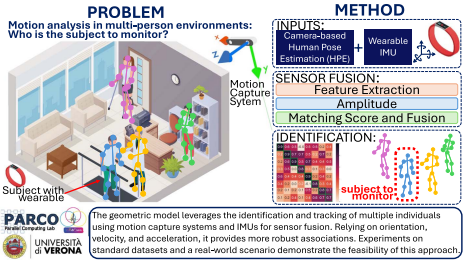
Sensor applications

Combining 3-D Human Pose Estimation and IMU Sensors for Human Identification and Tracking in Multiperson Environments

Mirco De Marchi, Cristian Turetta*, Graziano Pravadelli**, and Nicola Bombieri***

Department of Engineering for Innovation Medicine, University of Verona, 37129 Verona, Italy
**Graduate Student Member, IEEE*
***Senior Member, IEEE*
****Member, IEEE*

Manuscript received 13 March 2024; revised 1 May 2024; accepted 7 May 2024. Date of publication 13 May 2024; date of current version 22 May 2024.



Abstract—Human pose estimation (HPE) based on deep neural networks aims to predict the poses of human body in videos without needing markers. One of the main limitations in its applicability is consistently identifying and tracking the keypoints of an individual in multiperson scenarios. Despite various solutions based on image analysis being attempted, challenges, such as model accuracy, occlusions, or individuals, exiting the camera's field of view often result in the loss of the association between humans and their keypoints across video frames. In this letter, we propose a human identification and tracking methodology in multiperson environments based on data fusion between HPE software and wearable inertial measurement unit (IMU) sensors. We demonstrate how to align the data generated by these two sensor categories (camera-based HPE and IMUs) and assess the alignment between each skeleton of keypoints and IMU pair using a scoring system. In addition, we illustrate how to combine different metrics, such as orientation, acceleration, and velocity, to address alignment problems caused by inaccuracies in sensor data.

Index Terms—Sensor applications, body area network (BAN), data fusion, human pose estimation (HPE), human identification and tracking, inertial measurement unit (IMU), wearables.

I. INTRODUCTION

Human pose estimation (HPE) entails the estimation of geometric and kinetic data of the human body using data acquired through sensors, in particular videos. The advances in neural network architectures and to the fact that such a deep learning technology does not require markers attached to the subject's body have facilitated its expanding utilization across various domains, such as human–robot interaction, surveillance, healthcare, and telemedicine [1].

Several HPE solutions have shown high accuracy in estimating the pose of a single person from images and videos with both 2-D and 3-D pose annotations. However, they have shown accuracy limitations in highly occluded and multiperson scenarios. In addition to the accuracy limitations in these real-world application contexts, reliable *identification* and *tracking* of an individual within a multipeople enviroment monitored by a HPE system is very challenging [2]. Identification consists of detecting the instance of a specific subject in a video frame, while tracking is the process of temporally associating the human detection within a video sequence to generate persistent paths or trajectories. They are considered the first two processes in many computer video pipelines, and often feed into higher level reasoning modules, such as action recognition and dynamic scene analysis.

Different approaches based on computer vision techniques, such as face recognition, clothing color detection, and body shape analysis, represent the current state of the art for identifying individuals in multiperson scenarios. Nevertheless, these approaches are not applicable to HPE signals and tend to be unreliable when subjects are partially occluded [3].

A different approach consists of using wearable inertial measurement units (IMUs) to measure the motion, orientation, and position of a specific subject [4]. IMUs consist of accelerometers, gyroscopes, and magnetometers, and are the basic components of body area networks (BANs). Since these daily wearables contain an intrinsic identifier for individuals, they provide a practical solution to the human identification and tracking problem in crowded and dynamic environments [5]. State-of-the-art works use wearable IMU sensors for 3-D human motion analysis, developing portable and real-time pose estimation systems robust to sensor view occlusions and lighting issues [6]. Wearable IMUs are also used to track a person's movement by integrating acceleration and angular velocity measurements over time. However, they suffer from accumulative errors over extended periods, which lead to significant inaccuracies, thus making the identification and tracking of a person unreliable [7].

