.9323, 0.8950, 0.5676 respectively. We plot the scatter plots of the prediction of ground truth in Fig. 11. The depth direction is skipped due to the low R^2 value. Thus Proto.2 also have the similar statistical metrics with Proto.1 from every aspect.



(a) Proto.1. (b) Proto.2

Fig. 12. Comparisons between Proto.1 and Proto.2 of the capacitive signals and pose estimation of the same participant (P4) following the same video instruction. A period of relatively high error was chosen from each dataset to emphasise on the false estimations. The Y axes are relative scales after normalization.

Fig. 12 compares a segment of recording and prediction results from Proto.1 and Proto.2 with the same participant following the same instruction video. In the case of Proto.1, an obvious burst in one capacitive channel can be observed around t=86431, which causes a faulty pose prediction and peak in the MAE. Such signal artefacts could have been caused by short circuit of the sensing patch with other patches or conductors, which happened regularly in Proto.1 since the conductive patches were exposed. Proto.2 on the other hand, has addressed this issue by isolating the conductive fabric under thermotransfer foils in the layer stack shown in Fig. 4a, and we have not seen any similar signal artefacts. During the power consumption testing mentioned in Section 3.1, moving Proto.1 or touching the exposed sensing pads sometimes trigger the fuse due to transient power surge, which has never happened with Proto.2. The same phenomenon was likely the signal artefact seen with Proto.1, while the LiPo battery is more tolerable with transient power surge and the DAU kept on operating after the transient surge.

Also, one capacitive channel in Proto.1 is obviously drifting across the time window. Drifting happened very rarely in Proto.1 and this recording was the only one that the drift removal from Section 5.2 was not effective, which can be caused by combination of signal artefacts focused on part of the recording, thus dragging the linear trend-line. Since FDC2214 during our test is robust against drifting, we suspected the drifting of individual channels in Proto.1 was due to either unstable connections or exposed patches, as it was not observed in Proto.2. With two prototypes, we tested four different types of conductive fabrics, and all of them produced similar signal results. From the entire dataset of Proto.2, the signal and prediction quality were consistent with the segment shown in Fig. 12b, which was apparently better than those of Proto.1. The unchanged or even slightly higher median RMSE on some joints might be a result of the smaller dataset, and thus less training data.

Overall, considering only 6 participants, Proto.2 has shown comprehensive improvement from Proto.1 from garment design, sensing robustness, signal quality, and pose estimation results.

8 USABILITY EVALUATIONS

So far, we have developed a novel system for capacitive-based motion capturing which exhibits metrics on par with the motion tracking SOTA with other wearable sensors. However, the purpose of such a system is beyond the scope of motion capturing alone. Therefore, in this section we deep dive into several usability aspects including context recognition, performance of reduced channels, motion speed relevance and the interference from external factors.

8.1 Pose and Motion Recognition through the Natural Movement Common Ground

To explore the usability of our approach in classification tasks, we developed a workflow to define pseudo-classes of poses or motions in our unlabeled dataset of continuous movements using self-clustering. The workflow included an auto-encoder and unsupervised learning methods, and was implemented for both static poses and dynamic sequences of poses (motions), including five processes:

(1) Process 1 AEC uses a convolutional auto-encoder to deconstruct the input (poses or motions) into a latent vector with the size of 32, and then reconstructs the upper body poses or motions.
(2) Process 2 ULGM uses the unsupervised learning method Gaussian Mixture Model[80] with k-means initialization, to locate 10 classes (cluster centers) in the latent feature space. We used the default settings from scikit-learn for the Gaussian mixture model.
(3) Process 3 REG-ENC-CLGM uses the trained deep regressor we introduced in Section 5.3 to reconstruct the pose from capacitive signals. Then the sensor-reconstructed pose is passed through the trained encoder from Process 1 to extract the feature vector. Eventually, we perform inference with the optimized and locked Gaussian mixture model from Process 2 as a classifier to predict the pose or motion.
(4) Process 4 GT-ENC-CLGM is similar as the encoder and Gaussian mixture steps. However, the input of the process is the video-extracted pose or motion of the same data sample as in Process 3. This process thus sets the classification ground truth (GT).
(5) Process 5 CAP-CL is our baseline comparison that follows the traditional activity recognition from sensor signals. A deep classifier is created following the practices for similar capacitive sensors in [13–15].

Processes 1 and 2 constitute the class definition and training phase of the simulated classification workflow, during which we use only the poses extracted from the videos following the process in Section 5.1. The latent vector between the encoder and the decoder are unique values that can be used for decomposition and reconstruction of the poses or motions, thus we consider them as features suitable for classifying different types of poses or motion. The encoder part of the auto-encoder will then function as a feature extractor specialized for the poses or motions. Similar approach has been evaluated for sensor data in activity recognition [41].

Processes 3 and 4 are the testing phase, during which we lock all the models, including the deep convolutional regressor, encoder and decoder of the auto-encoder, and the Gaussian mixture model and use them for inference (prediction). The Gaussian mixture model functions as a classifier during prediction operations. We set the class ground truth of the samples by the locked encoder and the Gaussian mixture from video-extracted poses or motions in the testing data partition. Since our experiment contains continuous motions, the samples with a low confidence (below median) from the Gaussian mixture model were considered as transitioning motions between the cluster centers and were removed.

In Process 5, a 2D CNN deep classifier was developed representing the state-of-the-art activity recognition with capacitive sensors. Our pipeline from the capacitive signals via poses reconstructed by the MoCaPose regressor,

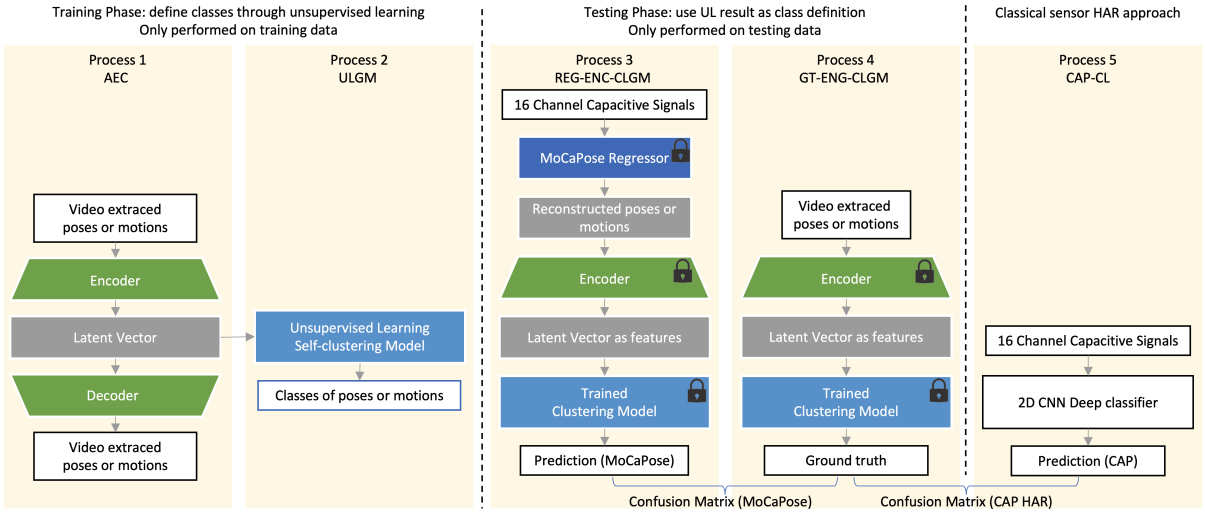


Fig. 13. The validation workflow of using unsupervised learning in our unlabeled data to simulate a classification task.

and features extracted by the encoder, then classification by the trained Gaussian Mixture model provides an alternative solution.

