2024 Design, Automation & Test in Europe Conference (DATE 2024)

Real-Time Multi-Person Identification and Tracking via HPE and IMU Data Fusion

Mirco De Marchi[†], Cristian Turetta[†], Graziano Pravadelli[*], and Nicola Bombieri[*]
[*]*Department of Engineering for Innovation Medicine, University of Verona, Italy,* name.surname@univr.it
[†]*Department of Computer Science, University of Verona, Italy,* name.surname@univr.it

***Abstract*—In the context of smart environments, crafting remote monitoring systems that are efficient, cost-effective, user-friendly, and respectful of privacy is crucial for many scenarios. Recognizing and tracing individuals via markerless motion capture systems in multi-person settings poses challenges due to obstructions, varying light conditions, and intricate interactions among subjects. In contrast, methods based on data gathered by Inertial Measurement Units (IMUs) located in wearables grapple with other issues, including the precision of the sensors and their optimal placement on the body. We claim that more accurate results can be achieved by mixing Human Pose Estimation (HPE) techniques with information collected by wearables. To do that, we introduce a real-time platform that fuses HPE and IMU data to track and identify people. It exploits a matching model that consists of two synergistic components: the first employs a geometric approach, correlating orientation, acceleration, and velocity readings from the input sources. The second utilizes a Convolutional Neural Network (CNN) to yield a correlation coefficient for each HPE and IMU data pair. The proposed platform achieves promising results in identification and tracking, with an accuracy rate of 96.9%.**

***Index Terms*—Human tracking, data fusion, HPE, wearables, IMU**

I. Introduction

In the evolving landscape of the smart environments, the design of remote monitoring systems has become paramount, accommodating a broad spectrum of applications, including industry, smart homes, smart cities, and healthcare. For instance, in industrial contexts, these systems can be exploited to verify workers' adherence to safety guidelines, prevent personnel entry into hazardous zones, and ensure that only unauthorized individuals access designated sectors [1]. Utilizing video, wearable, and radio-based technologies, it is possible to deduce the posture of individuals within a designated space, identify undertaken activities (e.g., walking, sitting, standing), and estimate the environmental status (e.g., the presence and total of occupants).

To achieve these, Human Pose Estimation (HPE) systems are widely used to extrapolate the position and orientation of a person within an environment [2]. Typically, HPE marker-based systems use markers attached to the subjects' bodies, but they tend to be time-consuming and invasive. In contrast, HPE markerless motion capture systems offer a more convenient and non-invasive solution, using computer vision techniques to estimate body movements directly from video data. Nevertheless, accurately identifying individual bodies in a multi-person scenario remains a crucial challenge in such markerless systems [3], [4]. On the other hand, wearables equipped with Inertial Measurement Units (IMU) comprise accelerometers, gyroscopes, and sometimes magnetometers, which can also be used to measure human motion, orientation, and position.

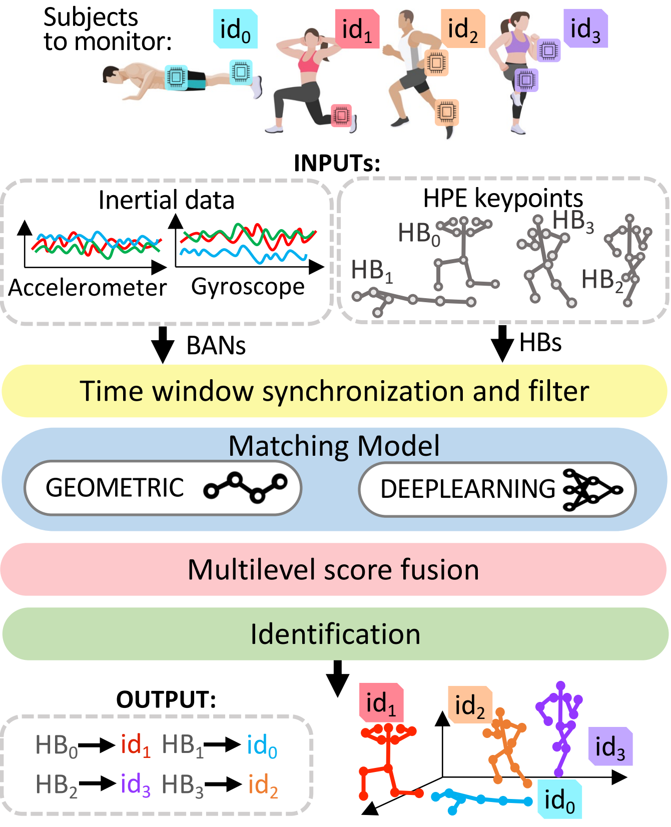


Fig. 1. Methodology overview: synchronized IMU and human bodies data fusion and matching models for multi-person pose estimation, identification, and tracking.

State-of-the-art works use wearable IMU sensors for 3D human motion analysis, developing portable and real-time pose estimation systems robust to sensor view occlusions and light [5], [6]. Besides, wearable IMUs can also be used to track a person's movement inside an environment by integrating acceleration and angular velocity measurements over time. However, accumulative errors over more extended periods can lead to significant inaccuracies, thus making the location and tracking of a person unreliable.

