Human Activity Recognition Framework in Monitored Environments

O. León[1], M. P. Cuéllar[1], M. Delgado[1], Y. Le Borgne[2] and G. Bontempi[2]
[1]*Dept. Computer Science and Artificial Intelligence, University of Granada,*
C/. Pdta. Daniel Saucedo Aranda s.n., Granada, Spain
[2]*Machine Learning Group, Universit Libre de Bruxelles, Campus LaPlaine, Bruxelles, Belgium*

Keywords: Human Activity Recognition, Ambient Assisted Living, Vision Computing, Data Mining.

Abstract: This work addresses the problem of the recognition of human activities in *Ambient Assisted Living (AAL)* scenarios. The ultimate goal of a good AAL system is to learn and recognise behaviours or routines of the person or people living at home, in order to help them if something unusual happens. In this paper, we explore the advances in unobstrusive depth camera-based technologies to detect human activities involving motion. We explore the benefits of a framework for gesture recognition in this field, in contrast to raw signal processing techniques. For the framework validation, Hidden Markov Models and Dynamic Time Warping have been implemented for the action learning and recognition modules as a baseline due to their well known results in the field. The results obtained after the experimentation suggest that the depth sensors are accurate enough and useful in this field, and also that the preprocessing framework studied may result in a suitable methodology.

1 INTRODUCTION

Currently, there are several approaches to monitor the movements of a person and their health (Hein and Kirste, 2008). Nevertheless, behaviour recognition, especially regards to most complex activities such as the ADL (Activities of Daily Living), is still a challenge. There are some proposals that try to deal with specific requirements, but a single solution that addresses all the problems is not known. Among some of the well known research works in this field, we emphatise the project *Tagged World* [1], which tries to provide services to users to make their lives safer and easier. The goal of this project is to develop new applications based on the use of RFID tags and the assumption that in a short time, all the objects that we can buy will include one. On the other hand, other projects focus in the pattern mining of the user's activity to detect abnormal behaviours. For instance, (Rantz et al., 2008), Rantz et al. developed a temporal clustering method based on detecting gradual changes as a result of deteriorating conditions in the person observed. This proposal is included in the project TigerPlace, where the authors propose a research and educational environment based on the concept of *Aging in Place*. In the work done by Storfer et al. (Storf et al., 2009), an activity recogniser based on a multi-agent framework is presented. Their approach to recognise complex activities (ADLs) is based on the technique *divide-and-conquer*, decomposing them into atomic actions, and each atomic and complex action is detected by a specialised agent with its own logical unit.

Another option to detect indoor human behaviours is by means of video-cameras. The cameras may provide all the necessary information about a scene, but the complexity arises when making the system capable of *understanding* what is happening in the images of the video stream. Further limitations of these techniques are occlusions, cluttered background, shadows, varying illuminations, and viewpoint changes. At the moment, there are no proposals that are able to learn and detect complex behaviours such as the ADLs using video data, although in the last few years there has been an increasing effort in the field of automatic gesture recognition as a first step. For example, the work done by Reifinger (Reifinger et al., 2007) is intended to recognise a person's hand gestures, whether static or dynamic, to be used in augmented reality applications. It is based on an infrared tracking system where the aims of the infrared device are targeted on the index and thumb fingers, collecting information about the position and orientation of each finger. As a result, the gesture detection is provided to any application that is connected to the back-end.

Depth cameras appeared on the market in the last

[1] http://taggedworld.jp/

León O., P. Cuellar M., Delgado M., Le Borgne Y. and Bontempi G..
Human Activity Recognition Framework in Monitored Environments.
DOI: 10.5220/0004755504870494
In *Proceedings of the 3rd International Conference on Pattern Recognition Applications and Methods* (ICPRAM-2014), pages 487-494
ISBN: 978-989-758-018-5

