sustainability

MDPI

Article

Introducing VTT-ConIot: A Realistic Dataset for Activity Recognition of Construction Workers Using IMU Devices

Satu-Marja Mäkela [1], Arttu Lämsä [1], Janne S. Keränen [1], Jussi Liikka [1], Jussi Ronkainen [1], Johannes Peltola [1], Juha Häikiö [1], Sari Järvinen [1] and Miguel Bordallo López [1,2,*]

[1] VTT Technical Research Centre of Finland Ltd., 90150 Oulu, Finland; satu-marja.makela@vtt.fi (S.-M.M.); arttu.lamsa@vtt.fi (A.L.); janne.s.keranen@vtt.fi (J.S.K.); jussi.liikka@vtt.fi (J.L.); jussi.ronkainen@vtt.fi (J.R.); johannes.peltola@vtt.fi (J.P.); juha.haikio@vtt.fi (J.H.); sari.jarvinen@vtt.fi (S.J.)

[2] Center for Machine Vision and Signal Analysis, University of Oulu, 90017 Oulu, Finland

* Correspondence: miguel.bordallo@vtt.fi

Abstract: Sustainable work aims at improving working conditions to allow workers to effectively extend their working life. In this context, occupational safety and well-being are major concerns, especially in labor-intensive fields, such as construction-related work. Internet of Things and wearable sensors provide for unobtrusive technology that could enhance safety using human activity recognition techniques, and has the potential of improving work conditions and health. However, the research community lacks commonly used standard datasets that provide for realistic and variating activities from multiple users. In this article, our contributions are threefold. First, we present VTT-ConIoT, a new publicly available dataset for the evaluation of HAR from inertial sensors in professional construction settings. The dataset, which contains data from 13 users and 16 different activities, is collected from three different wearable sensor locations.Second, we provide a benchmark baseline for human activity recognition that shows a classification accuracy of up to 89% for a six class setup and up to 78% for a sixteen class more granular one. Finally, we show an analysis of the representativity and usefulness of the dataset by comparing it with data collected in a pilot study made in a real construction environment with real workers.

Dataset License: CC-By 4.0

Keywords: IoT; human activity recognition; construction; IMU

Citation: Mäkela, S.-M.; Lämsä, A.; Keränen, J.S.; Liikka, J.; Ronkainen, J.; Peltola, J.; Häikiö, J.; Järvinen, S.; Bordallo López, M. Introducing VTT-ConIot: A Realistic Dataset for Activity Recognition of Construction Workers Using IMU Devices. *Sustainability* **2022**, *14*, 220. https://doi.org/10.3390/su14010220

Academic Editors: Marc A. Rosen and Jaime Lloret

Received: 1 December 2021
Accepted: 23 December 2021
Published: 26 December 2021

1. Introduction

Sustainable work is defined by EUROFUND [1] as *a set of practices that aims at achieving living and working conditions that meet the needs of the workers in a durable and lasting way that does not compromise their current or future working life*. In this context, the work conditions must be transformed to eliminate factors that prevent workers from staying in or entering the workforce.

In the labor-intensive construction industry, work safety and well-being are major concerns to achieve truly sustainable work. Some of the main causes are both nonfatal and fatal accidents at the construction work site [2], and musculoskeletal disorders that decrease the workers' ability to work effectively [3]. Their effects are very noticeable and result in prolonged absences, and even premature retirement.

In addition to the reduced well-being of individual employees, the negative effects can be significant for both the employer and society. The current status has a clear impact on the sustainability of the work at both economic and social levels. The cost of accidents in construction work is hundreds of billions of euros annually worldwide [4]. For example, the U.K. economic costs of workplace injuries and new cases of work-related ill health were estimated to be about GBP 1.2 billion in 2018/2019 [5].

The United Nations has defined two relevant sustainable development goals: (1) *to ensure healthy lives and promote well-being for all at all ages*, and (2) *to promote sustained,*

inclusive and sustainable economic growth, full and productive employment, and decent work for all [6]. Correspondingly, employee well-being and safety has emerged as a strategic priority in all organizations facing ongoing demographic and technological changes [7]. Healthy, skilled, and motivated employees are seen as the most important capital of construction company organizations in the increasingly fierce global competition and pace of work, and with extended working life [8]. To maintain sustainable work, organizations invest in various occupational safety programs, but such programs lack the capabilities to support and assess employees individually and measure the impact of their guidelines and recommendations [9].

It is suggested that most of the causes for absence are directly related to the activities of the workers and their recommended way of performing them [10,11]. However, the adoption of complex ergonomic solutions at construction sites requires long times and the involvement of a wide range of different stakeholders [12]. Hence, these solutions are not always easily deployed. Instead of periodic surveys and post-mortem reports, construction organizations are looking for well-accepted solutions to continuously measure employee safety and ergonomics, to minimize health risks, and to avoid adverse outcomes [13].

Novel sensor-based well-being technologies could facilitate the accurate assessment of employees' activities and ergonomics in real time. However, the severe effects on occupational safety and well-being and its remaining challenges call for automatic detection and monitoring methods. In this context, the rise of the Internet of Things (IoT) devices and wearable technologies provides for a relatively unobtrusive technology that could enhance safety by helping in the monitoring of the workers' activities and their adherence to the recommended ergonomic and risk-avoidance guidelines. In addition, the analysis of abnormal ways of performing activities, such as walking very slowly or around objects, could provide insight on the conditions of the construction site, allowing taking important measures that mitigate safety risks [11].

Human activity recognition (HAR) from wearable sensor data is a field of research with numerous applications, and their solutions can be applied in professional settings, such as construction work. HAR is mainly concerned with identifying movements or activities in relatively unconstrained environments, using data obtained from sensors worn by users while performing different activities. However, HAR in professional contexts is a challenging problem due to the lack of available datasets that are representative enough of the problem to study. The particularities of construction sites, which are highly regulated environments that change at a fast pace, are additional challenges for automatic HAR deployment and data collection. Our work aims at mitigating this shortcoming by presenting the first publicly available dataset for human activity recognition from wearable sensors specifically designed for the evaluation of activities related to construction work.

The key contributions of our work are threefold:

- A new multi-sensor and multi-modal dataset (the dataset is preliminarily available at Zenodo [14] https://zenodo.org/record/4683703 (accessed on 22 December 2021) collected in controlled conditions depicting realistic construction activities and designed in consensus with building construction relevant partners. The dataset is designed, employing a protocol directly usable to collect activity data from construction sites. The setup is based on IoT devices integrated in real work clothing, and complies with the needs and regulations of the construction sector.
- A baseline benchmark evaluation of HAR in the context of construction site activities, depicting a six-class setup with general tasks and a more granular sixteen-class setup with particular activities in both recommended and not recommended variants.
- A comparative analysis of the collected in-lab data with data collected in a pilot study made with real construction workers in a real construction environment that uses exactly the same setup as the provided dataset. The comparison proves the feasibility of recognizing construction site activities, including potentially dangerous ones from IoT inertial sensor data. The pilot study depicts the differences in the data

particularities related to the deployment of the system in real-world scenarios, as compared with typical in-lab experiments.

The rest of the paper is organized as follows. Section 2 describes the shortcomings of human activity recognition in construction work, depicting particularities and challenges. Section 3 summarizes the previous work in human activity recognition using IMUs and IoT devices, focusing on the few works related to professional activities and construction. A detailed description of the introduced dataset, including the technical setup, study protocol, and machine learning-based data analysis methods are detailed in Section 5. Section 6 presents baseline results on the accuracy of HAR classification on different dataset setups. A discussion of the applicability of the dataset in real scenarios, including comparative analysis of the activities in the feature space, is presented in Section 7. Finally, Section 8 concludes and summarizes the paper.

2. Motivation

Moving from conventional safety methods and tools to IoT-based data-driven safety solutions has the potential to change work safety in construction. Currently, IoT-based technologies, including sensors, predictive analytics, high capacity communication infrastructures, and cloud computing, are emerging widely in different industries. However, IoT-based systems for improving work safety in labor-intensive sectors, such as construction, present a few particular challenges. These can be traced to the complex nature of construction sites and inadequate knowledge regarding construction site-specific requirements for IoT-based solutions.

Improving occupational safety, in pursuit of more sustainable construction work, requires the collection and utilization of correct information and services in a way that helps to detect how agreed regulations and recommendations in terms of safety, security and ergonomics are followed in practice. Therefore, providing a broad view of safety conditions and their evolution in time as the building process progresses is of high importance. The required collection of information can be divided into three main types: First is the collection and real-time analysis of sensor data that are able to generate *real-time personal alarms* in case of a potentially hazardous or not recommended activity. Second is the collection and abstraction of activities for providing statistics that *anonymously track compliance* with agreed security, safety and ergonomic recommendations. Third is the use of recorded sensor data for the *forensic investigation of accidents*, allowing for a better understanding of the reasons.

Human activity recognition using inertial motion units (IMU) has the potential to accurately classify these activities of the construction workers in an automatic manner, providing continuous statistics with a granularity that currently does not exist.

