Recognition of Activities of Daily Living via Hierarchical Long-Short Term Memory Networks

Maxime Devanne[1], Panagiotis Papadakis[1] and Sao Mai Nguyen[1]

***Abstract*— In order to offer optimal and personalized assistance services to frail people, smart homes or assistive robots must be able to understand the context and activities of users. With this outlook, we propose a vision-based approach for understanding activities of daily living (ADL) through skeleton data captured using an RGB-D camera. Upon decomposition of a skeleton sequence into short temporal segments, activities are classified via a hierarchical two-layer Long-Short Term Memory Network (LSTM) allowing to analyse the sequence at different levels of temporal granularity. The proposed approach is evaluated on a very challenging daily activity dataset wherein we attain superior performance. Our main contribution is a multi-scale, temporal dependency model of activities, founded on a comparison of context features that characterize previous recognition results and a hierarchical representation with a low-level behaviour-unit recognition layer and a high-level units chaining layer.**

I. INTRODUCTION

The improvement in quality of life combined with a declining birth rate results in an increasingly ageing population, particularly in Europe. Together with seniors' need for staying longer in their homes, this drives the need for a development of personal assistance technologies to improve a person's autonomy. Relatedly, human behaviour understanding is essential in enabling a system to provide consistent assistance to people in a timely manner. This becomes nowadays feasible via smart sensors or cameras combined with powerful algorithms for real-time tracking of human body.

However, understanding human behaviour is still a difficult task due to the variability and the complexity of human activities of daily living (ADL) [36]. Indeed, ADL are characterized by various combinations of atomic movements which complicates their understanding at a higher level. Humans seem to organize their behaviour and the accompanying perception into small, compositional structures or building blocks of behaviour or behavioural primitives. Neuromodelling [15], [12] and behaviour learning [13], [32] studies have outlined a hierarchical representation of motion based on a low-level representation of action units, simpler activities, and a high-level structure to chain these unit blocks into more complex action sequences.

*The research work presented in this paper is partially supported by the Region Bretagne through the AMUSAAL project and by the European Regional Development Fund (ERDF) via the VITAAL Contrat Plan Etat Region.

[1]Authors are with IMT Atlantique, Lab-STICC, UMR CNRS 6285, F-29238 Brest, France, team IHSEV. maxime.devanne@gmail.com, panagiotis.papadakis@imt-atlantique.fr, nguyensmai@gmail.com

Based on theses considerations, we attempt to leverage this hierarchical representation in this study. Towards this goal, we propose a vision-based approach for daily activity recognition from skeleton features extracted from an RGB-D sensor, that allows to account for the previous constraints (see Fig. 1 for an overview of our approach). The contributions of this work are summarized as follows:

- We study how to decompose a continuous stream of body movement into action units. By comparing several methods, we conclude that a fixed-window decomposition is more stable and effective
- We propose two methods to chain action units: by augmenting the skeleton features with context features that capture previous recognition results, and by a hierarchical two-layer LSTM. In order to handle the hierarchical and temporally abstracted nature of daily activities characterized by varying composition of actions, the first layer captures the dynamics within each action unit, while the second layer models the evolution of these action units describing the entire daily activity. Our analysis shows that the hierarchical structure is more effective for recognition of complex activities.

II. STATE OF THE ART IN ACTIVITY RECOGNITION

Visual behaviour analysis and understanding has been widely investigated in the last two decades. In related literature, human behaviour may refer to different types of human movements from gestures and actions to daily activities. However, boundaries between different types of movements are not always easily identifiable. Hence, recent taxonomies have been proposed by considering motion complexity and duration, converging towards similar definitions for human movements [6], [35]. In this work, we are particularly interested in activities of daily living carried out by a person at home in an ambient assisted living context. Based on those taxonomies, such daily activities are mainly related to the term of 'activity'.

In the following we review some relevant vision-based approaches addressing the issue of human activity understanding. Assuming that extraction of relevant human body information has been previously done using approaches such as [31] or [5], in this study we focus on the way of characterizing human movement for the purpose of recognition. Such approaches can be grouped in three different categories: rule-based approaches, stochastic approaches and deep learning approaches.

