Continuous Human Activity Recognition in Logistics from Inertial Sensor Data using Temporal Convolutions in CNN

Abbas Shah Syed\textsuperscript{1}  
University of Louisville, USA  
Zafi Sherhan Syed\textsuperscript{2}  
Mehran University, Pakistan  
Areez Khalil Memon\textsuperscript{3}  
University of Electronic Science and Technology of China (UESTC), P. R. China

Abstract—Human activity recognition has been an important task for the research community. With the introduction of deep learning architectures, the performance of activity recognition algorithms has improved significantly. However, most of the research in this area has focused on activity recognition for health/assisted living with other applications being given less attention. This paper considers continuous activity recognition in logistics (order picking and packing operations) using a convolutional neural network with temporal convolutions on inertial measurement sensor data from the recently released LARa dataset. Four variants of the popular CNN-IMU are experimented upon and a discussion of the results is provided. The results indicate that temporal convolutions are able to achieve satisfactory performance for some activities (hand center and cart) whereas they perform poorly for the activities of stand and hand up.

Keywords—Convolutional Neural Networks; deep learning; Human Activity Recognition (HAR); inertial sensors; LARa dataset

I. INTRODUCTION

Activity recognition has been an important task for researchers in the field of gaming\textsuperscript{11}, assisted living\textsuperscript{2}, sports analysis\textsuperscript{3}, logistics and other industrial operations\textsuperscript{4} and for monitoring of patients diseases, such as Parkinsons to build activity profiles for therapeutic purposes\textsuperscript{5}. Eventhough activity recognition, generally speaking, can be performed in a variety of ways\textsuperscript{5},\textsuperscript{7},\textsuperscript{8}, inertial sensors have been far by the most popular modality to use for this purpose. This is due to the fact that they are mobile, less cumbersome to wear and cost less than sensing devices for other modalities. Moreover, with their incorporation in phones and smart watches etc, these sensors are usually easily available to the subject for use in activity recognition tasks. This ubiquitous presence combined with ease of data collection has resulted in large datasets being produced which has led to the use of deep learning for various aspects of activity recognition tasks as shown in\textsuperscript{9},\textsuperscript{10}.

As mentioned, the field of activity recognition has been of attention to researchers in various domains, a domain that has not received as much interest is activity recognition in industrial settings. In this paper, data from the LARa dataset\textsuperscript{11} is utilized to perform continuous recognition of activities in a logistics scenario using two different convolutional neural network (CNN) architectures, one is a typical convolutional network and the other is a modified version of the parallel CNN architecture called CNN-IMU suggested in\textsuperscript{12} which performs convolutions in the temporal domain. The experiments indicated that the parallel CNN architecture performed better than the considered typical CNN. The rest of the paper is organized as follows: section\textsuperscript{11} discusses previous work carried out for activity recognition, section\textsuperscript{12} presents an overview of the dataset used, section\textsuperscript{12} presents the methodology of the paper; the pre-processing steps and the discussion of the networks used in the experiments, section\textsuperscript{12} discusses the results for the current work with a conclusion being provided in section\textsuperscript{11} and future directions in section\textsuperscript{12}.

II. LITERATURE REVIEW

Activity recognition has been at the forefront of pervasive computing research and the development of cyber physical systems as this has enormous societal and economical impact potential\textsuperscript{13}. This section covers previous work in the direction of activity recognition using utilizing inertial measurement sensors (IMUs).

Industrial activity recognition using IMUs has been targeted by multiple research works for varying applications, these include, wood shops\textsuperscript{14}, construction\textsuperscript{15}, assembly line\textsuperscript{16}, process optimization\textsuperscript{17}. An early work using deep learning methods for activity recognition for industry was suggested by\textsuperscript{18} on the Skoda dataset\textsuperscript{19}. Their network consists of one convolutional layer, one pooling layer, two hidden layers and one softmax layer for classification. Moreover, the convolutional layer contains several convolutional blocks in parallel with partial weight sharing for the three axes of accelerometer sensor values. The pooling layer also pools convolutional blocks sharing their weights separately before the outputs are passed on to the later layers.