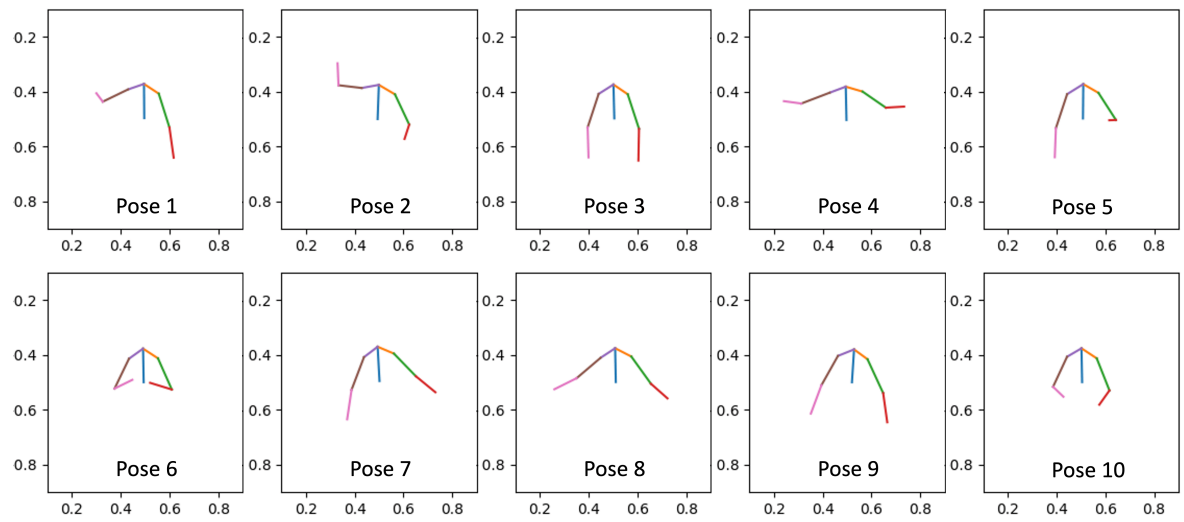
To avoid over-fitting, we observed several safeguarding rules:

- For the training and testing phases of the simulated classification, the data are strictly partitioned with the leave-sessions-out scheme, and there are no common recordings that exist in both phases.
- All the classification ground truth for classes are determined by video-extracted poses and motions, and all the prediction come from capacitive sensor signals.
- All the models from our proposed workflow, including the MoCaPose regressor, encoder and clustering model, are locked in the testing phase and only perform inference operations.

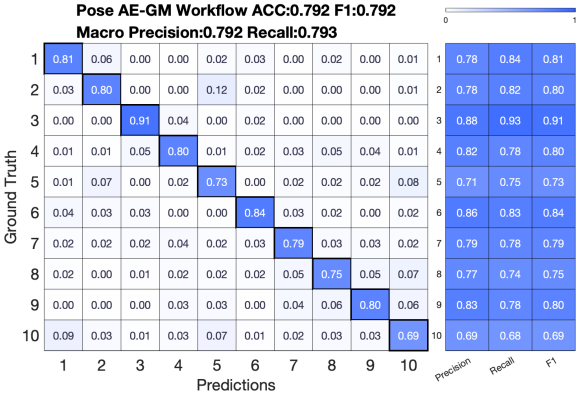
At first glance, the unsupervised class definition seems leaning towards pose and motions rather than raw sensor signals. However, this mimics the usual practice in HAR, where manual annotations rely on the natural movements observed with human experts' supervision. The classes in HAR categorizes the context of the activity rather than the sensor signal patterns. Thus, we believe this workflow is unbiased between the motions and sensor signals and is truthful to the consensus activity recognition practices.

Fig. 14 and Fig. 15 show the results of the simulated classification task. First, the determined classes from Processes 1 and 2 are shown in Fig. 14a and Fig. 15a, which are cluster centers of the Gaussian mixture model and reconstructed by the decoder. In other words, these poses or motions might not be from an actual experiment recording, but represent the most representative poses or motions occurring in the dataset. Although this is the probabilistic outcome, it might not cover all of the classes in the experiment and some classes might appear similar for a human observer, such as Motion 6, 8 and 9 in Fig. 15a.

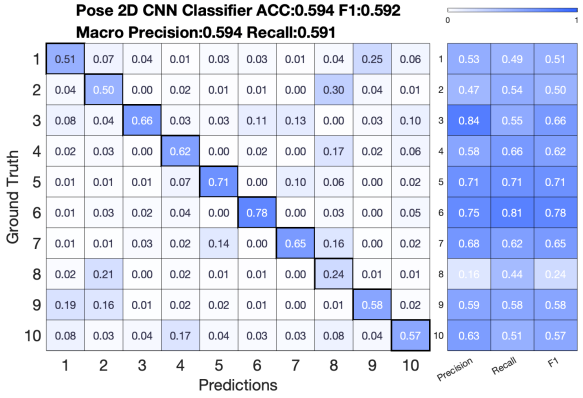
We then show the confusion matrix of Processes 3 and 4 in Fig. 14b and Fig. 15b. For comparison with classical sensor based HAR approach, the confusion matrix of Process 5 is presented in Fig. 14c and Fig. 15c. In both pose and motion classification tasks, going through the REG-ENC-CLGM process provides significant accuracy gains compared with directly classification from the sensor data from the CAP-CL process. This shows promising yet preliminary result that moving from the sensor signal level to the pose level might improve context recognition applications. The limitations of this approach will be further discussed in Section 9.



(a) The identified pose classes through unsupervised learning from the frame-based auto-encoder's latent feature space.



(b) the confusion matrix of classifying these pose classes in the test data with MoCaPose reconstructed poses.

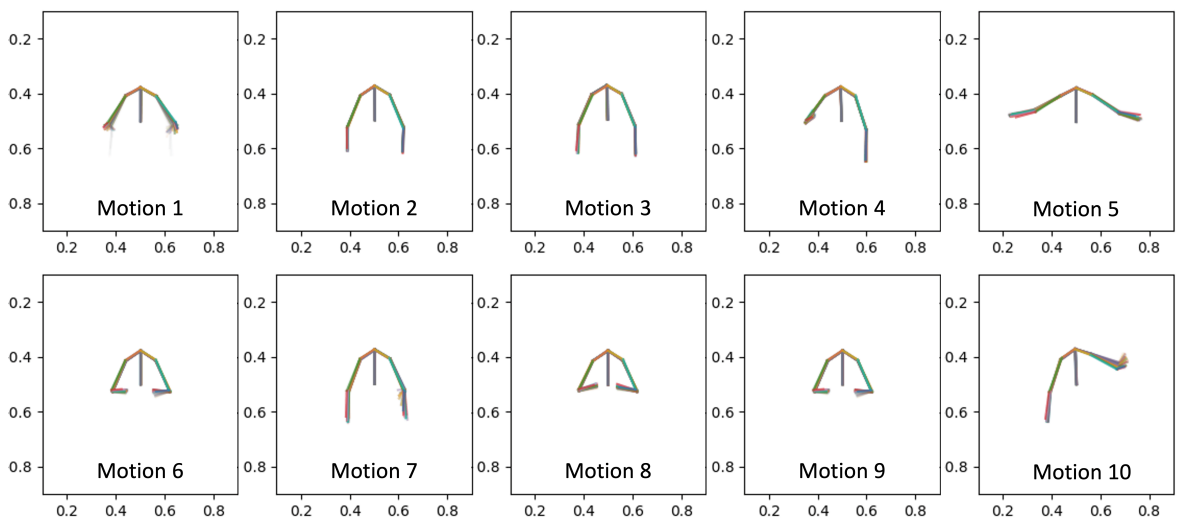


(c) the confusion matrix of classifying these pose classes in the test data following standard activity recognition process.

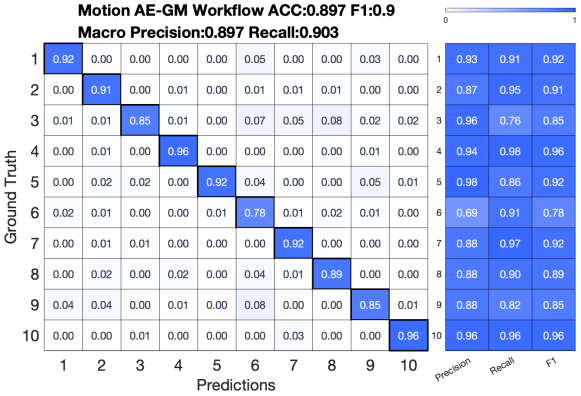
Fig. 14. The results of the simulated classification assisted by unsupervised learning for recognizing still poses and comparison with classical sensor-based activity recognition.

8.2 Reducing Sensor Channels

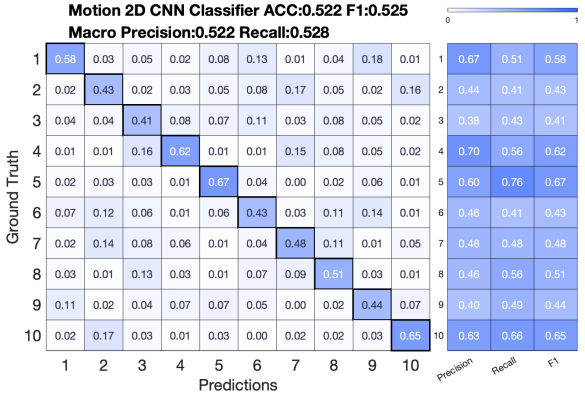
In the fashion-technology co-design process of textile integration, it is important for the technology to impose as little limitations on the garment design as possible. In our case, we investigated if less capacitive channels can achieve similar efficacy in predicting the wearer's pose. To determine which input channel has more significance in the deep regressor, we added a single locally connected layer (LocallyConnected2D in Keras) with 8 neurons and a kernel of (1,1), shared between the left and right side of the sensor patches. This locally connected layer essentially adds an individual weight to each of the channels independently. We follow the same leave-session-out



(a) The identified pose classes through unsupervised learning from the temporal-based auto-encoder's latent feature space. The time steps of the motion is represented by traces of the pose.