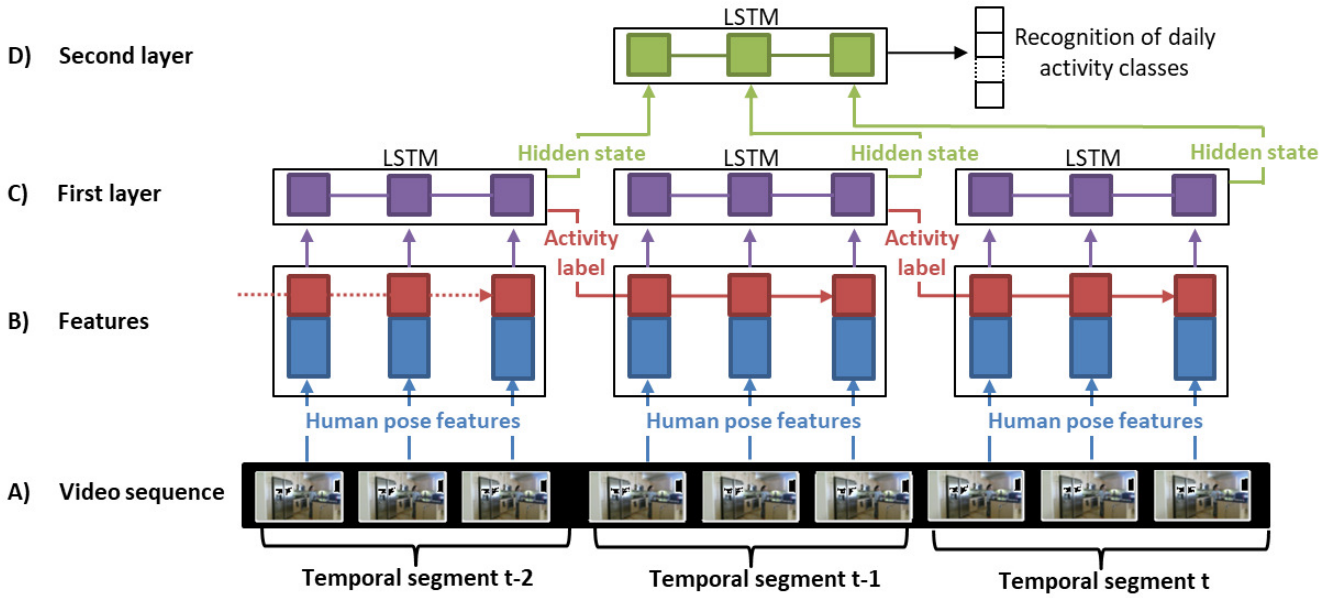

Fig. 1: Overview of our approach. A) A video sequence is decomposed into fixed-length temporal segments. B) Human pose features are extracted from each frame and combined with a context feature from the activity label of the previous temporal segment. C) A first LSTM layer captures the dynamics within each temporal segment. D) A second LSTM layer models the evolution of the temporal segments in order to recognize activities of daily living (ADL)

A. *Rule-based approaches*

Rule-based approaches are inspired by the natural language processing community and are based on the definition of rules or set of attributes that describe an activity. Commonly, the occurrence of some attributes is used to build semantic grammars describing activities. These grammars represent the co-occurrence of objects-actions or actions-activities. They are then used during testing to estimate the probability of activities from observed attributes. For instance, Yang et. al. [28] propose a rule-based approach to make a robot learn kitchen activities from human demonstrations. The authors employed visual features to characterize different types of objects as well as different grasping types. The grammar is built from occurrence of these features during kitchen activities. Conversely, Tayyub et. al. [33] proposed to automatically learn a stochastic grammar describing the hierarchical structure of complex activities from annotations acquired from multiple annotators. During testing, they recognized low-level actions using visual features and Support Vector Machines [34] and then inferred the most likely activity hierarchy that generates these actions.

B. *Stochastic Approaches*

Many researchers modelled human motion as a stochastic process where the movement can be seen as a sequence of successive states. Inspired by the efficiency of the popular Hidden Markov Model (HMM), a hierarchical extension has been proposed in [27] to handle the motion complexity of activities. The authors proposed a two-level HMM where actions are modelled as a succession of poses and activities as successive actions states. Stochastic processes have also been employed as a means for mapping the raw action sequences to a lower-dimensional latent space and enforcing intra-class similarity and inter-class dissimilarity, facilitating the recognition of simple actions [25], [24]. More complex models such as Dynamic Bayesian Networks have been proposed to mainly consider several variables per time, contrarily to HMM, so as to differentiate each joint [26] or the pose and the object [9]. A similar consideration of both human motion and possible objects within short time intervals is proposed in [18] where a Conditional Random Field (CRF) is proposed to anticipate further activities. While these stochastic approaches performed well for relatively simple activities, their ability to model long-term dependencies as required for daily activities has not been evaluated. Finally, more recent approaches drawing a parallel between activity understating from videos and document analysis are based on the unsupervised Latent Dirichlet Allocation (LDA). LDA models the probability that a video (document) is generated from a set of actions (topics) where each action (topic) is a distribution of words from a codebook. They are employed for activity recognition [37] and forgotten action detection [38].