An interesting approach for segmenting different types of activities for health risk assessment in an order picking process is presented in\textsuperscript{20} who utilize angles of human body joints from 17 IMUS placed on a workers body performing the order picking activity. Joint angles between body limbs are computed using an extended Kalman filter\textsuperscript{21} and these are used to segment the sub-activities within the picking process. Risk assessment is performed using the rapid entire body assessment (REBA) standard\textsuperscript{22}. The authors in\textsuperscript{23} propose using accelerometer and gyroscope signal data to perform activity recognition for worker performance assessment in the meat industry. After extracting segments, they compute several features from both sensors and test the performance of multiple classifiers for determining output activity. Following from their...
work in [24], the authors in [25] combine video and inertial measurement sensor data to determine grabbing actions in a picking process. They do this by extracting various time and frequency domain features from the IMU data as well as colour and descriptor features from the video data and then passing them on to three machine learning classifiers for prediction purposes.

The approach suggested in [12] makes use of CNNs on the data provided in [17]. They propose a CNN architecture called CNN-IMU based on the CNN proposed in [26]. Their network uses four CNN layers with parallel CNN blocks sharing their weights for each IMU, two pooling operations, fully connected layer before they are combined using a fully connected layer followed by a softmax for classification. The motivation to spread the network instead of making it deeper is that it becomes more descriptive. The authors in [27] make use of accelerometer sensor data in the dataset in [28] and CNNs to differentiate between activities in the industry. They achieve their best results when using raw signal data and sending it to the CNN for classification. The authors in [29] use the CNN-IMU on three different datasets of activity recognition, two consisting of various different activities in daily life and one of activities from a logistic scenario of order picking. Their experiments involved comparing two different CNN architectures, one a CNN-IMU and the other a baseline CNN with the same layerwise make up as the CNN-IMU. For the logistics scenario, their result indicates that the CNN-IMU outperforms the typical CNN in nearly all experiments. Their results indicate to the effectiveness of using wider networks consisting of parallel layers instead of using deeper ones. CNNs have also been used with other modalities for industrial processes, in [30] semantic representations have been used for activity recognition in a picking setup. Motion capture data from the MoCap dataset is used with the CNN architecture described in [31]. The authors in [11] use the Logistics Activity Recognition Challenge (LARA) dataset provided by them to determine activities in a logistics scenario. They carry this out by using a modified version of the t-CNN in [26] which consists of four convolutional layers, two fully connected layers followed by two separate softmax and sigmoid layers to determine the sub-activity being performed and the attribute from an activity attribute list on motion capture data from the dataset.

It can be observed from the literature review that convolutional neural networks have proved to be very useful for performing activity recognition in industrial scenarios. This paper utilizes a convolutional neural network for performing continuous activity recognition for logistics using inertial sensor data from the LARA dataset. This paper compares the performance of a modified version of the CNN-IMU network presented in [26] and used in [11] to a typical CNN. The CNN utilizes convolutions in the temporal axis to extract important features from time series sensor data and is well suited for use with inertial measurement sensor signals.

### III. DATASET

The LARa dataset provides data of multiple modalities from recordings in a logistics scenario. Video recordings, Motion Capture data and data from inertial measurement units is recorded from 14 people in the dataset. Each of the participants is asked to perform three tasks which are common in logistics operations, two of these are picking tasks and the third is packing. Motion capture data was captured using an Optical Marker-based Motion Capture (OMoCap) system which resulted in markers for the movements of the participants, moreover, several IMUs were used to record the movement patterns too along with RGB videos of the activities being performed. The total duration of the recorded data is 758 minutes which has been annotated in two ways, first, an annotation is provided for each intra-activity that comprises the picking and pacing tasks and second, binary semantic representations of a different type of representation for the picking and packing tasks. The first representation represents the activities in terms of eight intra-activities and is used in this work and the second type of representation as attributes to describe the task recordings. The eight annotations are standing, walking, cart (participant is walking with the cart), handling upwards (participant has atleast one hand raised upward to shoulder height), handling centered (participant can handle things without bending, lifting their arms or needing to kneel), handling downwards (participant has hands below his knees while kneeling or otherwise), synchronization (waving motion before each recording) and a set of samples which were unrecognizable by the annotators and have been marked as None.

