This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2017.2716112, IEEE Journal of Biomedical and Health Informatics

Matrix and Tensor Completion on a Human Activity Recognition Framework

Sofia Savvaki*†, Grigorios Tsagkatakis*, Athanasia Panousopoulou*, and Panagiotis Tsakalides*†

*Institute of Computer Science, Foundation for Research and Technology Hellas (FORTH)
†Department of Computer Science, University of Crete
savvaki@csd.uoc.gr, {greg, apanouso, tsakalid}@ics.forth.gr

***Abstract*—Sensor-based activity recognition is encountered in innumerable applications of the arena of pervasive healthcare and plays a crucial role in biomedical research. Nonetheless, the frequent situation of unobserved measurements impairs the ability of machine learning algorithms to efficiently extract context from raw streams of data. In this work, we study the problem of accurate estimation of missing multi-modal inertial data and we propose a classification framework that considers the reconstruction of sub-sampled data during the test phase. We introduce the concept of forming the available data streams into low-rank 2D and 3D Hankel structures, and we exploit data redundancies using sophisticated imputation techniques, namely Matrix and Tensor Completion. Moreover, we examine the impact of reconstruction on the classification performance by experimenting with several state-of-the-art classifiers. The system is evaluated with respect to different data structuring scenarios, the volume of data available for reconstruction, and various levels of missing values per device. Finally, the trade-off between sub-sampling accuracy and energy conservation in wearable platforms is examined. Our analysis relies on two public datasets containing inertial data, which extend to numerous activities, multiple sensing parameters, and body locations. The results highlight that robust classification accuracy can be achieved through recovery, even for extremely sub-sampled data streams.**

***Index Terms*—Matrix Completion, Tensor Completion, Data Imputation, Human Activity Recognition, Supervised Learning.**

I. Introduction

HUman activity recognition (HAR) is currently faced up with the challenge of enabling the interpretation of physical activities in complex spaces that rise beyond smart homes and medical centers [1], [2]. Wearable computing and machine learning, which are typically combined for extracting context from raw streams of inertial data [3], are therefore expected to pioneer the provision of daily activities analytics while in motion within pervasive environments [4], [5].

To make substantial steps towards this vision, innovative solutions are sought [6], [7] against the operational imperfections of the underlying sensor-rich infrastructure that may lead to a significant volume of unobserved measurements. From the perspective of data analytics, the resulting sub-populated datasets challenge the performance of the employed machine learning algorithms. As such, imputation methods, responsible for replacing missing data with statistically plausible values prior to any classification technique [8], are considered essential [9]. Recent bibliography considers imputation as an inseparable aspect of training a prediction model during a typical training/testing classification process. Specifically, the authors in [10] propose a technique that either calculates the posterior distribution of missing values using the training dataset, or stores the training dataset in memory in order to use it when needed. Yan et. al [11] consider a voting extreme learning scheme, that trains a group of classifiers based on the missing values of the training set, and uses the sub-classifiers with the optimal performance during the test phase. In a similar fashion, the authors in [12] propose an adaptive strategy that employs Self-Organizing Maps or K-Nearest Neighbors for reconstructing missing data from the training dataset, only when classification fails.

The aforementioned approaches share as a common assumption that the training set suffers from missing values, which is not viable in the HAR domain and the use of wearable sensors; training of predictive models is typically performed in well-controlled environments, wherein the sensing data acquisition, transmission and storage is performed under ideal conditions. Even so, the runtime, which corresponds to wearable activity recognition in a long-term and unattended fashion, is more prone to failures during data capture or storage, due to realistic factors such as battery depletion, poor Quality-of-Service [13], and sensors displacement [14].

In this work, we study the impact of imputation on human activity classification tasks, wherein the unobserved values are introduced during the evaluation phase. Towards this direction, we extend our recent work [15] on a human activity classification framework that considered the well-established method of Matrix Completion (MC) [16] for efficiently reconstructing missing values. We herein exploit in a systematic fashion the inherent correlations between heterogeneous inertial sensors, and we study how data streams corresponding to different sensing parameters and/or devices can be combined to 2D or 3D low-rank structures. For the data reconstruction, we formulate the use of a recently proposed Tensor Completion (TC) algorithm [17]. We provide direct comparison between MC and TC, emphasizing on their impact on the classification accuracy of landmark classifiers in the HAR domain, considering real world datasets. We study the performance of the proposed framework under the realistic condition of different levels of missing data per sensing device, and we examine the impact of sub-sampling on the lifetime of the underlying sensing infrastructure. The results highlight the efficacy of the proposed methods in achieving high classification accuracy with less than 50% of the available measurements.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2017.2716112, IEEE Journal of Biomedical and Health Informatics

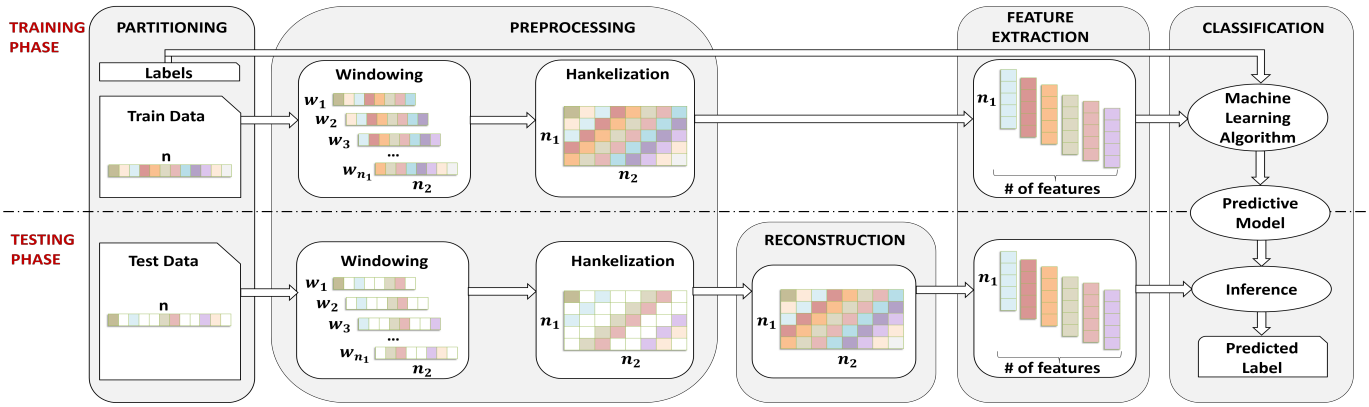

Fig. 1. The architecture of the proposed framework for human activity classification in the presence of missing values.