few years and are useful to overcome some limitations of raw video cameras, such as shadows, viewpoint changes and body detection. The release and popularity of the Microsoft Kinect provides RGB image and depth image streams (Raheja et al., 2011). Although targeted initially for the home entertainment market with the XBOX console, the Kinect has received increasing interest from the vision and robotics community due to its great potential (Giles, 2010). A good example of its possible application is to detect the presence of people in a scene. In the work of Salas (Salas and Tomasi, 2011), a strategy that combines colour and depth images (obtained by the Kinect device) by *Histograms-of-Oriented Gradients* (HOG) to detect people in indoors environments is presented. This strategy greatly improves previous work in this field obtaining high accuracy (up to 34,981 targets could be detected in an image). This people detector is efficient and accurate from the computational point of view, and was validated in a pilot phase. Another work that tries to solve the activity recognition problem can be found in (Shotton et al., 2013), where a method to quickly and accurately estimate 3D positions of the user's skeleton joints from a single depth image from Kinect was proposed. With this kind of information, we can address the problem of human action recognition in a simpler way compared to the use of classic RGB images.

In the last two years, there have been a wide number of approaches for human activity recognition using depth cameras. For instance, in proposals like (Yang et al., 2012; Yang and Tian, 2013), the authors develop a technique using EigenJoints (Yang and Tian, 2012) to find movement features. The learning and recognition is executed by means of Nave-Bayes-Nearest-Neighbor (NBNN) and SVM. Their classification rate results are between 71% and 97% depending on the dataset used in the experiments and on the configuration of the training stage. We also cite the article (Azary and Savakis, 2010), where the authors explored the possibility of creating a fingerprint for each action using radial distance and *Manifold Learning* to represent the action on a smaller dimensional space. Finally, they perform classification using lazy learning by means of the nearest neighbour technique.

In this paper, we present a human activity recognition framework based on depth image processing. The motivation of the current work is to propose a standard framework composed of 5 modules and steps ((a) body representation, (b) time series summarization, (c) posture clustering-quantization, (d) action learning, and (e) action recognition), to achieve an accurate model-based learning and recognition of human activities. This modular design eases the adaptation to different scenarios and techniques, so that each module can be superseded by the implementation of a different technique. For the framework validation, Hidden Markov Models have been implemented for the action learning and recognition modules as a validation technique due to its well known results in the field (Gao and Sun, 2013a; Gao and Sun, 2013b). The results are compared with standard technique in gesture recognition, such as Dynamic Time Warping (DTW) (Berndt and Clifford, 1994), in order to compare model-based learning and recognition, and raw signal processing techniques. Besides a detailed discussion of our recognition approach, we also present a comparison of the results with previous work done with normal video-cameras.

The paper is structured as follows: in Section 2, we introduce the general problem of activity recognition and then, our activity recognition framework approach for solving this problem. Section 3 reports and discusses the experimental results obtained as well as a comparison with previous related work. Finally, we provide some conclusions and future work.

2 ACTIVITY RECOGNITION FRAMEWORK

A central challenge faced by Ambient Assisted Living systems is to find a reasoning method for detection of human behaviour, based upon a continuous sequence of events (states and state changes) measured by the different sensing systems. This becomes even more complicated if the sensors are integrated into the environment and thus, cannot be directly correlated with an user. In our approach, we intend to use a single sensor that is able to provide us with data about how many users are in the scene and their body postures of these users; but the main problem is the complexity of how to process the data received by the sensor and how to make the system capable of understanding the user's behaviour through the actions he/she carries out. The device selected for this task is the Microsoft's Kinect camera, which offers depth images of the scene. To avoid privacy problems, our framework directly uses three-dimensional points of the user's body, obtained by an existing middleware called OpenNI [2], so that no images are saved. We have created a database, recording different actions done by different users (see section 3) to be able to train the models and test the applicability of the approach to detect accurately short-time actions.

[2] http://en.wikipedia.org/wiki/OpenNI

The framework model proposed consists of five steps: First, the data provided by the camera is processed to obtain the user's body skeleton representation as a temporal data series (section 2.1); secondly, the data series are summarized (section 2.2). Then we quantize each body posture of the series using clustering to make the representation of the whole action easier (section 2.3); then the training of each action by means of Hidden Markov Models is carried out (section 2.4). Finally, we obtain a trained HMM used for a later recognition of new instances (section 2.5). Our goal is to create a framework that learns, detects, and recognises different events performed by users at home. Each of these events or actions will correspond to a sequence of images obtained by the camera, in this case, a sequence of the user's body postures.