Recent works focus on fusing 3D human pose data extracted from IMU sensors and cameras to achieve improved accuracy, reduced temporal drift, and full-body tracking even in parts occluded from the camera view, while maintaining real-time

This work has been partially supported by the "PREPARE" project (n. F/310130/05/X56 - CUP: B39J23001730005) - D.M. MiSE 31/12/2021.

979-8-3503-4859-0/DATE24/© 2024 EDAA

and robustness requirements [7]–[10]. These studies reveal that data fusion from marker-less camera systems and IMU sensors results in enhanced performance compared to utilizing either IMUs or cameras separately. However, none of these studies focuses on identification and tracking within a multi-person context, thus sidestepping the challenge of correctly pairing HPE skeletons with IMU data.

A. Problem statement

HPE systems alone undertake neither tracking nor identification tasks. IMUs lead to more inaccuracies in motion analysis but provide an intrinsic identification. HPE and IMUs allow to leverate the motion capture system for identification and tracking individuals in a multi-person environment. The collection of HPE keypoints, representing a person's body joints, is tagged with a distinct identifier. Since this identifier bears no relation to the individual's identity and lacks consistency over time, we integrate data from HPE with data from IMU sensors. In this way, his/her identification and tracking becomes more reliable.

B. Paper contribution

The proposed approach, illustrated in Figure 1, takes a set of IMU sensors integrated into a Body Area Network (BAN) and HPE human body data (HB) as input. This data is then synchronized and concurrently cleaned through filtering and denoising. Subsequently, the data goes through a matching model, which embeds a geometric and a deep-learning. All the computed correlations pass through a multilevel score fusion, which leverages a voting technique to provide the best correlation score between HPE and IMU data. The correlation is then used to identify and track the target subject.

The main contributions of this paper are:

1) The design of synchronization and filtering solutions for merging BAN and HPE data.
2) The definition of a matching model consisting of both a geometric component and a deep learning component, capable of generating a correlation score for pairs of HPE and BAN data.
3) A voting system that intelligently performs score fusion by matching models on motion analysis to finally identify and track the target subject.

The structure of this paper is the following. Section II presents the background technologies. Section III exposes the proposed approach, including the geometric model, the Convolutional Neural Network (CNN), and the preprocessing pipeline tailored for human identification and tracking using HPE and IMU data. Section IV showcases the experimental results and Section V provides the concluding remarks on the research findings.

II. Background

A. Camera-based Motion Capture

HPE systems estimate a series of 3D coordinates indicating the position of the human joints over time. The main advantage of marker-based motion capture is the outstanding precision and adaptability in recording intricate movements. However, it has several drawbacks, such as, the time consuming initial setup, the sensitivity to occlusions, and the limited versatility.

Instead of relying on external markers, markerless motion capture systems reduce the time for setup and can be used in every environment [2]. Nevertheless, they grapple with challenges, especially when identifying individuals in multi-person settings. Intricate interactions, occlusions, alterations in the pose, fluctuating lighting conditions, variances in appearance, and multifaceted interactions among multiple participants can impede reliable identification, tracking, and pose estimation [11]. Contemporary methods for multi-person body identification frequently hinge on visual indicators like facial recognition, clothing color, or morphological details. However, these cues can fail in intricate settings or when individuals are only partially visible [12].

B. Inertial Measurement Unit and Wireless Body Area Network

IMU sensors offer a direct means to record body motion. These devices integrate accelerometers, gyroscopes, and magnetometers to measure linear and angular accelerations, rotational velocities, and magnetic fields. Typically attached to various body regions like the wrists, ankles, or torso, IMU sensors have the distinct advantage of directly capturing motion data, free from the dependence on external tracking mechanisms.

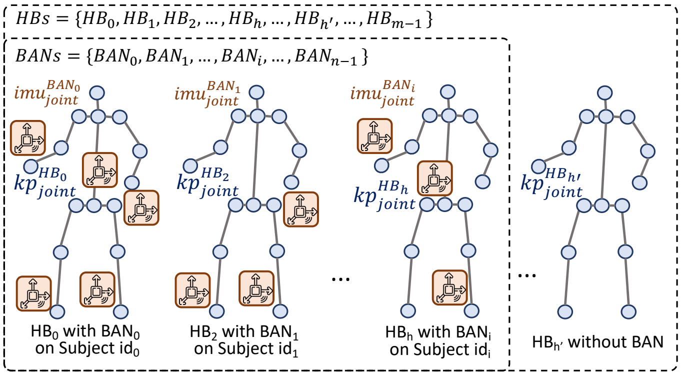
IMU sensors, strategically positioned on different body parts, are generally integrated into a BAN where they collaboratively gather data related to human movements. This architecture has become increasingly popular and available due to integrating IMU sensors into ubiquitous devices such as smartphones and smartwatches. BANs achieve outstanding results in realms like fitness monitoring, vital sign tracking, disease diagnosis, progression analysis, and, more broadly, in recognizing human activities and behaviors [13], [14] and also hold potential for estimating human pose.