II. Human Activity Classification in the Presence of Missing Values

We consider a Body Sensor Network (BSN) architecture [18], comprised of wearable sensing platforms which are deployed on several body locations and record a number of sensing modalities (*e.g.*, acceleration, angular velocity) during the execution of a-priori defined sets of physical activities. Each sensing modality may describe the performed activities in more than one sensing channels, *e.g.* the x, y, and z directions of a three-axial inertial sensor. The resulting multi-modal BSN architecture yields in total S sensing channels and provides an equivalent number of data streams in the form of time series. For the remaining of this paper, we will use the terms "sensing channel" and "data stream" interchangeably.

The objective is to accurately map the input data streams, collected using the same multi-modal BSN architecture deployed on a group of individuals, to specific activities. According to the training/testing learning paradigm [19], to address this challenge, a two-step procedure is employed; during the training phase, user annotated data from a subset of individuals are introduced into a learning architecture which estimates the parameters of the classifier. The trained classifier is in turn evaluated during the testing phase with respect to the identification of the activities that are performed by the individuals not included in the training set.

The problem at hand is the classification of the activities in the presence of missing values from the available sensing streams during the test phase. To both investigate as well as compensate the impact of data sub-sampling on accurately predicting the human activity, we propose the framework presented in Fig. 1. Similarly to parallel efforts on activity recognition [3], [20], our framework considers modules for feature extraction and classification for both the training and the test phases. The point of entry for both the training and testing phase is the set of raw values from the S channels, which are stored according to their time instance of capture. By taking into account that the data transmission rate is a-priori known, the lack of recorded data at specific time instances during run time reflects on the presence of missing values in our dataset. As such, our framework emphasizes on reconstructing these missing values during run time, by exploiting in a systematic fashion the correlations that characterize the heterogeneous data streams. Note that our framework is independent both on the number of wearable devices that comprise the BSN, as well as their mounting position on the human body. In fact, as elaborated below, it models the number of associated data streams and the number of channels per device as the key parameters for examining different scenarios for reconstructing the missing measurements. The recovery is performed in a two-fold manner that entails both the appropriate pre-processing of input data, as well as the adaptation of state-of-art imputation methods that exploit the multi-dimensional data format during reconstruction.

A. Data Pre-processing

Both for the training as well as for the testing phase, the objective of the pre-processing module is to transform 1D time series to the appropriate data structure that will reveal the multi-dimensional dependencies among sensing streams. As such, the herein proposed approach targets at enforcing a low-rank assumption on the available information, according to which, linear temporal dependencies over the sensed time series are exploited. This is achieved by the process of *Hankelization*, where consecutively lagged temporal windows of the input sensing stream are structured into low-rank Hankel matrices called trajectory matrices [21].

Specifically, let $X_s = \{x_1, x_2, \ldots, x_n\}$ be a sensor time series of length n. The defining parameter for the windowing of X_s into temporal data segments is the window size $n_2 < n$ that will dictate the structuring of the one-dimensional time-series into a two-dimensional matrix format. The result of the windowing process are

lagged vectors $X_s^i = \{x_i, x_{i+1}, \ldots, x_{i+n_2-1}\}$ where $i = 1 : n - n_2 + 1$. The temporal vectors X_s^i are subsequently concatenated, yielding the trajectory matrix $\mathbf{M}_s \in \mathbb{R}^{n_1 \times n_2}$ of the time series X_s where $n_1 = n - n_2 + 1$, composed of elements:

$$\mathbf{M}_s = \begin{bmatrix} x_1 & x_2 & x_3 & \ldots & x_{n-n_2} \\ x_2 & x_3 & x_4 & \ldots & x_{n-n_2+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n_2} & x_{n_2+1} & x_{n_2+2} & \ldots & x_n \end{bmatrix}$$

Applying the same procedure on all sensing channels yields s trajectory matrices $\mathbf{M}_s$, where $s = 1, 2, \ldots S$. To handle multiple sensing channels, we either consider concatenating individual matrices into large ones, or grouping them into

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2017.2716112, IEEE Journal of Biomedical and Health Informatics

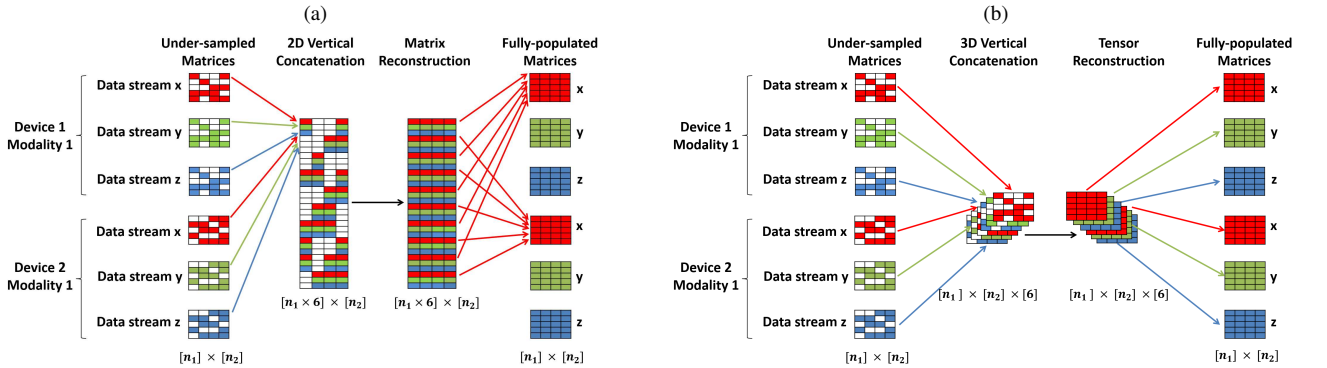

Fig. 2. 2D (a) and 3D (b) data structuring for a testbed that considers two sensing platforms and one sensing modality with three data streams, when the Collective Recovery per Modality (CRM) scenario is adopted.

higher order structures called tensors depending on the particular scenario, as discussed in the next section. While for the training stage the trajectory matrices and tensors are fully populated, during the test phase each matrix $\mathbf{M}_s$ conveys the missing entries of the s-th data stream X_s. Without loss of generality missing entries are modeled as zero values on each trajectory matrix and tensor. In addition, we consider that unobserved values at specific time instances for the s-th data channel (*e.g.*, the x-axis of the accelerometer) impose missing values on the remaining channels of the same modality (*e.g.*, the y- and the $z-$axes of the accelerometer). As explained below, the resulting sub-populated Hankel structures contain the missing values, which are subject to reconstruction.