Recent works investigated on the fusion of 3-D human pose data extracted by HPE software and IMU sensors to improve the pose accuracy [8]. These studies reveal that data fusion from HPE and IMU sensors results in higher pose accuracy compared to utilizing either HPE or IMUs separately. However, none of these studies investigated on the fusion of HPE and IMUs data for identification and tracking of the 3-D human skeleton associated with the individual wearing the IMUs in real and multiperson scenarios. Other works combine video and IMU to generate inertial data from human poses, with the aim of mitigating the lack of labeled training data [9], [10].

In this work, we tackle this challenge by proposing a data fusion methodology aimed at processing, matching, and assessing the alignment between the signals generated by these two sensor categories. The alignment is based on a geometric model alongside three different metrics (i.e., orientation, acceleration, and velocity) to capture various features of human movements. To assess the alignment between each skeleton of keypoints detected by the HPE and the IMUs signals, the methodology implements a scoring system based on different

Corresponding author: Mirco De Marchi (e-mail: mirco.demarchi@univr.it).
Associate Editor: Hamza Shakeel.
Digital Object Identifier 10.1109/LSENS.2024.3400614

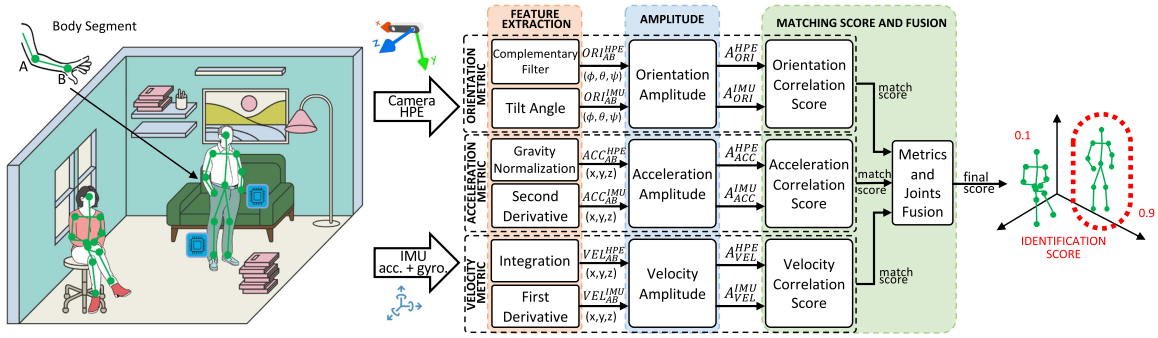


Fig. 1. Pipeline of the proposed identification and tracking methodology through sensor fusion.

correlation and similarity functions. We present the experimental evaluation of the proposed methodology, utilizing both a standard dataset (TotalCapture [11]) and a real-world case study. Our results demonstrate that the integration of diverse alignment metrics, the use of multiple IMUs positioned on joints keypoints, and the implementation of the scoring system collectively mitigate sensor inaccuracies. Consequently, this approach achieves a satisfactory level of accuracy in identifying and tracking human skeletons within crowded environments, even under realistic conditions.

II. METHODOLOGY

Fig. 1 depicts the pipeline of the proposed methodology. We start from the 3-D keypoints generated by an HPE platform (i.e., the 3-D human skeleton) alongside the acceleration and angular velocity data generated by each IMU within a BAN. We assume a shared operating frequency, where both HPE and BAN provide data simultaneously. For the sake of clarity, we consider multiple instances of 3-D human skeletons generated by the HPE and a single BAN. The objective is to determine the 3-D human skeleton associated with the individual wearing the BAN. The approach can be extended to identify the skeleton-to-BAN correspondence when multiple BANs are involved.