2.1 Step 1: Initial Processing of the 3D Data Provided by the Camera

We obtain the user's body skeleton directly from the camera, represented as a set of fifteen 3D points (body joints). Then, we process this data to obtain a body representation independent of the heading and distance from the camera. This representation is the set of angles between consecutive body joints with the diference of the height of the user's chest in respect to an initial value obtained from the first frame. The height difference is useful to detect actions such as bending, sitting down, etc...). As a result, the body posture on each frame received by the camera is represented by a set of eleven real values (ten angles plus the *height* variation).

2.2 Step 2: Compressing the Data Series

Video cameras, and also depth sensors, provide a large amount of data in a short amount of time. Therefore, it becomes necessary to simplify the processing of these data as much as possible to achieve a fast and real-time processing. In our approach we perform data summarization and dimensionality reduction in steps 2 and 3. In step 2, the multivariate data sequence obtained from the sensor is compressed to reduce the computational cost of later steps in recognition (see figure 1 for an illustrated example of time series summarization). Another advantage of summarization is that it also serves as a filter to prevent potential false sensor measures and to smooth the data series. The method selected for this task in the experimentation is *Piecewise Aggregate Approximation* (*PAA*) (Keogh et al., 2001) as a proof of concept.

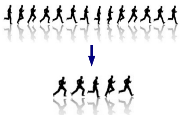

Figure 1: Illustrated example of compressing time series of postures.

2.3 Step 3: Quantizing each Posture

Step 3 aims at dimensionality reduction by means of quantization. The objective is to reduce a frame (posture) composed by a multivariate signal (11 real values in our case) to a single dimension so that the computational time of the recognition techniques can be decreased for real-time processing. To achieve this, we have applied clustering techniques, and more specifically the K-means (Mitsa, 2010) algorithm. The clustering is used to find template postures that can be tagged uniquely using the cluster center, so that during the recognition process each new frame is quantized and assigned with the label of its nearest template posture. Figure 2 illustrates this idea, where we show 5 different clusters with their central postures (left side) and different postures being tagged according to their nearest template (right side). For the example, we have used letters, i.e "G", "C", "T", "B", "A", as cluster labels for clarity.

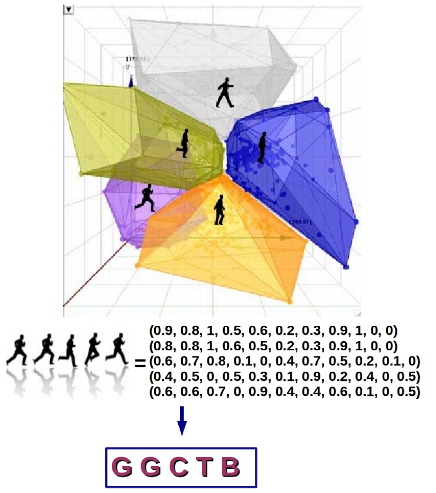

Figure 2: Example of posture quantization using clustering, where each posture of the series was initially a vector of 11 real values. After clustering, we obtain a single tag.

We illustrate the main advantages of steps 2 and 3 to justify their need: At first, a sample of an action that lasts 2 seconds is a data series of 60 frames, where each frame is composed of 11 real values (660 real values). After applying PAA (compression rate of 4) and K-means ($k = 21$), the data series is reduced to a sequence of 15 discrete labels. As stated before, this preprocessing is useful to achieve a real-time processing in our approach.

2.4 Step 4: Training the Model for Action Learning

In this work, we use Hidden Markov Models (HMM) and Dynamic Time Warping (DTW) to learn and recognise the actions performed by different participants, since both techniques have been widely used in previous research works regarding human activity recognition (Crandall and Cook, 2010; Xia et al., 2012; Corradini, 2001). In our case, the objective is to learn and recognise 7 different actions. To this end, we trained 7 different HMMs with the Baum-Welch (Rabiner and Juang, 2003) algorithm, each one matched with an activity. The number of states in each HMM is set to the number of clusters obtained from Step 3, and each state is matched with a template posture (cluster center). Training sequences of each activity were provided to find the optimal transition and a priori probabilities of each model. Thus, after we obtain a sequence of labels corresponding to an activity in the recognition state, a HMM may provide us with the probability that the sequence can be generated from the learned model. On the other hand, in the case of DTW we use instance-based learning. We select a subset of instances of the recorded actions as template activities. New instances acquired from the sensor data are compared to the templates and they are classified using the k-Nearest Neighbour method (k-NN).