III. Methodology

A. Problem Formulation

Formally, a group of individuals, denoted as $\mathbb{S}$ and consisting of n subjects, is characterized by unique names or identifiers: $\mathbb{S} = \{id_0, ..., id_i, ..., id_{n-1}\}$. Each subject's body is a collection of key points, represented as 3D coordinates within a global reference system, associated with various joints. We define a human body as follows: $HB_h = \{kp_j^{HB_h} \mid joint \in Joints\}$, where $Joints$ represents the set of joint identifiers monitored by the specific HPE system, such as head, left wrist (lwrist), right wrist (rwrist), and so on. The notation $kp_j^{HB_h}$ represents the coordinates (x, y, z) of a particular joint linked to the human body HB_h. For the sake of simplicity, we assume that the HPE system is tracking a total of m human bodies denoted as $HBs_t = \{HB_h \mid h \text{ in } [0, m)\}$ at each time frame t.

The BAN system employs wearable IMU sensors to record inertial data while maintaining a consistent identifier for each sensor. During the initial system setup, each BAN is linked to a specific subject, denoted as id_i, and registers a group of IMU sensors, each associated with a particular body joint. Formally, we define a BAN as $BAN_i = \{imu_j^{BAN_i} \mid joint \in Sub[Joints] \subseteq Joints\}$, with the condition that BAN_i corresponds to BAN_{id_i}. In this setup, every individual in the set $\mathbb{S}$ wears a BAN, and we represent the overall collection of BANs as $BANs = \{BAN_i \mid i \text{ in } [0, n)\}$. Figure 2 graphically represents these formulations.



HBs = {HB_0, HB_1, HB_2, ..., HB_h, ..., HB_h', ..., HB_m−1 }
BANs = {BAN_0, BAN_1, ..., BAN_i, ..., BAN_n−1 }
imu_joint^BAN_0
imu_joint^BAN_1
imu_joint^BAN_i
kp_joint^HB_0
kp_joint^HB_2
kp_joint^HB_h
kp_joint^HB_h'
...
...
HB_0 with BAN_0 on Subject id_0
HB_2 with BAN_1 on Subject id_1
HB_h with BAN_i on Subject id_i
HB_h' without BAN

Fig. 2. Formalizing the problem statement considering the BAN, comprised of IMU sensors affixed to specific joints and linked to individual subjects, and the HB, consisting of key points located on joints.

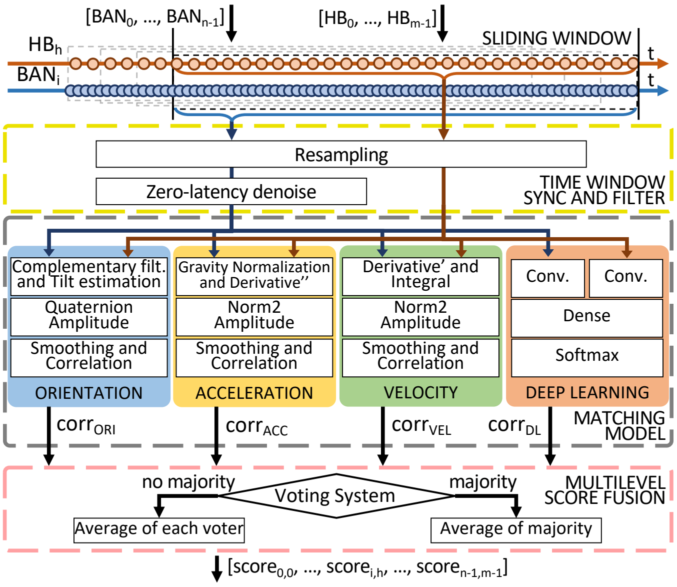


Fig. 3. The processing and fusion core pipeline for BANs and HBs signals consists of three key stages: *time window synchronization and filter*, collecting data from sensory sources and ensuring temporally alignment for comparison; *matching model*, correlation metrics using orientation, acceleration, velocity, and deep-learning techniques applied to raw data; *multi-level fusion*, utilizing a voting system to generate the most optimal matching score.

As the human bodies within the HPE system lack a consistent identifier, we leverage IMU sensors, which inherently possess unique identifiers, to establish associations between the sensors' BANs and human bodies. The goal is to maximize the following objective function: $id\_map(BANs, HBs) = \{(BAN_i, HB_h) \mid BAN_i \sim HB_h\}$, with the symbol $\sim$ denoting a similarity between the two representations of human motion.

The method operates without making any assumptions regarding the number of subjects to be identified (with a BAN) and the subjects currently present in the scene. We define *"intruder"* as a human body within the scene for which we have no interest of monitoring or identification (represented as $HB_{h'}$ in Figure 2).

B. Time window synchronization and filter

The initial real-time identification and tracking stage involves data acquisition, as depicted in Figure 3. Data is sourced from diverse sensors characterized by unique features, including varying frequencies, transmission methods, and signal noise profiles. We collect this information within a packet queue to facilitate data integration, enabling signal processing within defined time windows via windowing algorithms. This phase takes inertial data from each IMU sensor and human body traces generated by a 3D HPE system.