However, to realistically detect and analyze the activities performed in a construction site in an accurate and useful manner, several challenging points still remain:

The nature of the activities is not clear: In the application of construction activity recognition, the activity classes—the types of activities—are defined in a domain-specific manner. Hence, specific activities are not always easy to recognize since the possible list includes large varieties, even for single classes. Simple activities, such as painting or cleaning, are composed of numerous sub-activities that are shared with other tasks, such as pushing objects, walking or going up and down stairways. Moreover, such activities have imbalance qualities, such as the number of occurrences during a day, their typical duration and starting times. Because the traditional approach normally assumes that activity classes have similar probabilities of being performed, similar probabilities any time in a day, and similar duration, the way in which accuracy changes when we consider such imbalances is not known.

The application is not clear: In the application of construction activity analysis, we can set up clear goals, such as improving activities effectively in timing and duration, or by reducing hazardous activities. For such goals, the technical objective is not only improving the recognition accuracy each time, derived from the traditional recognition of the current time window or those in the vicinity (called local time windows), but also estimating the

segment—the range where the activity is performed continuously—attached with correct timestamps and duration. Thus, by clarifying the application aspects that are of importance, we can choose the recognition aspects to which to assign importance. This is not the case with the existing work.

No existing datasets with clear or specific enough goals: To overcome the aforementioned challenges, we require real data to evaluate our input into a machine learning algorithm. However, there is an extreme shortage of such open datasets obtained from multiple subjects, and a set of activities with densely annotated labels. In the literature, there are several datasets that provide data for activity recognition, but because they do not aim at the particular application in a construction setup, it is not clear what accuracy aspects to pursue. By contrast, the few small-scale datasets that have focused on construction work only look at one particular aspect of the construction, such as body postures in particular activities.

We aim at mitigating some of these challenges with the introduction of a mid-scale dataset that depicts construction activities that are deemed to be of particular interest for sectorial workers for the potential to reduce occupational hazards, mimicking them in a laboratory setup that shows that they can be translated to real scenarios through comparative evaluation.

3. Related Work

In recent years, human activity recognition (HAR) from wearable sensor data has become a field of research with numerous applications in both personal and professional settings. HAR is mainly concerned with identifying movements or activities in relatively unconstrained environments using data obtained from sensor-based wearable devices, worn by users while performing different activities. These devices use a combination of in-device processing and cloud services to produce information about the physical activity of the user, providing them with different context-adaptive services.

In specific or semicontrolled settings, HAR from sensors has shown to offer relatively accurate performance and high utility. However, developing robust classifiers for detecting multiple activities is a challenging task that requires large amounts of labeled training data, collected for the particular context of interest.

3.1. Activity Recognition Using Inertial Sensors

The previous work in HAR based on inertial measurement units (IMUs) has mainly focused on the use of acceleration signals [15,16]. Accelerometers are lightweight [17], inexpensive [18], and widely available [19], many times integrated in consumer products [16]. Previous work in IMU-based sensors usually follows a multistep approach that consists of the aggregation and annotation of a subset of a sensor signal, the summarization of the information in the subset using different signal features and an instantaneous classification of the physical activity using machine learning [20,21].

The methodologies across applications vary, but in general, they can be divided by their learning approach (supervised and semi-supervised) and by their response time (real-time and offline) [22]. According to the feature extraction and summarization techniques, HAR systems can be based on handcraft features, or learned features. Handcraft features are arbitrarily chosen and commonly include a wide range of statistical features, frequency-based features, or specific features based on human motion models, typically denoted as physical features [23]. On the other hand, learned features are usually based on feature selection schemes from a large number of features [23], or on the application of deep learning techniques [24]. Typical classifiers include support vector machines, Gaussian mixture models, tree-based models, such as random forests, or hidden Markov models [22]. The most recent approaches integrate multiple steps in end-to-end systems [25]. In our work, we provide baseline results based on statistical features and multiple machine learning classifiers, focusing on the comparative analysis across different configurations, signal modalities and sensor locations.

3.2. Activity Recognition in Professional Contexts

Across the years, HAR has been focusing on ordinary activity recognition in outdoor and home settings recognizing basic activities, such as walking, driving, sitting or laying down). Particular attention has been given to sports contexts where the type of activity coupled with accurate timing is of high interest. However, until recently, very little attention was given to the recognition of activities in professional contexts. Special interest seems to be rising for tracking the activities of medical practitioners, such as doctors and nurses [26] and, in a smaller scale, activities such as cooking [27].

When discussing about activity recognition on construction work in particular, only a few small-scale studies are available. Joshua and Varghese [28] studied masonry activities using accelerometers in a laboratory and small-scale setting. Their study showed up to a 80% classification accuracy in relatively unconstrained environments. Akhavian and Behzadan [29], simulated, using only two subjects, a three-class construction activity setup that included sawing, hammering and turning a wrench and loading and unloading activities. Their three-class model obtained accuracies close to 90%, with high variability on users and activities. The rest of the studies related to construction focused on complementary cases without human focus, such as tracking the activity of particular equipment [30], or machinery [31].

To the best of our knowledge, no dataset for human activity recognition in construction work is currently available. In this paper, we present a novel dataset for construction-related human activity recognition. The subjects (n = 13) were instructed to perform diverse activities, while wearing sensorized clothes in a similar manner as they would in a real construction site. The dataset contains high resolution motion data from several IMU sensors and complementary human pose and keypoints obtained from a fixed camera. The data of each subject were carefully annotated following both a six-class general protocol and a more granular sixteen class protocol. The dataset is well suited to benchmark and evaluate methods for human activity recognition in professional contexts, and more specifically in construction work. A first evaluation of the dataset is presented in this paper.

4. Data Collection

This section provides details on the subjects, employed sensors, sensor placement, the study protocol, and the protocol for annotation. The study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the Ethics Committee of VTT Technical Research Center of Finland and the ConIoT project. The study was approved by the ethical board and the data security officer of our research center. Individual consent forms regarding the data collection and its possible sharing and publication conditions were collected and stored. All subjects gave their informed consent for inclusion before they participated in the study.

4.1. Participants

Based on our defined protocol, we targeted different workers in our research center. The inclusion criteria asked them to be of working age in a range from 25 to 55 years old. Exclusion criteria, stated in the study invitation, were pregnancy, serious chronic or cardiovascular diseases and mental disorders. We aimed at collecting data from 13 users for 16 different activities performed for a duration of one minute each. Of them, 10 were men and 3 were women, a gender distribution similar to what can be seen in typical construction sites in Europe. Due to a camera malfunction, the data of one extra subject were not complete and we decided not to include this user's data in the final dataset. One other subject decided to not participate in the activities related to floor work. Except for these exceptions, the data collected are complete and have no missing parts.

4.2. Sensor Setup and Placement

This dataset consists of sessions where 13 persons performed 16 different construction work-related activities for one minute in an "in-lab" setup that mimics the activities observed

at a construction site. For the data collection, and as the results of surveys and interviews made with several sectorial actors [32], we decided to create a setup where the subjects were wearing sensorized clothes that incorporated three IMU sensors. The location of the sensors was selected in agreement with the construction workers and managers, and selected in a manner that would not impede any of their typical activities. The final setup consisted of one sensor located in the hip, and two additional ones located near the shoulder of the nondominant hand, more specifically on the upper part of the arm and on the back part of the shoulder. A depiction of the sensorized clothes and the sensor location can be seen in Figure 1. The clothes were specifically designed to hold this particular set of sensors, minimizing undesired wobbling or oscillations, although some minor ones could occur during activities with fast and sudden movement, such as jumping.

Figure 1. Depiction of the sensor locations. From left to right: planned locations, depiction of the actual sensors and example setup in the working clothes.

The sensors provided inertial movement data at various sampling rates. More concretely, we used Aistin iProxoxi sensors that integrate a 10-DOF IMU composed of a 3-axis accelerometer, a 3-axis gyroscope, a 3-axis magnetometer and a barometer. The hardware offered fixed sampling rates: the barometer sampled at 0.033 Hz, the gyroscope and magnetometer sampled at 97 Hz, and the accelerometer samples at 103 Hz. Although wireless transmission of the data is possible, for our setup, we decided to save the data in the on-device flash memory and to extract them after the recording session using a USB connection, minimizing possible problems related to interrupted connectivity.

4.3. Complementary Video Information and Pose Extraction

Each session was recorded, for reference, with a standard off-the-shelf video camera (full-HD 720p resolution, 25 fps), standing on a tripod in the same three locations. The locations were the same for activities 1 to 6, for activities 7 to 12, and for activities 13 to 15. Activity 16 was not recorded since it depicted the subject going up and down several flights of stairs.