B. Reconstruction of missing data

This module considers as input the under-sampled Hankel structures of the sensing channels, with the objective to recover their missing entries, by systematically exploiting the inherent correlations between the available data. A primary aspect that arises is the origin of the data streams involved during reconstruction. Specifically, we consider the following three scenarios for combining the partially populated data streams per sensing modalities and/or devices:

- *Collective Recovery per Device* (CRD) considers that reconstruction is made with respect to the device, and as such, it involves the data streams from all the modalities originating from the same sensing platform, *e.g.*, the accelerometer and the gyroscope data streams acquired from the device attached to the chest of the individual. The reconstruction instances for CRD are equivalent to the scale of the BSN.
- *Collective Recovery per Modality* (CRM) recovers missing values with respect to the type of modality. Specifically, it involves data streams originating from different platforms which are grouped with respect to modality, *e.g.*, the accelerometer data of all BSN nodes. As such, the instances of CRM are directly mapped to the number of modalities.
- *Collective Recovery* (CR) incorporates all available data streams originated from different sensing platforms and corresponding to different types of measurements.

For each of these scenarios the trajectory matrices $\mathbf{M}_s$ of the sensing channels involved are concatenated to either trajectory matrices $\mathbf{M} \in \mathbb{R}^{[n_1 \times n_3] \times n_2}$ (2D structuring), or the respective trajectory tensors $\boldsymbol{\mathcal{M}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ (3D structuring). For both structuring cases, n_3 defines the number of participating channels per scenario. An example of the 2D and 3D data structuring is presented in Fig. 2 for the CRM case, considering two sensing platforms, and one sensing modality with three channels (x, y, z), thus resulting to $S = 6$ sensing channels. The 2D data structuring (Fig. 2(a)) combines the input trajectory matrices $\mathbf{M}_s$, $s = \{1,2,\ldots,6\}$ to the vertically concatenated $\mathbf{M} \in \mathbb{R}^{[n_1 \times 6] \times n_2}$. The 3D data structuring (Fig. 2(b)) concatenates the same input along the third dimension, thereby yielding the tensor $\boldsymbol{\mathcal{M}} \in \mathbb{R}^{n_1 \times n_2 \times 6}$. Considering the same example, scenario CR would generate the same matrix and tensor for the 2D and 3D data structuring, respectively. Scenario CRD would produce the matrix (tensor) $\mathbf{M} \in \mathbb{R}^{[n_1 \times 3] \times n_2}$ ($\boldsymbol{\mathcal{M}} \in \mathbb{R}^{n_1 \times n_2 \times 3}$) for each device.

The recovery of the missing entries of the partially populated matrices or tensors, is performed by the means of either MC or TC. MC has been successfully applied in various datasets including vital signs [22], as well as human activity recognition in videos [23]. Similarly, TC has already delivered promising results on the recovery of both synthetic and real-world MRI and hyper-spectral data [17] [24]. The background on MC and its extension to TC is elaborated below.

2D Data Reconstruction. The reconstruction of the partially observed matrix $\mathbf{M} \in \mathbb{R}^{[n_1 \times n_3] \times n_2}$ relies on the theory of MC, which considers the availability of $k << [n_1 \times n_3] \times n_2$ entries in $\mathbf{M}$. While this is an ill-posed problem, it was recently shown [16] that exact recovery is possible from a limited set of k sampled entries, given that $\mathbf{M}$ is characterized by a much lower rank compared to its ambient dimensions. A straightforward approach for filling in the missing values is to estimate the lowest-rank matrix $\mathbf{M}^*$ which agrees with the given data [25]. While rank-minimization is an NP-hard problem [26], exact matrix recovery is possible by replacing rank with the computationally tractable nuclear norm given by $\|\mathbf{M}\|_* = \sum_{i=1}^{k} \|\sigma(i)\|$, where $\sigma(1) \ldots \sigma(k)$ are the singular values of matrix $\mathbf{M}$ [27]. The problem of matrix recovery can thus be formulated as:

$$\begin{aligned} &\text{minimize} \quad \|\mathbf{Z}\|_* \\ &\text{subject to} \quad \mathcal{A}(\mathbf{Z}_{ij}) = \mathcal{A}(\mathbf{M}_{ij}), \quad (i,j) \in \Omega \ , \end{aligned} \tag{1}$$

where Ω is the set of indices (i,j) corresponding to the observed measurements from $\mathbf{M}$, and $\mathcal{A}$ is a linear operator defined as $\mathcal{A}(\mathbf{M}_{ij}) = \{m_{ij}$ if $(i,j) \in \Omega \mid 0$ otherwise$\}$. Recovery of the input matrix is

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2017.2716112, IEEE Journal of Biomedical and Health Informatics

achieved, provided that $\mathbf{M}$ is sampled uniformly and $k \geq Cr(max(n_1 \times n_3, n_2))^{6/5} \log((max(n_1 \times n_3, n_2)))$, C is a constant, and r is the rank of the matrix [16]. In order to solve Eq. (1) we herein adopt the Augmented Lagrange Multipliers (ALM) method [28], due to its performance in terms of recovery of missing entries and computational complexity.

3D Data Reconstruction. Similar to the 2D case, the reconstruction of the sub-populated tensor $\boldsymbol{\mathcal{M}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ relies on the availability of $k << n_1 \times n_2 \times n_3$ entries in $\boldsymbol{\mathcal{M}}$. State-of-the-art methods [29], [30] involve unfolding the tensor into a matrix and the formulation of a matrix recovery problem, which is solved by the means of Singular Value Decomposition. However, this approach is impractical for large-scale problems, and more importantly, it exploits the low-rank characteristics of a single dimension of the underlying tensor during each unfolding. To address these issues, we herein employ the recently proposed low-rank TC using Parallel Matrix Factorization [17], which offers lower computational complexity and better performance compared to the unfolding family of solutions.

Specifically, each mode (*i.e.*, dimension) of tensor $\boldsymbol{\mathcal{M}}$ is unfolded to a set of matrix factors $\mathbf{X}_n$ $\mathbf{Y}_n$, such that $\mathbf{M}_n \approx \mathbf{X}_n \mathbf{Y}_n$, where n =1,2,3 is the mode considered. Introducing one common variable $\boldsymbol{\mathcal{Z}}$ to relate these matrix factorizations, we solve the following problem to recover tensor $\boldsymbol{\mathcal{M}}$:

$$\begin{aligned} \underset{\mathbf{X},\mathbf{Y},\mathbf{Z}}{\text{minimize}} \quad & \sum_{n=1}^{3} \frac{\alpha_n}{2} \|\mathbf{X}_n \mathbf{Y}_n - \mathbf{Z}_n\|_F^2 \\ \text{subject to} \quad & \mathcal{A}(\boldsymbol{\mathcal{Z}}_{ijl}) = \mathcal{A}(\boldsymbol{\mathcal{M}}_{ijl}), \quad (i,j,l) \in \Omega \ , \end{aligned} \tag{2}$$

where $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\}$, $\mathbf{Y} = \{\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3\}$ and $\mathbf{Z} = \{\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3\}$ correspond to the unfolding of the three-way tensors we consider in this work. The parameters α_n are weights which satisfy $\sum_n \alpha_n = 1$, Ω is the index set (i,j,l) of observed entries, and $\mathcal{A}$ is defined similarity to the matrix case as $\mathcal{A}(\boldsymbol{\mathcal{M}}_{ijl}) = \{m_{ijl}$ if $(i,j,l) \in \Omega \mid 0$ otherwise$\}$.