Given the collected signals' diverse origins, transmission frequency variations subsist. Specifically, wireless IMU sensors typically operate at significantly higher frequencies (e.g., 100-300 Hz), whereas camera sensors operate at lower frequencies (e.g., 15-60 Hz). IMU sensors are more susceptible to packet loss due to wireless transmission, while camera sensors exhibit greater robustness. Consequently, the proposed platform implements a time synchronization process, which involves (a) verifying the completeness of signal queues for comparison, (b) confirming that collected packets enter the predefined time window interval, and (c) downsampling the signals to the lowest frequency. This ensures that resulting signals in HBs and $BANs$ share a uniform frequency without any interpolation resulting from up-sampling.

Since IMU and HPE signals exhibit distinct types of noise, we introduce diverse filtering approaches. We apply a zero-latency denoising technique, specifically a Kalman filter, incorporating rigid body dynamics and kinematic constraints to the HPE signal [15]. We also also apply a low-pass frequency filter to the IMU accelerometer signals, retaining only the human motion signal (3 Hz). The zero-latency filter relies solely on the previous sample, while the frequency filter considers the entire time window.

C. Matching model

Each matching model takes a pair $BAN_i \in BANs$ and $HB_h \in HBs$ as input. For each $i \in [0, n)$ and $h \in [0, m)$, we iterate over each $kp_j^{HB_h} \in HB_h$ and each $imu_j^{BAN_i} \in BAN_i$, and computes a set of correlation scores $corr_{\rho,j,h}^{i}$ relative to BAN_i, calculated with respect to HB_h and joint j, using the ρ model presented below.

1) Geometrical model: The geometrical model processes the 3D positions of key points $kp_j^{HB_h}$ of a human body and the acceleration, angular velocity signals of a $imu_j^{BAN_i}$. The output is the orientation, acceleration, and velocity correlation scores at a specific joint (j).

We describe the orientation of a body segment using Euler angles, denoted as (ϕ, θ, ψ), where ψ_j and $\theta_{j'}$ represent the segment's tilt based on the 3D positions of two human body joints (j and j'). The IMU system employs a *complementary filter* [16] to determine orientation relative to the world system, utilizing angular velocity and accelerometer data. We discard the magnetometer signal due to interference from the indoor environment.

We calculate the acceleration of a human body joint as the second derivative of the 3D position over time. In the IMU system, this acceleration includes the gravity vector, estimated and subtracted to obtain normalized IMU acceleration. We calculate the gravity vector ($g(t)$) by applying the rotation matrix ($R(t)$) from the complementary filter to the nominal gravity vector, generally represented as $[0, 0, 9.81]$: $g(t) = R(t)^{-1} g_{nom}^{T}$. We calculate the velocity of a human body joint from the gravity-normalized acceleration data through a first derivative and the velocity of an IMU joint through integration.

Both the HPE and IMU orientations are converted in *quaternions*, defined as a quadruple (q_x, q_y, q_z, q_w), which are transformed into modulus values through the quaternion magnitude: $magnitude_q = 2 \arctan \left(\frac{\sqrt{q_x^2 + q_y^2 + q_z^2}}{q_w} \right)$. Similarly, we convert acceleration and velocity into modulus by computing the norm 2 of their corresponding vectors.

We apply a moving-average smoothing to the modulus signals to enhance the signal quality, and then we compute the Pearson correlation between the corresponding signals from the two systems.

2) Deep-learning model: The proposed matching model for comparing HPE and IMU signals utilizes a CNN, which adaptively learns spatial and temporal patterns from input data. This characteristic makes CNNs well-suited for tasks involving the analysis of signal patterns. By transforming HPE and IMU signals into a high-dimensional feature space using CNN, we calculate their similarity in this new space.

The CNN architecture consists of the following components. HPE and IMU input data goes through two separate branches, featuring the same architecture composed of a convolutional layer comprising 128 filters, each with a length of 3, with a Rectified Linear Unit (ReLU) serving as the activation function. This layer employs a convolutional operation on human key points and IMU input data, preserving the original signal length by zero-padding as needed. The output of the convolutional layer undergoes a max pooling layer, reducing the dimensionality and computational complexity while retaining essential features. This combination of convolutional and pooling is repeated twice, except in the last repetition, where the global average pooling replaces the max pooling. This convolution-pooling combination repeats twice, with global average pooling replacing max pooling in the final repetition. Global average pooling averages the feature map's values across its entire length, preventing overfitting and resulting in a more compact model. After the global average pooling layer, the two branches are concatenated, and then the model features four fully connected layers with 256, 256, 512, and 128 units, respectively. To mitigate overfitting, we apply a dropout of 20% during training on each fully connected layer. The output layer is a binary softmax, whose confidence value is used as the correlation score between the inputs BAN_i and HB_h.

D. Multilevel score fusion

The score fusion involves aggregating multilevel correlation scores generated by individual matching models for various joints within the BANs and HBs to produce a final composite score. This process employs a voting mechanism to determine the combinations of models and joints that contribute most significantly to achieving accurate matches.

