Human Activity Recognition using Inertial Sensors with Invariance to Sensor Orientation

Blanca Florentino-Liaño, Niamh O’Mahony, Antonio Artés-Rodríguez

Department of Signal and Communications Theory, Universidad Carlos III de Madrid, Avenida de la Universidad, 30, Leganés, 28911, Spain
Email: {blanca, niamh, antonio}@tsc.uc3m.es,

Abstract—This work deals with the task of human daily activity recognition using miniature inertial sensors. The proposed method reduces sensitivity to the position and orientation of the sensor on the body, which is inherent in traditional methods, by transforming the observed signals to a “virtual” sensor orientation. By means of this computationally low-cost transform, the inputs to the classification algorithm are made invariant to sensor orientation, despite the signals being recorded from arbitrary sensor placements. Classification results show that improved performance, in terms of both precision and recall, is achieved with the transformed signals, relative to classification using raw sensor signals, and the algorithm performs competitively compared to the state-of-the-art. Activity recognition using data from a sensor with completely unknown orientation is shown to perform very well over a long term recording in a real-life setting.

I. INTRODUCTION

Human activity recognition has recently become a popular topic of research interest due to the growth of applications based on context-aware monitoring, including home-based rehabilitation, independent living solutions for the elderly and ambulatory monitoring of patients with psychiatric or other disorders. Knowing the activity being carried out by the patient throughout their day-to-day life provides context-awareness for the physiological or other measurements that are being monitored, allowing a more accurate analysis of the measurements than in a stand-alone monitoring system.

The two main methods for human activity recognition are vision-based, e.g. [1], and inertial sensor-based, e.g. [2]. Vision-based systems suffer from limitations such as only being usable in a confined space, interfering with the privacy of the individual and producing an excessive amount of information that is costly to process. On the other hand, due to recent advances in sensor technologies, inertial sensor devices have become compact and portable enough to be unobtrusively attached to the human body. For this reason, wearable miniature inertial sensors, incorporating accelerometers and gyroscopes, have become the ideal platform for human movement monitoring [2], falls detection [3], medical diagnosis and treatment [4], and tele-rehabilitation [5].

Recently, the authors presented a novel algorithm for the classification of human activities based on a hierarchical dynamic model (HDM) with hidden Markov models (HMM) [6]. This method was shown to give competitive classification results, compared to state-of-the-art methods, whilst avoiding the computational bottleneck of traditional feature extraction methods, by basing the entire algorithm on the raw signals measured by the sensors. One drawback of this method, due, in part, to not extracting features, is that the raw sensor signals are sensitive to the placement of the sensor on the subject’s body, in terms of position and orientation. For applications in real life situations, control of the exact placement of the sensor is not feasible and adverse effects of variations due to body shape, clothing and other factors must be eliminated from classification algorithms.

This work proposes a novel transformation of sensor measurements, before classification, which renders the collected signals insensitive to the position and orientation of the sensor on the subject’s body. The proposed algorithm allows the sensor to be placed in any fixed location within a region approximately bounded by a belt at the waist and a trouser pocket, as illustrated in Fig. 1. The only assumption is that the sensor is fixed such that its movement during the day is limited to a few millimeters. All measurements are transformed to a ‘virtual’ sensor placement, defined at the approximate center of mass of the subject’s body and with a known orientation, with respect to the body in a standing position, as illustrated in Fig. 1.

![Fig. 1. Allowable Sensor Placement and Virtual Sensor Position](image-url)
using raw sensor signals. These results are based on a database available in [7]; as such, the authors have no information regarding the sensor position and orientation for each subject, other than the side of the hip on which it was placed, and this information is not made available to the algorithm.

The rest of this paper is laid out as follows: in Section II, a brief introduction to the background theory of inertial sensors, coordinate systems and the HDM is provided. Section III describes the proposed transform. The experimental procedure is outlined in Section IV and classification results are presented in Section V. These results are followed by a discussion in Section VI and finally conclusions and future work are outlined in Section VII.