(b) the confusion matrix of classifying these pose classes in the test data with MoCaPose reconstructed motions.



(c) the confusion matrix of classifying these pose classes in the test data following standard activity recognition process.

Fig. 15. The results of the simulated classification assisted by unsupervised learning for recognizing short time window motions and comparison with classical sensor-based activity recognition.

process of training the deep regressor, and then take the weight values of the locally connected layer as the factor that indicates the importance of each channel. From such method, the ranking from the most to the least important channels are determined as: {CH0, CH1, CH2, CH7, CH5, CH6, CH4, CH3}, the numbers match those in Fig. 3a.

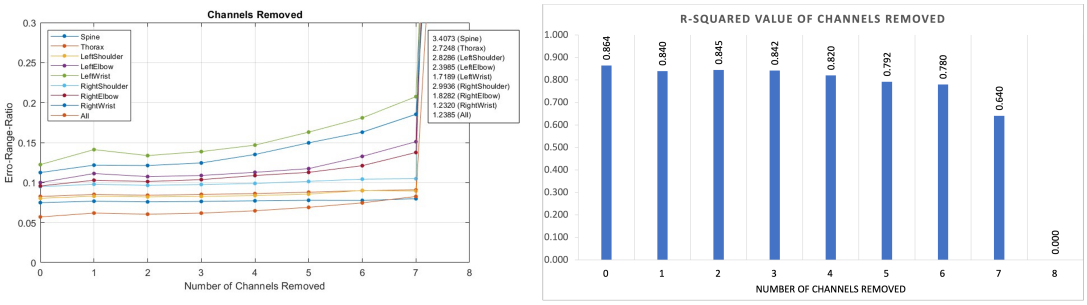
We then remove one channel at a time, from no channels removed to all channels removed, and replace the sensor values of the removed channels with zero. With each iteration, a new regressor is trained from a blank model and validated following the leave-sessions-out scheme. Fig. 16 shows an example of a time period of the

Fig. 16. Example of predicted poses while gradually removing capacitive channels.

predicted poses, from which we can see a gradual degradation of the details of the predicted poses as less channels are used. In the case of using zero channels, the model outputs constant zeros. To represent all the test data, we plot the error-range ratio in Fig. 17b and R-squared values between the prediction and ground truth in Fig. 17b. These results indicate that:

- Removing a single channel already degrades the prediction performance of pose prediction.
- The prediction performance degrades further as less channels are used.
- However, while keeping only half of the channels (4 channels each side), the R-squared values are still above 80%, indicating strong correlations between the prediction and ground truth.
- Even with a single channel (CH0 at the front shoulder), the regressor can give some rudimentary pose predictions, which might be useful for certain applications.

However, the reduction of channels should be considered together with the specific usecase and garment design, which is beyond the scope of this study. Also, the limitation of our data-level channel reduction approach is that: we cannot remove the physical sensor patch, and thus the influence of other active sensor patches cannot be eliminated. However, the FDC2214 module has strong cross-channel isolation both from the sensing principle and our test. Therefore, our data-level channel reduction discussion still offers useful guidance.



(a) The Error-to-Range ratio of reducing capacitive channels from the data level.

(b) The R-squared value of predicted poses as capacitive channels are reduced from the data level.

Fig. 17. The metrics comparison of gradually removing pairs of channels.

8.3 Influence of Different Motion Speeds

We further investigate the influence of the motion speed on the pose prediction results. We take the LPO results (plotted in Fig. 9) and calculate the sum of the range of every joint's coordinates of the ground truth within the 1 second window centered at every sample. This is thus defined as the transient range of motion. We then rank the transient speed of motion and calculate the R^2 value between the pose ground truth and prediction of the 3D vector length for every 5 percentile from slow to fast speed of motion. The result is shown in Fig. 18. We can observe that the prediction is more reliable in the lower speed of motion, with an R^2 value of over 0.9 at the lowest 5 percentile. A majority of 60% of the data from the lower speed of motion has over 0.85 R^2 which is generally considered very strong correlation.

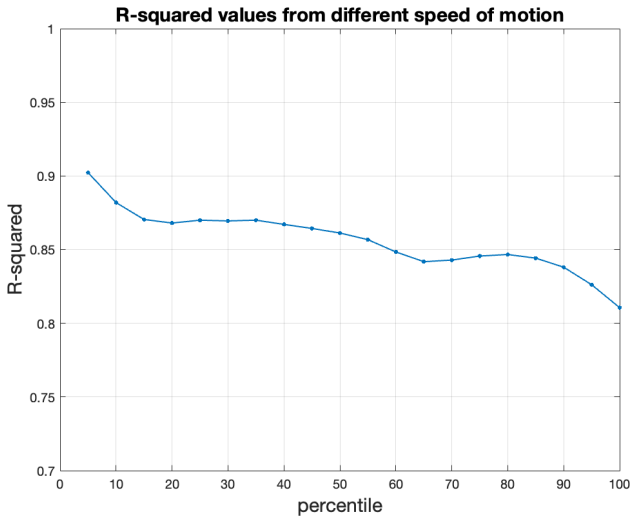


Fig. 18. R-squared values between MoCaPose predicted poses and ground truth in different motion speeds sorted by every 5 percentile of the data in leave-sessions-out. The lower percentiles correspond to slower speed of motion

8.4 Interference

Through discussions among the electronics engineers and smart textile engineers, we tested the possible interference vulnerabilities of our prototypes that are of practical concerns for such smart garments.

In Proto.1, bending the wiring from the sensor patches generated small signal changes compared with bending the patches. This is because the equivalent capacitor consisted mainly of the larger conductive patch surface. When an external person or conductor, tested with a palm from another person and an ATX PC power supply module, approached the sensing patch within 1cm distance, there was a signal change correlated with the proximity; we did not observe more signal distortion for further than 1cm distances. Practically, this is smaller than the socially acceptable personal distances within normal conditions. However, both the distortion by bending the wire and by external conductor proximity were on less than 10% magnitude compared with the wearer moving the body part with the sensing patch.

In Proto.2, the proximity of external conductors no longer caused any signal distortion. Distortion only occurred when the external rigid object touched the patch surface regardless of the conductivity property of the external object, and had much smaller magnitude compared to Proto.1. We also turned on both prototypes with the DAUs in full operation mode, and we did not observe any interference between the two prototypes while they were worn by two persons standing next to each other with 1cm distance. It is worth pointing out that all 32 channels on all 4 DAUs were configured with the same capacitor and inductor values, setting the excitation frequency to the similar range within the passive components' factory tolerance.

While physical contact on the sensing patches generate large peaks, if the contact is from the user itself, and the corresponding pose ground truth is provided to the regressor, the model could also learn those self-contact poses. For example, the synchronization points involve the participants touching the patches on the belly, and the regressor model could successfully predict those movements with the sensor signals that include these peaks.

9 LIMITATIONS AND FURTHER DISCUSSION

9.1 Results Comparison with Related Works

While the SOTA we listed in Table 1 use different metrics, we have converted our own results to all of those metrics to provide a comprehensive comparison as detailed in Section 6 and especially in Table 3.

First of all, R^2 of up to 0.915 with Proto.1 and 0.932 with Proto.2 shows strong correlation between the predicted and ground truth coordinates according to the broader literature consensus of R^2 standards. While the work from [10] showed R^2 close to 1, it was limited to only a single joint and 1 participant while the complexity and dataset of our work is orders of magnitudes higher. MPJPE and RMSE(P) provide metrics of the coordinate position precision in distance units. While our results of 86mm is larger than some results reported in the SOTA such as 43.3mm from information fusion of 4 IMUs and 1 camera in [56], 31.8mm with 12 on body transmitters in [48]; our results have truthfully included the ground truth pipeline error of 46mm in Eq. (2) according to appropriate measurement instrumentation practices [42], which cannot be said for the SOTA results. This inherent ground truth error accounts for than 50% of our distance results, which we will further discuss in Section 9.2. The EM-Pose [48] is sensitive to external persons and electronics within 1.5m range, while our Proto.1 is only sensitive to those within 1cm and Proto.2 is not influenced by external non-contact conductive bodies. Many of the results reported in the SOTA use angles as the metric. Our RMSE(A) of 12.8° is well in line with those reported in [34, 43, 47, 48, 95]. Teufl, et al. [87] has reported very small RMSE(A) of 2.28°; however, the IMU sensors were tightly fixed to the skin in the lower body, and we also argue that the lower body has less ranges of motion during the activities described in [87] of motion compared to the upper body in our experiments.