This dataset provides the opportunity to develop algorithms for both the picking and packing operations in logistics by containing recordings of multiple modalities to researchers. From these modalities, this paper focuses on the data from the IMUs collected in these experiments. Three types of IMUs were used in the trials with 14 people in total performing the said tasks with data being collected from five points on the body, both the arms, legs and the chest/mid-body. The sampling frequency for the IMUs is 100 Hz. A summary of the recordings present in the dataset are given in Table I. Readers interested in more detail are referred to [11].

<table>
<thead>
<tr>
<th>Subject ID</th>
<th>Gender</th>
<th>Age</th>
<th>Scenario 1</th>
<th>Scenario 2</th>
<th>Scenario 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>007</td>
<td>M</td>
<td>23</td>
<td>13</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>008</td>
<td>F</td>
<td>51</td>
<td>13</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>009</td>
<td>M</td>
<td>35</td>
<td>14</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>010</td>
<td>M</td>
<td>49</td>
<td>13</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>011</td>
<td>M</td>
<td>47</td>
<td>12</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>012</td>
<td>F</td>
<td>23</td>
<td>6</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>013</td>
<td>F</td>
<td>25</td>
<td>14</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>014</td>
<td>M</td>
<td>54</td>
<td>14</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td><strong>99</strong></td>
<td><strong>95</strong></td>
<td></td>
</tr>
</tbody>
</table>

### IV. METHODOLOGY

To perform continuous monitoring of activities in logistics, we use a two step process. Segments are first extracted from the IMU sensor data for each trial which are then passed to the CNNs to test their performance. For the first two experiments, segmentation is performed for all IMUs together whereas for the last two experiments data segmentation takes place for each of the five IMUs individually. These are then fed to the CNN networks as inputs.

#### A. Preprocessing Stage

Windows of 100 samples are extracted from the recording for each sensor and position with a step size of 25 samples
Moreover, training was performed for 12 epochs with early stopping utilized to retain the best model. The data from the sensors was split into train, validation and test sets and the network was trained using an Adam Optimizer with cross entropy as the loss function. A learning rate of $1 \times 10^{-6}$ was used along with a batch size of 400. Moreover, training was performed for 12 epochs with early stopping utilized to retain the best model.

(75% overlap) for successive windowed segments. An overlap is used to ensure that enough samples are generated to develop a large enough dataset for training of deep learning networks. Furthermore, since the annotations in the dataset are present on a sample by sample basis for each value of each sensor, an extracted segment is assigned a segment annotation by majority voting of the annotations of its samples. We then use the segment annotations as the appropriate annotation labels, a similar approach has been used in [29]. Once segments have been extracted from all the trials for all subjects, the segments belonging to the synchronization and None class are removed. The rest of the segments are used for classification with the convolutional neural network.

### A. Experiment with Typical CNN-1

In this experiment we use the typical CNN-1 architecture that consists of 4 convolution layers, 2 max pooling layers, 2 fully connected layers and one softmax layer. The input to this network were segments/frames consisting of the combined information from all the IMUs together. The results for the classification are provided in Table III. It can be observed from the table that the CNN has performed poorly for the activities Hand up and Stand. Moreover, as can be observed, the best performing activities were Cart and Hand Center.

### B. Experiment with Typical CNN-2

For the second experiment, we use the typical CNN-2 architecture consisting of 4 convolution layers, 2 fully connected layers and one softmax layer. The input to this network too were segments/frames consisting of the combined information from all the IMUs together. The results of the classification are presented in Table IV. It can be observed that the performance for this network is very similar to the Typical CNN-1 network of experiment 1 which used maxpooling layers, there has been some degradation in performance for some activities. In this experiment too, the network was able to best recognize the activities of Cart and Hand Center while poor performance was observed for the activities of Stand and Hand up.

### V. Experimentation, Results and Discussion

In order to test the efficacy of the four CNNs considered in this work, we perform experiments for each of the four networks individually and report on the results obtained. The results for each experiment are reported in terms of the Precision, Recall and F1 score for each activity class.
Fig. 1. CNN Architectures used in this Work (a) Typical CNN (b) CNN-IMU.
C. Experiment with CNN-IMU-1

This experiment involved the usage of the CNN-IMU-1 network to determine activity classes. This architecture consists of four convolutional layers with five convolutional blocks, one for each IMU data and these blocks share their weights. The IMU segments are individually fed to the blocks resulting in the classification scores shown in Table VII. This network has also been able to recognize the activities of Hand Center and Cart well but the activity Stand is still the worst performing activity among the six activity classes being considered.