II. BACKGROUND

A. Inertial Measurement Units

The inertial measurement units (IMUs) used in this work are each equipped with a triaxial accelerometer and triaxial gyroscope. The accelerometers measure, in m/s², the total inertial force acting on the sensor. This inertial force includes both linear accelerations in each of the three sensor axes and a gravitational force component in each axis. The gyroscopes measure the angular velocity of the sensor in rad/s.

The signals recorded by the sensors can be modeled as follows: the accelerometer measurement vector, \( \alpha_t \), at time, \( t \), is given by:

\[
\alpha_t = a_t + g_t + \mu_{a,t},
\]

where \( a_t \) is the linear acceleration component due to sensor motion, \( g_t \) is the component due to the Earth’s gravitational force and \( \mu_{a,t} \) is a noise term. Similarly, the gyroscope signal, \( \omega_t \), is given by:

\[
\omega_t = w_t + \mu_{w,t},
\]

where \( w_t \) is the angular rotation and \( \mu_{w,t} \) is a noise term. The measurement noise terms, \( \mu_{a,t} \) and \( \mu_{w,t} \), are assumed to be zero mean Gaussian random processes. Signals from both accelerometers and gyroscopes also contain bias components but, for the purposes of classification, these can be considered negligible.

B. Coordinate Systems

The signals recorded by the inertial sensors are measured in a three-dimensional coordinate system which is fixed to and moves with the sensor, both linearly and rotationally. This frame is referred to as the sensor frame (\( S \)) and is defined by an orthogonal set of unit vectors, \( \{ \vec{x}^S, \vec{y}^S, \vec{z}^S \} \). A fixed frame (\( F \)) can also be defined, in which the gravitational component of the force of acceleration is constant; for example, in the frame, \( \{ \vec{x}^F, \vec{y}^F, \vec{z}^F \} = \{ \text{North, West, Up} \} \), acceleration due to gravity is given by \( G_F = [0, 0, 9.81] \) m/s².

During epochs of little or no linear acceleration (i.e. \( a_t \approx 0 \) for \( t_1 \leq t \leq t_2 \)), comparing the mean of the measured acceleration vector in the sensor frame, \( \vec{a}^S = \text{mean}(\alpha_{t_1:t_2}) \), to the gravitational vector in the fixed frame, \( G_F \), allows the orientation of the sensor frame, relative to the fixed frame, to be partially resolved. The inclination of the sensor, given by the angles of roll, \( \theta_x \), (the angle between the \( y^S \)-axis and the \( x^F \)-\( y^F \) plane) and pitch, \( \theta_y \), (the angle between the \( x^S \)-axis and the \( x^F \)-\( y^F \) plane), can be estimated by means of ratios of the gravitational acceleration component in each sensor axis [8]:

\[
\theta_x = \arctan\left( \frac{\mu_y^S}{\mu_z^S} \right),
\]

\[
\theta_y = \arcsin\left( \frac{-\mu_z^S}{\sqrt{(\mu_y^S)^2 + (\mu_z^S)^2 + (\mu_z^S)^2}} \right).
\]

With just accelerometers and gyroscopes, there is not sufficient information to resolve the angle of yaw, \( \theta_z \), (the angle between the projection of the \( x^S \)-axis onto the \( x^F \)-\( y^F \) plane and the \( x^F \)-axis), since the gravitational component in both the \( x^F \)- and \( y^F \)-axes is zero, resulting in an infinite number of possible solutions in the range \( \{0, 2\pi\} \). Fortunately, for the purposes of activity recognition, it is irrelevant whether a subject is facing due North or in any other direction whilst carrying out a particular activity and it will be seen in Section III that a yaw estimate is not required for the proposed method.