The concept of merging correlation levels from different models and joints originates in motion analysis. Firstly, including additional IMUs in the BAN has benefits, as they augment the available motion data sources used to assess the similarity with key points, thereby enhancing the matching process. Furthermore, each model focuses on distinct characteristics of the actions and joints under consideration. Relying solely on orientation data is insufficient. Incorporating acceleration and velocity data is essential for capturing rapid and precise movements, enabling the detection of similar actions. Also, including acceleration, a double derivative of positions, may introduce noise into the signal. Hence, velocity data is often more robust and dependable, albeit at the cost of inertial signal integral. Given the complexity of determining which information exerts the most significant influence on the matching between the traces of the two data sources, we leverage on the proposed deep-learning model that, through a dedicated training phase, learns to identify the most crucial features for establishing similarity effectively.

Given $corr^i_{\rho,j,h}$, representing the correlations of model ρ between joint j of the human body HB_h and the body area network BAN_i, the multilevel score fusion process constructs a score matrix denoted as $Score[i, h]$. This matrix encapsulates the final correlation score between the subject with ID id_i wearing IMUs and the human body HB_h. The procedure, outlined in Algorithm 1, delineates the extrapolation of correlations from the matching models and joints (rows 1 to 9), followed by the execution of a voting system to determine the optimal joint model pairs for integration into the final score (rows 10 to 20).

Algorithm 1 Score matrix extrapolation algorithm from BAN and HB traces through multilevel voting score fusion

```
Input: $BANs$ (size $n$), $HBs$ (size $m$), $Models = \{ori, acc, vel, dl\}$
Output: Score ($n \times m$)
 1: for $i = 0$ to $n - 1$ do
 2:     for $\rho \in Models$ do
 3:         for $imu_j^{BAN_i} \in BAN_i$ do
 4:             for $h = 0$ to $m - 1$ do
 5:                 $corr^i_{\rho,j,h} \leftarrow \rho(imu_j^{BAN_i}, HB_h[j])$
 6:             end for
 7:             $match^i_{\rho,j} \leftarrow argmax_h(corr^i_{\rho,j,h})$
 8:         end for
 9:     end for
10:     $best\_match^i \leftarrow max_{\rho,j}(match^i_{\rho,j})$
11:     $best\_voters^i \leftarrow \{(\rho, j) \mid match^i_{\rho,j} = best\_match^i\}$
12:     for $h = 0$ to $m - 1$ do
13:         if $|\{m \in bincount_{\rho,j}(match^i_{\rho,j}) \mid m = best\_match^i\}| > 1$
            then
14:             $Score[i, h] = mean_{p,j}(\{corr^i_{\rho,j,h}\})$
15:         else
16:             $Score[i, h] = mean_{p,j}(\{corr^i_{\rho,j,h} \mid (p, j) \in best\_voters^i\})$
17:         end if
18:     end for
19: end for
20: return Score
```

The voting system determines the optimal human body HB_h that best associates with BAN_i by keeping a record of the *voters*, represented by the pairs (ρ, j), which contribute to the majority in the matching process (rows 10 and 11). We examine whether a majority consensus has been reached (row 13). If a majority is achieved, we calculate the final scores by considering only the average of the pairs (p, j) that form the majority (row 16). Otherwise, we calculate the average score for all pairs (row 14).

E. Identification

Identification is an assignment problem solved with a combinatorial optimization algorithm on a weighted bipartite graph. The weighted bipartite graph is defined by the set of subjects $\mathbb{S}$

and the set of human bodies HBs. The weights are the Score matrix of size $n \times m$, where each cell $[i, h]$ is the matching level between subject id_i and human body HB_h.

The algorithm relies on the linear sum assignment, a maximum weight-matching optimization. Let X be an $n \times m$ binary assignment matrix, where $X[i][h] = 1$ if subject i is assigned to the human body h and $X[i][h] = 0$ otherwise. The goal is to find the assignment matrix X that maximizes the total cost: $\sum_{i=0}^{n-1} \sum_{h=0}^{m-1} Score[i][h] \cdot X[i][h]$.

The optimization algorithm has the following constraints: (a) $\sum_{h=0}^{m-1} X[i][h] = 1$, for $i = 1, \ldots, n-1$, i.e. each subject is assigned to precisely one human body; (b) $\sum_{i=0}^{n-1} X[i][h] = 1$, for $h = 1, \ldots, n-1$, i.e. each human body is assigned to at most one subject; (c) $X[i][h] \in \{0, 1\}$, for $i = [0, n), h = [0, m)$, i.e. $X[i][h]$ is a binary variable.

Identification is finally done by iterating the X matrix and finding the indices that contain 1. Specifically, we construct the map of subjects and human bodies, formally defined as $id\_map = \{(id_i \in \mathbb{S}, HB_h \in HBs) | X[i][h] = 1 \text{ for } i = [0, n) \wedge h = [0, m)\}$.

IV. Experimental Results

A. Datasets

We adopted the TotalCapture dataset [17], which contains a comprehensive collection of synchronized multi-view video, IMUs, and accurate 3D skeletal joint ground-truth data. The subjects are recorded in a controlled environment with eight calibrated HD RGB cameras and 13 high-precision Xsens IMUs. The configurations involve five subjects performing four activities, i.e., acting, freestyle, range-of-motion, and walking, and each repeated three times.

Despite the optimal conditions of the publicly available dataset, we took our evaluation a step further by testing our approach on a dataset that we constructed under real-world conditions, using less precise sensors. In this real-world scenario, we recorded data using StereoLabs ZED2 stereo cameras and Thingy Nordic IMU sensors. Our dataset captures two subjects engaging in various activities such as, walking, sitting, discussing, eating, and waiting.