Since video data are personal and privacy sensitive, we decided to preprocess them and incorporate in the final dataset only the human poses and body keypoints extracted from them. In this context, the video preprocessing starts by first extracting the human poses from each individual frame using a state-of-the-art human pose detector, in this case, a pretrained version of Detectron2 [33]. This stage provides us with poses composed by the coordinates of 17 keypoints that represent different articulations in the human body. The keypoints are extracted according to the typical format as provided in the Common Objects in Context (COCO) dataset and protocol [34]. These poses are directly offered in the dataset, and its visualization provides a simple way of observing the performed human activity in a visual way.

4.4. Study Protocols

The goal of the study was to reproduce as accurately as possible different construction activities that are relevant for construction safety and that depict realistically the activities performed at a construction site. In this context, we identified, in collaboration with sectorial actors, 6 of the most important and typical tasks performed at a construction site. Each one of the 6 tasks includes 2 or 3 activities for a total of 16. Among them, there is also

the depiction of a few activities that could be relatively usual in casual or nonprofessional settings but that are not recommended in the construction site due to poor ergonomic or safety concerns.

Detailed descriptions of the 16 activities are as follows:

(1) Roll painting: a subject uses a paint roll on a wall; *(2) Spraying paint:* a subject uses a tube that mimics a machine to perform movements depicting the spraying of paint on a wall; *(3) Leveling paint:* a subject uses a tool to mimic the spreading of screed or paint on a wall; *(4) Vacuum cleaning:* a subject uses a vacuum cleaner on the floor; *(5) Picking objects:* a subject picks objects from the floor with their hands and throws them into a bin; *(6) Climbing stairs:* a subject goes up 3 steps on a staircase, turns around and goes down 3 steps; *(7) Jumping down:* a subject goes up 3 steps on a staircase, turns around and jumps down the 3 steps; *(8) Laying back:* a subject mimics working with their hands up while laying back on a mid-level surface; *(9) HandsUp high:* a subject mimics working on tubes with their hands up high above the head; *(10) HandsUp low:* a subject mimics working on tubes with their hands up at the head or shoulder level; *(11) Crouch floor:* a subject works on the floor, placing tiles while crouching; *(12) Kneel floor:* a subject works on the floor placing tiles while kneeling; *(13) Pushing cart:* a subject walks on a corridor 20 m pushing a cart, turns around and pushes it back; *(14) Walk straight:* a subject walks straight on a corridor 20 m, turns around and walks back; *(15) Walk winding:* a subject walks winding around 7 cones, then 20 m, turns around, and walks back; *(16) Stairs up-down:* a subject climbs up stairs for 30 s, turns around and climbs them back down.

A depiction of the first 15 activities as they were recorded by the complementary video data can be seen in Figure 2. Activities numbered 7, 9, 12 and 15 are examples of not recommended activities at a construction site due to poor ergonomy or safety risks.

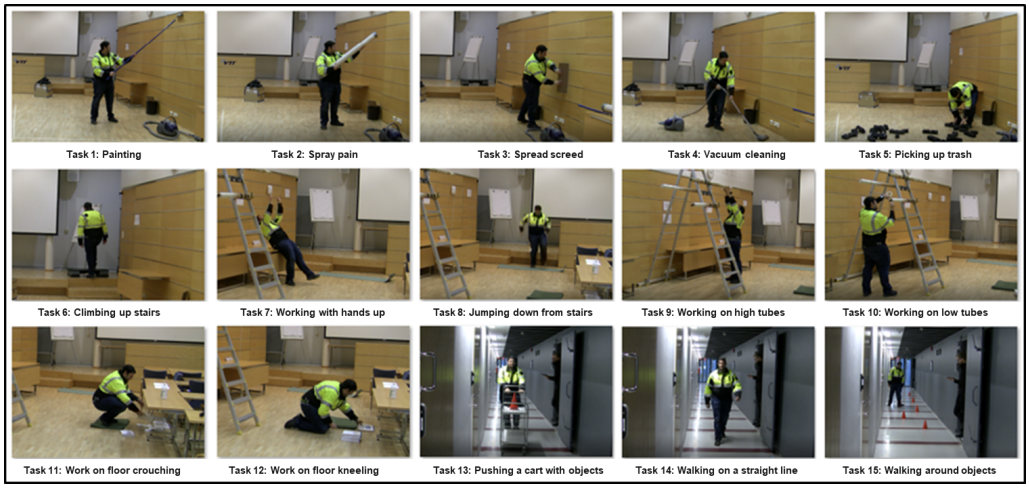

Figure 2. Example of the 16-activity setup from the VTT-ConIot dataset. Activity 16 (stairs) not depicted.

The six different tasks that group activities according to their tasks are as follows: **Painting**, that includes Roll-Painting, Spraying-Paint, and Leveling-Paint; **Cleaning**, that includes Vacuum-cleaning and Picking-objects; **Climbing**, that includes, *Climbing-stairs*, *Jumping-down* and *Stairs-Up-Down*; **HandsUp**, that includes *Laying-back*, *HandsUp-high* and *HandsUp-low*; **FloorWork** that includes *Crouch-floor* and *Kneel-floor*; and **WalkingDisplacements** that includes *Walk-straight*, *Walk-winding* and *Pushing-cart*. This setup corresponds to the activity classification in 6 classes.

Finally, in order to provide for measurement that could improve work well-being and reduce accidents, the tasks are also binary labeled based on their recommendability for the professional and casual workers in terms of ergonomics or safety. Non-recommended activities include *Jumping down* due to potential risks of falling, *Laying back*, *HandsUp high* and *Kneel floor* due to ergonomics and *Walk winding*, which could be indicative of

unwanted obstacles or dangerous features in the walking path. The rest of the activities were considered recommended.

4.5. Subject Training and Annotation

When arriving to the study setup, each subject was instructed briefly on the type of activity they must perform by showing them a few example images and videos of workers performing the activity. We collected the data from the 16 tasks in a sequential manner for each one of the subjects (mock workers) by following a carefully designed protocol and scheme, as depicted in Figure 3. This protocol simplified the annotation of the activities since we just needed a separate device recording accurate timestamps where the subject started the activities as instructed. The complementary timestamped video data allowed for possible corrections on the annotations made a posteriori in case of possible errors.

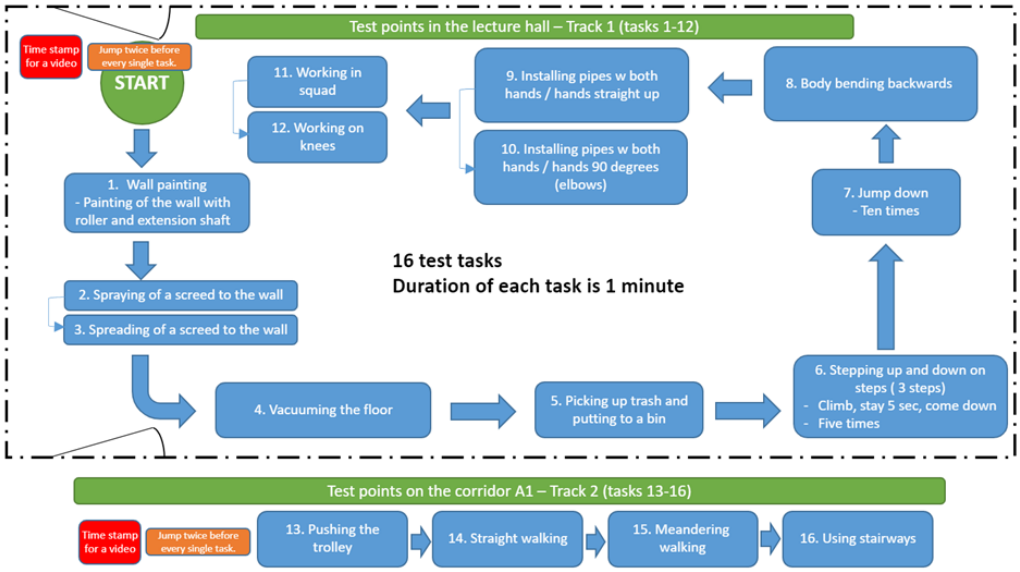

Figure 3. The data collection protocol, designed so the users can perform activities in a sequential manner.

5. Methods

5.1. Preprocessing the Datasets

VTT-ConIoT raw sensor data were preprocessed to ensure the quality and coherence among different measurements. In this context, we first applied a check of the sensor scales and orientations of the sensor inside the pocket (4 different orientations are possible, with 2 of them being much more likely). This ensures that the data format is coincident for all sensors and recording sessions. Based on the direction of the gravity vector for a static person and sensor, as seen in the IMU signal and its value, all sensors readings were set so that they reflect the same axis and dynamic ranges. This checkup is specially relevant for the data collected in the real construction site.

Due to the nature of the hardware, the clocks of the sensors were not accurately synchronized between the sensors located in different parts of the body, and the sampling rates among different signal modalities were not identical. We manually adjusted the clock offset between different sensors using timestamps and synchronization signals. We resampled the signals using linear interpolation, aiming at synchronizing the data offered by the magnetometer and gyroscope (97 Hz) to the accelerometer data (103 Hz), by matching both signals to the higher sampling rate. Although this could have some small effects during classification, especially when using frequential features, we argue that the effect is minimal for signals with a similar sampling rate that is significantly higher than the important frequential components of human activity recognition, considered to be well below 6 Hz. Since the IMU signals were used directly as acceleration vectors and not

integrated further to obtain absolute positions, the possible noise has only instantaneous properties and is not accumulated further as error.