The concept aims to align the data produced by the two sensors (HPE and BAN) and assess the alignment between each skeleton-BAN pair using a scoring system. This process is calibration-free, i.e., it does not require calibrating the cameras and tolerates slightly position changes of the wearable sensors. It involves three main steps: A feature extraction module converts the positional data from HPE and the vector data from IMUs into three reference metrics: orientation, acceleration, and velocity. An amplitude module condenses the multidimensional signals into a single magnitude signal for each metric. A matching score and fusion module evaluates the correlation between the amplitudes of the HPE and IMU signals for each metric, generating a final score. At each time step, the skeleton with the highest score signifies the individual wearing the BAN.

A. Feature Extraction

Starting from the set of 3-D keypoints generated by the HPE, we define a *body segment AB* as a couple of keypoints A and B in the HPE reference system, where $A = (x_A, y_A, z_A)$ and $B = (x_B, y_B, z_B)$.

For each IMU within the BAN, such as the wrist IMU or ankle IMU, we designate a distinct body segment (e.g., elbow–wrist for the wrist IMU and knee–ankle for the ankle IMU). This process entails extrapolating the orientation of a body segment AB into an Euler angle triple (roll, pitch, and yaw): $ORI_{AB}^{\text{HPE}} = \phi_{AB}^{\text{HPE}}, \theta_{AB}^{\text{HPE}}$, and ψ_{AB}^{HPE}, where $\theta_{AB}^{\text{HPE}} = \arctan(-\frac{|x_A - x_B|}{\sqrt{(y_A-y_B)^2+(z_A-z_B)^2}})$, $\psi_{AB}^{\text{HPE}} = \arctan(\frac{|z_A-z_B|}{|y_A-y_B|})$, and $\phi_{AB}^{\text{HPE}} = 0$.

The module extrapolates the orientation of the IMU linked to the body segment AB, $\text{ORI}_{AB}^{\text{IMU}} = \phi_{AB}^{\text{IMU}}, \theta_{AB}^{\text{IMU}}$, and ψ_{AB}^{IMU} by employing a complementary filter [12] on the acceleration and angular velocity concerning the world reference system. To prevent interference in indoor settings, the magnetometer signal is disregarded ($\psi_{AB}^{\text{IMU}} = 0$).

For the acceleration and velocity metrics, we link a single keypoint with each IMU (e.g., the wrist 3-D keypoint corresponds to the wrist IMU, and so forth). The module extrapolates the acceleration of the HPE keypoints by utilizing a second derivative of the keypoint's movement over time. The result is a 3-D array of values (i.e., the acceleration over x, y, and z). Conversely, the IMU sensor provides raw acceleration data, which includes the gravity vector distributed across the three Cartesian axes. To derive the inertial acceleration over time, the module subtracts the gravity vector from the raw data. To compute the gravity vector $g(t)$, it constructs the rotation matrix $R(t)$ from the quaternion acquired through the complementary filter. Subsequently, it applies the rotation to the nominal gravity vector $g(t) = R(t)^{-1} g_{\text{nominal}}^{T}$, where the nominal gravity (i.e., $[0, 0, 9.81]$) depends on the local reference system of the IMU sensor. The outcome is the acceleration over the three Cartesian axes, normalized relative to gravity.

Similarly, the module calculates the velocity of an HPE keypoint through a first derivative of the keypoint movement across the time. It extrapolates the velocity of the IMU sensor from the raw acceleration data, normalized by the gravity vector. It extrapolates the inertial velocity through the integral of the normalized acceleration.

B. Amplitude

The amplitude module translates the Euler angle triple representing the orientation of both HPE and IMU into a 4-D signal (i.e., a quaternion) $[q_w, q_x, q_y, q_z]$ for both HPE and IMU ($A_{\text{ORI}}^{\text{HPE}}$ and $A_{\text{ORI}}^{\text{IMU}}$): $A_{\text{ORI}}^{*} = 2\arctan(\frac{\sqrt{q_x^2+q_y^2+q_z^2}}{q_w})$. In such a translation, it considers the pitch angle as the only information provided by both HPE and IMU feature extraction modules.