For the classification tasks, our result of 90% F1 for 10 pseudo-classes of motions is competitive with the SOTA. While in works such as [14] and [96], above 97% accuracy values were reported, there is a significant distinction that in those annotated activity recognition tasks, the data of different classes are well separated in the timeline of

the recording with start and end timestamps of the activity, and the transitioning motions were discarded, which is a common practice in supervised learning and HAR. However, in our process described in Section 8.1, we discover classes from continuous video recordings where there were no disconnected motions within a recording, thus all the transitioning motions were included. The only mechanism to reduce the transitioning motions were removing those data points there were too far away from the cluster centers in the pose-specific feature space, which is also an unsupervised process and might not be as effective as instructed and annotated data from the SOTA. We thus refer to the baseline of Process 5 CAP-CL, where a deep classifier was used following the SOTA model architecture reported in [14] for multi-channel capacitive signals. And the pipeline of first converting the capacitive signals to pose through MoCaPose and then perform classification in the pose-specific feature space shows 20% and 37% improvement for still poses and dynamic motions. This could be a reflection of the precise pose estimation results discussed above: once the obscure capacitive signals are converted to pose, recognizing the pose and motion are then trivial tasks.

Furthermore, MoCaPose only relies on the capacitive sensing modality integrated into a smart jacket that deeply fuses technology and design, which cannot be said for any of the systems from the SOTA.

9.2 Ground Truth Limitations

From our results, the error was correlated with the speed of the activity, where slower speed generates better results. It could be partially caused by imperfect synchronization between the video source and the DAUs, and small errors in the sampling rate might accumulate to substantiated drift. While the DAUs data has a unix timestamp for every sample, the video lacks such and was taken as a stable 30 frames-per-second (fps) data stream and matched with the sensor data at the beginning of every recording. Although according to the video pose extraction pipeline and the media metadata, the video footage was of 30 fps, a professional video editor BlackMagic® DaVinci Resolve 18 reads some videos with not exactly 30 fps (e.g. 29.97 fps). Another contributing factor might be delays introduced by the temporal convolutions in the deep convolutional regressor, which takes a short time window of capacitive signals to produce a single pose frame. Such process can be sensitive to imperfectly aligned input and output training samples. This can be further supported by the observation of the MAE values from Fig. 6, Fig. 12, which exhibit patterns matching the motion. A shift by one time sample between the ground truth and prediction, for example, would result in similar behavior. Although signal aligning methods such as dynamic time warping may further improve our median RMSE results, we argue this is not truthful nor practically useful for such a pose prediction system, as in runtime there will not be any ground truth template to match. Moreover, state-of-the-art signal processing algorithms can easily handle such sample level temporal shifts, from temporal neural network layers to temporal self-similarity [32].

The pose ground truth pipeline in our study utilizes the state-of-the-art markerless pose estimation models from a low-cost smartphone camera, which simplifies the reproducibility of our work. However, the pipeline introduces an inherent error of 0.046 meters. While marker-based motion capturing systems are not ideal for garment design since rigid reflective markers need to be attached to the garment surface; to show the true potential of pose estimation based on capacitive sensing, it would be necessary to reduce the ground truth error and thus have the same golden standard as other pose estimation studies. Also the depth information of the pose estimation might be improved by multi-angle cameras or with calibration background patterns from open source projects such as EasyMocap[1] and freemocap[66]. Nevertheless, the motion tracking precision cannot be directly translated to the usability in specific applications. Hence, we have explored the HAR classification potential with our limited study in Section 8.1.

9.3 Dataset Limitations

Another limitation in our dataset is that the participants were mostly in an upright standing posture, and thus the deep regerssor was trained with an inherent bias towards such postures. This might explain the high torso prediction accuracy. Similar dataset bias was observed in the computer vision society, for example, the MS-COCO [57] dataset consisted mostly of upright positions as well, and vision-based pose estimation models trained with it usually struggle with atypical postures [44].

It would also be interesting to develop a pipeline to investigate if larger errors are associated with certain poses, which might help informing further patch geometry design improvements and validate the usefulness for certain applications. For example, if an application considers a defined set of poses or motions, the garment does not need to have high prediction performance on other poses or motions. This may be possible by leveraging the autoencoder and the latent pose feature space from our usability study on simulated classification in Section 8.1.

Fig. 19. Proto.1 and Proto.2 side by side detailed comparisons.

9.4 Full-textile Sensing and Scalability

The placement of the textile sensor patches in our work were designed empirically by considering how the motions would affect body poses. It would be interesting to see how other placement designs would compare with our results. While the two prototypes followed the same patch placement, they were not precisely replicated. The sensor patches in Proto.2 did not follow the exact measurement of Proto.1; and the position of the electronics and the tracing are also completely different. Yet two prototypes produced comparable results as we demonstrated in Section 7.

While the textile layer stack in Proto.2 provides a strong and reliable integration solution, it was more prone to handmade production errors due to the heat press process. As detailed in Table 5 and Fig. 19, the patches in the left and right side of Proto.2 were not exactly symmetric with up to 3cm of displacement error. However, the best result shown in Fig. 10 (Proto.2 Vanilla LSO) exhibits symmetry between the left and right joints. Thus, the deep regressor has already compensated the production error. These hint towards the scalability of our approach to other designs that welcome creativity. On the other hand, it would also be interesting to see if the same design can be precisely reproduced, i.e. whether the deep convolution regressor would produce reliable inference estimations without adapting to every new replica, which would open the gate towards market scale production.

9.5 The Common Ground of Pose in Wearable Sensing and HAR

Our approach is not specific to certain applications but rather focuses on the fundamental motion tracking. Our contributions also emphasize on the decoupling between sensor placement and body joints, which is supported by the common ground of the pose-level motions instead of raw sensor signals. By providing such a common ground, our method can not only be adapted to different applications, but also creates a space for knowledge transfer between designs and applications.

The state-of-the-art has explored more advanced unsupervised learning approaches in activity discovery such as [38, 81]. However, since those works were mostly validated on data and features specific to wearable sensors, mostly IMUs, the effectiveness for features specific to pose and motions were not clear. We thus followed the rudimentary but well-established Gaussian mixture with K-means self-clustering approach. In the future it would be interesting to see how the other algorithms can be adapted from sensor-specific data to the common pose and motion representation. By discovering classes in the pose-specific feature space, we can also visualize the context for human experts. It also provides a link with computer vision methods, such as image based self-clustering [92] or image and video captioning [26, 55, 55, 99, 101]. For example, the class can be defined from the video recording to train activity recognition models with MoCaPose enabled smart garments.

10 CONCLUSION

In conclusion, we have proven our hypothesis, and thus demonstrated that multi-channel capacitive sensors made of conductive fabrics can be used for continuous upper body pose estimation. Two iterations of prototypes designed with capacitive sensing technology beyond the SOTA in the wearable field and textile integration techniques for robust and reliable sensing performance. We have collected a dataset of 21 participants and 38 hours, where they followed video instructions of various motions. A deep convolutional regressor was designed to predict the 3D joints coordinates from independent short time windows of capacitive signals. The pose estimation results were analysed according to statistical science, motion tracking metrics and usability in HAR and smart wearable design. The R-square value between prediction and ground truth was up to 0.823 for the 3D space, 0.915 and 0.834 for the horizontal and vertical directions in leave-person-out (LPO) validation, demonstrating strong statistical correlation. The error-to-range ratio was below 5% for every joint, and approximated MPJPE of below 90mm for both leave-session-out (LSO) and LPO validations, indicating high tracking precision comparing with the SOTA with other modalities. To validate the potential contribution to activity recognition, we also used unsupervised learning and auto-encoder to discover representative pseudo-classes of pose or motion from our dataset. In the case of 10-class classification, through the pipeline of capacitive sensor via reconstructed poses and pose-specific features yielded 0.792 F1 score for still poses, and 0.9 F1 score for motions in short time windows, outperforming the traditional pipeline of classification from sensor signals following state-of-the-art classification models of the same modality, which gave 0.592 and 0.525 F1 score. Apart from sensing technology and deep learning pose reconstruction, MoCaPose also emphasizes the scalability towards design-centric smart garments where the technology can adapt to new designs. Instead of dictating the garment design process by imposing strict technology requirements, MoCaPose promotes and inspires aesthetic and styling creativity by using fully textile sensing materials and seamlessly adapting to new design implementations.

By converting from abstract sensor data to the natural movements, we can create a common ground across physical sensing modalities and also a bridge between sensor based wearable activity recognition and computer vision methods such as pose extraction, animation, and video captioning [22].

REFERENCES

[1] 2022. EasyMoCap - Make human motion capture easier. Github. https://github.com/zju3dv/EasyMocap
[2] 2022. Shieldex Bremen PW. https://www.shieldex.de/products/shieldex-bremen-pw/

[3] Fatemeh Abyarjoo, Armando Barreto, Jonathan Cofino, and Francisco R Ortega. 2015. Implementing a sensor fusion algorithm for 3D orientation detection with inertial/magnetic sensors. In *Innovations and advances in computing, informatics, systems sciences, networking and engineering*. Springer, 305–310.