The drawback of this approach is that all three ranks of the tensor must be a-priori defined, an assumption which is impractical to consider. To tackle this issue, we adopt the rank increasing scheme introduced in [17], according to which the rank estimates are iteratively increased, if the algorithm detects slow progress in the updates of the singular values of their corresponding factor matrices. Thus, Eq. (2) is solved by cyclically updating $\mathbf{X}_n, \mathbf{Y}_n$, and $\mathbf{Z}_n$. The procedure is performed for all modes in parallel, thereby improving both the speed of execution and its effectiveness, since it exploits the low-rank characteristics of tensor modes at the same time.

C. Feature Extraction and Classification

During the training phase, the S trajectory matrices $\mathbf{M}_s$ are injected into the feature extraction module for extracting a set of representative attributes. For each temporal window and sensing channel, the resulting feature vector is comprised of N statistical attributes. In total n_1 feature vectors are produced for each channel. The classification module considers this input along with the labels of the activities for calculating the activity recognition models according to the adopted machine learning algorithm.

During the test phase, the 2D or 3D reconstruction process yields a fully populated matrix or tensor respectively, which is decomposed to n_3 trajectory matrices for each instance of the adopted data structuring scenario (CRM, CRD, CR). Similarly to the training phase, the feature extraction module applies on the reconstructed trajectory matrices for the total number of channels, producing a feature matrix $\mathbb{R}^{n_1 \times [S \times N]}$. The classification module employs this input and the trained models for inferring the predicted activity label.

III. Evaluation Studies

The proposed framework has been evaluated on two popular activity recognition datasets, namely the HAR [31], and the MHEALTH [32]. The HAR dataset employs a conventional smartphone, deployed on the waist. The dataset contains filtered 3-axial inertial data (acceleration, angular velocity), providing in total $S = 6$ sensing channels, from thirty healthy individuals who are performing five physical activities. The MHEALTH dataset utilizes Shimmer2 [33] wearable sensors that are deployed on three body locations, i.e.chest (CH), right wrist (RW), left ankle (LA). It contains three different types of 3-axial modalities (acceleration, angular velocity, variations of magnetic field) and $S = 21$ sensing channels, which describe twelve daily activities, ranging from body postures to combined motions performed by ten volunteers.

For the purposes of our study, both datasets were randomly partitioned to 70% of subjects for training and 30% of subjects for testing, respectively. For the pre-processing of the available data streams, we consider a window size of n_2 =128 with a 50% overlap to form the aforementioned Hankel structures, in accordance to the bibliography [34], [35]. These parameters yield n_1 equal to 406 and 544 for the HAR and MHEALTH datasets, correspondingly. The existence of missing measurements was simulated by randomly taking a certain amount of data out from the original set for testing. For selecting which values to omit, we employed a uniform distribution, in compliance to the related bibliography [16]. In total, ten different cases of data partitioning were considered, resulting to an equivalent number of independent experiments for the reconstruction and classification processes.

The efficacy of the MC and the TC reconstruction methods is evaluated against the state-of-the-art Regularized Expectation Maximization (RegEM) [36] imputation method, which considers an iterative process for finding maximum likelihood estimates of parameters and has been applied for the 2D reconstruction of vital signs [22]. With regard to the feature extraction process, we consider the $N = 22$ statistical features per sensing channel described in [15]. To study the effect of reconstruction on the classification accuracy we employ two state-of-art families of classifiers, which are extensively used in the HAR domain, as well as reasonably time responsive [37], namely: (a) the K-Nearest Neighbours (KNN) considering both the Euclidean and the Cosine metrics (K=10) and (b) the Support Vector Machines (SVM), adopting both Gaussian and Quadratic kernels [19].

A key experimental parameter is the *fill ratio* $f \in [0,1]$ defined as the percentage of measurements available for reconstruction. When $f \rightarrow 0$ (1), the percentage of available

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2017.2716112, IEEE Journal of Biomedical and Health Informatics

measurements decreases (increases), while $f = 1$ corresponds to a fully populated set of measurements, containing no missing entries. The performance metrics employed are: (a) the Normalized Mean Square Error (NMSE) defined as the mean squared error between the fully populated and the reconstructed structures normalized with respect to the l_2 norm; (b) the running time for each reconstruction method; and (c) the classification accuracy (CA), which quantifies the ability of each classifier to predict the correct activity label. Finally, for the purposes of this study our framework has been implemented on MATLAB®, considering a conventional CPU processor (Intel Core™ i5-4590 CPU@3.30GHz, 8GB RAM).

Evaluation Results

Error of missing data reconstruction. Figure 3 illustrates the cumulative NMSE of RegEM, MC, and TC both on HAR and MHEALTH datasets. The data structuring is made according to all of the engaged scenarios and it involves the totality of the data streams available per dataset. It is essential to note that when the HAR dataset is considered, CRD is identical to CR, since in this dataset there is only one sensing device.

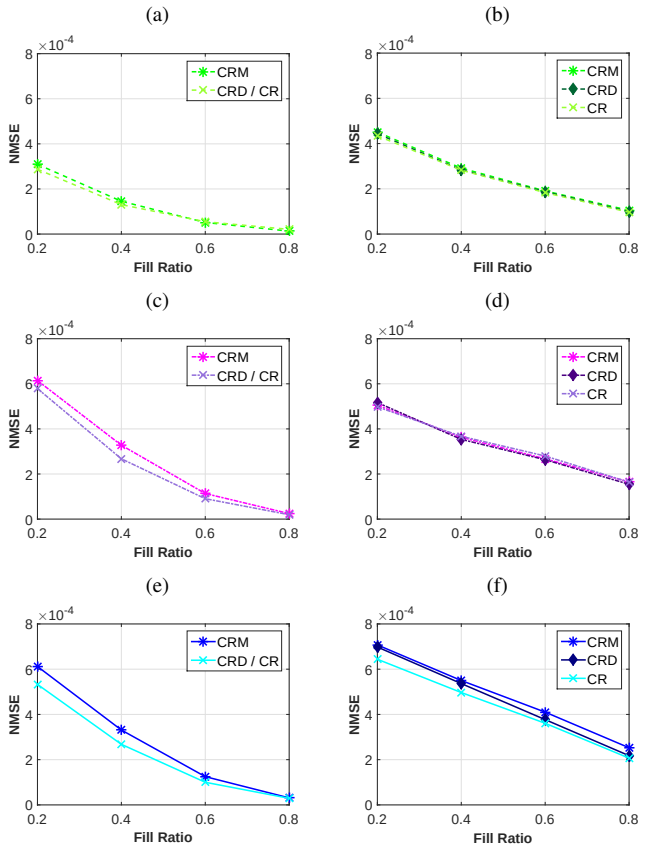

Fig. 3. Cumulative NMSE of RegEM (a-b), MC (c-d), and TC (e-f), for HAR (left) and mHealth (right) w.r.t. f, considering CRM, CRD, and CR. (Mean of 10 independent runs).