The files recorded with the IMU sensors were continuous recordings that included data from all the subjects participating in the data collection. Hence, the data files needed to be split into separate chunks per user and per activity. The activities depicted in the VTT dataset all start with a control signal that consists of two consecutive jumps, a feature that that could be easily seen in the IMU signals as two spikes and also very clearly in the videos. Using these spikes, we synchronized the signals for each activity and user and removed the signal windows containing such jumps.

Splitting per user and per activity was done using a developed GUI tool that used the annotations made during the recording of the activities as an additional input. Using the GUI, the annotations were then fine-tuned to match the exact start time positions of each activity segment. Based on these fine-tuned annotations, the data were split into CSV files depicting 1-min activities for each user (e.g., activity-2-user-10.csv).

We computed the L2 norm of the 3 axes of the accelerometer and included it in the set as a separate signal since it showed discriminatory capabilities in HAR. All resulting 7 signals were also low-pass filtered with a cutoff frequency of 25 Hz. This bandwidth reduction reduces high frequency noise and was shown to not affect the HAR classification process since human activities do not present frequencies over 20 Hz, while most discriminative components are well below 6Hz [35,36].

5.2. Segmentation and Feature Extraction

The segmentation of each preprocessed 1-minute signal is done by sliding 2-, 5- or 10-s windows, with a sliding shift of one second. That means that two consecutive windows overlap in all signals except one second. Each one of these windows is used as one "activity sample" directly for both the training and test sets, depending on the setup. The window sizes are selected based on their broad application in acceleration- and IMU-based context recognition. The literature usually recommends window sizes from 5 to 10 s, but shows feasible recognition with as little as 2 s of sensor data [37,38].

5.3. Feature Extraction

To feed the classification algorithms, we compute individual features for each one of the activity samples, i.e., for each 2-, 5- or 10-s signal segment. We select a set of seven well-known statistical features: average *(avg)*, median *(med)*, variance *(var)*, 25th percentile *(lq)*, 75th percentile *(uq)*, minimum *(min)* and maximum *(avg)*. We compute each feature separately in each different axis (*x, y, z* for both the gyroscope and the accelerometer. In addition, we compute the features also in the signal depicting the total acceleration *(tot_acc)*, defined as the L2 norm of the values of the three axes of the accelerometer) and name the resulting 4-axis signals as *(all_acc)*. These features are roughly based on the prior knowledge of their usability in human activity.

We compute the features separately for each one of the three sensors. This results in setups that range from 7 features per sample for the simpler case (total acceleration in one sensor) to 147 features per sample in the most complex one (all 7 signals for all three sensors).

Although we are aware that other features are possible and might be more discriminative, our feature set is selected by keeping the computational limitations of wearable devices in mind, ensuring that the classifier inference can be performed on-device and in real time. In addition, as the first baseline results for the VTT-ConIoT dataset, these simple features and their associated results are easy to reproduce.

5.4. Classification Algorithms

The previously described extracted features are directly used in the input of machine learning-based classifiers. We decided to employ six different classifiers, following its standard implementation in Python scikit-learn. The classifiers were random forest (RF), extra trees (ET), XGBoost (XGB), linear discriminant analysis (LDA), support vector machines

(SVM) and logistic regression (LR). The selection of the classifiers was based on a mix of examples in the literature, the variety of their nature, and performance analysis (i.e., running a few setups using a larger number of regressors, and checking their performance). The selection of parameters was based on a coarse grid search and heuristics. Varying the parameters did not significantly increase the performance of the classifiers and yielded very similar results. Thus, we decided to use the standard configuration, for the sake of reproducibility. XGB, RF and ET were used with 100 estimators, with a not-defined max depth and a minimum sample split of 2. LDA was used employing the SVD solver with no shrinkage, while SVC used a penalty value c equal to 1.0 and a RBF kernel.

5.5. Evaluation Metrics

As the main evaluation metrics, we use the mean accuracy of the classification. In particular, accuracy represents the number of correctly classified instances out of all samples. Since for the more complex case, 16-class classification, the number of samples per class is approximately the same, we believe it is a simple metric that is representative of the task at hand. We evaluate all our models using a cross-validation scheme based on LOSO, where for N subjects, N different models are trained. Each model is trained using data from all the subjects except one, which is then used for testing and computing the performance of the model. Thirteen models per classifier and modality are evaluated. Although sustainable work calls for considerations for individual workers, we show that the averaged results demonstrate the generalization capabilities of the model to unseen "new" subjects, which were not previously modeled. As a metric to depict the variability across different subjects, for each classification task, we compute and show the standard deviation of the accuracy across models tested in different subjects.

6. Results

This section provides, first, a simple protocol and experiment to provide for a simple baseline to compare more sophisticated results. Then, detailed results on the evaluation of the different sensor locations, signals obtained from the sensors and different classifiers are depicted. A comparison between 6-class and 16-class classification is shown. For the results shown in here, we only consider the signals obtained from the accelerometer and the gyroscope since the classification using just raw magnetometer signals could be biased due to the fixed location and orientation of the task setup. In this particular evaluation, the barometer data are also discarded since it only collects a sample of 30 s each, so their results might be distorted due to the placement in a pocket, and since it offers no meaningful information for short time signal windows. We provide comparisons of different methods and classifiers, which are depicted in this novel database for the first time.

6.1. A Simple Baseline Protocol

In order to facilitate jump-starting with the dataset, we define a simple baseline protocol that uses a simplified approach for processing and classification and provides insight for the difficulty of the task. For this simple baseline, we utilize the study protocol defined by six classes. We divide each 1-minute signal by segmenting it into sliding 2-second windows that are used as "activity samples" in both training and test sets. The validation is made using a LOSO approach.

In this baseline, we include only the features provided by acceleration signals from one single sensor, placed on the heap (3-axis accelerometer, x, y, z and total acceleration tot). We compute, for each signal, the above-described 7 different statistical features, and use them to train only one classifier, in this case, random forest in its standard configuration as provided by Python *sklearn*. We provide the results in terms of mean accuracy, using LOSO validation. We detail the results for each one of the subjects and the average of all of them in Table 1. The results show that the classification of the activities in six classes using only short samples is possible and yields results that are significantly better than random guessing. More specifically, the classifier achieves a mean accuracy of 52% compared with the 17% expected accuracies when using just random guessing. However,

classifying this dataset with a simple approach is not a trivial task since misclassifications in this simple setup occur in almost half of the cases. Additionally, when analyzing the results across different subjects, we can see that the performance varies from 39% to 62%, showing the variability among different persons and their different styles for performing the activities. In the rest of the section, we use these results as the simple baseline for comparing and improving.

Table 1. LOSO (per-subject) validation results of random forest classifier in a 6-class simple setup.

Subject	Accuracy (tot_acc)
S1	0.495
S2	0.461
S3	0.415
S4	0.466
S5	0.493
S6	0.518
S7	0.623
S8	0.567
S9	0.392
S10	0.589
S11	0.560
S12	0.582
S13	0.621
Mean (std)	**0.521 (0.075)**
Random guess	*0.167*

6.2. Evaluation of the Sensor Locations, Modalities, and Extracted Features

Based on the construction-related tasks and activities defined in the protocols, we distinguish two classification tasks. First, a six-class problem is defined, where the main task groups related to construction are discerned. The results of this classification task are presented in Table 2. Second, a 16-class classification task is defined by further separating the six-class problem into subtasks, again according to the protocol. The results of this classification task are presented in Table 3. For both classification tasks, 16 different sensor and modality combinations are evaluated. The combinations are depicted as four different signals *(i.e., acceleration, total acceleration, gyroscope and all-combined)* in four different sensor location configurations *(i.e., hip, back, shoulder and all-combined)*. Each sensor and modality configuration is evaluated using the six machine learning classifiers specified previously. The validation is made using a LOSO approach, and for each configuration, we report the mean accuracy and (in brackets) the standard deviation across the 13 different subjects. We report these detailed results for the configuration using 5 s windows, as it provides better accuracy with a reasonable response time. In addition, we report the results of training an RF ensemble of classifiers, using as inputs the classification results of different modalities.

Table 2. Evaluation of the given modalities, sensor locations, and classifiers on a 6-class setup for 5 s windows. The results show the mean accuracy and standard deviation among different subjects in a leave-one-subject-out validation scheme. Abbreviations: RF—random forest, ET—extra trees, LDA—linear discriminant analysis, LR—logistic regression, SVM—support vector machine, XGB—XGBoost extreme gradient boosting.