[4] Talha Agcayazi, Murat A Yokus, Max Gordon, Tushar Ghosh, and Alper Bozkurt. 2017. A stitched textile-based capacitive respiration sensor. In *2017 IEEE SENSORS*. IEEE, 1–3.

[5] Karan Ahuja, Paul Streli, and Christian Holz. 2021. TouchPose: Hand Pose Prediction, Depth Estimation, and Touch Classification from Capacitive Images. In *The 34th Annual ACM Symposium on User Interface Software and Technology*. 997–1009.

[6] Mazen Al Borno, Johanna O'Day, Vanessa Ibarra, James Dunne, Ajay Seth, Ayman Habib, Carmichael Ong, Jennifer Hicks, Scott Uhlrich, and Scott Delp. 2022. OpenSense: An open-source toolbox for inertial-measurement-unit-based measurement of lower extremity kinematics over long durations. *Journal of neuroengineering and rehabilitation* 19, 1 (2022), 1–11.

[7] I Al-Nasri, Aaron David Price, Ana Luisa Trejos, and David M Walton. 2019. A commercially available capacitive stretch-sensitive sensor for measurement of rotational neck movement in healthy people: Proof of concept. In *2019 IEEE 16th International Conference on Rehabilitation Robotics (ICORR)*. IEEE, 163–168.

[8] Anindya Das Antar, Masud Ahmed, and Md Atiqur Rahman Ahad. 2019. Challenges in sensor-based human activity recognition and a comparative analysis of benchmark datasets: a review. In *2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*. IEEE, 134–139.

[9] Abdul Hakeem Anwer, Nishat Khan, Mohd Zahid Ansari, Sang-Soo Baek, Hoon Yi, Soeun Kim, Seung Man Noh, and Changyoon Jeong. 2022. Recent Advances in Touch Sensors for Flexible Wearable Devices. *Sensors* 22, 12 (2022), 4460.

[10] Ozgur Atalay. 2018. Textile-based, interdigital, capacitive, soft-strain sensor for wearable applications. *Materials* 11, 5 (2018), 768.

[11] Chaiyawan Auepanwiriyakul, Sigourney Waibel, Joanna Songa, Paul Bentley, and A Aldo Faisal. 2020. Accuracy and acceptability of wearable motion tracking for inpatient monitoring using smartwatches. *Sensors* 20, 24 (2020), 7313.

[12] Christopher A Bailey, Thomas K Uchida, Julie Nantel, and Ryan B Graham. 2021. Validity and sensitivity of an inertial measurement unit-driven biomechanical model of motor variability for gait. *Sensors* 21, 22 (2021), 7690.

[13] Hymalai Bello, Bo Zhou, Sungho Suh, and Paul Lukowicz. 2021. Mocapaci: Posture and gesture detection in loose garments using textile cables as capacitive antennas. In *2021 International Symposium on Wearable Computers*. 78–83.

[14] Hymalai Bello, Bo Zhou, Sungho Suh, Luis Alfredo Sanchez Marin, and Paul Lukowicz. 2022. Move With the Theremin: Body Posture and Gesture Recognition Using the Theremin in Loose-Garment With Embedded Textile Cables as Antennas. *Frontiers in Computer Science* (2022), 72.

[15] Sizhen Bian and Paul Lukowicz. 2021. Capacitive sensing based on-board hand gesture recognition with TinyML. In *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers*. 4–5.

[16] Sizhen Bian and Paul Lukowicz. 2021. A systematic study of the influence of various user specific and environmental factors on wearable human body capacitance sensing. In *EAI International Conference on Body Area Networks*. Springer, 247–274.

[17] Sizhen Bian, Vitor F Rey, Peter Hevesi, and Paul Lukowicz. 2019. Passive capacitive based approach for full body gym workout recognition and counting. In *2019 IEEE International Conference on Pervasive Computing and Communications (PerCom*. IEEE, 1–10.

[18] Sizhen Bian, Vitor F Rey, Junaid Younas, and Paul Lukowicz. 2019. Wrist-worn capacitive sensor for activity and physical collaboration recognition. In *2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE, 261–266.

[19] Kyle J Boddy, Joseph A Marsh, Alex Caravan, Kyle E Lindley, John O Scheffey, and Michael E O'Connell. 2019. Exploring wearable sensors as an alternative to marker-based motion capture in the pitching delivery. *PeerJ* 7 (2019), e6365.

[20] Sameh Neili Boualia and Najoua Essoukri Ben Amara. 2019. Pose-based human activity recognition: a review. In *2019 15th international wireless communications & mobile computing conference (IWCMC)*. IEEE, 1468–1475.

[21] Arianna Carnevale, Umile Giuseppe Longo, Emiliano Schena, Carlo Massaroni, Daniela Lo Presti, Alessandra Berton, Vincenzo Candela, and Vincenzo Denaro. 2019. Wearable systems for shoulder kinematics assessment: A systematic review. *BMC musculoskeletal disorders* 20, 1 (2019), 1–24.

[22] Shaoxiang Chen, Ting Yao, and Yu-Gang Jiang. 2019. Deep Learning for Video Captioning: A Review.. In *IJCAI*, Vol. 1. 2.

[23] Jingyuan Cheng, Bo Zhou, Kai Kunze, Carl Christian Rheinländer, Sebastian Wille, Norbert Wehn, Jens Weppner, and Paul Lukowicz. 2013. Activity recognition and nutrition monitoring in every day situations with a textile capacitive neckband. In *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication*. 155–158.

[24] Kunigunde Cherenack and Liesbeth Van Pieterson. 2012. Smart textiles: Challenges and opportunities. *Journal of Applied Physics* 112, 9 (2012), 091301.

[25] P Chinmilli, Sangram Redkar, Wenlong Zhang, and Tom Sugar. 2017. A review on wearable inertial tracking based human gait analysis and control strategies of lower-limb exoskeletons. *Int. Robot. Autom. J* 3, 7 (2017), 00080.

[26] Jaemin Cho, Seunghyun Yoon, Ajinkya Kale, Franck Dernoncourt, Trung Bui, and Mohit Bansal. 2022. Fine-grained image captioning with clip reward. *arXiv preprint arXiv:2205.13115* (2022).

[27] Minho Choi and Sang Woo Kim. 2017. Driver's movement monitoring system using capacitive ECG sensors. In *2017 IEEE 6th Global Conference on Consumer Electronics (GCCE)*. IEEE, 1–2.

[28] Gabe Cohn, Sidhant Gupta, Tien-Jui Lee, Dan Morris, Joshua R Smith, Matthew S Reynolds, Desney S Tan, and Shwetak N Patel. 2012. An ultra-low-power human body motion sensor using static electric field sensing. In *Proceedings of the 2012 ACM conference on ubiquitous computing*. 99–102.

[29] Gabe Cohn, Daniel Morris, Shwetak N Patel, and Desney S Tan. 2011. Your noise is my command: sensing gestures using the body as an antenna. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 791–800.

[30] Kaya de Barbaro. 2019. Automated sensing of daily activity: A new lens into development. *Developmental psychobiology* 61, 3 (2019), 444–464.

[31] Junting Dong, Qi Fang, Wen Jiang, Yurou Yang, Hujun Bao, and Xiaowei Zhou. 2021. Fast and Robust Multi-Person 3D Pose Estimation and Tracking from Multiple Views. In *T-PAMI*.

[32] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. 2020. Counting out time: Class agnostic video repetition counting in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10387–10396.

[33] Qi Fang, Qing Shuai, Junting Dong, Hujun Bao, and Xiaowei Zhou. 2021. Reconstructing 3D Human Pose by Watching Humans in the Mirror. In *CVPR*.

[34] Gabriele Frediani, Federica Vannetti, Leonardo Bocchi, Giovanni Zonfrillo, and Federico Carpi. 2021. Monitoring Flexions and Torsions of the Trunk via Gyroscope-Calibrated Capacitive Elastomeric Wearable Sensors. *Sensors* 21, 20 (2021), 6706.

[35] A Garinei and R Marsili. 2014. Development of a new capacitive matrix for a steering wheel's pressure distribution measurement. *International Journal of Industrial Ergonomics* 44, 1 (2014), 114–119.

[36] Kirill Gavrilyuk, Ryan Sanford, Mehrsan Javan, and Cees GM Snoek. 2020. Actor-transformers for group activity recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 839–848.

[37] John Ghattas and Danielle N Jarvis. 2021. Validity of inertial measurement units for tracking human motion: a systematic review. *Sports Biomechanics* (2021), 1–14.

[38] Hristijan Gjoreski and Daniel Roggen. 2017. Unsupervised online activity discovery using temporal behaviour assumption. In *Proceedings of the 2017 ACM International Symposium on Wearable Computers*. 42–49.