The module computes the acceleration and velocity amplitude ($A_{\text{ACC}}^{\text{HPE}}$, $A_{\text{ACC}}^{\text{IMU}}$, and $A_{\text{VEL}}^{\text{HPE}}$, $A_{\text{VEL}}^{\text{IMU}}$) as the modulus of the corresponding 3-D signal provided by the feature extraction module. We implemented and compared three different solutions to compute the amplitude, i.e., norm of order one and two, and the modulus of the sum of the vector components (*l1-norm* and *l2-norm*).

C. Matching Score and Fusion

The last module assesses the similarity between each HPE and IMU pairs for each metric. We implemented and compared different solutions to compute the similarity score. They include correlation functions, such as cross-correlation (*ccorr*), normalized cross-correlation (*nccorr*), and Pearson correlation (*pcorr*). They also include distance-based similarity functions, such as mean absolute error (*mae*), mean squared error (*mse*), root-mean-squared error (*rmse*), and dynamic time warping (*dtw*).

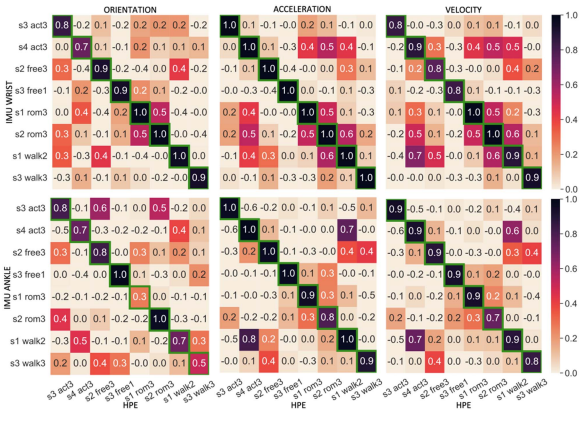


Fig. 2. Correlation matrix between IMUs and marker-based motion capture ground truth with the TotalCapture dataset.

Finally, the module combines the matching score of each metric and of different human joints to generate a final score. The fusion process computes the average of all single matching scores by discarding those with low correlation (i.e., under 0.5 Pearson correlation [13]).

III. EXPERIMENTAL RESULTS

We evaluated the methodology through both a standard dataset, TotalCapture [11], and a real-case study. The TotalCapture dataset comprises several videos taken by eight RGB camera points, which can be processed by a markerless HPE software platform. It also provides the ground-truth HPE results extrapolated through a marker-based motion capture system. Each video represents one action (among *walk*, *run*, *act*, and *free*style) performed by one subject (among a total of five) wearing 13 high-precision Xsens IMU sensors on key body joints, repeated three times. For the real case of study, we gathered a dataset featuring two subjects and five distinct actions (*walk*, *wait*, *eat*, *sit*, and *talk*) employing low-cost and fewer precise devices: a StereoLabs ZED2 RGB-D camera that provides a stereo matching algorithm for depth estimation and the Thingy Nordic IMU sensor.

We present the results of the identification a tracking through the proposed methodology by means of correlation matrices. The rows and columns represent the data collected by IMU and HPE, respectively, of a subject performing a specific action. Each matrix cell reports the final matching score (on average along all video instants) between the corresponding IMU and HPE signals. Consequently, the best matching score should be on the diagonal, with the theoretical maximum matching score equal to one. In this letter, we show the most representative correlation matrices, since we observed equivalent results for all combinations of subjects and actions of the datasets.

A. Results With TotalCapture

We started the analysis by examining the correlation between the IMU signals and the HPE data generated by the ground-truth marker-based system (see Fig. 2). The results show that under these optimal conditions, where both the HPE data and IMU signals exhibit exceptional quality, the diagonal consistently stands out. The cells bordered in green represent the outcome of the identification process, denoting the highest matching score among the eight pairs of signals. In these cases, each of the three distinct metrics (orientation, acceleration, and velocity) demonstrates perfect discrimination between matching and nonmatching subjects. In some cases, the final matching score falls short of the maximum value due to either noise in the IMU signals or the stationary nature of individuals in the corresponding videos. However, the significant disparity between the final matching score along the diagonal and other pairs enables the methodology to confidently identify the subject in any scenario.