2.5 Step 5: Action Recognition

For the recognition stage, we follow the same three steps of the process explained before to reduce the data sequence. After that, the resulting reduced sequence is used as input for all the models trained. In the case of HMM, each model returns a probability, reflecting the likelihood of the input sequence that conforms to the model. The recognised activity is the one whose HMM provides the highest probability. In the case of DTW, the new instance is tagged with the activity of the nearest template using 1-NN. Figure 3 illustrates this procedure with HMM.

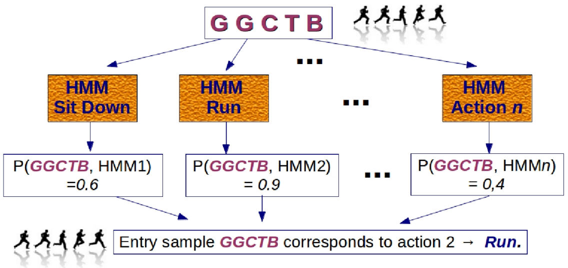

Figure 3: Example of action recognition with the trained HMMs.

3 EXPERIMENTAL RESULTS

3.1 Dataset and Data Acquisition

The human actions selected in this study were chosen because they are present in many of the existing video databases on this topic (for example: (Mokhber et al., 2008), (Mat, 2007), (Laptev, 2005)). In our case, we wanted to test the approach in real-time scenarios under controlled environmental conditions, and decided to create our own dataset. This was an experimental design decision to be able to move the camera, change the lighting of the room, and the background, to test the performance of the approach under a well-known work environment with changing conditions to describe the user's posture independently of his/her position. Moreover, this decision was useful to test real-time processing.

Therefore, a first step to perform the experimentation was to record the dataset with some volunteers from the university. The 7 selected actions were: *walk, sit down, stand up, bend down, bend up, twist right and twist left*, since these activities are the most frequent ones in the literature. For the experiments, we got 17 different participants that consented to recording the activities (array of 3D joint locations of the body, no image recording to preserve privacy). Each participant performed 10 repetitions of each activity, so the final dataset is composed of 1190 samples. The duration of the actions was chosen so that every participant had enough time to perform the requested tasks, and we recorded 60 frames (2 seconds) of each activity during its execution except for *walk*, which required 90 frames because it is a longer action. We decided that the duration of the recordings should not be longer than these values since our objective is to test the approach in real-time scenarios. Finally, we have decided to make the dataset available to other researchers for reproducibility[3].

[3] Dataset available at: http://decsai.ugr.es/ manupc/presens

3.2 Parameters and Experimental Settings

The algorithm used to summarize the number of frames of each action sample was PAA. We have used three different compression rates to compare which one performs better. The rates chosen were 2, 4, and 6, which reduce the frame samples to half, one quarter, and one sixth of the total size respectively. The best results were obtained with $PAA = 2$ and, for space limitations, we only show the results with this parameter value in the article. After executing the algorithm, the summarized data is stored in the database. On the other hand, for the clustering stage we selected the K-means algorithm due to its simplicity and good average results in multiple problems. The distance metric we selected to compare different postures was Chebisher distance, i.e. the maximum absolute difference of the 11 values of the two postures, since it provided us with better average results than other metrics such as Euclidean distance or City Block. Also, the data series were normalized before clustering to avoid effects of scale/translation. Finally, the values of k chosen for the experiments with k-Means were 14, 21, 28, 35, 49, 70, 84, 98, 112 and 126. In this stage, we implemented a *multistart* technique [4] to obtain the best set of clusters that separate the postures, specifically 100 times for each configuration. The resulting preliminar experimentation provided us with an optimal value of $k = 112$.