<table>
<thead>
<tr>
<th>Activity</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>F1 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stand</td>
<td>72.68</td>
<td>66.85</td>
<td>70.15</td>
</tr>
<tr>
<td>Walk</td>
<td>69.42</td>
<td>67.03</td>
<td>68.20</td>
</tr>
<tr>
<td>Cart</td>
<td>87.95</td>
<td>79.67</td>
<td>83.61</td>
</tr>
<tr>
<td>Hand up</td>
<td>43.62</td>
<td>60.18</td>
<td>50.58</td>
</tr>
<tr>
<td>Hand Center</td>
<td>90.34</td>
<td>80.04</td>
<td>84.88</td>
</tr>
<tr>
<td>Hand Down</td>
<td>71.93</td>
<td>75.68</td>
<td>73.76</td>
</tr>
</tbody>
</table>

D. Experiment with CNN-IMU-2

In this experiment, we used the CNN-IMU-2 architecture which consists of four convolution layers, two fully connected layers, and one softmax layer. The input to this network were segments/frames consisting of the individual IMUs. The classification results are presented in Table VI. The omission of the pooling layer has impacted network performance positively as was observed in other works. Similar to the previous cases, this network produces the best results for the activities Cart and Hand Center.

<table>
<thead>
<tr>
<th>Activity</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>F1 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stand</td>
<td>40.97</td>
<td>47.11</td>
<td>43.31</td>
</tr>
<tr>
<td>Walk</td>
<td>49.08</td>
<td>82.73</td>
<td>61.60</td>
</tr>
<tr>
<td>Cart</td>
<td>81.79</td>
<td>83.40</td>
<td>82.59</td>
</tr>
<tr>
<td>Hand up</td>
<td>26.04</td>
<td>73.36</td>
<td>38.43</td>
</tr>
<tr>
<td>Hand Center</td>
<td>91.87</td>
<td>77.88</td>
<td>82.82</td>
</tr>
<tr>
<td>Hand Down</td>
<td>71.46</td>
<td>75.37</td>
<td>73.37</td>
</tr>
</tbody>
</table>

Following from the experiments conducted in this work with the considered CNN architectures, the most suitable network for continuous activity recognition from inertial sensor data was found to be the Typical CNN-1 architecture which involves pooling operations. The best scores are achieved for the activities Cart and Hand Center whereas the worst scores have been produced for the activities Hand Up and Stand, this was the case for all the networks considered in this research work. The F1 scores for each of the networks for the six activities are listed in Table VII.

VI. Conclusion

This paper explores the usage of temporal convolutions in a CNN for the problem of continuous activity recognition in a logistics scenario using inertial measurement sensor data. Data from the LARa dataset which consists of video, OMOCap and IMU signal recordings from seven different people performing three different tasks concerning picking and packing has been used in this work. To accomplish the aims of this work, four CNN architectures, have been tested which take windowed segments of IMU recordings. From the experiments conducted, the typical CNN-1 architecture involving pooling operations was found to be the best performing model. High scores were achieved for the activities Hand Center and Cart; however, scores for the activity Stand and Hand Up weren’t satisfactory. While satisfactory performance was achieved for the former activities, the performance of the considered networks for the latter activities was poor. Therefore, modifications need to be made for improvement of the network for such activities.

VII. Future Work

This work presents experimental work for the continuous recognition of activities for logistics using the LARa dataset. In this paper, only CNN architectures have been considered, for future attempts at this task, other deep learning architectures could be considered to improve activity recognition such as Recurrent Neural Networks with attention, etc. Moreover, sensor fusion could also be used, especially OMOCap data from LARa dataset could be fused with IMU data and used with various deep learning networks to check for performance. Video data could also combined to create a multimodal solution for activity recognition as suggested in [25]. Dependable activity recognition systems will help in the optimization of industrial processes as well as be used for health assessment purposes.

REFERENCES