We then considered the HPE signal generated by a standard markerless pose estimation software, i.e., OpenPose [14], instead of the

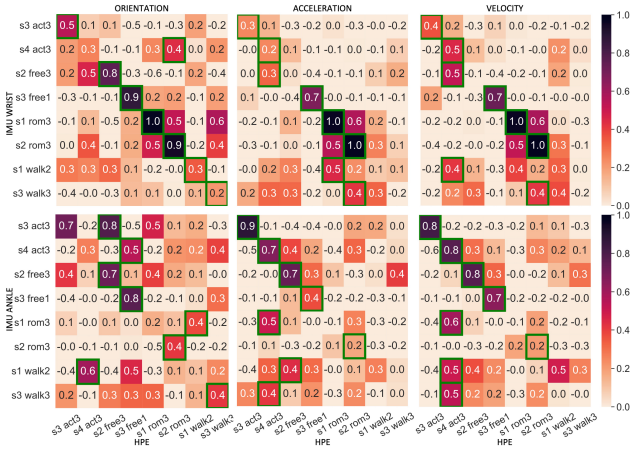


Fig. 3. Correlation matrix between IMU and OpenPose with the TotalCapture dataset.

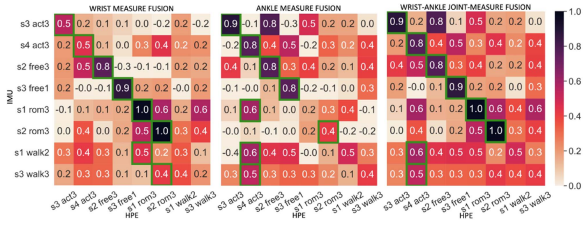


Fig. 4. Correlation matrix combining different metrics and human joints with the TotalCapture dataset.

ground-truth marker-based system. As TotalCapture exclusively includes RGB videos, we simulated typical depth errors for 2-D-to-3-D lifting by integrating depth information from ground-truth alongside Gaussian noise (mean 0, standard deviation 0.2 m) and 5 m random artifacts. The outcomes are depicted in Fig. 3. During actions characterized by high activity and substantial movements of the keypoints associated with an IMU, such as wrist movements in range of motion (*rom*) exercises or ankle movements during *free* actions, the orientation metric proves effective in accurately identifying the correct subject. However, neither orientation nor any other metric alone is consistently adequate, especially in scenarios where movements are limited, and the actions of the target individual closely resemble those of others. In such cases, the traces of each metric tend to overlap with one another, as observed in *act* or *walk* actions.

Fig. 4 illustrates the outcomes achieved by combining the matching scores of the three metrics with two IMU signals. For activities like *rom*, characterized by predominant arm movements, wrist IMUs demonstrate higher accuracy. Conversely, ankle IMUs prove more efficient in actions like *act*, primarily involving leg movements. Overall, we noted a remarkably high accuracy of the proposed method when combining the three metrics with two IMUs, positioned on the wrist and ankle (depicted on the right-most side of Fig. 4). However, in this configuration, the performance for the *walk* action is comparatively poorer due to the presence of highly similar and stationary movements. Incorporating matching scores associated with additional IMUs may provide a potential solution to this challenge.

In general, we found that the matching score and fusion step enable the integration of individual assessments to capture various aspects of human movements. This enhances the efficiency of determining the correlation between HPE and IMU signals by incorporating contributions from each joint of the BAN. Combining acceleration and velocity metrics with orientation improves the discrimination between actions, particularly during rapid and subtle movements. However, it is crucial to note that the computation of derivatives introduces noise into the signal, and this noise becomes more pronounced when computing the second derivative on the HPE 3-D signals, resulting in