The initial observation to make is that for both datasets RegEM achieves the best reconstruction quality in NMSE terms, followed by MC. Specifically, when the HAR dataset is considered (Fig. 3 (a),(c),(e)), RegEM and MC have an almost identical NMSE performance for $f \geq 0.6$, while RegEM achieves a smaller reconstruction error for $f \leq 0.5$. On the other hand, TC is inferior with respect to the value of the cumulative NMSE, which remains greater than 10^{-4} for $f < 0.6$. Similar observations are made for the case of the MHEALTH dataset (Fig. 3 (b),(d),(f)), as RegEM is shown to be the most efficient reconstruction method for all values of f. Indicatively, RegEM achieves a NMSE lower than 5×10^{-4} for $f = 0.2$, as opposed to the MC and TC techniques which yield a NMSE equal to 5×10^{-4} and 7×10^{-4}, respectively.

With regard to the recovery scenarios, CR is reported as the most NMSE effective. This is more evident for the MC and TC techniques on the HAR dataset (Fig. 3 (c),(e)), as well as for the TC applied on MHEALTH (Fig. 3 (f)). The remaining plots demonstrate similar NMSE performance for all scenarios. The associated reasoning is related to the number of data streams per dataset and their corresponding rank, leading to marginal reconstruction quality per sensing channel.

Execution Time per Scenario. Figure 4 presents the cumulative running time for the reconstruction of all data streams when different scenarios of data structuring are considered.

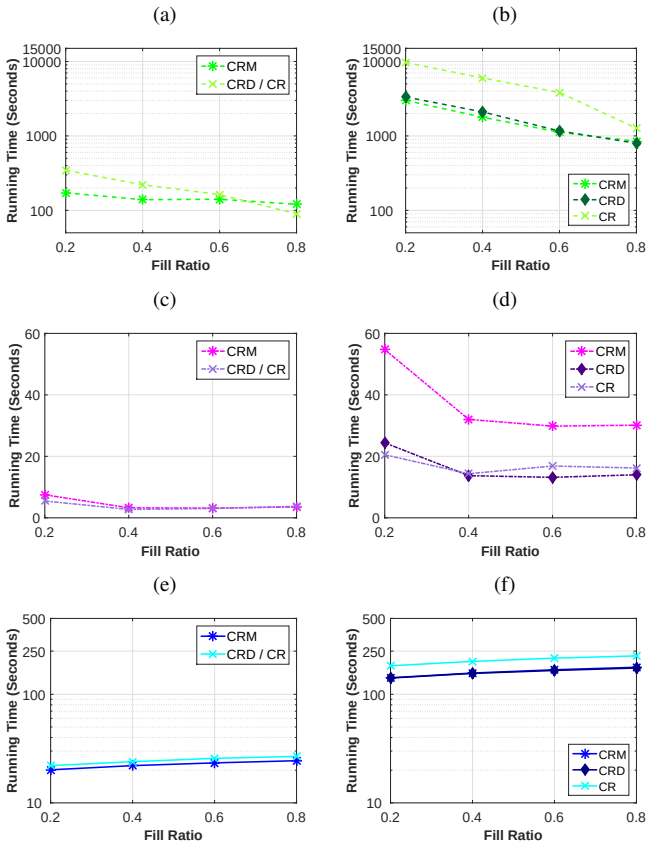

Fig. 4. The cumulative execution time required by RegEM (a-b), MC (c-d), and TC (e-f), for HAR (left) and mHealth (right) w.r.t. f, considering CRM, CRD, and CR. (Mean of 10 independent runs).

The results presented in Fig. 4 indicate that the superiority of RegEM in terms of NMSE comes with a price: RegEM's running time increases exponentially when the CR scenario, involving larger-scale structures, is considered. Indicatively, the cumulative running time when CR is adopted at $f = 0.2$, is approximately 6 minutes for the HAR dataset, while for the more extensive MHEALTH dataset it increases to 2.7 hours.

The advantage of the MC technique in terms of its computational complexity is shown in Fig. 4(c)-4(d) for the HAR and MHEALTH datasets, respectively. For both datasets and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2017.2716112, IEEE Journal of Biomedical and Health Informatics

all cases of structuring scenarios, the running time decreases as the value of the fill ratio increases, thereby highlighting the dependency of the convergence of the MC algorithm on the amount of missing data. Moreover, we observe that it is more time consuming for MC to perform recovery per modality (CRM) rather than per device (CRD) and/or collectively from all data (CR). This is more evident for the case of the MHEALTH dataset (Fig. 4(d)), wherein 21 channels representing three modalities are collected from three devices. It is related to the rank of the under-complete structures, defined by their associated data streams, and the total number of MC reconstruction instances per structuring scenario; including a rather uncorrelated data stream, among the ones combined per scenario, could lead to structures that tend to be of full rank. This results in a dramatic delay regarding the convergence of the specific recovery call, thus increasing the overall running time for the corresponding scenario. This is the case for the CRM scenario when MC is adopted. By contrast when the CR policy is employed, one MC-based reconstruction instance of high rank is applied, which is overall less time-costly.

Interesting observations are derived when the TC method is employed for the reconstruction of the HAR (Fig. 4(e)) and the MHEALTH (Fig. 4(f)) dataset. The execution time remains similar for the entire range of values of f, with a slight upward trend that follows the increase of f. This outcome suggests that the computational complexity of the TC algorithm is not directly related to the volume of missing data to be recovered. Instead, it is associated to the size of the underlying tensor, since CR requires slightly longer execution times than CRD and CRM. The difference in the size and structuring of each tensor per scenario is dictated by its third dimension, which is equivalent to the number of sensing channels; considering the MHEALTH scenario, the value of n_3 ranges from 3 channels (acceleration along the x,y,z axes) from the sensor attached on the subject's chest (for an instance of CRD) to 21 (for CR), which is the total number of data channels in this case.

Impact of reconstruction on the Classification Accuracy. The CA achieved when all classifiers and data structuring scenarios are employed is presented in Fig. 5 and Fig. 6 for the HAR and MHEALTH datasets, respectively. We therein also present the classification performance achieved when fully populated matrices are fed to the classification module ($f = 1$).