The final coordinate system to be introduced is the body frame (\( B \)), which is fixed to and moves with the center of mass of the subject’s body (approximately located at the waist). This is the frame of the virtual sensor to which all of the sensor measurements will be transformed and is shown in Fig. 1. The directions of each axis, relative to the subject’s body in a standing position, can be described as: \( \{ \vec{x}^B, \vec{y}^B, \vec{z}^B \} = \{ \text{Forward, Left, Up} \} \). It should be remembered that as the subject changes position, these directions will change with respect to the fixed frame; for example, if the subject is lying down, the \( z^B \)-axis will no longer point upwards, but along the \( x^F \)-\( y^F \) plane. This is the key to the operation of the transform.

C. HDM with HMM

In previous work by the authors [6], a HDM with HMM was proposed for the task of activity recognition. In that paper, the raw signals of the sensor were directly used as the inputs of the activity recognition algorithm. In order to evaluate the effectiveness of the transformation proposed, the same model is used in this work, but in this case, transforming the input signals. This section briefly reviews the HDM with HMM, for details see [6].

The HDM with HMM constructs a model taking into account two levels of dynamics: inter-activity and intra-activity. The inter-activity dynamics refer to the temporal dependency among activities. This level of dynamics helps in the recognition task because the current activity depends on which activity the subject was doing in the previous time step. This is modeled by transition probabilities among activities.

On the other hand, the amplitude of the signals and how they evolve in time, both give valuable information for the recognition of the activity. This level of dynamics is referred
to as intra-activity dynamics. This is modeled by constructing a HMM for each of the activities. Various different dynamic models are proposed, each with a different topology. For example, for stationary activities like standing, sitting and lying, a left-right model with three states is defined. The first and the last states are transient states and the state in the middle models the permanent state of, for example, being seated. This can be seen within the activity “LYING” in Fig. 2.

The final result of the HDM is a single HMM built up of “sub”-HMMs, one for each activity, which are interconnected by means of their transient states according to the transition probabilities defined by the inter-activity level. Fig. 2 shows an example of these interconnections.

Fig. 2. Example of HDM hierarchy with “sub”-HMMs.

### III. Transform

Knowledge of the orientation of a body segment can be very useful for activity recognition, especially for distinguishing between different “low-motion” activities, such as standing still, sitting and lying down. For all of these activities, the linear acceleration and angular rotation are close to zero throughout the activity and, so, the classification algorithm must rely on only the acceleration due to gravity to recognize the activities by determining the orientation or pose of the subject’s body. The measurement of the gravitational acceleration component, $g^S$, directly in the sensor frame, does not give any information about the orientation of the subject’s body, relative to the fixed frame, because it depends on the initial placement of the sensor. This can be observed in Fig. 3a and 4a, which show accelerometer observations recorded during two similar sequences of activities, carried out by the same subject, but with the sensor at different positions and orientations, namely, attached to a belt at the left hip and in the right trouser pocket. The values of the acceleration in each axis in the sensor frame, during, for example, standing, can be seen to be highly dependent on sensor orientation and, thus, not comparable across multiple sequences or epochs of the same activity.

Using the roll and pitch, estimated by (3) and (4), respectively, and assuming a yaw of zero, a rotation matrix can be defined to partially transform the measurements from the sensor frame to the fixed frame, such that the $z$-component of the transformed measurement is aligned with the $z^F$-axis and the $x^S$-$y^S$ plane is aligned with the $x^F$-$y^F$ plane. Once the initial orientation has been calculated in this manner, the angular velocity could be integrated over time to update the rotation matrix between the sensor and the fixed frame at each time instant [2]. In the fixed frame, the gravitational component of acceleration is independent of sensor orientation. However, it is always contained in the same axis ($z^F$), regardless of whether the subject is lying down, sitting, standing, etc. Thus, always transforming the data to an Earth fixed frame does not help to distinguish between the low-motion activities.