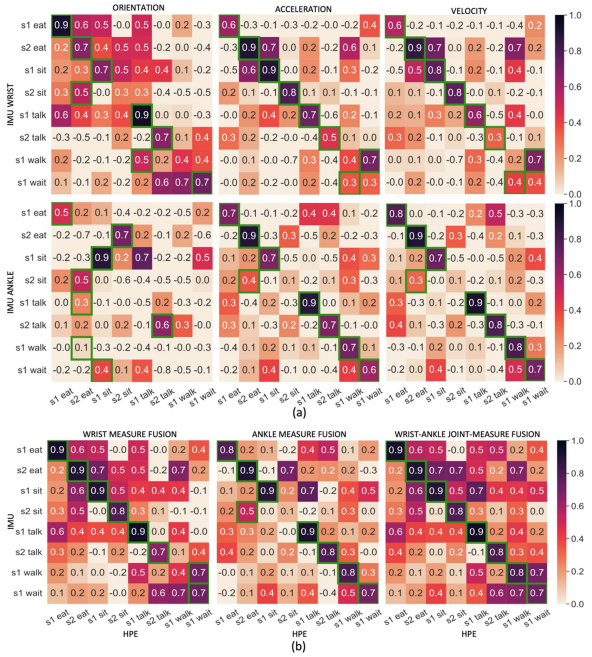


Fig. 5. Correlation matrix between low-cost IMUs and OpenPose with real case dataset using (a) individual metrics and combining metrics and (b) human joint information.

TABLE 1. Comparison Between Matching Functions in 3-D HPE With Noise Depth Calculated From Marker-Based Motion Capture Ground Truth

	Corr. right match (%)							Corr. wrong match (%)						
	ccorr	nccorr	pcorr	mae	mse	rmse	dtw	ccorr	nccorr	pcorr	mae	mse	rmse	dtw
ACC	7.67	0.69	0.69	0.91	2.56	1.37	52.26	5.37	-0.00	-0.01	1.93	7.17	2.50	84.26
VEL	0.07	0.70	0.70	0.70	0.67	0.77	38.56	0.05	-0.00	-0.00	0.70	0.71	0.79	39.11
ORI	1.01	0.63	0.63	0.15	0.04	0.18	4.66	0.95	0.04	0.04	0.32	0.16	0.39	10.53
Tot	2.92	0.67	0.67	0.59	1.09	0.77	31.83	2.12	0.01	0.01	0.98	2.68	1.22	44.63

Distance-based functions: mae, mse, rmse, dtw; correlation-based functions: ccorr, nccorr, pcorr.

degraded estimation. In these scenarios, velocity gives a more reliable and resilient matching though a single derivative of the position signal, at the cost of an integral on the IMU acceleration.

B. Results With the Real Case Real-World Scenario

With the real case dataset, it is even more evident that the combination of multiple metrics and different IMUs is necessary to achieve sufficient precision (see Fig. 5). For ankle IMUs, the orientation metric exhibits limited effectiveness, while acceleration and velocity metrics show a more discernible diagonal. In contrast, orientation contributes significantly to actions characterized by large movements, for instance, wrist movements on eating action or ankles movements on sitting action performed by subject s1. Consequently, integrating multiple metrics yields an enhancement in overall performance [see Fig. 5(b)]. Fusing data from wrist and ankle joints provides the best results.

Table 1 tabulates the average correlation score ($p \in [0, 1]$) in correct and wrong matches. It compares the matching functions employed, categorized into correlation-based and distance-based methods. All distance-based functions except *dtw* give similar results, being adapt to correctly identify data alignments. Among the correlation-based methods, *pcorr* or *nccorr* are preferable to *ccorr* as they provide bounded values, ensuring a more reliable matching process.

We evaluated different amplitude functions. For both acceleration and velocity, *l2-norm* shows higher discrimination capabilities. *l2-norm* leads to better results, with an average score of 0.70 for correct matches and −0.01 for wrong matches, instead *l1-norm* yields lower performance, with correct and wrong average scores of 0.50 and 0.00. In terms of orientation, the *quaternion magnitude* achieves an average score of 0.63 and 0.04 for correct and wrong matches. While it suggests effective discrimination capabilities, it also indicates lower efficacy compared to acceleration and velocity amplitudes.