Position	Modality	RF	ET	LDA	LR	SVM	XGB
hip	acc	0.52 ± (0.08)	0.50 ± (0.09)	0.60 ± (0.07)	**0.62** ± (0.06)	0.56 ± (0.12)	0.52 ± (0.09)
hip	all_acc	0.57 ± (0.08)	0.55 ± (0.08)	0.62 ± (0.07)	**0.64** ± (0.05)	0.60 ± (0.07)	0.57 ± (0.08)
hip	gyro	0.69 ± (0.09)	**0.70** ± (0.09)	0.56 ± (0.07)	0.59 ± (0.08)	0.67 ± (0.08)	0.68 ± (0.08)
hip	all	0.69 ± (0.08)	0.70 ± (0.05)	0.65 ± (0.06)	0.68 ± (0.07)	0.69 ± (0.08)	**0.71** ± (0.08)
back	acc	0.76 ± (0.07)	0.75 ± (0.06)	0.69 ± (0.06)	0.71 ± (0.06)	**0.78** ± (0.05)	0.76 ± (0.06)
back	all_acc	0.80 ± (0.07)	0.80 ± (0.06)	0.72 ± (0.07)	0.77 ± (0.05)	**0.82** ± (0.04)	0.80 ± (0.06)
back	gyro	0.58 ± (0.05)	**0.58** ± (0.04)	0.46 ± (0.03)	0.47 ± (0.04)	0.57 ± (0.05)	0.56 ± (0.05)
back	all	0.83 ± (0.06)	0.84 ± (0.06)	0.76 ± (0.05)	0.79 ± (0.05)	**0.85** ± (0.05)	0.84 ± (0.05)
shoulder	acc	0.69 ± (0.08)	0.68 ± (0.07)	0.68 ± (0.05)	0.70 ± (0.05)	**0.71** ± (0.07)	0.70 ± (0.08)
shoulder	all_acc	0.72 ± (0.07)	0.72 ± (0.07)	0.70 ± (0.06)	0.73 ± (0.05)	**0.76** ± (0.07)	0.73 ± (0.07)
shoulder	gyro	0.59 ± (0.05)	**0.60** ± (0.05)	0.42 ± (0.05)	0.43 ± (0.05)	0.58 ± (0.06)	0.57 ± (0.06)
shoulder	all	0.77 ± (0.08)	0.78 ± (0.07)	0.73 ± (0.04)	0.75 ± (0.05)	**0.80** ± (0.06)	0.79 ± (0.07)
all	acc	0.84 ± (0.07)	0.83 ± (0.07)	0.83 ± (0.04)	0.82 ± (0.06)	**0.84** ± (0.06)	0.82 ± (0.07)
all	all_acc	0.86 ± (0.06)	0.85 ± (0.06)	0.84 ± (0.04)	0.84 ± (0.07)	**0.86** ± (0.05)	0.85 ± (0.06)
all	gyro	0.79 ± (0.07)	**0.79** ± (0.06)	0.67 ± (0.06)	0.68 ± (0.07)	0.77 ± (0.06)	0.80 ± (0.07)
all	all	0.88 ± (0.06)	0.89 ± (0.06)	0.86 ± (0.05)	0.84 ± (0.08)	**0.89** ± (0.06)	0.89 ± (0.04)
ensemble	all	0.81 ± (0.07)	0.82 ± (0.06)	0.83 ± (0.04)	**0.84** ± (0.05)	0.80 ± (0.05)	0.80 ± (0.06)

Accuracy of random guessing: 0.17.

Table 3. Evaluation of the given modalities, sensor locations, and classifiers on a 16-class setup for 5 s windows. The results show the mean accuracy and standard deviation among different subjects in a leave-one-subject-out validation scheme.

Position	Modality	RF	ET	LDA	LR	SVM	XGB
hip	acc	0.34 ± (0.08)	0.33 ± (0.08)	0.41 ± (0.07)	**0.44** ± (0.07)	0.36 ± (0.09)	0.33 ± (0.08)
hip	all_acc	0.38 ± (0.09)	0.36 ± (0.07)	0.43 ± (0.07)	**0.46** ± (0.07)	0.41 ± (0.09)	0.38 ± (0.08)
hip	gyro	0.51 ± (0.08)	**0.51** ± (0.07)	0.38 ± (0.07)	0.41 ± (0.07)	0.48 ± (0.08)	0.50 ± (0.08)
hip	all	**0.55** ± (0.10)	0.53 ± (0.09)	0.50 ± (0.08)	0.53 ± (0.09)	0.53 ± (0.09)	0.53 ± (0.10)
back	acc	0.53 ± (0.08)	0.51 ± (0.08)	0.50 ± (0.06)	0.52 ± (0.05)	**0.58** ± (0.08)	0.54 ± (0.08)
back	all_acc	0.60 ± (0.10)	0.59 ± (0.08)	0.53 ± (0.05)	0.59 ± (0.04)	**0.62** ± (0.08)	0.59 ± (0.10)
back	gyro	0.44 ± (0.05)	**0.44** ± (0.05)	0.31 ± (0.07)	0.35 ± (0.06)	0.43 ± (0.06)	0.43 ± (0.04)
back	all	0.67 ± (0.09)	0.67 ± (0.09)	0.60 ± (0.07)	0.63 ± (0.08)	**0.67** ± (0.08)	0.67 ± (0.08)
shoulder	acc	0.54 ± (0.07)	0.53 ± (0.07)	0.50 ± (0.08)	**0.55** ± (0.07)	0.54 ± (0.08)	0.54 ± (0.06)
shoulder	all_acc	0.58 ± (0.07)	0.58 ± (0.07)	0.53 ± (0.07)	0.57 ± (0.06)	**0.59** ± (0.08)	0.57 ± (0.06)
shoulder	gyro	0.45 ± (0.06)	**0.46** ± (0.06)	0.33 ± (0.05)	0.38 ± (0.05)	0.44 ± (0.06)	0.43 ± (0.06)
shoulder	all	0.65 ± (0.10)	0.65 ± (0.10)	0.61 ± (0.06)	0.63 ± (0.07)	**0.67** ± (0.09)	0.66 ± (0.09)
all	acc	0.68 ± (0.10)	0.68 ± (0.09)	0.65 ± (0.07)	0.64 ± (0.08)	**0.68** ± (0.07)	0.68 ± (0.09)
all	all_acc	0.70 ± (0.09)	0.71 ± (0.08)	0.67 ± (0.06)	0.67 ± (0.09)	**0.72** ± (0.07)	0.71 ± (0.08)
all	gyro	0.67 ± (0.08)	**0.67** ± (0.06)	0.55 ± (0.07)	0.59 ± (0.06)	0.64 ± (0.06)	0.66 ± (0.06)
all	all	0.77 ± (0.08)	0.77 ± (0.08)	0.73 ± (0.08)	0.71 ± (0.09)	**0.78** ± (0.08)	0.76 ± (0.08)
ensemble	all	0.67 ± (0.08)	0.68 ± (0.08)	0.70 ± (0.07)	**0.71** ± (0.07)	0.67 ± (0.07)	0.66 ± (0.07)

Accuracy of random guessing: 0.06.

The data considered in this paper, belonging to periods where the users were performing the activities of interest, amount to approximately 16 min per subject for a total

of 208 min of data. In this case, we divide each 1 min signal by segmenting it into sliding 5 s windows that are used as "activity samples" in both training and test sets. Using a sliding period of one second, this amounts to approximately 12,700 windows, uniformly distributed along the 16 tasks/activities.

This configuration would yield a random guessing accuracy of about 6% for the 16-class problem and about 17% for the 6 class problem. Compared to random guessing, the best classification accuracies are 78% for the 16-class problem and 89% for the 6-class problem, showing that an accurate classification of the activities is possible with a much better accuracy than random chance or simple setups. The multi-signal, longer window configuration shows also a much expected increased accuracy when compared with the simple setup described above.

When comparing the performance of the different classifiers, it becomes apparent that a support vector machine-based classifier shows superior performance for most of the cases, including those that combine more information. The extra trees classifier seems to obtain relatively better accuracies on the analysis of the gyroscope data. The logistic regression-based classifier shows comparable performance to SVM and seems to be especially good for the hip sensor location.

When observing sensor locations, it can be seen that overall, the sensor placed on the back offers the best individual performance for both classification configurations (6 and 16 classes), while the sensor on the hip shows the worst performance. This suggest that activity recognition in a construction context is more discernible by the movement occurring in the upper part of the body. The combination of several sensors and locations results in increased accuracy in all cases, suggesting that different sensor locations provide indeed complementary information.

When observing the different individual sensor modalities, we can see that the total acceleration, defined as a nonlinear combination of the acceleration signals, shows the best individual modality performance in a manner that is very consistent across different sensor locations. This observation is in line with previous results in the literature that speculated that this particular transformation reduces the impact of variant sensor positioning. The gyroscope alone shows better accuracy than the acceleration-based signals for the hip sensor, while being significantly worse for both the back and shoulder sensors. In any case, the combination of signals shows, as expected, increased accuracy, suggesting that the information provided by each modality is indeed complementary and not purely redundant.