With this in mind, the virtual sensor in the body frame is introduced. When the subject is in a standing position (a duration of two seconds was found to be sufficient), the roll and pitch are estimated using (3) and (4) and the rotation matrix is calculated once by:

$$R(\theta_x, \theta_y) = 
\begin{bmatrix}
\cos(\theta_y) & \sin(\theta_x)\sin(\theta_y) & \cos(\theta_x)\sin(\theta_y) \\
0 & \cos(\theta_x) & -\sin(\theta_x) \\
-\sin(\theta_y) & \sin(\theta_x)\cos(\theta_y) & \cos(\theta_x)\cos(\theta_y)
\end{bmatrix}.$$  \hfill (5)

Using (5), the measured signals are all transformed by the same constant rotation at each time instant, such that it appears that all measurements have been recorded from the virtual sensor position. The transformation of the acceleration, for example, is given by:

$$\alpha^B_t = R(\theta_x, \theta_y)\alpha^S_t,$$  \hfill (6)

where the frame, $B'$, denotes the body frame with an arbitrary yaw angle. The gyroscope signals are transformed in the same way. One of the benefits of using a constant rotation matrix is that the transform is insensitive to the accumulation of biases in the sensors as the orientation does not need to be updated by integrating the gyroscope signals over time. Furthermore, the computational load requirement for multiplication of the signals by a constant matrix is low.

Fig. 3b and 4b show that the $z^B$-components of the transformed acceleration signals are very similar despite the significant differences in sensor placement and orientation, observed in Fig. 3a and 4a. Clearly, there are some remaining variations from one sequence to the next, which depend on the subject’s exact behavior, whether they are seated upright or slouching, whether they are lying face down or on their side, among many other variable factors. The $x^B$- and $y^B$-components remain dependent on the initial yaw. However, the modulus of the acceleration in the $x^B$-$y^B$ plane, $\sqrt{(\alpha_x^B)^2 + (\alpha_y^B)^2}$, can be seen to behave similarly for both sequences, i.e. independently of sensor orientation, suggesting that it may be a suitable signal for classification.

In the body frame, the orientation of the gravitational acceleration is the distinguishing signal characteristic for low-motion activities, whilst specific periodic patterns in acceleration and angular velocity characterize movement activities, such as walking, running, jumping, etc. This periodicity can be observed independently of the sensor orientation, as shown in Fig. 3 and 4. More importantly, the modulus of the $x^B$- and $y^B$-components of acceleration and angular velocity conserves the periodicity of the signals while making their joint magnitude invariant to the yaw angle.
**IV. EXPERIMENTAL PROCEDURE**

A. Database Description

In order to evaluate the effectiveness of the proposed transformation, the results obtained with the transformed signals will be compared with the results obtained without any transformation. We have used the database found in [7], which used Xsens MTx-28A53G25 IMUs.

The IMU was placed on a belt, either on the right or left hip, providing 3-axis acceleration and 3-axis angular velocity signals at a sampling rate of 100 Hz. The activities labelled in the database are running, walking, standing, sitting, lying, jumping, falling, and transient activities. In the database, there are both training sequences and benchmark sequences, recorded in semi-naturalistic conditions. There are three benchmark sequences from three different subjects. Two of the subjects have the IMU placed on the lefthand side of the waist and the third, on the righthand side. More details of the data collection and labeling can be found in [7].

In addition, to further test the generality of the method, data was collected with an APDM [9] Opal IMU placed at the subject’s waist, during their normal work day. More than six hours of data were collected, including mainly the following activities: sitting, standing and walking. The subject was requested to very roughly report their activities, in terms of time of day and what they were doing (working at desk, walking to a meeting, eating lunch, etc.) These data were collected for the purposes of algorithm evaluation only and were not used for training.

B. Training

As every training sequence in the database starts with the activity, “standing”, the rotation matrix for transformation was computed using the two first seconds of the acceleration signals. Sequences corresponding to one single activity were extracted from the database, to be used as training data for that activity. Each HMM learned its parameters using the Baum-Welsh algorithm [10]. The emission distributions were defined as mixtures of two Gaussian distributions with diagonal covariance matrix. Training was carried out, in the same way, for both transformed and raw data sequences, separately.