IV. CONCLUSION

This work investigated on the fusion of HPE and IMUs data for identification and tracking of the 3-D human skeleton associated with the individual wearing the IMUs in real and multiperson scenarios. It presented a methodology to process and assess the alignment between HPE and IMU signals based on different metrics and on a scoring system. The results showed that the integration of multiple alignment metrics and multiple IMUs positioned correctly can mitigate the sensor inaccuracy and provide reliable identification and tracking results.

ACKNOWLEDGMENT

This work was supported in part by the "PREPARE" project n. F/310130/05/X56 - CUP: B39J23001730005 - D.M. MiSE 31/12/2021, and in part by the "UNISCO" project CUP: 1695-0026-553-2023 - PR Veneto FSE+ 2021-2027.

REFERENCES

[1] B. Scott et al., "Healthcare applications of single camera markerless motion capture: A scoping review," *PeerJ*, vol. 10, May 2022, Art. no. e13517.

[2] W. W. T. Lam et al., "A systematic review of the applications of markerless motion capture (MMC) technology for clinical measurement in rehabilitation," *J. NeuroEng. Rehabil.*, vol. 20, no. 1, May 2023, Art. no. 57, doi: 10.1186/s12984-023-01186-9.

[3] A. Zahra et al., "Person re-identification: A retrospective on domain specific open challenges and future trends," *Pattern Recognit.*, vol. 142, 2023, Art. no. 109669.

[4] Y. Li, Z. Meng, N. Gao, and Z. Zhang, "Heading angle correction with building map model constraint for single IMU-based pedestrian wearable localization," *IEEE Sens. Lett.*, to be published, doi: 10.1109/LSENS.2024.3370805.

[5] G. Retsinas et al., "Person identification using deep convolutional neural networks on short-term signals from wearable sensors," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 3657–3661.

[6] F. M. Calatrava-Nicolás and O. M. Mozos, "Light residual network for human activity recognition using wearable sensor data," *IEEE Sens. Lett.*, vol. 7, no. 10, Oct. 2023, Art. no. 7005304.

[7] Y. Huang et al., "Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time," *ACM Trans. Graph.*, vol. 37, no. 6, pp. 1–15, Dec. 2018, doi: 10.1145/3272127.3275108.

[8] F. Li, L. Wei, J. Chen, X. Huang, and K. Wang, "MSFusion: Multilayer sensor fusion-based robust motion estimation," *IEEE Sens. Lett.*, vol. 7, no. 2, Feb. 2023, Art. no. 6001004.

[9] A. Lämsä et al., "Video2IMU: Realistic IMU features and signals from videos," in *Proc. Int. Conf. Wearable Implantable Body Sensor Naetw.*, 2022, pp. 1–5.

[10] V. F. Rey et al., "Let there be IMU data: Generating training data for wearable, motion sensor based activity recognition from monocular rgb videos," in *Proc. Adjunct Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput. Proc. ACM Int. Symp. Wearable Comput.*, 2019, pp. 699–708, doi: 10.1145/3341162.3345590.

[11] M. Trumble et al., "Total capture: 3D human pose estimation fusing video and inertial sensors," in *Proc. Brit. Mach. Vis. Conf.*, 2017, pp. 1–13.

[12] H. Rong, C. Peng, Y. Chen, J. Lv, and L. Zou, "A time-efficient complementary Kalman gain filter derived from extended Kalman filter and used for magnetic and inertial measurement units," *IEEE Sensors J.*, vol. 22, no. 23, pp. 23077–23087, Dec. 2022.

[13] H. Xiong et al., "Exploiting a support-based upper bound of Pearson's correlation coefficient for efficiently identifying strongly correlated pairs," in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2004, pp. 334–343, doi: 10.1145/1014052.1014090.

[14] Z. Cao, Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, Jan. 2021.

Open Access funding provided by 'Università degli Studi di Verona' within the CRUI CARE Agreement