It is interesting to notice that the best results in terms of sensor location, signal modalities, and ML classifiers are very consistent for both 6- and 16-class configurations, while the variability across different subjects is also relatively low and does not increase significantly with an added number of tasks or with higher accuracies.

To provide an insight of the type of activities that are more often misclassified, we present classification matrices for both setups. The matrices, shown in in Tables 4 and 5, are computed for the best classifier (SVM) in the best sensor and signal configuration (all signals and all sensors combined).

Table 4. Classification matrix for a 6-class setup using SVM classifier and all sensor locations and signal modalities.

		Painting	**Cleaning**	**Climbing**	**Handsup**	**Floorwork**	**Walking**
True Class	Painting	1750	105	76	172	34	8
	Cleaning	171	1172	18	2	39	28
	Climbing	75	26	1761	43	0	175
	Handsup	236	2	68	1828	1	10
	Floorwork	25	42	20	5	1280	3
	Walking	23	43	236	4	0	1774

Predicted class.

Table 5. Classification matrix for a 16-class setup using SVM classifier and all sensor locations and signal modalities. Class description: C1—Roll painting, C2—Spraying paint, C3—Levelling paint, C4—Vacuum cleaning, C5—Picking objects, C6—Climbing stairs, C7—Jumping down, C8—Laying back, C9—HandsUp high, C10—HandsUp low, C11—Crouch floor, C12—Kneel floor, C13—Pushing cart, C14—Walk straight, C15—Walk winding, C16—Stairs up-down.

True class	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16
C1	618	46	28	9	0	1	8	1	1	0	0	0	0	0	0	3
C2	44	488	13	44	2	65	4	1	0	48	0	2	2	0	2	0
C3	48	52	492	40	11	0	3	0	34	27	1	7	0	0	0	0
C4	33	17	34	597	30	0	0	0	0	0	0	3	1	0	0	0
C5	0	1	5	10	651	10	4	1	0	1	5	8	5	0	12	1
C6	3	21	1	0	10	626	21	0	0	5	0	0	15	9	2	4
C7	0	0	0	0	1	26	673	0	0	0	0	0	4	2	4	5
C8	8	11	4	1	7	14	3	559	76	22	7	0	0	0	0	2
C9	5	10	15	1	2	10	3	125	516	77	0	0	3	0	2	2
C10	33	59	56	0	0	4	4	69	120	412	0	0	5	0	0	0
C11	1	1	4	2	4	11	2	22	0	0	506	112	2	0	0	13
C12	0	1	7	7	11	0	0	2	0	0	126	563	0	0	0	0
C13	0	8	4	2	11	9	8	0	0	0	0	0	665	108	0	1
C14	0	0	0	0	13	9	4	0	0	0	0	0	85	394	100	171
C15	0	0	0	0	7	7	15	0	0	0	0	0	1	44	517	40
C16	0	2	0	0	5	44	36	0	0	0	0	0	21	43	20	479

Predicted class.

The classification matrix of the 6-class configuration shows that although most of the classes are recognized properly in the majority of cases, a few class pairs are more prone to confusion. Painting and cleaning are relatively often misclassified. We argue that this is due to the similar repetitive patterns shown in both activities, with similar periodic movements. In a similar fashion, painting and working with the hands up are also misclassified often, and we argue that it is most likely due to the similar arm–hand positions when performing, for example, painting with the roll and working on the mid- and high tubes. The last pair that shows an abnormal classification pattern is walking and climbing. This was an expected pair since both activities focus on the displacement and show similar periodic patterns and even similar segments, where the subject was just walking normally before, e.g., a jump down. We believe that the use of more accurate information of the altitude variations of the subject in a given window time could mitigate these misclassifications.

For the 16-class configuration, the observations are relatively similar. The right activity or task is recognized correctly in the majority of cases. Most of the misclassifications occur for task pairs that belong to the same general activity. The worst cases seem to be related to discerning what types of hands-up work the subject is performing, e.g., if the floor work is performed kneeling or crouching, or among different types of walking, such as pushing a cart or walking straight. This classifier behavior is expected for activities where the subjects remain in a static position with small displacements and similar to each other, which is especially relevant for sensors placed only on the upper part of the body. This shows that recognizing activities with such a level of granularity might require specific classifiers trained with these particular classes and it is an argument for, for example, hierarchical classification models.

6.3. Evaluation of Window Size

To evaluate the effect of the window size on the classification accuracy, we train classifiers segmenting the activities in 2, 5, and 10 s windows. A summary of the results including the best classifier per setup and its accuracy is depicted in Table 6.

Table 6. LOSO validation results of the best classifier for different window sizes.

Setup	2 s Window	5 s Window	10 s Window
6-class hip	0.66 (RF)	0.71 (XGB)	0.79 (SVM)
6-class back	0.80 (SVM)	0.85 (SVM)	0.93 (SVM)
6-class shoulder	0.73 (SVM)	0.80 (SVM)	0.86 (ET)
6-class all	0.86 (SVM)	0.89 (SVM)	0.94 (SVM)
16-class hip	0.46 (RF)	0.55 (RF)	0.67 (LR)
16-class back	0.61 (SVM)	0.67 (SVM)	0.77 (ET)
16-class shoulder	0.59 (SVM)	0.67 (SVM)	0.76 (RF)
16-class all	0.72 (SVM)	0.78 (SVM)	0.84 (RF)

As expected, it can be seen that longer window sizes show better accuracies, while the results across sensor locations are consistent for all window sizes. The best performing classifiers overall are those based on SVM, but tree-based classifiers such as random forests or extra trees might offer some small advantages when dealing with longer window sizes. After all, the selection of the optimum window size is a tradeoff between higher accuracy and expected latency. While offline computations might be better by selecting longer window sizes, real-time implementations might require setting shorter windows to be able to respond on a reasonable time.

6.4. Real-Time Classification of Construction Activities

To test the feasibility of the real-time detection of construction activities, we deployed one of the models in an IoT sensor platform that includes IMU and a microcontroller unit (MCU), in this case, a RuuviTag. The platform includes a 3-axis accelerometer, Bluetooth 5 connectivity, 512 kB of memory, and an ARM Cortex M4F CPU. For demonstration purposes, we deployed a simpler model based on linear regression, and only included one sensor on the heap and the acceleration signals. The activity classification was run using 5 s windows with 1 s intervals entirely on the device, and only the results of the inference were transmitted via Bluetooth for further preprocessing. The implementation was able to compute and transmit the inference results in real time without any dropped window. The model is a direct translation with identical weights as the model described before, so it showed similar accuracy when classifying similar data (up to 65% in a 6-class model). The computation times are in the range of 100–200 ms and thus well below the sliding period of 1 s, enabling real-time operation. Figure 4 depicts one of the workers performing activities with their hands up. The obtained signal and the classification score for the particular activity are overlaid on the picture.

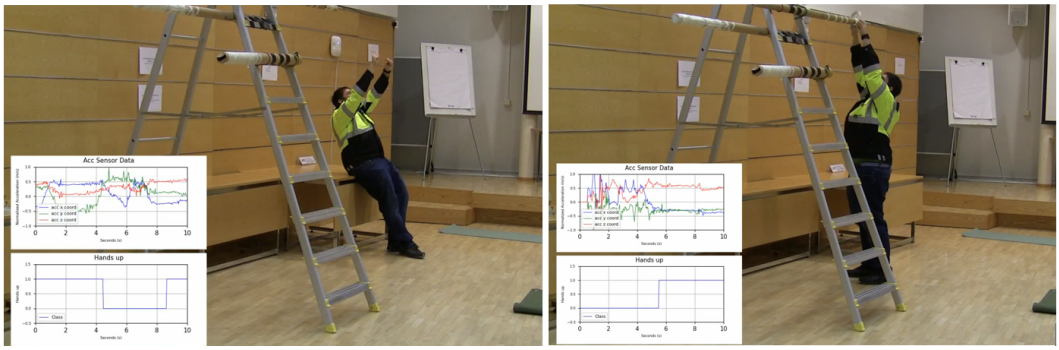

Figure 4. Example images of a worker with their hands up, synchronized signals and real-time classification.

7. Discussion on the Application to Realistic Data

When collecting the VTT-ConIoT dataset, our aim was to obtain a dataset, an evaluation protocol, and a HAR model that is indeed representative of the activities performed in the construction site. To this extent, we collected in collaboration with sectorial actors, an unannotated set of data belonging to real workers in a real construction site. In practice, we deployed a set of sensorized clothes, each including two 10-DOF IMU sensors, one placed on the hip, and the other one on the back or the shoulder depending on the particular set of clothes. The sensors were worn by 15 workers in a construction site in some periods during their regular working hours, deciding when on their own will. The sensors were available to them for four days a week, for a total of 30 work days over a period of 7 weeks. We aimed at keeping the sensors available to the workers during approximately 200 h per subject, for a maximum of 3000 h of data. As expected [32], the adherence to wearing the sensors was relatively low. At the end of the pilot period, we obtained over 300 h of data that contain activities. We assessed the validity of the collected data both algorithmically through simple signal quality analysis, ensuring that the averages, variance, and standard deviation of the activity windows are in similar ranges as the in-lab data. We then further ensured the validity by visual inspection in collaboration with domain experts, checking that the data, timing and predicted results show the realistic distributions that one could expect in that particular setting (e.g., that the hours of activity and rest corresponds to work timetables).