With regard to the impact of the MC reconstruction method on the CA, the Quadratic and Gaussian SVM exhibit superior performance for both datasets when $f = 1$. Nevertheless, when $f < 1$, the optimal CA results are provided when MC is combined with the KNN family of classifiers. For example, considering the HAR dataset at $f = 0.2$ and the CRD/CR scenario, the CA achieved by the Euclidean KNN equals to 73.0%, as opposed to 63.8% obtained by the Gaussian SVM. This behavior is even more prominent when the MHEALTH dataset (Fig. 6) is examined at $f = 0.2$; the data structuring according to CRD yields a CA equal to 63.8% for the Gaussian SVM method. By contrast, when the Euclidean KNN classifier is employed, the mean value of the CA increases to 84.0%. From these results, we can get three major insights about the MC reconstruction method: (a) its impact on the classification performance depends heavily on the employed

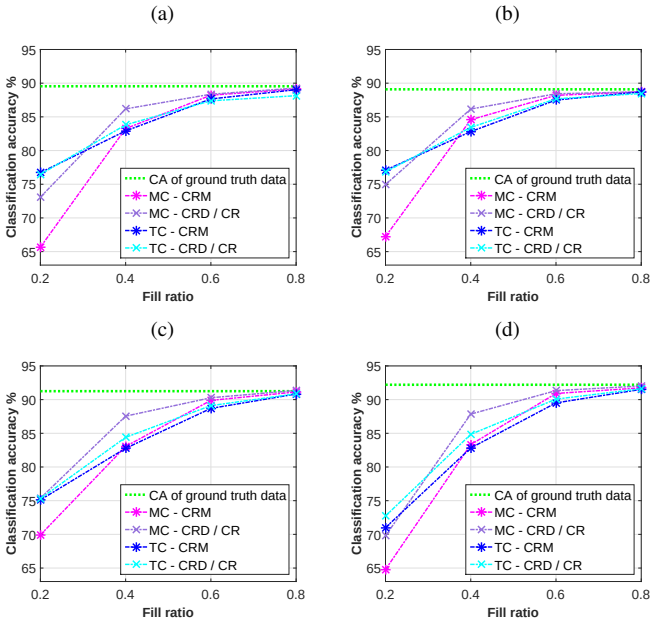

Fig. 5. CA of (a) Euclidean KNN, (b) Cosine KNN, (c) Gaussian SVM, and (d) Quadratic SVM on HAR, considering MC and TC recovery for all scenarios (Mean of 10 independent runs).

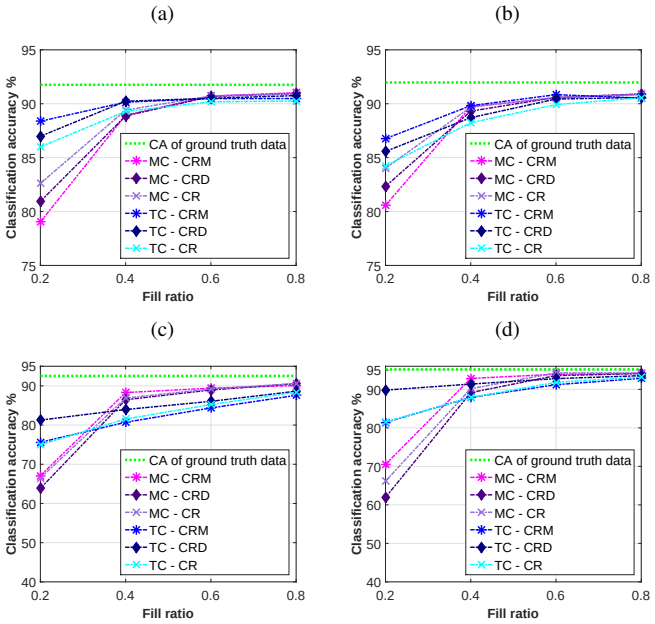

Fig. 6. CA of (a) Euclidean KNN, (b) Cosine KNN, (c) Gaussian SVM, and (d) Quadratic SVM on MHEALTH, considering MC and TC recovery for all scenarios (Mean of 10 independent runs).

classifier; (b) when the percentage of missing values is high (i.e., for low f values), the correlations that appear in the reconstructed data fade out at the presence of the feature space transformation (*e.g.*, due to the kernel trick), which is a distinct characteristic for the SVM family of classifiers; (c) with the appropriate classifier (*i.e.*, KNN family), MC can fully utilize the correlations that exist among diverse modalities of the same sensing device, even if such correlations are not explicitly encoded into the recovery process (CRD and CR versus CRM).

When the TC method is applied we observe that the CA performance achieved by the SVM classifiers when $f \leq 0.4$ is consistent to the optimal performance provided when $f = 1$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2017.2716112, IEEE Journal of Biomedical and Health Informatics

As such, when 20% of the measurements of the HAR dataset is available, the combination of CDR/CR and the Gaussian SVM yields 81.2% CA, which is the highest achieved among all classifiers. Similarly, for the MHEALTH dataset and $f = 0.2$, when the data are structured according to CRD (CR), the value of CA for the Quadratic SVM remains optimal and equal to 89.8% (81.5%). This outcome initially contradicts our aforementioned intuition on the performance of TC, with respect to the cumulative NMSE (Fig. 3). Such a result can be explained, though, by examining the correlations among the data, since some modalities (e.g., acceleration) are low-rank, while others are not. TC can recover the strongly correlated streams with high precision, but fails for higher-ranked data, resulting in a larger total NMSE than that of MC.

Different percentage of missing values per device. In this experiment, we address the realistic case of each device being affected by a different value of f [6]. To this end, we employ the MHEALTH dataset and the MC/TC recovery methods. Missing values are introduced by applying different sampling rates per device, which are centered at three values of the fill ratio, namely $\bar{f}$ =0.2, $\bar{f}$=0.5, and $\bar{f}$=0.8. Considering the three different body locations, i.e. CH, LA, and RW, 9 different combinations of values of f per device are extracted, ranging from 0.1 to 1. For each of these combinations, we apply the Gaussian SVM classifier, which has the optimal performance in terms of CA (92.5%, Fig. 6) when f =1.

TABLE I
CA OF GAUSSIAN SVM AT DIFFERENT VALUES OF f PER DEVICE.