B. Experimental setup

We adopted TRTPose [18] as 2D HPE system. It is based on the DenseNet architecture, which has been optimized through pruning and quantization. We performed the Direct Linear Transform (DLT) algorithm from two adjacent views for TotalCapture and disparity map triangulation to reconstruct the three-dimensional pose information for our use-case study.

We assessed several deep-learning models, including a multilayer perceptron, an LSTM-based recurrent neural network, and a CNN. Based on our evaluations and experiments, the CNN outperformed the other models, demonstrating superior performance. Although not as complex or accurate as the state-of-the-art models, the neural network architectures were deliberately designed to be lightweight. This choice ensured the system can operate in real-time conditions without compromising efficiency.

In our study, we individually assessed the effectiveness of the geometric approach (Geo) and the deep-learning (DL) approach. We introduced a hybrid approach combining geometric and deep-learning methods. Additionally, we implemented straightforward fusion techniques for comparison against our advanced multilevel score fusion approach.

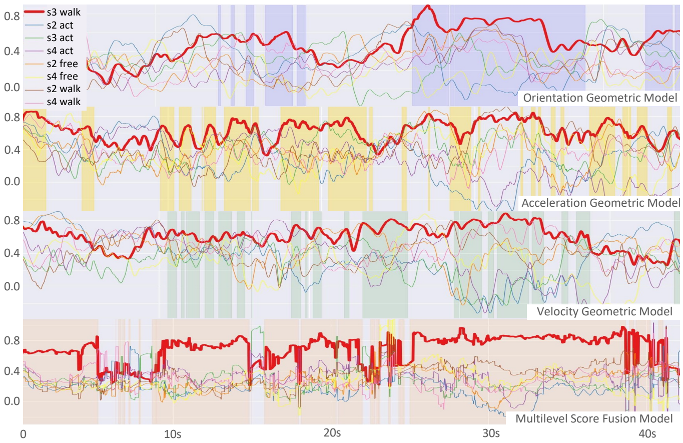


Fig. 4. Correlation level over time of 8 people doing different activities. The $s3walk$ red trace represents the target subject. The highlighted areas indicated when the red trace has the highest correlation with other subjects. The multilevel score fusion model performs better than the other models.

The fusion methods include: The mean fusion, which computes the average of all correlations among the voters; The max fusion, which retains the maximum; The threshold fusion, which selection is based on predefined thresholds (e.g., we consider orientation if it exceeds 80% and speed if it is above 70%, otherwise, we consider acceleration).

Furthermore, the voting system within our multilevel score fusion framework operates in two modes: The first involves placing only the models into the voting process, where the contributions made by various joints are averaged. In the second mode, both models and joint combinations participate in the voting process. The joint fusion involves a BAN composed of the wrists, ankles, and hip joints.

C. Result analysis

Figure 4 shows that merging the contributions of various models is critical to achieve high levels of accuracy The graphs show the level of correlation over time with eight people in the scene. The red line represents the target subject to be identified and tracked. The first three graphs show the contribution of orientation, acceleration, and speed, respectively, while the last one shows the proposed approach merged with CNN. The Multilevel Score Fusion Model obtains the more accurate correlation. The tables in this section show a quantitative measure of the performance of the presented models. For each model, we show the percentage of correctly matched frames and evaluate the identification of a single subject while multiple people interact in the scene (1 to 7 additional people).

The TotalCapture dataset provides a split in the trainset and testset. We evaluated the last 30% of the trainset (Table I) and the testset (Table II and Table III). The test set differs from the trainset in subjects, action, and repetition number. Concerning our *real-world* dataset, we considered the last 30% of the whole dataset (Table IV). The geometric model with CNN and the voting system provide the best results. To solve the identification problem, it is essential to merge information

TABLE I
ACCURACY OF TOTALCAPTURE TRAINING EVALUATION DATASET BY COMPARING SUBJECTS (S) AND ACTIONS (A).

		Geo. (%)			Geo Fusion (%)			Geo Multilevel Fusion (%)		DL (%)	Geo + DL Fusion (%)		Geo + DL Multilevel Fusion (%)	
Subj Action		acc.	vel.	ori.	mean fusion	thr. fusion	max fusion	model vote fusion	m+joint vote fusion	CNN	mean fusion	max fusion	model vote fusion	m+joint vote fusion
S	s2	81.9	80.4	85.3	91.3	80.4	89.3	92.0	93.6	80.0	93.0	85.4	**98.0**	96.8
	s3	99.8	94.8	77.3	**99.9**	94.8	99.6	90.0	92.9	61.6	88.8	85.1	84.4	94.5
	s1	87.2	86.4	94.7	95.8	86.4	86.7	96.3	96.5	64.0	98.1	90.6	**98.9**	98.6
A	act	96.7	79.7	93.8	99.0	79.7	99.1	92.1	93.0	95.3	**99.2**	97.4	94.7	98.9
	free	83.3	91.7	87.5	95.6	91.7	89.4	95.7	94.7	51.1	87.4	72.6	**97.1**	95.6
	rom	99.8	94.1	80.4	**99.9**	94.1	99.3	94.7	97.1	31.2	89.1	79.2	89.1	93.5
	walk	73.5	79.5	85.5	86.0	79.5	75.3	89.9	93.4	99.7	**99.8**	**99.8**	99.0	99.6
Avg.		88.4	86.2	86.8	95.1	86.2	90.9	93.1	94.5	69.4	93.9	87.3	95.0	**96.9**

TABLE II
ACCURACY OF TOTALCAPTURE TESTING DATASET BY COMPARING SUBJECTS (S) AND ACTIONS (A).