[39] Tobias Grosse-Puppendahl, Eugen Berlin, and Marko Borazio. 2012. Enhancing accelerometer-based activity recognition with capacitive proximity sensing. In *International Joint Conference on Ambient Intelligence*. Springer, 17–32.

[40] Tobias Grosse-Puppendahl, Christian Holz, Gabe Cohn, Raphael Wimmer, Oskar Bechtold, Steve Hodges, Matthew S Reynolds, and Joshua R Smith. 2017. Finding common ground: A survey of capacitive sensing in human-computer interaction. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. 3293–3315.

[41] Harish Haresamudram, David V Anderson, and Thomas Plötz. 2019. On the role of features in human activity recognition. In *Proceedings of the 23rd International symposium on wearable computers*. 78–88.

[42] Edward F Harris and Richard N Smith. 2009. Accounting for measurement error: a critical but often overlooked process. *Archives of oral biology* 54 (2009), S107–S117.

[43] Yinghao Huang, Manuel Kaufmann, Emre Aksan, Michael J Black, Otmar Hilliges, and Gerard Pons-Moll. 2018. Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time. *ACM Transactions on Graphics (TOG)* 37, 6 (2018), 1–15.

[44] Jihye Hwang, John Yang, and Nojun Kwak. 2020. Exploring Rare Pose in Human Pose Estimation. *IEEE Access* 8 (2020), 194964–194977.

[45] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 2013. 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.

[46] Yifeng Jiang, Yuting Ye, Deepak Gopinath, Jungdam Won, Alexander W Winkler, and C Karen Liu. 2022. Transformer Inertial Poser: Attention-based Real-time Human Motion Reconstruction from Sparse IMUs. *arXiv preprint arXiv:2203.15720* (2022).

[47] Haojian Jin, Zhijian Yang, Swarun Kumar, and Jason I Hong. 2018. Towards wearable everyday body-frame tracking using passive RFIDs. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 1–23.

[48] Manuel Kaufmann, Yi Zhao, Chengcheng Tang, Lingling Tao, Christopher Twigg, Jie Song, Robert Wang, and Otmar Hilliges. 2021. Em-pose: 3d human pose estimation from sparse electromagnetic trackers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11510–11520.

[49] Karly Kudrinko, Emile Flavin, Xiaodan Zhu, and Qingguo Li. 2020. Wearable sensor-based sign language recognition: A comprehensive review. *IEEE Reviews in Biomedical Engineering* 14 (2020), 82–97.

[50] Kai Kunze and Paul Lukowicz. 2008. Dealing with sensor displacement in motion-based onbody activity recognition systems. In *Proceedings of the 10th international conference on Ubiquitous computing*. 20–29.

[51] Chang June Lee and Jung Keun Lee. 2022. Inertial Motion Capture-Based Wearable Systems for Estimation of Joint Kinetics: A Systematic Review. *Sensors* 22, 7 (2022), 2507.

[52] Frédéric Li, Kimiaki Shirahama, Muhammad Adeel Nisar, Xinyu Huang, and Marcin Grzegorzek. 2020. Deep transfer learning for time series data based on sensor modality classification. *Sensors* 20, 15 (2020), 4271.

[53] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. 2021. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13401–13412.

[54] Siming Li, Ruiqing Li, Tianjiao Chen, and Xueliang Xiao. 2020. Highly sensitive and flexible capacitive pressure sensor enhanced by weaving of pyramidal concavities staggered in honeycomb matrix. *IEEE Sensors Journal* 20, 23 (2020), 14436–14443.

[55] Sheng Li, Zhiqiang Tao, Kang Li, and Yun Fu. 2019. Visual to text: Survey of image and video captioning. *IEEE Transactions on Emerging Topics in Computational Intelligence* 3, 4 (2019), 297–312.

[56] Han Liang, Yannan He, Chengfeng Zhao, Mutian Li, Jingya Wang, Jingyi Yu, and Lan Xu. 2022. HybridCap: Inertia-aid Monocular Capture of Challenging Human Motions. *arXiv preprint arXiv:2203.09287* (2022).

[57] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[58] Mengmeng Liu, Xiong Pu, Chunyan Jiang, Ting Liu, Xin Huang, Libo Chen, Chunhua Du, Jiangman Sun, Weiguo Hu, and Zhong Lin Wang. 2017. Large-area all-textile pressure sensors for monitoring human motion and physiological signals. *Advanced materials* 29, 41 (2017), 1703700.

[59] Shiqiang Liu, Junchang Zhang, Yuzhong Zhang, and Rong Zhu. 2020. A wearable motion capture device able to detect dynamic motion of human limbs. *Nature communications* 11, 1 (2020), 1–12.

[60] Shi Qiang Liu, Jun Chang Zhang, Guo Zhen Li, and Rong Zhu. 2020. A wearable flow-MIMU device for monitoring human dynamic motion. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 28, 3 (2020), 637–645.

[61] Michael Lorenz, Gabriele Bleser, Takayuki Akiyama, Takehiro Niikura, Didier Stricker, and Bertram Taetz. 2022. Towards Artefact Aware Human Motion Capture using Inertial Sensors Integrated into Loose Clothing. In *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 1682–1688.

[62] Zhuyu Ma, Yang Zhang, Kaiyi Zhang, Hua Deng, and Qiang Fu. 2022. Recent progress in flexible capacitive sensors: Structures and properties. *Nano Materials Science* (2022).

[63] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. 2019. AMASS: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*. 5442–5451.

[64] Denys JC Matthies, Chamod Weerasinghe, Bodo Urban, and Suranga Nanayakkara. 2021. Capglasses: Untethered capacitive sensing with smart glasses. In *Augmented Humans Conference 2021*. 121–130.

[65] Denys JC Matthies, Alex Woodall, and Bodo Urban. 2021. Prototyping Smart Eyewear with Capacitive Sensing for Facial and Head Gesture Detection. In *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers*. 476–480.

[66] Jonathan Samir Matthis and Aaron Cherian. 2022. *FreeMoCap: A free, open source markerless motion capture system.* https://github.com/freemocap/freemocap/

[67] Denisa Qori McDonald, Richard Vallett, Erin Solovey, Geneviève Dion, and Ali Shokoufandeh. 2020. Knitted Sensors: Designs and Novel Approaches for Real-Time, Real-World Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–25.

[68] Zhaozong Meng, Mingxing Zhang, Changxin Guo, Qirui Fan, Hao Zhang, Nan Gao, and Zonghua Zhang. 2020. Recent progress in sensing and computing techniques for human activity recognition and motion analysis. *Electronics* 9, 9 (2020), 1357.

[69] Matteo Menolotto, Dimitrios-Sokratis Komaris, Salvatore Tedesco, Brendan O'Flynn, and Michael Walsh. 2020. Motion capture technology in industrial applications: A systematic review. *Sensors* 20, 19 (2020), 5687.

[70] Jieming Pan, Yuxuan Luo, Yida Li, Chen-Khong Tham, Chun-Huat Heng, and Aaron Voon-Yew Thean. 2020. A wireless multi-channel capacitive sensor system for efficient glove-based gesture recognition with AI at the edge. *IEEE Transactions on Circuits and Systems II: Express Briefs* 67, 9 (2020), 1624–1628.

[71] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 2019. 3D human pose estimation in video with temporal convolutions and semi-supervised training. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[72] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. 2021. Neural Body: Implicit Neural Representations with Structured Latent Codes for Novel View Synthesis of Dynamic Humans. In *CVPR*.

[73] Lukasz Piwek, David A Ellis, Sally Andrews, and Adam Joinson. 2016. The rise of consumer health wearables: promises and barriers. *PLoS medicine* 13, 2 (2016), e1001953.

[74] Chiara Plizzari, Marco Cannici, and Matteo Matteucci. 2021. Skeleton-based action recognition via spatial and temporal transformer networks. *Computer Vision and Image Understanding* 208 (2021), 103219.

[75] Franchino Porciuncula, Anna Virginia Roto, Deepak Kumar, Irene Davis, Serge Roy, Conor J Walsh, and Louis N Awad. 2018. Wearable movement sensors for rehabilitation: a focused review of technological and clinical advances. *Pm&r* 10, 9 (2018), S220–S232.

[76] E Ramanujam, Thinagaran Perumal, and S Padmavathi. 2021. Human activity recognition with smartphone and wearable sensors using deep learning techniques: A review. *IEEE Sensors Journal* 21, 12 (2021), 13029–13040.

[77] Manju Rana and Vikas Mittal. 2020. Wearable sensors for real-time kinematics analysis in sports: a review. *IEEE Sensors Journal* 21, 2 (2020), 1187–1207.

[78] Christopher Reining, Friedrich Niemann, Fernando Moya Rueda, Gernot A Fink, and Michael ten Hompel. 2019. Human activity recognition for production and logistics—a systematic literature review. *Information* 10, 8 (2019), 245.