C. *Deep Learning Approaches*

Deep learning and neural networks became very popular in computer vision for pattern recognition and especially action/activity recognition. Based on the success of Convolutional Neural Networks (CNN) for extracting relevant features from images [29], several researches proposed to employ such features for the task of action/activity recognition. For instance, Cho et. al. [7] employed CNN to first compute features on each frame and then to recognize actions from the word embedding representation of the entire

video. As CNN perform well on images, Liu et. al. [21] first proposed to represent 3D skeleton sequences as large 2D images by replacing the three channels of images by the three coordinates of joints and then employed CNN to perform action recognition. Another family of deep learning techniques called Recurrent Neural Networks (RNN) has also been employed for human motion analysis as such networks are particularly efficient to model temporal sequences. Especially, the RNN variant called Long-Short Term Memory model [16] is able to model time series while keeping track of previous observations which is useful for the task of human motion analysis. As a result, several works addressed the task of action/activity recognition by using skeleton features and multi-layers (deep) LSTM [39], [8]. Differently, Du et. al. [11] proposed a spatial hierarchical LSTM so as to consider different combinations of body parts separately. Instead of considering different spatial scales, Lee et. al. [19] investigated different temporal scales trough sliding windows and ensemble LSTM. Another extension of LSTM has been proposed in [22] by the inclusion of an attention mechanism in order to learn which joints are more relevant as a function of activities. Finally, a combination of CNN for extracting features and LSTM for temporal modelling has also been investigated [20]. The majority of these approaches have been efficiently applied on relatively simple and short activities, i.e. actions including interactions with objects, from the NTU RGB+D dataset [30]. However, the understanding of more complex activities such as daily activities and the issue of continuous analysis are not addressed in these works.

III. Proposed Approach

A. Human Skeleton Features

In this work, we focus on activities of daily living (ADL) observed using RGB-D sensors. Such sensors allow the use of powerful algorithms to extract in real-time a 3D human skeleton representing the position and the orientation of different body parts. For instance, the Microsoft Kinect v2 [23] provides the 3D position of 25 joints forming a human skeleton. Fig. 2 shows the structure of the skeleton detected using Kinect. In this work, we focus our analysis on upper body joints, using $J = 12$ joints (we discard the neck joint because it is not very informative as well as finger joints because they are very noisy).

To allow invariance to scale and translation (i.e. subjects' size and position), we employ normalized relative positions. For a given joint j, its normalized Cartesian position P_j is computed relatively to the *Spine Base* absolute position p_1 and normalized using the length L_{spine} of the spine bone (between *Spine Shoulder* joint and *Spine Mid* joint): $P_j = (p_j - p_{sb})/L_{spine}$. This row feature vector is computed for all joints except the *Spine Base* joint (j=1). In addition, we compute velocity features by computing the difference between the current joint position at frame t and its position in the previous frame $t-1$: $V_j = P_j^t - P_j^{t-1}$. The row velocity feature vector is computed for all joints. As a result, a skeleton pose x_t at frame t can be represented as:

$$x_t = [V_1, V_2, P_2, \ldots, V_J, P_J]^T \quad (1)$$

which is a 69-dimensional feature vector. Note that the data extracted from the Kinect sensor are first processed using a Butterworth filter to remove noise before computing the human skeleton features described above.

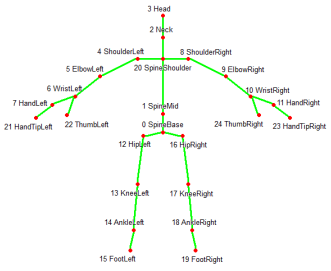

Fig. 2: Structure of the skeleton captured by Kinect v2 [23]

B. Temporal Segments

Human activities are mainly characterized by motion whose complexity is much higher compared to actions [6]. Indeed, activities are characterized by combinations of atomic actions in various order. To address this issue, a finer temporal analysis is often required [35]. In this work, we propose to decompose the entire activity sequence into a set of short temporal segments instead of treating an activity as an integral, inseparable instance. This facilitates the understanding of motion and allows to handle variations in the execution of an activity. In addition, it is more suitable for an online continuous analysis required in the context of daily activity analysis where the temporal limits of an activity are never known. We investigate two different approaches based on the use of fixed-length temporal windows and on an automatic segmentation into motion units.

For the first approach, the entire activity sequence is decomposed into short temporal windows of size WS. Since our goal is to represent the complex motion sequence as a more comprehensible set of segments and thus reduce the number of features used for activity recognition, we consider successive and non overlapping segments. Knowing the ground truth label of each training sequence of an activity, the frames of the segmented windows are appointed the same label. As a result of the segmentation, the last segment of an activity is often smaller than the window size WS, as illustrated in Fig. 3. For a test sequence, we simply segment it into successive windows of equal WS.