		Geo. (%)			Geo Fusion (%)			Geo Multilevel Fusion (%)		DL (%)	Geo + DL Fusion (%)		Geo + DL Multilevel Fusion (%)	
Subj Action		acc.	vel.	ori.	mean fusion	thr. fusion	max fusion	model vote fusion	m+joint vote fusion	CNN	mean fusion	max fusion	model vote fusion	m+joint vote fusion
S	s2	94.8	87.8	77.6	95.2	92.8	93.9	94.5	93.3	75.8	93.7	81.1	94.6	**95.4**
	s3	90.7	78.0	80.3	93.0	82.3	86.4	90.1	90.0	72.2	94.5	77.7	93.7	**95.9**
	s4	86.4	81.9	65.9	89.3	87.7	90.4	88.5	80.7	80.3	87.4	86.1	**91.2**	83.2
A	act	92.7	86.4	81.2	**95.9**	92.2	94.4	93.7	91.5	78.4	94.6	86.9	95.4	95.3
	free	85.7	79.2	68.1	86.2	82.4	88.1	87.6	84.2	82.2	86.2	82.8	**89.1**	85.8
	walk	91.9	82.5	70.5	93.1	88.2	88.7	90.9	86.3	71.1	92.0	77.0	**93.6**	90.1
Avg.		90.6	83.2	73.9	92.4	88.2	90.7	91.2	87.7	76.6	91.5	82.1	**93.1**	91.0

TABLE III
ACCURACY OF TOTALCAPTURE TESTING SET BY VARYING THE NUMBER OF SUBJECTS PRESENT IN THE SCENE AND IDENTIFYING A SPECIFIC INDIVIDUAL

	Geo. (%)			Geo Fusion (%)			Geo Multilevel Fusion (%)		DL (%)	Geo + DL Fusion (%)		Geo + DL Multilevel Fusion (%)	
#S	acc.	vel.	ori.	mean fusion	thr. fusion	max fusion	model vote fusion	m+joint vote fusion	CNN	mean fusion	max fusion	model vote fusion	m+joint vote fusion
2	90.6	83.2	73.9	92.4	88.2	90.7	91.2	87.7	76.6	91.5	82.1	**93.1**	91.0
3	85.7	78.6	60.3	89.5	83.9	86.2	89.2	84.9	60.7	86.3	73.3	**90.4**	89.7
4	80.7	69.1	52.1	84.8	76.0	79.0	84.5	82.7	54.6	82.7	61.8	84.8	**88.1**
5	75.0	59.5	44.2	79.6	68.5	71.2	77.4	78.5	50.8	77.8	54.5	79.6	**82.4**
8	66.6	50.7	36.2	72.7	61.3	64.0	71.0	70.4	46.6	68.9	49.2	73.7	**74.4**

TABLE IV
ACCURACY OF REAL-CASE STUDY DATASET BY VARYING THE NUMBER OF SUBJECTS PRESENT IN THE SCENE AND IDENTIFYING A SPECIFIC INDIVIDUAL

	Geo. (%)			Geo Fusion (%)			Geo Multilevel Fusion (%)		DL (%)	Geo + DL Fusion (%)		Geo + DL Multilevel Fusion (%)	
#S	acc.	vel.	ori.	mean fusion	thr. fusion	max fusion	model vote fusion	m+joint vote fusion	CNN	mean fusion	max fusion	model vote fusion	m+joint vote fusion
2	69.6	76.1	77.2	78.7	77.5	77.7	86.5	85.6	89.8	97.1	94.7	96.2	**98.8**
3	55.2	64.8	72.5	67.7	66.6	67.2	80.5	76.5	81.0	96.1	93.8	95.5	**99.1**
4	48.9	59.3	64.8	63.0	61.8	63.9	76.7	74.2	81.0	93.7	88.7	93.7	**98.7**
5	45.1	56.3	60.4	60.5	57.2	59.9	74.3	71.4	80.9	92.8	86.7	90.8	**97.7**
8	40.8	52.5	52.9	56.9	52.9	56.5	66.1	65.5	75.6	85.4	78.0	78.5	**93.9**

from different matching models, especially when the sensors and systems are not accurate. Under these conditions, each singular matching model fails to produce acceptable results, while the multilevel score fusion achieves accuracy levels often higher than 90%. As the number of IMU sensors in the BAN increases, the correct matching and identification rates are higher. The tables indicate that a straightforward mean fusion approach achieves better results than the other individual models. On the other hand, when the individual models struggle with identification, the voting system yields better results.