Analogously to the *multistart* technique implemented in K-Means, we performed multiple runs to obtain the 7 models (HMM) that best differentiate the actions. For this task, we implemented the *cross validation* technique (*10-Fold* and 80% of training data). The set of models that made the least classification errors was chosen. In total, we performed *cross-validation* 100 times for each combination *number of clusters-compression rate*, and the best training data set for each action was stored.

3.3 Results

We tested the approach in two different scenarios: First, training and test of data recorded from each participant separately. This experimentation has been done because in real AAL environments, mainly focused on elderly people, the average household will be comprised of one or two users. The second scenario considers all participants together, and is aimed to test the degree of abstraction that can be achieved to learn each activity independently of the participant performing it. In this case, HMM approach is compared with the results obtained by an implementation of a well known technique, as it is Dynamic Time Warping (DTW) (Berndt and Clifford, 1994), used to find patterns in time series and works directly with the raw data for the inference and recognition of the actions. The experiments were carried out on an average personal PC with Pentium Dual Core processor, CPU E5700, 3,00 GHz, 800 MHz FSB, 2 GB RAM, running Ubuntu Linux 12.10.

3.3.1 Testing Participants Data

In this test we have processed the data of each participant separately. Each participant performed the 7 actions 10 times, so we are provided with 70 samples for each one. To perform the training, we used the *10-Fold cross validation* technique using 80% of training data, uniformly distributed between positive and negative samples for each activity. Table 1 shows the results obtained for the classification of each participant (success rate). The first column of the table shows the participant's *id*; and the second and third columns describe the success rate obtained after the classification of all actions in the training and test datasets for each participant, respectively. Finally, the last column represents the total success rate for that participant's data. The last row of the table shows the average success rate for all participants, with a value of 99.3%. We have not found significant differences between the recognition of different activities, which means that the proposal is robust and performs similarly for all actions in the dataset.

According to these results, we may conclude that the proposed methodology is suitable to distinguish actions when the same user is being monitored in the environment. The average computational time to achieve the recognition of actions is **0.635** seconds for the activity recordings of 60 frames and **0.938** seconds for those of 90 frames, which in turn means that the approach is also suitable for near real-time processing scenarios with a frequency of 17fps for sensor data acquisition. As stated before, we find these results specially relevant for Ambient Assisted Living scenarios, where the number of inhabitants at home is generally low (up to 4 people, usually 1-2 people). For this reason, we are considering the inclusion of this system in a global solution to detect human behaviours, where the approach presented in this work would be able to recognise actions whose performance usually takes a small amount of time, and another upper system could be able to infer and recognise longer and more complex activities such as *morning routine*, *lunch time*, etc., using not only a single

[4] How GlobalSearch and MultiStart Work. http://www.mathworks.es/help/toolbox/gads/bsc59ag-2.html

depth camera, but also further sensors located in the environment that could be necessary depending of the scenario.

Table 1: Percentage of success rate obtained by each participant's dataset using HMM.

Participant	% Hits Training Set	% Hits Test Set	% Hits Total
1	100	92.8	**98.6**
2	100	100	**100**
3	100	92.8	**98.6**
4	100	92.8	**98.6**
5	98.2	100	**98.6**
6	100	92.8	**98.6**
7	100	100	**100**
8	98.2	100	**98.6**
9	100	100	**100**
10	100	92.8	**98.6**
11	100	100	**100**
12	100	100	**100**
13	100	100	**100**
14	96.4	100	**97.1**
15	100	100	**100**
16	100	100	**100**
17	100	100	**100**
Average All participants			**99.3**

3.3.2 Testing Activity Data

This section addresses the capability of the approach to infer the main features of an action performed by different users, so that the trained model could recognise the action in future instances. Since this is the usual scenario of previous approaches in gesture recognition, we compare the approach with an implementation of DTW. The DTW implementation finds patterns over the time series of the 3D joints positions directly, without the need of data preprocessing. We choose this technique for the comparison since it has provided good results in this research field previously, and its simplicity does not require computational time for preprocessing. Moreover, we provide the results of DTW as a lazy learning and recognition technique within the proposed framework to validate our approach. The study of HMM outside the framework was not possible since it requires discrete symbols as input.