Sampling Rate per Device	CH, LA, RW	Missing Values	CRD MC / TC	CRM MC / TC	CR MC / TC
$\bar{f}$ =0.2	CH=0.1, LA=0.2, RW=0.3	**46.2**	60.3 / **82.0**	**64.2** / 75.2	59.8 / 74.7
	CH=0.3, LA=0.2, RW=0.1	45.7	58.4 / 80.4	61.7 / 74.5	60.7 / 75.1
	CH=0.1, LA=0.1, RW=0.4	42.1	44.5 / 79.5	44.4 / 69.0	38.8 / 69.8
$\bar{f}$ =0.5	CH=0.3, LA=0.5, RW=0.7	**66.5**	**88.4** / 84.8	84.7 / 81.2	83.8 / 81.9
	CH=0.7, LA=0.5, RW=0.3	65.2	86.5 / **84.7**	84.7 / 83.6	85.6 / 84.2
	CH=0.4, LA=0.4, RW=0.7	58.2	87.0 / 83.8	87.1 / 80.7	87.0 / 80.9
$\bar{f}$ =0.8	CH=0.6, LA=0.8, RW=1	**87.6**	89.9 / 87.0	90.8 / 87.2	90.7 / 86.8
	CH=1, LA=0.8, RW=0.6	86.4	90.1 / **88.6**	90.9 / 88.1	**91.2** / 88.4
	CH=0.7, LA=0.7, RW=1	85.3	89.5 / 86.7	90.6 / 85.8	90.5 / 85.7

The results in terms of CA are presented in Table I, for the MC- and the TC-based reconstruction, against the CA when no recovery method is applied on the dataset. The first observation to make is that the CA for all cases increases along with the increase of the value of the fill ratio, in compliance to the results presented above when the same value of fill ratio applies to all devices. With respect to the data structuring scenarios when the TC reconstruction method is applied, CRD is the most promising, since for all different cases of mean value of f it yields the highest value of CA compared to the remaining structuring scenarios: 82%, 84.7%, and 88.6% for mean values of f equal to 0.2, 0.5, and 0.8, respectively. Similar performance in observed for the MC method.

Moreover, it is shown in Table I that for each sampling set considered, there is at least one reconstruction scheme that significantly improves the CA achieved in the case of non-recovered data. For instance, when CRD is considered and the mean value of f is 0.2, the TC-based reconstruction yields accuracy which is approximately 35% higher than the one resulting when no recovery method is applied (46.2%). Notably, there is not a specific sampling set per mean value of fill ratio that produces the optimal accuracy. When the mean value of f is 0.2 or 0.5, the occurrence of a higher percentage of available data from the RW yields the optimal performance for both the MC and TC reconstruction methods. By contrast, when the mean value of f is 0.8, it is more advantageous to provide a richer volume of data from the CH location.

Impact of Sub-sampling on Battery Consumption. We finally examine the impact of data sub-sampling and reconstruction on the lifetime of the BSN responsible for the data acquisition for recognizing human activities. Towards this direction, we employ the Shimmer2 sensing platforms for capturing 3-axial accelerometer, gyroscope, magnetometer, and battery level data, in accordance to the available physical data streams provided within the MHEALTH dataset. We consider four different cases of sampling rates, *i.e.* {50, 35, 20, 10}Hz, for continuously streaming data within a time-span of 9 hours. Without loss of generality, the data and a transmission sampling rate at 50Hz corresponds to a fully populated matrix, i.e., f =1. Subsequently, reducing the sampling rate to {35, 20, 10}Hz would produce sub-sampled datasets corresponding to $f = \{0.7, 0.4, 0.2\}$.

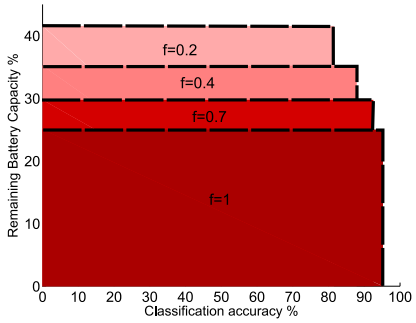

Fig. 7. The remaining battery capacity of a Shimmer3 platform after 9 hours of use, w.r.t. the CA of Quadratic SVM for TC-based reconstruction and data structuring according to CR.

Figure 7 presents the percentage of the remaining battery with respect to the CA achieved when the Quadratic SVM classifier is applied. As expected, reducing the value f, and subsequently the sampling and transmission rate, increases the remaining battery capacity, thereby improving the power autonomy of the device. Notably, prolonging the lifetime of the device is accompanied by a relatively slight degradation on the CA. For instance, when the value of sampling rate drops from 50 Hz (f =1) to 10 Hz (f =0.2) Hz, the capacity of the remaining battery increases by a factor equal to 82.1% (25.0% $\rightarrow$ 41.5%), at the expense of reducing the SVM CA by a factor of 14.4% (95.3% $\rightarrow$ 81.5%). This significant outcome indicates that the TC reconstruction method can be exploited for a new line of energy-efficient data acquisition protocols for standalone BSN architectures.

IV. CONCLUSIONS

In this work we have addressed the problem of missing values in the HAR domain, considering: (a) the inherent correlations that characterize heterogeneous inertial data, and (b) the use of MC and TC as an inseparable part of a modular human activity classification framework for reconstructing sub-populated 2D and 3D structures, respectively. This is part of a fundamental change in data understanding for HAR problems, where machine learning algorithms are

autonomously trained using label data for the automated characterization of observations. Our studies elaborate on the impact of imputation on the reconstruction error and CA of landmark classifiers in the HAR domain, considering different scenarios for the combination of the inertial sensing channels. The results demonstrate that MC yields the optimal performance in terms of time response, while the combination of TC with a SVM classifier achieves the most promising results when multi-modal data are acquired by multiple devices. Results also outline the efficacy of our proposed framework in the case of different data loss percentages per device, and the improvement of the lifetime of the underlying BSN at minimum loss of CA.

Extensions of this work consider: (a) distributed aspects for the reconstruction and classification procedures, while additionally considering the trade-off between the achieved energy gain and the computational overhead per device; (b) the real-time implementation of the proposed framework, taking into consideration both the necessary memory modules for enabling data streaming alongside data reconstruction, as well as the computational/energy overhead introduced by the recovery process on conventional low-complexity and low-power receivers (e.g. smart phones).

ACKNOWLEDGMENT

This work was funded by the DEDALE project, contract no. 665044, within the H2020 Framework Program of the European Commission.

REFERENCES

[1] A. Solanas, C. Patsakis, M. Conti, I. S. Vlachos, V. Ramos, F. Falcone, O. Postolache, P. A. Perez-martinez, R. D. Pietro, D. N. Perrea, and A. Martinez-Balleste, "Smart health: A context-aware health paradigm within smart cities," *IEEE Communications Magazine*, vol. 52, no. 8, pp. 74–81, Aug 2014.

[2] E. Berlin and K. Van Laerhoven, "Detecting leisure activities with dense motif discovery," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. ACM, 2012, pp. 250–259.

[3] A. Bulling, U. Blanke, and B. Schiele, "A tutorial on human activity recognition using body-worn inertial sensors," *ACM Computing Surveys (CSUR)*, vol. 46, no. 3, p. 33, 2014.