[79] Yili Ren, Zi Wang, Sheng Tan, Yingying Chen, and Jie Yang. 2021. Winect: 3d human pose tracking for free-form activity using commodity wifi. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 4 (2021), 1–29.

[80] Douglas A Reynolds. 2009. Gaussian mixture models. *Encyclopedia of biometrics* 741, 659-663 (2009).

[81] Aaqib Saeed, Tanir Ozcelebi, and Johan Lukkien. 2019. Multi-task self-supervised learning for human activity detection. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 2 (2019), 1–30.

[82] Nikolaos Sarafianos, Bogdan Boteanu, Bogdan Ionescu, and Ioannis A Kakadiaris. 2016. 3d human pose estimation: A review of the literature and analysis of covariates. *Computer Vision and Image Understanding* 152 (2016), 1–20.

[83] Fatemeh Serpush, Mohammad Bagher Menhaj, Behrooz Masoumi, and Babak Karasfi. 2022. Wearable Sensor-Based Human Activity Recognition in the Smart Healthcare System. *Computational Intelligence and Neuroscience* 2022 (2022).

[84] Young-Eun Shin, Jeong-Eun Lee, Yoojeong Park, Sang-Ha Hwang, Han Gi Chae, and Hyunhyub Ko. 2018. Sewing machine stitching of polyvinylidene fluoride fibers: programmable textile patterns for wearable triboelectric sensors. *Journal of Materials Chemistry A* 6, 45 (2018), 22879–22888.

[85] Monit Shah Singh, Vinaychandran Pondenkandath, Bo Zhou, Paul Lukowicz, and Marcus Liwickit. 2017. Transforming sensor data to the image domain for deep learning—An application to footstep detection. In *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2665–2672.

[86] Patrick Slade, Ayman Habib, Jennifer L Hicks, and Scott L Delp. 2021. An open-source and wearable system for measuring 3D human motion in real-time. *IEEE Transactions on Biomedical Engineering* 69, 2 (2021), 678–688.

[87] Wolfgang Teufl, Markus Miezal, Bertram Taetz, Michael Fröhlich, and Gabriele Bleser. 2019. Validity of inertial sensor based 3D joint kinematics of static and dynamic sport and physiotherapy specific movements. *PloS one* 14, 2 (2019), e0213064.

[88] Texas Instruments 2015. *FDC2x1x EMI-Resistant 28-Bit,12-Bit Capacitance-to-Digital Converter for Proximity and Level Sensing Applications*. Texas Instruments.

[89] Matthew Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John Collomosse. 2017. Total capture: 3d human pose estimation fusing video and inertial sensors. In *Proceedings of 28th British Machine Vision Conference*. 1–13.

[90] Shuhei Tsuchida, Satoru Fukayama, Masahiro Hamasaki, and Masataka Goto. 2019. AIST Dance Video Database: Multi-Genre, Multi-Dancer, and Multi-Camera Database for Dance Information Processing.. In *ISMIR*, Vol. 1. 6.

[91] Hanyue Tu, Chunyu Wang, and Wenjun Zeng. 2020. Voxelpose: Towards multi-camera 3d human pose estimation in wild environment. In *European Conference on Computer Vision*. Springer, 197–212.

[92] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. 2020. Scan: Learning to classify images without labels. In *European conference on computer vision*. Springer, 268–285.

[93] Jinbao Wang, Shujie Tan, Xiantong Zhen, Shuo Xu, Feng Zheng, Zhenyu He, and Ling Shao. 2021. Deep 3D human pose estimation: A review. *Computer Vision and Image Understanding* 210 (2021), 103225.

[94] Lin Wang, Hristijan Gjoreski, Mathias Ciliberto, Paula Lago, Kazuya Murao, Tsuyoshi Okita, and Daniel Roggen. 2021. Three-year review of the 2018–2020 SHL challenge on transportation and locomotion mode recognition from mobile sensors. *Frontiers in Computer Science* (2021).

[95] Mathias Wilhelm, Jan-Peter Lechler, Daniel Krakowczyk, and Sahin Albayrak. 2020. Ring-based finger tracking using capacitive sensors and long short-term memory. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 551–555.

[96] WK Wong, Filbert H Juwono, and Brendan Teng Thiam Khoo. 2021. Multi-features capacitive hand gesture recognition sensor: A machine learning approach. *IEEE Sensors Journal* 21, 6 (2021), 8441–8450.

[97] Emil Woop, Esther Friederike Zahn, Rahel Flechtner, and Gesche Joost. 2020. Demonstrating a Modular Construction Toolkit for Interactive Textile Applications. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*. 1–4.

[98] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. 2019. Detectron2. https://github.com/facebookresearch/detectron2.

[99] Chenggang Yan, Yunbin Tu, Xingzheng Wang, Yongbing Zhang, Xinhong Hao, Yongdong Zhang, and Qionghai Dai. 2019. STAT: Spatial-temporal attention mechanism for video captioning. *IEEE transactions on multimedia* 22, 1 (2019), 229–241.

[100] Ruiyang Yin, Depeng Wang, Shufang Zhao, Zheng Lou, and Guozhen Shen. 2021. Wearable sensors-enabled human–machine interaction systems: from design to application. *Advanced Functional Materials* 31, 11 (2021), 2008936.

[101] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4651–4659.

[102] Qi Zhang, Yu Lu Wang, Yun Xia, Timothy Vernon Kirk, and Xiao Dong Chen. 2020. Textile-Only capacitive sensors with a lockstitch structure for facile integration in any areas of a fabric. *ACS sensors* 5, 6 (2020), 1535–1540.

[103] Qi Zhang, Yu Lu Wang, Yun Xia, Peng Fei Zhang, Timothy V Kirk, and Xiao Dong Chen. 2019. Textile-only capacitive sensors for facile fabric integration without compromise of wearability. *Advanced Materials Technologies* 4, 10 (2019), 1900485.
[104] Boyu Zhao, Zhijia Dong, and Honglian Cong. 2022. A wearable and fully-textile capacitive sensor based on flat-knitted spacing fabric for human motions detection. *Sensors and Actuators A: Physical* 340 (2022), 113558.
[105] Enhao Zheng, Jingeng Mai, Yuxiang Liu, and Qining Wang. 2018. Forearm motion recognition with noncontact capacitive sensing. *Frontiers in Neurorobotics* 12 (2018), 47.
[106] Huiyu Zhou, Thomas Stone, Huosheng Hu, and Nigel Harris. 2008. Use of multiple wearable inertial sensors in upper limb motion tracking. *Medical engineering & physics* 30, 1 (2008), 123–133.

A EXPERIMENT PROTOCOL DETAILS

The ground truth video data was recorded with an iPhone SE (MX9R2ZD/A) front facing the participant. We recorded each participant in landscape mode to catch all movements within the field of view of the camera. Further, each participant was asked to stand on a marked section of the floor to keep a distance to the camera of 4 meters and also a distance of 2.5 meters to the Mac Pro 2013 which was used to gather the capacitive data. Both DAUs of the jacket were connected simultaneously via BLE to the Mac Pro, of which the clock was synchronized through the internet. The recording itself was realized under the usage of a Javascript web application. The BLE connection was implemented with the Web Bluetooth API (experimental at the time of writing and only supported in the Google Chrome browser). We used two instances within a Google Chrome browser to receive data from the left and the right side of the jacket concurrently. The capacitive data as well as the camera footage was recorded with a sampling rate of approximately 30 Hertz. Each recording involved the same start and end procedure with the order of starting or stopping the camera, afterwards the capacitive recording and the instructing video at last. Further, five consecutive touches of the left and right belly antenna by the participant at the start and the end of each recording generated a unique signal pattern to simplify the later synchronization between the capacitive data and the video footage.

B PROTOTYPE SENSING PATCH DIMENSIONS

The measurements of all the sensing patches are listed in Table 5. We set three origin points on each prototype jacket as shown in Fig. 3. OG1 is a floating origin in the middle of the three trapezoid-shaped patches; OG2 is at the inner terminal of the sleeve; OG3 is at the bottom of the front zipper. W and H stand for the width and height of the sensor patch. The distance from OG1 to the edge of the sleeve is 67cm in Proto.1, and 61cm in Proto.2. For the trapezoid-shaped patches (CH0, CH1, CH2), the minimum and maximum width were measured. The distance from the edge of the patch to the closest origin is marked as $OGn(x, y)$, where x and y are the horizontal and vertical directions as indicated in Fig. 3. Sinch CH7 is tilted, we measured both the top and bottom ends of the patch to OG3.

C DEEP LEARNING DETAILS

The deep convolutional regressor model from Fig. 5 and Table 6 was trained on a GPU High Performanc Computing (HPC) cluster with NVIDIA® RTXA6000 GPUs and AMD® EPYC™ 7002 CPUs and the nvcr.io/nvidia/tensorflow:22.07-tf2-py3 container. The batch size was set to 2048. It was trained with the Adam optimizer with an initial learning rate of 0.01, which is reduced by half after every 50 epochs. The validation split was 0.2. The loss function was Mean Absolute Error and the early stopping was set to track the minimum validation MAE with 100 patience. We further list the details of the deep learning models for the temporal pose sequences (motion) used in Section 8.1 in Table 7 and Table 8.