V. CONCLUSION

The paper presented a platform that combines HPE and BAN data for people identification and tracking. The platform relies on a matching model the consists of a geometrical component and a deep learning component. It applies a voting system to perform score fusion among the geometrical and deep learning components to identify and track the terget subject. The paper presented the results obtained by applying the proposed platform to a standard dataset as well as a dataset generated with low-cost and less accurate sensors to reproduce real-world application conditions.

REFERENCES

[1] M. Boldo, N. Bombieri, M. De Marchi, L. Geretti, S. Germiniani, and G. Pravadelli, “Risk assessment and prediction in human-robot interaction through assertion mining and pose estimation,” in *2022 IEEE 23rd Latin American Test Symposium (LATS)*. IEEE, 2022, pp. 1–5.

[2] E. Martini, M. Boldo, S. Aldegheri, M. De Marchi, N. Valè, M. Filippetti, N. Smania, M. Bertucco, A. Picelli, and N. Bombieri, “Real-time human pose estimation at the edge for gait analysis at a distance,” in *2022 18th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, 2022, pp. 45–48.

[3] B. Scott, M. Seyres, F. Philp, E. K. Chadwick, and D. Blana, “Healthcare applications of single camera markerless motion capture: a scoping review,” *PeerJ*, vol. 10, p. e13517, May 2022.

[4] W. W. T. Lam, Y. M. Tang, and K. N. K. Fong, “A systematic review of the applications of markerless motion capture (mmc) technology for clinical measurement in rehabilitation,” *Journal of NeuroEngineering and Rehabilitation*, vol. 20, no. 1, p. 57, May 2023. [Online]. Available: https://doi.org/10.1186/s12984-023-01186-9

[5] P. Slade, A. Habib, J. L. Hicks, and S. L. Delp, “An open-source and wearable system for measuring 3d human motion in real-time,” *bioRxiv*, 2021. [Online]. Available: https://www.biorxiv.org/content/early/2021/03/24/2021.03.24.436725

[6] M. Kim and S. Lee, “Fusion poser: 3d human pose estimation using sparse imus and head trackers in real time,” *Sensors*, vol. 22, no. 13, 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/13/4846

[7] M. Yamamoto, K. Shimatani, Y. Ishige, and H. Takemura, “Verification of gait analysis method fusing camera-based pose estimation and an imu sensor in various gait conditions,” *Scientific Reports*, vol. 12, no. 1, p. 17719, Oct 2022. [Online]. Available: https://doi.org/10.1038/s41598-022-22246-5

[8] T. Tan, D. Wang, P. B. Shull, and E. Halilaj, “Imu and smartphone camera fusion for knee adduction and knee flexion moment estimation during walking,” *IEEE Transactions on Industrial Informatics*, vol. 19, no. 2, pp. 1445–1455, Feb 2023.

[9] M. Palermo, S. M. Cerqueira, J. André, A. Pereira, and C. P. Santos, “From raw measurements to human pose - a dataset with low-cost and high-end inertial-magnetic sensor data,” *Scientific Data*, vol. 9, no. 1, p. 591, Sep 2022. [Online]. Available: https://doi.org/10.1038/s41597-022-01690-y

[10] M. Boldo, N. Bombieri, S. Centomo, M. De Marchi, F. Demrozi, G. Pravadelli, D. Quaglia, and C. Turetta, “Integrating wearable and camera based monitoring in the digital twin for safety assessment in the industry 4.0 era,” in *International Symposium on Leveraging Applications of Formal Methods*. Springer, 2022, pp. 184–194.

[11] A. Zahra, N. Perwaiz, M. Shahzad, and M. M. Fraz, “Person re-identification: A retrospective on domain specific open challenges and future trends,” *Pattern Recognition*, vol. 142, p. 109669, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320323003709

[12] S. Alam and M. Yeasin, “Person identification with visual summary for a safe access to a smart home,” 2019.

[13] I. López-Nava and A. Muñoz-Meléndez, “Wearable inertial sensors for human motion analysis: A review,” *IEEE Sensors Journal*, vol. PP, pp. 1–1, 09 2016.

[14] F. Demrozi, C. Turetta, P. H. Kindt, F. Chiarani, R. Bacchin, N. Valè, F. Pascucci, P. Cesari, N. Smania, S. Tamburin *et al.*, “A low-cost wireless body area network for human activity recognition in healthy life and medical applications,” *IEEE Transactions on Emerging Topics in Computing*, 2023.

[15] E. Martini, A. Calanca, and N. Bombieri, “Denoising and Completion Filters for Human Motion Software: a Survey with Code,” 5 2023.

[16] P. Narkhede, S. Poddar, R. Walambe, G. Ghinea, and K. Kotecha, “Cascaded complementary filter architecture for sensor fusion in attitude estimation,” *Sensors*, vol. 21, no. 6, 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/6/1937

[17] M. Trumble, A. Gilbert, C. Malleson, A. Hilton, and J. Collomosse, “Total capture: 3d human pose estimation fusing video and inertial sensors,” in *2017 British Machine Vision Conference (BMVC)*, 2017.

[18] NVIDIA AI IoT, “Tensor RT Pose Estimation,” 2020, https://github.com/NVIDIA-AI-IOT/trt_pose.