Tables 2, 3 and 4 shows the success rate obtained with the whole dataset for each implementation, separating the results by action (*Walk (W), Sit Down (SD), Stand Up (SU), Bend Down (BD), Bend Up (BU), Twist Right (TR)* and *Twist Left (TL)*). Each cell contains the success in both training (above) and test (below) data sets. In the last column and row of both tables, the overall percentage of success rate for the corresponding action are shown.

Table 2: Contingency Table of errors obtained in training and test with classic DTW technique.

Action	W	SD	SU	BD	BU	TR	TL	%
W	**133** **32**	1 0	0 0	0 0	0 0	2 1	0 1	97,8 94,1
SD	0 0	**121** **26**	0 1	6 3	0 0	5 3	4 1	89 76,5
SU	0 0	3 0	**131** **30**	0 1	2 3	0 0	0 0	96,3 88,2
BD	0 0	21 4	0 0	**92** **17**	0 1	13 7	10 5	67,6 50,0
BU	0 0	0 0	4 4	1 0	**131** **30**	0 0	0 0	96,3 88,2
TR	0 0	0 0	3 1	12 6	2 1	**110** **18**	9 8	80,9 52,9
TL	0 0	11 5	0 0	7 1	0 0	7 9	**111** **19**	81,6 55,9
%	100 100	77,1 74,3	94,9 83,3	78 60,7	97 85,7	80,3 47,4	82,8 55,9	**87,1** **72,3**

Table 3: Contingency Table of errors obtained in training and test with the proposed framework and HMM.

Action	W	SD	SU	BD	BU	TR	TL	%
W	**129** **27**	0 0	0 0	0 0	0 0	0 **2**	**8** **4**	94,2 81,8
SD	0 0	**131** **27**	0 0	**6** **6**	0 0	0 0	0 0	95,6 81,82
SU	0 0	0 0	**127** **25**	0 0	**10** **8**	0 0	0 0	92,7 75,76
BD	0 0	**9** **8**	**4** 0	**119** **24**	**5** **1**	0 0	0 0	86,9 72,7
BU	0 0	0 0	**1** **5**	0 **1**	**136** **27**	0 0	0 0	99,3 81,8
TR	0 **3**	0 0	0 0	**1** **1**	0 0	**136** **28**	0 **1**	99,3 84,9
TL	**1** **3**	0 0	0 0	0 0	**1** 0	**1** 0	**134** **30**	97,8 90,9
%	99,2 77,8	93,1 70,4	96,1 80	94,1 66,7	88,2 66,7	99,7 92,9	94 83,3	**95,1** **81,4**

Table 4: Contingency Table of errors obtained in training and test with the proposed framework and DTW.

Action	W	SD	SU	BD	BU	TR	TL	%
W	**136** **34**	0 0	0 0	0 0	0 0	0 0	0 0	100 100
SD	0 0	**136** **34**	0 0	0 0	0 0	0 0	0 0	100 100
SU	0 0	0 0	**136** **33**	0 0	0 1	0 0	0 0	100 97,1
BD	0 0	0 0	0 0	**136** **34**	0 0	0 0	0 0	100 100
BU	0 0	0 0	0 2	0 0	**136** **32**	0 0	0 0	100 94,1
TR	0 0	0 0	0 0	0 0	0 0	**135** **33**	1 1	99,3 97,1
TL	0 0	0 0	0 0	0 0	0 0	0 0	**136** **34**	100 100
%	100 100	100 100	100 94,3	100 100	100 97	100 100	99,3 97,1	**99,9** **98,3**

We may notice that the overall results obtained by the techniques inside our framework outperformed the one with standard DTW over the time series signal. Overall success rate obtained by our framework with Hidden Markov Models is 95,1% for training data and 81,4% for test data, and 99,9% for training and 98,3% for test for DTW. As it may be expected, the highest rate of false positives is caused by actions with similar movements, such as SD/BD or SU/BU. In the opposite case, actions whose performance differs significantly from the remaining ones also have high success rates, such as W.