Fig. 3: Illustration of the fixed-length window segmentation

As an alternative to the fixed-window segmentation scheme, we divide the continuous sequence into motion units by automatically detecting salient motion changes. We consider two existing approaches based on Principal Component Analysis [2] and standard deviation within an sliding window [10]. While these approaches are efficient for simple movements like actions, their efficiency on more complex

movements like daily activities is not straightforward. More details about the two comparison of the two approaches can be found in Section IV-B.2.

C. Activity Classification using LSTM

The temporal segmentation of training sequences described above results in a set of temporal segments characterizing a part of an activity, each segment being associated with a ground truth activity label. In order to consider the dynamics of temporal segments and perform activity recognition, we propose to employ a Long-Short Term Memory Recurrent Neural Network (LSTM) [16]. LSTMs are a special kind of RNN that are able to learn long-term dependencies through the addition of memory cells and have proven their effectiveness for temporal sequences. It has been successfully used for action recognition using skeleton data. In this work we propose to adapt it for activity recognition.

For the sake of completeness, we briefly recall the main idea underlying the functionality of an LSTM. In particular, the memory cell contains several parameters and gates including an input gate i_t, a forget gate f_t, an input modulation gate g_t, an output gate o_t, an internal state c_t and an output state h_t. The LSTM transition equations are:

$$i_t = \sigma(U_i x_t + W_i h_{t-1} + b_i) \quad (2)$$

$$f_t = \sigma(U_f x_t + W_f x_{t-1} + b_f) \quad (3)$$

$$g_t = tanh(U_g x_t + W_g h_{t-1} + b_g) \quad (4)$$

$$o_t = \sigma(U_o x_t + W_o h_{t-1} + b_o) \quad (5)$$

$$c_t = f_t \odot c_{t-1} + g_t \odot i_t \quad (6)$$

$$h_t = o_t \odot tanh(c_t) \quad (7)$$

where $\odot$ indicates element-wise product, x_t is the input to the network at time step t and h_{t-1} is the hidden state at time step $t-1$. Also note that σ corresponds to a sigmoid function and $tanh$ to the hyperbolic tangent function. During training, the parameters b, U and W of cell gates are optimized.

We employ a conventional LSTM architecture with a LSTM layer, a fully connected layer and a softmax layer. In this work, we want to take advantage of LSTM to capture the dynamics of temporal segments and classify each temporal segment as an activity. As a result, the input of our LSTM is a temporal segment and the output is the activity label.

D. Temporal Dependency between Segments

By representing temporal windows with human motion features, the LSTM classifies each segment independently to what happened earlier in the sequence. For activity recognition, such information could be very helpful. For instance, the activity "putting back to fridge" should not occur before the activity "taking from fridge". To integrate such implicit information, we propose two strategies, (i) adding classification information of the previous segment in the feature of the current segment (see section III-D.1) and (ii) adding a second LSTM layer to model the dynamics of temporal segments (see section III-D.2).

1) Context Features: Inspired by the work of Dupont et.al on document analysis [14], we propose to consider information of the previous temporal segment. We perform this by increasing the dimension of the human pose features of the current segment s by concatenating it with a *one-hot* context vector C^{s-1} corresponding to the classification of the previous temporal segment $s-1$, as illustrated in Fig. 1.

The size of the context vector C^{s-1} is equal to the total number of activity classes. If the previous temporal segment is labelled as the n-th class, then the n-th element of C^{s-1} is equal to 1 and 0 elsewhere. In the case where the frame belongs to the first temporal window ($s = 1$), the context vector C^1 is a vector of zeros signifying that this activity occurs in the beginning of the sequence. The resulting feature vector at frame t belonging to segment s is $F_t^s = [x_t^s, C^{s-1}]$.

This process is done for all training sequences as we know the ground truth activity labels. During testing, the context vector is the output of the softmax layer corresponding to class probabilities of the previous temporal segment.

2) Two Layers Hierarchical LSTM: The second proposed strategy amounts to the addition of a second LSTM network in order to capture and model the dynamics of consecutive temporal segments characterizing an activity. The inputs of the second LSTM^H are the last hidden states of the first LSTM for each temporal segment, as illustrated in Fig. 1. The superscript H is used to distinguish the second LSTM. Thus the transition equations (2–5) of the second LSTM are:

$$i_t^H = \sigma(U_i^H h_t + W_i^H h_{t-1}^H + b_i^H) \quad (8)$$

$$f_t^H = \sigma(U_f^H h_t + W_f^H x_{t-1}^H + b_f^H) \quad (9)$$

$$g_t^H = tanh(U_g^H h_t + W_g^H h_{t-1}^H + b_g^H) \quad (10)$$

$$o_t^H = \sigma(U_o^H h_t + W_o^H h_{t-1}^H + b_o^H) \quad (11)$$

where h_t corresponds to the last hidden state of the first LSTM after a forward pass of the entire temporal segment.