To prove the usefulness of the ConIoT dataset, we aimed at comparing the similarity of the data collected in our tailor-made in-lab protocol with the unannotated pilot data from the real construction site. For this, we represented VTT-ConIoT data in a two-dimensional figure by applying a transformation using uniform manifold approximation and projection for dimension reduction (UMAP) [39]. We show the representation in Figure 5, depicting, in different colors, the activity windows belonging to different classes in our dataset. Using the same transformation and projection parameters obtained with the in-lab dataset, we overlay in the figure the representations of the signal windows pertaining to the data acquired during the real construction site pilot, and depict them in black. Although the amount of in-the-wild data is much larger and it is represented in a very dense manner, the figure shows that over 80% of the activity windows belonging to the pilot are projected in the same areas as the controlled tasks belonging to the classes present in the in-lab VTT-ConIoT dataset. Approximately 20% of the activity windows are projected, in a spread manner, outside of the areas where VTT-ConIoT activities are. This is expected, as some fraction of the activities in the construction site are likely to be either not related to construction (e.g., eating, and resting) or related to construction but not related to the main tasks depicted in the dataset. This distribution suggests that our dataset is indeed representative of the majority of the activities performed in this particular construction site, and has the potential to be generalized to others.

In addition to the joint projection of the activity windows, we analyzed the distribution of the individual activities for each week of each worker. Figure 6 shows the distribution of the activities related to displacements as the hours and days progress during the week. The distribution is calculated by analyzing the real construction site data using models created solely from the in-lab VTT-ConIoT dataset. Each individual square represents 15 min of activities and has an increased brightness proportional to the ratio of activity windows that are classified as displacements, among all the activity windows available in the period. The figure shows how the windows classified as displacements occur, as expected, only during the working hours. A smaller number of displacements can be also observed in a period in the middle of the day, coinciding with the designated lunch break time. What is especially interesting to see is that the displacements are more prevalent just before the end of the work shift, which varies a bit from day to day. This representation shows that a realistic distribution of the tasks during a construction worker shift can be obtained by using only the in-lab dataset for creating the classification model.

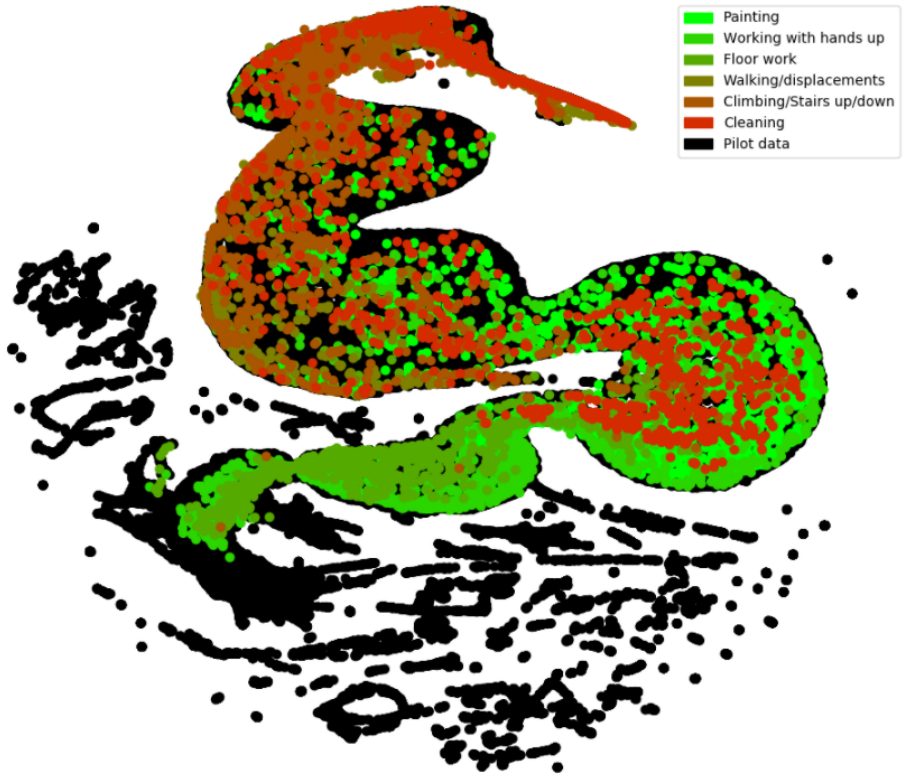

Figure 5. UMAP-based projection of both real pilot data (black) and lab dataset (colored). Most of the activity windows (>80%) of the real construction site pilot data are projected on the same areas as the controlled activities of the dataset.

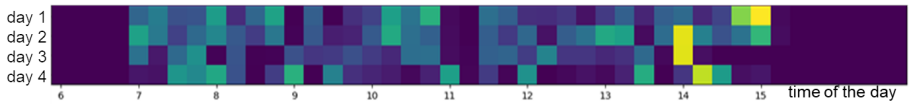

Figure 6. A distribution of the displacement activities of one real construction worker during four different days of work. Brighter squares mean a greater percentage of windows identified as the interesting activity.

8. Conclusions

We presented VTT-ConIoT, a realistic IMU-based dataset for activity recognition of construction workers, that aims at improving work safety, ergonomics and well-being by depicting activities performed in both recommended and unrecommended fashion. To the best of our knowledge this is the first sensor-based dataset that focuses on the particular activities related to construction. In contrast with other similar datasets for other diverse professional contexts, the usefulness of VTT-ConIoT was validated in a real construction site, ensuring the similarity of the activities depicted both in the dataset and in real conditions. The dataset includes data from several sensor locations and modalities.

For benchmarking, we used standard statistical features and common machine learning classifiers. On a balanced six-class setup, we achieved classification accuracies of 89%. For the more granular setup including 16 different classes, we achieved an accuracy of 78%. These results should still be interpreted with caution due to to the limitations of VTT-ConIoT regarding the number of subjects and their lack of professional construction skills. However, these effects are mitigated by the analysis of the tasks, activities and protocols that were selected by trying to ensure that they closely resemble real-world cases, in collaboration with professional construction workers and managers. The resemblance of the resulting signals is tested by statistically comparing the activities of the dataset with those collected from real workers and showing that a vast majority of the activities that pertain to a real-world setting show similar characteristics when compared to those

depicted in the dataset. This results suggest the usefulness of a sensor-based approach to achieve the recognition of recommended and unrecommended activities.

Our baseline results follow a LOSO evaluation scheme, and despite the relative simplicity of the approach, they indicate that generalization to different individuals is possible. We reported a detailed analysis of different sensor locations and modalities, showing that the best locations and modalities are consistent across settings. Our results suggest that a combination of sensors is the best choice. However, the results obtained by an individual sensor installed in the back of the shoulder are also comparable.

Further work is required to collect annotated data in real conditions on several construction site to be able to create fully personalized models. This would, in turn, allow to evaluate the effects of such a setup in long-term safety, ergonomics and well-being. Currently, the collection in real conditions is only possible for unannotated data, due to the highly regulated and changing environment that is particularly challenging for data collection [32]. It is possible that variations across sites and types of workers show very different distributions of both seen and unseen activities.

The dataset in this paper is publicly available (the dataset is preliminarily available at Zenodo [14] https://zenodo.org/record/4683703 accessed on 22 December 2021) and can be downloaded from [14] in both processed and unprocessed forms. We invite the research community to consider it for preprocessing techniques, algorithm development and benchmarking of HAR in professional contexts in the aim to improve and achieve sustainable work in the construction field.