Now, we analyze the robustness of the approach. For the DTW baseline method, we identified two classes of actions using non-parametric Kruskal-Wallis test with 95% of confidence level: those that can be predicted with higher performance (no statistical differences between members in the group) and those with lower performance (with statistical differences between members inside the group). Actions in the first group are W, SD, SU, BU. This analysis suggests that classic DTW performance may vary depending on the learned action. On the other hand, we applied the same analysis regarding the results in the test set of the framework approach with HMM and DTW, and we obtained no statistical relevance between the results of each action recognition. This means that, for the set of actions selected, the framework performs the same independently of the action to be learned. Thus, this fact suggests that the robustness of the framework, for both DTW and HMM, is higher than the baseline method.

Regarding execution time, the recognition of an action using the proposed framework is **0.635** seconds (60 frame processing) in average with HMM and **0.299** seconds in average with DTW. In respect to the classic DTW, which spends **1.118** seconds, we validate a clear improvement that makes our proposal not only more accurate in the classification success rate, but also more efficient in time and suitable for near real-time use of the approach.

Another aspect of interest is the improvement in performance of the DTW method with respect to the framework using DTW. If we compare Tables 2 and 4, we notice the increase in the success rate after using the framework. This can be explained because data sequences in our method are compressed and reduced, and accumulated errors during the recognition stage are not as relevant as in bigger data series. In addition, the compression in Step 2 also serves as a filtering process, which in turn removes outliers produced by the sensor and make the data signal smoother.

Finally, we are also interested in the scalability of both methods. DTW is an instance-based learning method, which means that it achieves a good performance when there are enough template instances to compare with new data. Thus, when the number of activities increases, so does the number of instances. This increase could slow down the computing time of the method, and therefore make more difficult its application in real-time tasks. On the other hand, model-based techniques have shown a poorer performance to abstract relevant activity features in this work, although they perform similarly to DTW when adapted to a single user. In this case, as the number of activities increases, the number of models grows in a relationship 1-to-1 (one model per activity), which suggests that this approach could be more scalable. Nevertheless, experiments must be carried out to test this hypothesis, and the framework should also be compared with different recognition techniques to give support to the results of this work. For now, the current work has served as a feasibility study of the proposal and, following the quality of the results obtained, we aim to improve the different aspects that make up the system to make it more competitive and applicable in commercial environments.

In respect to previous proposals, our approach shows promising results. For example, in Mokhber et al.(Mokhber et al., 2008), an accuracy rate of 90% based on their own database of 1614 sequences, divided into 8 actions, performed by 7 different people was obtained. They used a simple classic camera to record the video samples. Another proposal is the work of Azary and Savakis (Azary and Savakis, 2010), tested with an existent 2D video database of human actions which contains 10 basic actions. In this work, a hit rate of 92% is achieved. In Minhas et al.(Minhas et al., 2010), the system was tested with two known 2D video databases: the *Weizmann human action dataset* (Mat, 2007) (consisting of 9 human actions) and the *KTH dataset* (Laptev, 2005) (consisting of 6 actions), obtaining a hit rate of 98% and 94%, respectively.

4 CONCLUSIONS

In this paper, we have presented a framework for learning and recognising human actions by means of a depth camera as single sensor. The framework is modular so that each module can be superseded by another implementation of a different technique, and is easily adaptable to different contexts. For the implementation, we have tested learning and recognition based on Hidden Markov Models and Dynamic Time Warping. Furthermore, the results have been compared with raw DTW, which operates directly with the time series data. The experiments have shown that the framework is useful for both improving the accuracy in the recognition process and reducing the computational time to achieve an effective near real-time recognition. Our final goal is to create an AAL environment that learns and recognises complex behaviours or routines of people at home. The experiments carried out suggest that the approach may be useful in this scenario thanks to its flexibility, simplicity and robustness.

ACKNOWLEDGEMENTS

This work has been partially supported by the CIE BioTIC CEI2013-P-3, the Fellow FPU program of the Ministry of Education (Spain) AP2010-6075, the Contrato-Programa 2013-2014 of the Faculty of Humanities and Education of Ceuta, and Innoviris, the Brussels Institute for Research and Innovation through the ICT4Rehab project (http://www.ict4rehab.org).

REFERENCES

(2007). Actions as Space-Time Shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2247–2253.

Azary, S. and Savakis, A. E. (2010). View invariant activity recognition with manifold learning. volume 6454 of *Lecture Notes in Computer Science*, pages 606–615. Springer.