This hierarchical model allows to handle the complexity of an activity by considering both the dynamics characterizing motion units (first layer) and the evolution of such motion units characterizing the whole activity (second layer). We trained each layer independently. The first layer is trained as described in section III-C in order to optimize the recognition of each temporal segment being part of an activity. The second layer is trained in the same way using the ground truth label of each temporal segment.

IV. Experimental Results

We evaluate our method for ADL recognition and compare its effectiveness against state-of-the-art approaches.

A. Experimental Dataset and Protocol

We evaluate our approach on the challenging human complex activity RGB-D dataset Watch-n-Patch [37], consisting of 458 videos of about 230 minutes in total recorded by the Kinect v2. Each video in the dataset contains 2 to 7 activities that may involve interaction with different objects. 7 subjects perform daily activities in 8 offices and 5 kitchens with backgrounds recorded from different viewpoints. It is

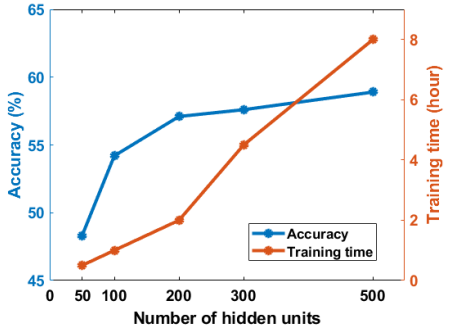

Fig. 4: Dependency on the number of hidden units in LSTM cells

composed of fully annotated 21 types of activities (10 in the office, 11 in the kitchen) interacting with 23 types of objects. The main difficulty of this dataset is that complex tasks can be different combinations of activities and with varying order, and some activities occur simultaneously such as fill-kettle and boil-water, while other activities occur in a fixed order such as turn-on-monitor and turn-off-monitor and other occur in random order.

We follow the same experimental set-up of [37] using the same selection of training and testing sequences. In order to evaluate the performance of our approach, two measures are used: a frame-level accuracy (Frame Acc) and a segment-level accuracy (Segment Acc). An activity segment is considered as correctly recognized if the intersection/union of the ground truth and the recognition is greater than 40%.

B. Model Settings Evaluation

First, we evaluate the effectiveness of the proposed approach with respect to different model parameters, segmentation strategies and model architectures. All evaluations are performed on the kitchen dataset using frame-based accuracy.

1) Latent space dimensionality: We evaluate the dimensionality of activities by measuring the performance of our model by varying the number of hidden units in the LSTM cell. In Fig. 4, we plot both the accuracy as well as the training time with respect to the number of hidden units.

We can observe that we obtain the best accuracy with 500 hidden units. However, if we analyse the training duration, we can observe that choosing a smaller number of hidden units results in a large reduction of training time while maintaining comparable performance. Using 200 hidden units could therefore be a good trade-off between accuracy and computational burden.

2) Temporal segment evaluation: This evaluation aims to automatically determine a timespan for activity recognition. We compare the fixed-length window strategy with automatic segmentation approaches, as described in section III-B, and analyse which method is more appropriate for segmenting movements for daily activity recognition. Such evaluation is performed on both training and testing accuracy. To highlight the impact of the segmentation strategy, we use our one-layer model without hierarchy. For the fixed-length windows method, we use $WS = 15$. Results are reported in Table I.

TABLE I: Evaluation of temporal segmentation methods

Method	Training acc. (%)	Test acc. (%)
PCA	23.9	10.4
STD	87.3	37.3
Windows	**100**	**49.0**

TABLE II: Evaluation of the window size parameter

WS	5	10	15	20	25	30
Acc. (%)	43.5	45.0	**49.0**	47.9	45.8	42.3
Time	≈6,5h	≈6,8h	≈7,5h	≈7,8h	≈8,4h	≈7,3h

We can see that the fixed-length windows method is more effective for both training and testing than automatic segmentation methods. We empirically observed that for similar activity sequences, the automatic segmentation methods may result in very different segmentations. For instance, the activity "pouring" can be segmented into "take bottle", "pour water" and "put bottle" for one sequence. For the same activity from another sequence, the segments "take bottle" and "pour water" may be merged together due to the noisy skeleton data. While automatic methods performed well for simpler movements like actions, they seemed less appropriate for the task of daily activities.

Finally, to complete the evaluation on the temporal segmentation, we evaluate the impact of the window size parameter. The choice of this parameter is crucial and depends on the task, as discussed in [1]. We compute the accuracy of our one-layer model with respect to various windows lengths. Results are reported in Table II.