Table 4. Experiment Instruction Video Contents

Task	Task Type	Detailed Descriptions
Instruction Video I : Daily and casual movements		
1	Upper body gestures from related work [13]	20 gestures including leaning or turning to different sides, shrugging, clapping, swinging...
2	Basic stretching exercises	19 body movements based on morning stretching exercises for school kids in the Asian region [1]
3	Beginner sign language tutorials	Various basic signs in standard and slow motion speed including for instance "Hello", "Yes", "No" and names like "Ashley" [2][2]
4	Dance videos from social media platform (TikTok)	Four selected viral TikTok dances to add fluent and uncommon movements to the instructions, including for instance the Macarena dance move
Instruction Video II : Controlled movements		
1	Shoulder movements	Moving the shoulder joint up and down, front and back, forward rotation and backward rotation
2	Arms down movements	With arms down, curl individual elbow and then both elbows from the inside, neural and outside tracks; rotate forearms clock wise and counter clock wise.
3	Arms flat movements	with each arm flat while the other arm down, and then both arms flat, repeat the previous elbow movements from Task 2
4	Arms up movements 1	with each arm raised while the other arm down, and then both arms flat, repeat the previous elbow movements from Task 2
5	Arms up movements 2	with each arm raised while the other arm flat, repeat the previous elbow movements from Task 2

[1] Link: https://datanews.caixin.com/interactive/2019/guangboticao/

[2] Link: Video Link: https://www.youtube.com/watch?v=Raa0vBXA8OQ

Table 5. The dimensions and positions of the patches from Proto.1 and Proto.2. (units= cm)

	Proto.1 Left	Proto.1 Right	Proto.2 Left	Proto.2 Right
CH0	W=12~4, H=7, OG1(0, 2)	W=12~4, H=6, OG1(0, 2)	W=13~4, H=7, OG1(0, 5)	W=13~4, H=7, OG1(0, 4)
CH1	W=12~4, H=7, OG1(4, 0)	W=12~4, H=7, OG1(4, 0)	W=13~4, H=7, OG1(3, 0)	W=13~4, H=7, OG1(4, 0)
CH2	W=12~4, H=7, OG1(0, 2)	W=12~4, H=7, OG1(0, 2)	W=13~4, H=7, OG1(0, 7)	W=13~4, H=7, OG1(0, 4)
CH3	W=1.5, H=39, OG2(11, 14)	W=1.5, H=39, OG2(11, 13)	W=3, H=38, OG2(12, 12)	W=3, H=38, OG2(10, 10)
CH4	W=2, H=38, OG3(7, 7)	W=2, H=38, OG3(7, 7)	W=3.5, H=30, OG3(8, 16)	W=3.5, H=30, OG3(7, 16)
CH5	W=12, H=7, OG2(10, 0)	W=12, H=7, OG2(9, 0)	W=12, H=7, OG2(7, 3)	W=12, H=7, OG2(8, 4)
CH6	W=12, H=7, OG2(31, 0)	W=12, H=7, OG2(30, 0)	W=12, H=7, OG2(28, 3)	W=12, H=7, OG2(29, 4)
CH7	W=7, H=46, OG3-bottom(43, 16), OG3-top(38, 62)	W=7, H=46, OG3-bottom(43, 16), OG3-top(38, 62)	W=6, H=31.5, OG3-bottom(46, 20), OG3-top(34, 48)	W=6, H=31.5, OG3-bottom(46, 20), OG3-top(36, 50)

Table 6. Model Summary of the MoCaPose Deep Convolutional Regressor

Layer (type)	Output Shape	Kernel	Parameters
Input layer	(None, 32, 8, 2)		
	sequence of 32 (1s) 16-ch capacitive signals (8,2)		
1: conv2d	(None, 32, 8, 32)	(5, 8)	2592
2: average pooling2d	(None, 16, 8, 32)	(2, 1)	0
3: dropout	(None, 16, 8, 32)		0
4: batch normalization	(None, 16, 8, 32)		128
5: conv2d	(None, 16, 8, 64)	(5, 8)	81984
6: average pooling2d	(None, 8, 8, 64)	(2,1)	0
7: dropout	(None, 8, 8, 64)		0
8: batch normalization	(None, 8, 8, 64)		256
9: conv2d	(None, 8, 8, 96)	(5,8)	245856
10: average pooling2d	(None, 4, 4, 96)	(2,2)	0
11: dropout	(None, 4, 4, 96)		0
12: batch normalization	(None, 4, 4, 96)		384
13: conv2d	(None, 4, 4, 64)	(4,4)	98368
14: dropout	(None, 4, 4, 64)		0
15: batch normalization	(None, 4, 4, 64)		256
16: reshape	(None, 16, 64)		0
17: conv1d	(None, 14, 32)	(3) no padding	6176
18: conv1d	(None, 12, 27)	(3) no padding	2619
19: conv1d	(None, 10, 9)	(3) no padding	738
20: conv1d (output)	(None, 8, 3)	(3) no padding	84
Total params			439,441
Trainable params			438,929
Non-trainable params			512
All activation functions are 'relu' when applicable			
'None' indicates batch size			

Table 7. Model Summary of the Autoencoder for Extracting Motion-specific Features

Layer (type)	Output Shape	Kernel	Parameters
Encoder - decompose pose sequences			
Input layer	(None, 32, 8, 3)		
	sequence of 32 (2s) 3D pose coordinates (8,3)		
1: conv2d	(None, 32, 8, 32)	(4, 8)	3104
2: average pooling2d	(None, 16, 8, 32)	(2, 1)	0
3: dropout	(None, 16, 8, 32)		0
4: batch normalization	(None, 16, 8, 32)		128
5: conv2d	(None, 16, 8, 32)	(4, 8)	32800
6: average pooling2d	(None, 8, 8, 32)	(2, 1)	0
7: dropout	(None, 8, 8, 32)		0
8: batch normalization	(None, 8, 8, 32)		128
9: conv2d	(None, 1, 1, 32)	(8, 8) no padding	65568
10: dropout	(None, 1, 1, 32)		0
11: batch normalization	(None, 1, 1, 32)		128
12: flatten	(None, 32)		0
13: dense	(None, 32)		1056
Latent motion-specific feature vector (None, 32)			
Decoder - reconstruct pose sequences			
14: dense	(None, 64)		2112
15: reshape	(None, 8, 8, 1)		0
16: conv2d transpose	(None, 16, 8, 32)	(4, 4) stride (2, 1)	544
17: conv2d transpose	(None, 32, 8, 32)	(4, 4) stride (2, 1)	16416
18: conv2d	(None, 32, 8, 32)	(4, 4)	16416
19: conv2d	(None, 32, 8, 3)	(4, 4)	1539
Total params			139,939
Trainable params			139,747
Non-trainable params			192
All activation functions are 'relu' when applicable			
'None' indicates batch size			

Table 8. Model Summary of the Deep Classifier from Capacitive Signals in Section 8.1 Process 5 CAP-CL

Layer (type)	Output Shape	Kernel / Activation	Parameters
Input layer	(None, 64, 8, 2)		
	sequence of 64 (2s) 16-ch capacitive signals (8,2)		
1: conv2d	(None, 64, 8, 20)	(8, 8)	2580
3: dropout	(None, 64, 8, 20)		0
4: batch normalization	(None, 64, 8, 20)		80
5: conv2d	(None, 64, 8, 40)	(8, 8)	51240
6: average pooling2d	(None, 16, 4, 40)	(4, 2)	0
7: dropout	(None, 16, 4, 40)		0
8: batch normalization	(None, 16, 4, 40)		160
9: conv2d	(None, 16, 4, 80)	(4, 4)	51280
10: average pooling2d	(None, 4, 2, 80)	(4, 2)	0
11: dropout	(None, 4, 2, 80)		0
12: batch normalization	(None, 4, 2, 80)		320
13: conv2d	(None, 4, 2, 160)	(2, 2)	51360
14: average pooling2d	(None, 2, 1, 160)	(2, 2)	0
15: dropout	(None, 2, 1, 160)		0
16: batch normalization	(None, 2, 1, 80)		640
17: conv2d	(None, 2, 1, 20)	(2, 1)	12820
18: flatten	(None, 40)		0
19: dropout	(None, 40)		0
20: batch normalization	(None, 40)		160
21: dense	(None, 40)		1640
22: dropout	(None,40)		0
23: batch normalization	(None, 40)		160
24: dense	(None, 10)	*sofmax	410
Total params			172,850
Trainable params			172,090
Non-trainable params			760
All activation functions are 'relu' when applicable except for Layer 24			
'None' indicates batch size			