C. Evaluation

Activity estimates were calculated for each benchmark sequence from [7], after resampling the signals to a frequency of 25 Hz. Each benchmark sequence was evaluated using the models for the HDM with HMM, obtained in the training phase for both transformed and raw data. In the case of the transformed model, acceleration and angular velocity signals of the benchmark sequences were transformed using the same process as described for training. In order to compute the most likely sequence of activities, given the observation sequence, the Viterbi algorithm [11] was used.
The algorithms reported by Frank et al. [7] produce activity estimates at a rate of 4 Hz. To provide a “like-with-like” comparison, the final outputs of our system are given by the mode of each set of 6 consecutive activity estimates (i.e. at 4.17 Hz). The data from the APDM sensor were evaluated in the same way, using the models trained by the Xsens sensor data.

V. CLASSIFICATION RESULTS

Tables I and II show the recall and precision values of each activity for the benchmark sequences of Emil, Sinja and Paula from [7], with and without transformation. The average recall and average precision over activities are also shown as well as a weighted average that takes into account the frequency of each activity. The precision of an activity, \( Z \), is measured as the number of samples classified correctly as \( Z \), divided by the total number of samples with inferred label equal to \( Z \). The recall parameter is the number of samples correctly classified as \( Z \), divided by the number of samples whose true label is \( Z \).

<table>
<thead>
<tr>
<th>TABLE I</th>
<th>RECALL RESULTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transformation</td>
<td>Emil</td>
</tr>
<tr>
<td>Running</td>
<td>100%</td>
</tr>
<tr>
<td>Walking</td>
<td>100%</td>
</tr>
<tr>
<td>Standing</td>
<td>96%</td>
</tr>
<tr>
<td>Sitting</td>
<td>100%</td>
</tr>
<tr>
<td>Lying</td>
<td>100%</td>
</tr>
<tr>
<td>Jump</td>
<td>75%</td>
</tr>
<tr>
<td>Fall</td>
<td>100%</td>
</tr>
<tr>
<td>Average</td>
<td>96%</td>
</tr>
<tr>
<td>Weighted Ave.</td>
<td>98%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TABLE II</th>
<th>PRECISION RESULTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transformation</td>
<td>Emil</td>
</tr>
<tr>
<td>Running</td>
<td>100%</td>
</tr>
<tr>
<td>Walking</td>
<td>97%</td>
</tr>
<tr>
<td>Standing</td>
<td>100%</td>
</tr>
<tr>
<td>Sitting</td>
<td>100%</td>
</tr>
<tr>
<td>Lying</td>
<td>100%</td>
</tr>
<tr>
<td>Jump</td>
<td>100%</td>
</tr>
<tr>
<td>Fall</td>
<td>100%</td>
</tr>
<tr>
<td>Average</td>
<td>100%</td>
</tr>
<tr>
<td>Weighted Ave.</td>
<td>98%</td>
</tr>
</tbody>
</table>

Fig. 5 shows the estimated activities obtained by evaluating approximately 1 hour of the long-term data collected with the APDM sensor (the time interval with most activities was chosen). The figure shows estimates obtained both with and without transformation. The time periods identified on the graph contain the following activities (not labelled but described by the subject): 1. Walk, Stand, Descend Stairs; 2. Stand, Lean, Walk, Ascend Stairs; 3. Sit; 4. Walk, Ascend Stairs, Stand; 5. Sit, Stand.

VI. DISCUSSION

As Table I and II show, we have achieved equal or higher average and weighted average of recall and precision using transformed signals, compared to the raw sensor signals. For activities, such as running and walking, the HDM with HMM, even without transformation, was capable of achieving good results, especially for the benchmark sequences of Emil and Sinja. However, without transformation, sitting suffered from a precision and recall of zero for Paula because, in this case, low motion activities are easily confused and the sensor orientation for Paula appears to be very different from the training data. With transformation, the average recall and precision show an improvement, relative to the results without transformation, of 15% and 32%, respectively, for Paula, and are equal or up to 7% better for both Emil and Sinja. Without transformation, sitting and lying for Paula are mostly classified as falling, resulting in a low precision and an unrealistically high recall for fall. However, with transformation, sitting and lying are well classified, at the expense of some reduction in recall for fall. However, the overall performance for Paula in these activities has improved with transformation.