We can notice that we obtain better results with a window's length of 15 frames which according to Watch-n-Patch dataset information [37] corresponds to 2.7 seconds. Despite the fact that activities are complex and the understanding of movements depends on the past and the context, this evaluation seems to indicate that taking into account more information from the past decreases the performance of our system. We believe this to be caused by the overwhelming amount of information provided to the LSTM.

3) Model architecture evaluation: Since activity recognition is highly non Markovian and depends on the context that can trace back further than 2.7 seconds, we propose to take into account this contextual information through two alternatives: (i) reusing the activity label of the previous time segment as input information for our recognition system, and (ii) using a hierarchical 2-layer architecture to represent a high-level representation of chaining action units. The impact of modelling temporal dependencies between segments and the comparative results of variants of our model that differ by the number of layers used (either 1-layer or 2-layer) and the consideration of the context feature are reported in Table III.

TABLE III: Comparison of variants of our model

Method	Accuracy (%)	Training time
1-layer without context feature	45.2	≈ 7.5h
1-layer with context feature	49.0	≈ 7.5h
2-layers without context feature	58.0	≈ 8h
2-layers with context feature	58.9	≈ 8h

By analysing these results, we first observe that using the

hierarchical model with two layers allows to significantly increase performance compared to the 1-layer model. This shows that the hierarchical model is indeed relevant in capturing motion complexity of daily activities. In addition, we can observe that the use of context feature results in a small but not negligible improvement of the performance when we use a model with only one layer. This demonstrates that considering temporal dependencies between successive segments is important for recognizing daily activities. When a hierarchical model is employed, the improvement of using context feature is trivial. As the hierarchical model captures dynamics of successive temporal segments and thus temporal dependencies, providing context information does not appear to contribute with additional discriminative information.

To emphasize the difference between our 1-layer and 2-layer models, we report the confusion matrices in Fig. 5. We can observe a confusion between the opposite activities 'fetch-from-fridge'/'put-back-to-fridge' and 'fetch-from-oven'/'microwaving' when the 1-layer model is employed (Fig. 5a). As temporal segments are analysed independently, the model is not able to use the context from the past to differentiate these opposite activities. Conversely, the 2-layers model differentiates these activities (Fig. 5b). This shows the importance of tackling non-Markovian properties of daily activities recognition. In our proposal, we modelled the temporal dependencies between movement segments mainly through our hierarchical LSTM model.

C. Comparison with State-of-the-art

We quantitatively compare our hierarchical LSTM model (H-LSTM) with state-of-the-art approaches on the *Watch-n-Patch* dataset. To the best of our knowledge the existing methods evaluated on this dataset are: Hidden Markov Model (HMM) [3], Latent Dirichlet Allocation (LDA) [4], Causal Topic Model (CaTM) [37] and Watch-Bot Topic Model (WBTM) [38]. All results are reported from [38]. For our method, we use a fixed-length segmentation with a window size $WS = 15$ and 500 units of the LSTM cell. We train our model using the ADAM optimization algorithm [17] with 1000 epochs. With a non optimized Matlab code and a computer with an Intel Core i5 CPU of 2,6 GHz and 8 Gb of RAM, training time is approximately 8 hours. For testing, we run our recognition model at 110 frames per second. We compute both the frame and segment accuracies. The comparative evaluation provided in Table IV shows a clear overall advantage of the proposed method.

V. Conclusion

In this paper, we proposed a vision-based approach for recognizing activities of daily living. We employ a temporally hierarchical model composed of two LSTM layers to analyse human body movement and recognize complex activities. To jointly consider the increased motion complexity of activities and allow online continuous observation, we decompose the entire skeleton sequence into short temporal segments. We employ a hierarchical model based on LSTM networks to analyse human body evolution and recognize the performed

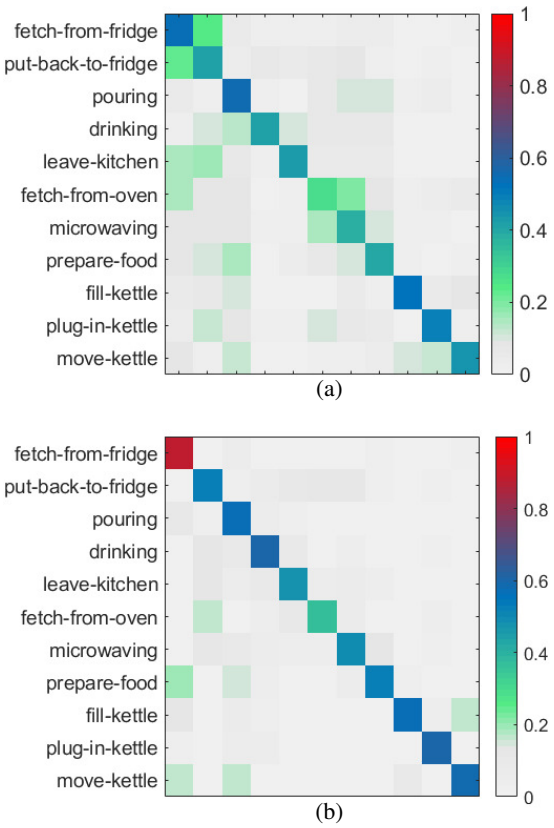

Fig. 5: Confusion matrix obtained with the proposed method when using a 1-layer model (5a) and a 2-layers model (5b)