[4] N. Alshurafa, W. Xu, J. J. Liu, M. C. Huang, B. Mortazavi, C. K. Roberts, and M. Sarrafzadeh, "Designing a robust activity recognition framework for health and exergaming using wearable sensors," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 5, pp. 1636–1646, Sept 2014.

[5] S. Chen, J. Lach, B. Lo, and G. Z. Yang, "Toward pervasive gait analysis with wearable sensors: A systematic review," *IEEE Journal of Biomedical and Health Informatics*, vol. 20, no. 6, pp. 1521–1537, Nov 2016.

[6] S. Movassaghi, M. Abolhasan, J. Lipman, D. Smith, and A. Jamalipour, "Wireless body area networks: A survey," *IEEE Communications Surveys Tutorials*, vol. 16, no. 3, pp. 1658–1686, Third 2014.

[7] C. C. Y. Poon, B. P. L. Lo, M. R. Yuce, A. Alomainy, and Y. Hao, "Body sensor networks: In the era of big data and beyond," *IEEE Reviews in Biomedical Engineering*, vol. 8, pp. 4–16, 2015.

[8] J. Luengo, S. García, and F. Herrera, "On the choice of the best imputation methods for missing values considering three groups of classification methods," *Knowledge and Information Systems*, vol. 32, no. 1, pp. 77–108, 2012.

[9] A. Farhangfar, L. Kurgan, and J. Dy, "Impact of imputation of missing values on classification error for discrete data," *Pattern Recognition*, vol. 41, no. 12, pp. 3692–3705, 2008.

[10] I. A. Gheyas and L. S. Smith, "A neural network-based framework for the reconstruction of incomplete data sets," *Neurocomputing*, vol. 73, no. 16, pp. 3039–3065, 2010.

[11] Y.-T. Yan, Y.-P. Zhang, J. Chen, and Y.-W. Zhang, "Incomplete data classification with voting based extreme learning machine," *Neurocomputing*, vol. 193, pp. 167 – 175, 2016.

[12] Z. Liu, Q. Pan, J. Dezert, and A. Martin, "Adaptive imputation of missing values for incomplete pattern classification," *Pattern Recognition*, vol. 52, pp. 85 – 95, 2016.

[13] Y. L. Zheng, X. R. Ding, C. C. Y. Poon, B. P. L. Lo, H. Zhang, X. L. Zhou, G. Z. Yang, N. Zhao, and Y. T. Zhang, "Unobtrusive sensing and wearable devices for health informatics," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 5, pp. 1538–1554, May 2014.

[14] O. Banos, M. A. Toth, M. Damas, H. Pomares, and I. Rojas, "Dealing with the effects of sensor displacement in wearable activity recognition," *Sensors*, vol. 14, no. 6, p. 9995, 2014.

[15] S. Savvaki, G. Tsagkatakis, A. Panousopoulou, and P. Tsakalides, "Effects of matrix completion on the classification of undersampled human activity data streams," in *EUSIPCO*, Aug 2016, pp. 2010–2014.

[16] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational mathematics*, vol. 9, no. 6, pp. 717–772, 2009.

[17] Y. Xu, R. Hao, W. Yin, and Z. Su, "Parallel matrix factorization for low-rank tensor completion," *arXiv preprint arXiv:1312.1254*, 2013.

[18] G.-Z. Yang, Ed., *Body Sensor Networks*, 2nd ed. Springer-Verlag London, ISBN 978-1-4471-6373-2, 2014.

[19] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer series in statistics Springer, Berlin, 2001, vol. 1.

[20] M. Zhang and A. A. Sawchuk, "Human daily activity recognition with sparse representation using wearable sensors," *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 3, pp. 553–560, 2013.

[21] A. L. Rukhin, "Analysis of time series structure ssa and related techniques," *Technometrics*, vol. 44, no. 3, pp. 290–290, 2002.

[22] S. Yang, K. Kalpakis, C. F. Mackenzie, L. G. Stansbury, D. M. Stein, T. M. Scalea, and P. F. Hu, "Online recovery of missing values in vital signs data streams using low-rank matrix completion," in *ICMLA, 2012 11th International Conference on*, vol. 1. IEEE, 2012, pp. 281–287.

[23] E. Adeli-Mosabbeb and M. Fathy, "Non-negative matrix completion for action detection," *Image and Vision Computing*, vol. 39, pp. 38–51, 2015.

[24] B. Huang, C. Mu, D. Goldfarb, and J. Wright, "Provable models for robust low-rank tensor completion," *Pacific Journal of Optimization*, vol. 11, no. 2, pp. 339–364, 2015.

[25] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM review*, vol. 52, no. 3, pp. 471–501, 2010.

[26] R. Meka, P. Jain, C. Caramanis, and I. S. Dhillon, "Rank minimization via online learning," in *ICML*. ACM, 2008, pp. 656–663.

[27] E. J. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *Information Theory, IEEE Transactions on*, vol. 56, no. 5, pp. 2053–2080, 2010.

[28] Z. Lin, M. Chen, and Y. Ma, "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices," *arXiv preprint arXiv:1009.5055*, 2010.

[29] S. Gandy, B. Recht, and I. Yamada, "Tensor completion and low-n-rank tensor recovery via convex optimization," *Inverse Problems*, vol. 27, no. 2, p. 025010, 2011.

[30] J. Liu, P. Musialski, P. Wonka, and J. Ye, "Tensor completion for estimating missing values in visual data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 208–220, 2013.

[31] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "A public domain dataset for human activity recognition using smartphones." in *ESANN*, 2013.

[32] O. Banos, R. Garcia, J. A. Holgado-Terriza, M. Damas, H. Pomares, I. Rojas, A. Saez, and C. Villalonga, "mhealthdroid: a novel framework for agile development of mobile health applications," in *Ambient Assisted Living and Daily Activities*. Springer, 2014, pp. 91–98.

[33] Shimmer user manual. [Online]. Available: http://www.shimmersensing.com/images/uploads/docs/Shimmer_User_Manual_rev3m.pdf

[34] L. Bao and S. S. Intille, "Activity recognition from user-annotated acceleration data," in *International Conference on Pervasive Computing*. Springer, 2004, pp. 1–17.

[35] S. Savvaki, "Novel techniques for the estimation of multi - modal missing data in wireless sensor networks," Master's thesis, University of Crete, 2016.

[36] T. Schneider, "Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values," *Journal of Climate*, vol. 14, no. 5, pp. 853–871, 2001.

[37] M. Shoaib, S. Bosch, O. D. Incel, H. Scholten, and P. J. Havinga, "A survey of online activity recognition using mobile phones," *Sensors*, vol. 15, no. 1, pp. 2059–2085, 2015.