Author Contributions: Data curation, A.L. and J.L. and J.H.; Formal analysis, A.L. and J.H. and M.B.L.; Funding acquisition, S.-M.M. and J.P. and J.H.; Methodology, M.B.L.; Project administration, S.-M.M. and J.P. and S.J. and J.H.; Resources, J.P. and S.J. Software, A.L. and J.S.K. and J.L. and J.R.; Validation, A.L. and J.R.; Visualization, J.L.; Writing—original draft, M.B.L.; Writing—review & editing, S.-M.M. and J.S.K. and J.R. and J.P. and S.J. and M.B.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Business Finland, under grant number 5432/31/2018 ConIoT project.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Ethics Committee of VTT Technical Research Centre of Finland Ltd. (November 2019).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data is avaiable doi:10.5281/zenodo.4683703 accessed on 22 December 2021.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Eurofund. Sustainable Work. Available online: https://www.eurofound.europa.eu/topic/sustainable-work (accessed on 22 December 2021).
2. Winge, S.; Albrechtsen, E. Accident types and barrier failures in the construction industry. *Saf. Sci.* **2018**, *105*, 158–166. [CrossRef]
3. Wang, X.; Dong, X.S.; Choi, S.D.; Dement, J. Work-related musculoskeletal disorders among construction workers in the United States from 1992 to 2014. *Occup. Environ. Med.* **2017**, *74*, 374–380. [CrossRef] [PubMed]
4. Yang, K.; Kim, K.; Go, S. Towards Effective Safety Cost Budgeting for Apartment Construction: A Case Study of Occupational Safety and Health Expenses in South Korea. *Sustainability* **2021**, *13*, 1335. [CrossRef]
5. Health and Safety Executive. Construction statistics in Great Britain. Available online: https://www.hse.gov.uk/Statistics/industry/construction.pdf (accessed on 11 April 2021).
6. Nations, U. Sustainable Development Goals. Available online: https://sdgs.un.org/goals (accessed on 22 December 2021).
7. Chen, W.T.; Merrett, H.C.; Huang, Y.H.; Bria, T.A.; Lin, Y.H. Exploring the Relationship between Safety Climate and Worker Safety Behavior on Building Construction Sites in Taiwan. *Sustainability* **2021**, *13*, 3326. [CrossRef]
8. Deloitte. The Rise of the Social Enterprise: 2018 Deloitte Global Human Capital Trends. Available online: https://www2.deloitte.com/content/dam/Deloitte/at/Documents/human-capital/at-deloitte-insights-the-rise-of-the-social-enterprise.pdf (accessed on 22 December 2021).

9. Hoła, B.; Nowobilski, T. Analysis of the influence of socio-economic factors on occupational safety in the construction industry. *Sustainability* **2019**, *11*, 4469. [CrossRef]
10. Kim, J.M.; Son, K.; Yum, S.G.; Ahn, S. Analyzing the risk of safety accidents: The relative risks of migrant workers in construction industry. *Sustainability* **2020**, *12*, 5430. [CrossRef]
11. Kim, J.; Youm, S.; Shan, Y.; Kim, J. Analysis of Fire Accident Factors on Construction Sites Using Web Crawling and Deep Learning Approach. *Sustainability* **2021**, *13*, 11694. [CrossRef]
12. Dale, A.M.; Jaegers, L.; Welch, L.; Barnidge, E.; Weaver, N.; Evanoff, B.A. Facilitators and barriers to the adoption of ergonomic solutions in construction. *Am. J. Ind. Med.* **2017**, *60*, 295–305. [CrossRef]
13. Park, I.; Kim, J.; Han, S.; Hyun, C. Analysis of fatal accidents and their causes in the Korean construction industry. *Sustainability* **2020**, *12*, 3120. [CrossRef]
14. Makela, S.M.; Lamsa, A.; Keranen, J.S.; Liikka, J.; Ronkainen, J.; Peltola, J.; Haikio, J.; Bordallo Lopez, M. VTT-ConIot: A realistic dataset for activity recognition of construction workers using IMU devices. *Zenodo Repos.* **2021**. [CrossRef]
15. Gupta, P.; Dallas, T. Feature Selection and Activity Recognition System Using a Single Triaxial Accelerometer. *IEEE Trans. Biomed. Eng.* **2014**, *61*, 1780–1786. [PubMed]
16. Kwapisz, J.; Weiss, G.; Moore, S. Activity Recognition Using Cell Phone Accelerometers. *SIGKDD Explor. Newsl.* **2011**, *12*, 74–82. [CrossRef]
17. Petersen, J.; Austin, D.; Sack, R.; Hayes, T. Actigraphy-Based Scratch Detection Using Logistic Regression. *IEEE J. Biomed. Health Inform.* **2013**, *17*, 277–283. [CrossRef] [PubMed]
18. Cheng, W.; Jhan, D. Triaxial Accelerometer-Based Fall Detection Method Using a Self-Constructing Cascade-AdaBoost-SVM Classifier. *IEEE J. Biomed. Health Inform.* **2013**, *17*, 411–419. [PubMed]
19. Matic, A.; Osmani, V.; Mayora, O. Speech activity detection using accelerometer. In Proceedings of the 2012 IEEE Annual International Conference on Engineering in Medicine and Biology Society, San Diego, CA, USA, 28 August–1 September 2012; pp. 2112–2115.
20. Pärkkä, J.; Ermes, M.; Korpipää, P.; Mäntyjärvi, J.; Peltola, J.; Korhonen, I. Activity Classification Using Realistic Data From Wearable Sensors. *IEEE Trans. Inf. Technol. Biomed.* **2006**, *10*, 119–128. [CrossRef] [PubMed]
21. Ermes, M.; Pärkkä, J.; Mäntyjärvi, J.; Korhonen, I. Detection of Daily Activities and Sports With Wearable Sensors in Controlled and Uncontrolled Conditions. *IEEE Trans. Inf. Technol. Biomed.* **2008**, *12*, 20–26. [CrossRef]
22. Attal, F.; Mohammed, S.; Dedabrishvili, M.; Chamroukhi, F.; Oukhellou, L.; Amirat, Y. Physical human activity recognition using wearable sensors. *Sensors* **2015**, *15*, 31314–31338. [CrossRef]
23. Zhang, M.; Sawchuk, A.A. *A Feature Selection-Based Framework for Human Activity Recognition Using Wearable Multimodal Sensors*; BodyNets: Florence, Italy, 2011; pp. 92–98.
24. Li, F.; Shirahama, K.; Nisar, M.A.; Köping, L.; Grzegorzek, M. Comparison of feature learning methods for human activity recognition using wearable sensors. *Sensors* **2018**, *18*, 679. [CrossRef] [PubMed]
25. Hassan, M.M.; Ullah, S.; Hossain, M.S.; Alelaiwi, A. An end-to-end deep learning model for human activity recognition from highly sparse body sensor data in Internet of Medical Things environment. *J. Supercomput.* **2020**, *77*, 2237–2250. [CrossRef]
26. Inoue, S.; Ueda, N.; Nohara, Y.; Nakashima, N. Mobile activity recognition for a whole day: Recognizing real nursing activities with big dataset. In Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, Umeda, Japan, 7–11 September 2015; pp. 1269–1280.
27. Alia, S.S.; Lago, P.; Takeda, S.; Adachi, K.; Benaissa, B.; Ahad, M.A.R.; Inoue, S. Summary of the cooking activity recognition challenge. In *Human Activity Recognition Challenge*; Springer: Berlin, Germany, 2021; pp. 1–13.
28. Joshua, L.; Varghese, K. Accelerometer-based activity recognition in construction. *J. Comput. Civ. Eng.* **2011**, *25*, 370–379. [CrossRef]
29. Akhavian, R.; Behzadan, A.H. Smartphone-based construction workers' activity recognition and classification. *Autom. Constr.* **2016**, *71*, 198–209. [CrossRef]
30. Rashid, K.M.; Louis, J. Times-series data augmentation and deep learning for construction equipment activity recognition. *Adv. Eng. Inform.* **2019**, *42*, 100944. [CrossRef]
31. Sherafat, B.; Ahn, C.R.; Akhavian, R.; Behzadan, A.H.; Golparvar-Fard, M.; Kim, H.; Lee, Y.C.; Rashidi, A.; Azar, E.R. Automated methods for activity recognition of construction workers and equipment: State-of-the-art review. *J. Constr. Eng. Manag.* **2020**, *146*, 03120002. [CrossRef]
32. Häikiö, J.; Kallio, J.; Mäkelä, S.M.; Keränen, J. IoT-based safety monitoring from the perspective of construction site workers. *Int. J. Occup. Environ. Saf.* **2020**, *4*, 1–14. [CrossRef]
33. Wu, Y.; Kirillov, A.; Massa, F.; Lo, W.Y.; Girshick, R. *Detectron2*. 2019. Available online: https://github.com/facebookresearch/detectron2 (accessed on 22 December 2021).
34. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft Coco: Common Objects in Context. In *Proceedings of the European Conference in Computer Vision (ECCV)*; Springer: Cham, Switzerland, 2014; pp. 740–755.
35. Kang, K.; Park, K.J.; Wang, Q.; Xu, W. Advanced technologies for mobile IoT and cyber-physical systems. *Mob. Inf. Syst.* **2016**. [CrossRef]
36. Bayat, A.; Pomplun, M.; Tran, D.A. A study on human activity recognition using accelerometer data from smartphones. *Procedia Comput. Sci.* **2014**, *34*, 450–457. [CrossRef]

37. Reiss, A.; Stricker, D. Introducing a new benchmarked dataset for activity monitoring. In Proceedings of the 2012 16th International Symposium on Wearable Computers, Newcastle, UK, 18–22 June 2012; pp. 108–109.
38. Schmidt, P.; Reiss, A.; Duerichen, R.; Marberger, C.; Van Laerhoven, K. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In Proceedings of the 20th ACM International Conference on Multimodal Interaction, Boulder, CO, USA, 16–20 October 2018; pp. 400–408.
39. McInnes, L.; Healy, J.; Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv* **2018**, arXiv:1802.03426.