TABLE IV: Evaluation of our hierarchichal LSTM model (H-LSTM) in comparison with state-of-the-art.

Method	Frame Acc (%)	Segment Acc (%)
kitchen		
HMM [3]	20.3	17.2
LDA [4]	14.0	6.7
CaTM [37]	34.0	29.0
WBTM [38]	39.2	33.2
H-LSTM	**58,9**	**38.8**
office		
HMM [3]	27.3	19.4
LDA [4]	18.4	12.2
CaTM [37]	38.5	32.9
WBTM [38]	41.2	35.2
H-LSTM	**58.0**	**40.2**

activities. Experiments on a very challenging dataset demonstrate the effectiveness of our approach in comparison to state-of-the-art approaches. Due to an unavoidable heterogeneity among the compared approaches (e.g. consideration of objects, supervision, etc), our future work will consist in a more systematic evaluation of each factor.

Our main contribution through the proposal of this hierarchical LSTM is to highlight the importance of modelling temporal dependencies between behaviour units through the exploration of two methods. Firstly, we considered a context input to take into account the activity label of the previous time segment and longer time-scale recognition layer. Secondly, we proposed a hierarchical representation of behaviours based on a low-level representation of behaviour

units and a higher-level representation of chaining behaviour units. Our analysis suggests that the effects of the high-level layer are more important than the effects of the context. In the future, we also plan to extend our model with an attention mechanism as its effectiveness has been proven for simple movements such as actions. Finally, we further wish to explore objects in the vicinity of human activities and human-object interactions, as an additional context feature.

REFERENCES

[1] Oresti Banos, Juan-Manuel Galvez, Miguel Damas, Hector Pomares, and Ignacio Rojas. Window size impact in human activity recognition. *Sensors*, 14(4):6474–6499, 2014.

[2] Jernej Barbič, Alla Safonova, Jia-Yu Pan, Christos Faloutsos, Jessica K. Hodgins, and Nancy S. Pollard. Segmenting motion capture data into distinct behaviors. In *Proceedings of Graphics Interface*, pages 185–194, 2004.

[3] Leonard E Baum and Ted Petrie. Statistical inference for probabilistic functions of finite state markov chains. *The annals of mathematical statistics*, 37(6):1554–1563, 1966.

[4] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3:993–1022, 2003.

[5] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, pages 7291–7299, 2017.

[6] Alexandros Andr Chaaraoui, Pau Climent-Prez, and Francisco Flrez-Revuelta. A review on vision techniques applied to human behaviour analysis for ambient-assisted living. *Expert Systems with Applications*, 39(12):1087310888, 2012.

[7] Sangwoo Cho and Hassan Foroosh. A temporal sequence learning for action recognition and prediction. In *IEEE Winter Conf. on Applications of Computer Vision*, pages 352–361, 2018.

[8] Srijan Das, Michal Koperski, Francois Bremond, and Gianpiero Francesca. Deep-temporal lstm for daily living action recognition. In *IEEE Int. Conf. on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, 2018.

[9] Maxime Devanne, Stefano Berretti, Pietro Pala, Hazem Wannous, Mohamed Daoudi, and Alberto Del Bimbo. Motion segment decomposition of rgb-d sequences for human behavior understanding. *Pattern Recognition*, 61:222–233, 2017.

[10] Maxime Devanne, Hazem Wannous, Mohamed Daoudi, Stefano Berretti, Alberto Del Bimbo, and Pietro Pala. Learning shape variations of motion trajectories for gait analysis. In *IEEE Int. Conf. on Pattern Recognition*, pages 895–900, 2016.

[11] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, pages 1110–1118, 2015.

[12] Yadin Dudai, Avi Karni, and Jan Born. The consolidation and transformation of memory. *Neuron*, 88(1):20 – 32, 2015.

[13] Nicolas Duminy, Sao Mai Nguyen, and Dominique Duhaut. Learning a set of interrelated tasks by using a succession of motor policies for a socially guided intrinsically motivated learner. *Frontiers in Neurorobotics*, 12:87, 2019.