The result of transforming the signals to the virtual sensor orientation was shown, in Section III, to improve the uniformity of the signals prior to classification. With the original training sequences, some of which were recorded on the right side and some on the left, the classification algorithm essentially has to learn two models for each activity. By transforming the signals to the virtual sensor orientation, all of the training sequences contribute to a single unified model, hence providing better classification results.

It should be noted that the database used for the experiments in this work contained only a small amount of data for the short-term activities, jumping and falling. This may be part of the reason why the classification of such activities performs worse than the others. Another factor, which was taken into account by [7], is human error in labeling, which will have a more significant effect in activities with a very short duration. However, it may also be the case that the models for short-term activities need to be modified to better capture the dynamics of the activities. This remains as future work.

To compare our results with those reported by Frank et al., the average precision and recall obtained by our method (with transformation) for Emil and Sinja were calculated (results for Paula were not included in the average, as results for this subject were not mentioned in [7]). Weighted average precision for our method is 97.5%, compared to 95% in [7], whilst weighted average recall for our method is 96.5%, compared to 97.9% in [7]. Thus, it can be seen that our method performs better, in terms of precision, with a small loss in recall performance. It should also be remembered that our method is computationally very fast, consisting of only a constant matrix multiplication of each signal, whilst the feature extraction and dynamic unrestricted Bayesian network recognition algorithm reported by [7] will be computationally more expensive.
For the data collected using the ADPM sensor (Fig. 5), it can be seen that the estimated activities with transformation represent quite well the activities described by the subject. For example, epoch 1 consists mainly of standing and walking, as would be expected in this case. The jumping and running estimates are thought to occur because the subject was descending stairs. Similarly, during epoch 3, sitting is predominant with some samples estimated as lying - possibly due to slouching in the seat. Throughout the entire period shown in the figure, the estimates without transformation have no relation to the labels, as would be expected given that the sensor orientation was the subject’s arbitrary choice, and not one of the options used in the training sequences. These results are very promising as they indicate invariance, not only to sensor orientation, but also to the brand of sensor used.

VII. FUTURE WORK

In this work, in order to compare, fairly, the results using transformed signals and non-transformed signals, exactly the same classification model, the HDM with HMM, was used for both sets of results. For this purpose each activity has been modeled in the same way as in the previous work by the authors [6], i.e. with the same number of states for each activity and with emission distributions defined as mixtures of two Gaussians. Nevertheless, thanks to the transformation proposed in this work, very similar signal amplitudes for each particular activity, have been achieved, independently of sensor orientation. With this in mind, it might be interesting to identify more representative topologies for some activities, modeling the emission distributions as a single Gaussian distribution, instead of a mixture model.

Other tasks which remain for future work include the removal of the restriction that the sensor must not move more than a few millimetres, by updating the sensor orientation, automating the orientation estimation to remove the need for an initial calibration stage (standing for two seconds) and implementing a “tuning” process, whereby models could be “personalised” for each subject. Furthermore, the transformation proposed in this work could also be used in approaches in which there is a previous stage of feature extraction.

Finally, in order to improve the estimation of inter-activity dynamics in real life situations, a database will be constructed by the authors. It is intended to collect data in more naturalistic circumstances than in existing databases and the database will contain data from a very large number of subjects for both training and test purposes.

ACKNOWLEDGEMENTS

This work has been partly supported by Ministerio de Educación de España (projects ‘DEIPRO’, id. TEC2009-14504-C02-01, and ‘COMONSENS’, id. CSD2008-00010).

REFERENCES