Berndt, D. J. and Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA.

Corradini, A. (2001). Dynamic time warping for off-line recognition of a small gesture vocabulary. In *Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, 2001. Proceedings. IEEE ICCV Workshop on*, pages 82–89.

Crandall, A. S. and Cook, D. J. (2010). Using a hidden markov model for resident identification. In *Proceedings of the 6th Int. Conf. on Intelligent Environments*, IE '10, pages 74–79.

Gao, Q. and Sun, S. (2013a). Human activity recognition with beta process hidden markov models. In *Proceedings of the International Conference on Machine Learning and Cybernetics*, pages 1–6.

Gao, Q. and Sun, S. (2013b). Trajectory-based human activity recognition with hierarchical dirichlet process hidden markov models. In *Proceedings of the 1st IEEE China Summit and International Conference on Signal and Information Processing*, pages 1–5.

Giles, J. (2010). Inside the race to hack the kinect. *New Scientist*, 208(2789):22–23.

Hein, A. and Kirste, T. (2008). Activity recognition for ambient assisted living : Potential and challenges. *Sensors Peterborough NH*, pages 263–268.

Keogh, E., Chakrabarti, K., Pazzani, M., and Mehrotra, S. (2001). Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases. *Knowledge and Information Systems*, 3(3):263–286.

Laptev, I. (2005). On space-time interest points. *Int. J. Comput. Vision*, 64:107–123.

Minhas, R., Baradarani, A., Seifzadeh, S., and Jonathan Wu, Q. M. (2010). Human action recognition using extreme learning machine based on visual vocabularies. *Neurocomput.*, 73:1906–1917.

Mitsa, T. (2010). *Temporal Data Mining (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series)*. Chapman and Hall/CRC.

Mokhber, A., Achard, C., and Milgram, M. (2008). Recognition of human behavior by space-time silhouette characterization. *Pattern Recognition Letters*, 29(1):81–89.

Rabiner, L. and Juang, B. (2003). An introduction to hidden Markov models. *ASSP Magazine, IEEE*, 3(1):4–16.

Raheja, J. L., Chaudhary, A., and Singal, K. (2011). Tracking of fingertips and centers of palm using kinect. In *Computational Intelligence, Modelling and Simulation (CIMSiM), 2011 Third International Conference on*, pages 248–252. IEEE.

Rantz, M., A., G., A., Oliver, D., M., M., S., J., K., Z., H., M., P., G., D., and S., M. (2008). An innovative educational and research environment. volume 17, page 8491.

Reifinger, S., Wallhoff, F., Ablassmeier, M., Poitschke, T., and Rigoll, G. (2007). Static and dynamic hand-gesture recognition for augmented reality applications. In *Int. Conf. on Human-computer interaction*, HCI'07, pages 728–737. Springer-Verlag.

Salas, J. and Tomasi, C. (2011). People detection using color and depth images. In *Proc. of the 3rd Mexican conference on Pattern recognition*, MCPR'11, pages 127–135. Springer-Verlag.

Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., Cook, M., and Moore, R. (2013). Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 56(1):116–124.

Storf, H., Becker, M., and Riedl, M. (2009). Rule-based activity recognition framework: Challenges, technique and learning. In *Pervasive Computing Technologies for Healthcare, 2009. PervasiveHealth 2009. 3rd International Conference on*, pages 1–7.

Xia, L., Chen, C.-C., and Aggarwal, J. K. (2012). View invariant human action recognition using histograms of 3d joints. In *CVPR Workshops*, pages 20–27. IEEE.

Yang, X. and Tian, Y. (2012). Eigenjoints-based action recognition using naive-bayes-nearest-neighbor. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pages 14–19. IEEE.

Yang, X. and Tian, Y. (2013). Effective 3d action recognition using eigenjoints. *Journal of Visual Communication and Image Representation*.

Yang, X., Zhang, C., and Tian, Y. (2012). Recognizing actions using depth motion maps-based histograms of oriented gradients. In Babaguchi, N., Aizawa, K., Smith, J. R., Satoh, S., Plagemann, T., Hua, X.-S., and Yan, R., editors, *ACM Multimedia*, pages 1057–1060. ACM.