[14] Yoann Dupont, Marco Dinarelli, and Isabelle Tellier. Label-dependencies aware recurrent neural networks. In *Int. Conf. on Computational Linguistics and Intelligent Text Processing*, pages 44–66, 2017.

[15] Okihide Hikosaka, Hiroyuki Nakahara, Miya K. Rand, Katsuyuki Sakai, Xiaofeng Lu, Kae Nakamura, Shigehiro Miyachi, and Kenji Doya. Parallel neural networks for learning sequential procedures. *Trends in Neurosciences*, 22(10):464 – 471, 1999.

[16] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

[18] H. S. Koppula and A. Saxena. Learning spatio-temporal structure from rgb-d videos for human activity detection and anticipation. In *Int. Conf. on Machine Learning*, 2013.

[19] Inwoong Lee, Doyoung Kim, Seoungyoon Kang, and Sanghoon Lee. Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks. In *IEEE Int. Conf. on Computer Vision*, pages 1012–1020, 2017.

[20] Chuankun Li, Pichao Wang, Shuang Wang, Yonghong Hou, and Wanqing Li. Skeleton-based action recognition using lstm and cnn. In *IEEE Int. Conf. on Multimedia & Expo Workshops*, pages 585–590, 2017.

[21] Jian Liu, Naveed Akhtar, and Ajmal Mian. Skepxels: Spatio-temporal image representation of human skeleton joints for action recognition. *arXiv:1711.05941*, 2017.

[22] Jun Liu, Gang Wang, Ling-Yu Duan, Kamila Abdiyeva, and Alex C Kot. Skeleton-based human action recognition with global context-aware attention lstm networks. *IEEE Transactions on Image Processing*, 27(4):1586–1599, 2018.

[23] Microsoft Kinect v2. https://developer.microsoft.com/en-us/windows/kinect, 2014.

[24] Valsamis Ntouskos, Panagiotis Papadakis, and Fiora Pirri. A comprehensive analysis of human motion capture data for action recognition. In *Int. Conf. on Computer Vision Theory and Applications*, pages 647–652, 2012.

[25] Valsamis Ntouskos, Panagiotis Papadakis, and Fiora Pirri. Probabilistic discriminative dimensionality reduction for pose-based action recognition. In *Pattern Recognition Applications and Methods*, pages 137–152, 2015.

[26] Lasitha Piyathilaka and Sarath Kodagoda. Human activity recognition for domestic robots. In *Field and Service Robotics*, pages 395–408, 2015.

[27] Natraj Raman and Stephen J Maybank. Activity recognition using a supervised non-parametric hierarchical hmm. *Neurocomputing*, 199:163–177, 2016.

[28] Karinne Ramirez-Amaro, Michael Beetz, and Gordon Cheng. Transferring skills to humanoid robots by extracting semantic representations from observations of human activities. *Artificial Intelligence*, 247:95–118, 2017.

[29] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[30] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, pages 1010–1019, 2016.

[31] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. Real-time human pose recognition in parts from single depth images. In *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, pages 1–8, 2011.

[32] Richard S Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2), 1999.

[33] Jawad Tayyub, Majd Hawasly, David C Hogg, and Anthony G Cohn. Learning hierarchical models of complex daily activities from annotated videos. In *IEEE Winter Conf. on Applications of Computer Vision*, pages 1633–1641, 2018.

[34] Jawad Tayyub, Aryana Tavanai, Yiannis Gatsoulis, Anthony G Cohn, and David C Hogg. Qualitative and quantitative spatio-temporal relations in daily living activity recognition. In *Asian Conf. on Computer Vision*, pages 115–130, 2014.

[35] Michalis Vrigkas, Christophoros Nikou, and Ioannis A Kakadiaris. A review of human activity recognition methods. *Frontiers in Robotics and AI*, 2:28, 2015.

[36] Joshua M Wiener, Raymond J Hanley, Robert Clark, and Joan F Van Nostrand. Measuring the activities of daily living: Comparisons across national surveys. *Journal of gerontology*, 45(6):229–237, 1990.

[37] Chenxia Wu, Jiemi Zhang, Silvio Savarese, and Ashutosh Saxena. Watch-n-patch: Unsupervised understanding of actions and relations. In *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2015.

[38] Chenxia Wu, Jiemi Zhang, Ozan Sener, Bart Selman, Silvio Savarese, and Ashutosh Saxena. Watch-n-patch: unsupervised learning of actions and relations. *IEEE transactions on pattern analysis and machine intelligence*, 40(2):467–481, 2018.

[39] Songyang Zhang, Xiaoming Liu, and Jun Xiao. On geometric features for skeleton-based action recognition using multilayer lstm networks. In *IEEE Winter Conf. on Applications of Computer Vision*, pages 148–157, 2017.