In: European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), pp. 465-470, Bruges, Belgium (2016)

Activity recognition with echo state networks using 3D body joints and objects category

Luiza Mici, Xavier Hinaut and Stefan Wermter *

University of Hamburg - Department of Informatics,
Vogt-Koelln-Strasse 30, 22527 Hamburg - Germany
http://www.informatik.uni-hamburg.de/WTM/

Abstract. In this paper we present our experiments with an echo state network (ESN) for the task of classifying high-level human activities from video data. ESNs are recurrent neural networks which are biologically plausible, fast to train and they perform well in processing arbitrary sequential data. We focus on the integration of body motion with the information on objects manipulated during the activity, in order to overcome the visual ambiguities introduced by the processing of articulated body motion. We investigate the outputs learned and the accuracy of classification obtained with ESNs by using a challenging dataset of long high-level activities. We finally report the results achieved on this dataset.

1 Introduction

The importance of recognizing articulated human activities is exhibited in a wide range of computer vision applications such as surveillance and multimedia retrieval and complex robotic applications such as human-robot communication and learning from demonstration. High-level human activities processed by computer vision are temporal sequences including a number of simple atomic actions such as standing, walking, reaching for an object, picking up an object etc. The major challenges faced when dealing with activity recognition tasks are the large variability of style and velocity of execution with which actions are performed by different subjects [1]. Therefore, recent studies [2, 3, 4, 6] are concentrating their efforts on integrating additional contextual cues (e.g. manipulated objects, spatial analysis of the scene etc.) with the body motion and pose information in order to enhance performance and reliability of the recognition despite highly probable visual ambiguities. While these approaches make use of machine learning algorithms for classification, we aim at investigating applications of bio-inspired methods based on recurrent neural networks (RNN), given the high efficiency of the mammalian brain in processing temporal data. The echo state networks (ESN)[8] are novel recurrent networks that have demonstrated advantages over traditional RNN due to their simplicity, computational efficiency and promising results in demanding tasks such as language acquisition [10]. Moreover they are suitable for both online and offline learning and their

*This research was partially supported by a Marie Curie Intra European Fellowship within the 7th European Community Framework Programme: EchoRob project (PIEF-GA-2013-627156).

In: European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), pp. 465-470, Bruges, Belgium (2016)

biological plausibility has caught the interest of studies on modelling canonical neural circuits of primate prefrontal cortex [11].

Our work focuses on the study and evaluation of the performance of ESNs for the task of activity recognition, by using a challenging benchmark dataset of RGB-D videos, Cornell Activity Dataset-120 (CAD-120)[1]. We are mainly interested in the use of depth, which leads to robustness under varying light conditions and differing viewpoints of observation and reduced computational costs [12]. We have set up experiments seeking to investigate the outputs obtained with an ESN having as input three dimensional coordinates of articulated body skeleton joints as well as category labels of objects being manipulated by the subject during the activity. In Section 2 we describe the ESN algorithm implemented for our experiments. In Section 3 the set-up of the experiments and implementation details are reported and in Section 4 we provide experimental results and comparison to the state-of-the-art recognition rates for the CAD-120 dataset.

2 Echo State Networks

We implemented ESNs with leaky integrator neurons. The state of the reservoir units is driven by the n-dimensional input sequences u_t where t is the time step, and updates of the states are computed through the following equation:

$$\mathrm{x}_{t+1} = (1-\alpha)\mathrm{x}_t + \alpha \tanh(W^{in}[1;\mathrm{u}_{t+1}] + W^{res}\mathrm{x}_t), \tag{1}$$

where x_t is the state of the reservoir neurons at time step t, tanh(.) is the activation function applied element-wise, W^{res} is the reservoir weight matrix, W^{in} is the input weight matrix and α is the leaking rate. The learning of the output weights W^{out} is typically performed through ridge regression with Tikhonov regularization:

$$W^{out} = Y^{target}X^T(XX^T + \beta I)^{-1}, \tag{2}$$

where Y^{target} are the desired outputs, X is the matrix of reservoir state sequences, β is a regularization parameter and I is the identity matrix. After the training, the computed outputs y_t are obtained using:

$$\mathrm{y}_t = W^{out}[1;\mathrm{u}_t;\mathrm{x}_t], \tag{3}$$

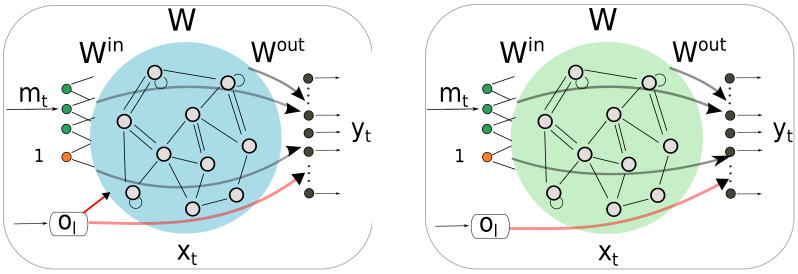
where W^{out} are the weights of connections between readout units and the reservoir and input units plus the constant bias.

3 Experiments

All experiments were carried out using CAD-120 benchmark dataset, which comprises RGB-D videos of 10 long daily activities: *arranging objects*, *cleaning objects*, *having meal*, *making cereal*, *microwaving food*, *picking objects*, *stacking*

[1]Cornell Activity Dataset-120. http://pr.cs.cornell.edu/humanactivities/data.php

In: European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), pp. 465-470, Bruges, Belgium (2016)



(a) $model_{jo}$ (b) $model_j$

Fig. 1: Two separate experiments run with two echo state networks: (a) object labels o_l are fed in input to the reservoir together with the skeleton joints coordinates m_t; (b) object labels o_l are kept out of the reservoir state sequence computations and used only for learning output weights W^{out}.

objects, *taking food*, *taking medicine*, *unstacking objects*. The activities are performed by 4 different subjects repeating each action three to four times. The dataset provides three dimensional coordinates of 15 joints and ground-truth category labels of manipulated objects for each activity. We used the whole skeletal data including erroneous sequences taking advantage of the robustness of ESN towards noisy real world inputs [9].

The simplest way to incorporate the objects' information (i.e. presence and type of object) into the problem is by directly giving ground-truth labels as input to the ESN model. The number and type of objects vary between different activities as well as for each activity repetition. Therefore, in order to have a fixed dimension input data, we encoded the object labels using one-of-k-encoding, i.e. all elements were set to 0 except the one with the index corresponding to the object category. The matrix of skeleton joints coordinates was normalized in the range $[-1, 1]$ for each activity sequence. The vector of object labels o_l was concatenated with the vector of skeleton joints coordinates m_t at each time step $t = 1, \ldots, T$ obtaining the augmented input vector $\hat{u}_t^i = [1; o_l; m_t]$, where 1 represents the bias term.

We tested two models with different set-ups of inputs (see Figure 1): (i) full connection between augmented inputs and the reservoir neurons, and (ii) zero weights for connections between o_l and the reservoir neurons. While in the former case the object labels influence the reservoir internal representations, in the latter case they are used only for training the output weights W^{out}. Matrix weights of W^{in} and W^{res} were initialized randomly with a uniform distribution in [-0.5, 0.5]. Since our task is multi-label classification of long sequences and the teacher signal is given during the whole activity sequence, we performed the readout of the outputs in three different ways: (i) by recording the activity of the readout units and averaging over the whole length of the sequence, (ii) by recording and averaging the activity of the readout units for the last half of the sequence or (iii) by keeping track of the activity of the readout units at the last

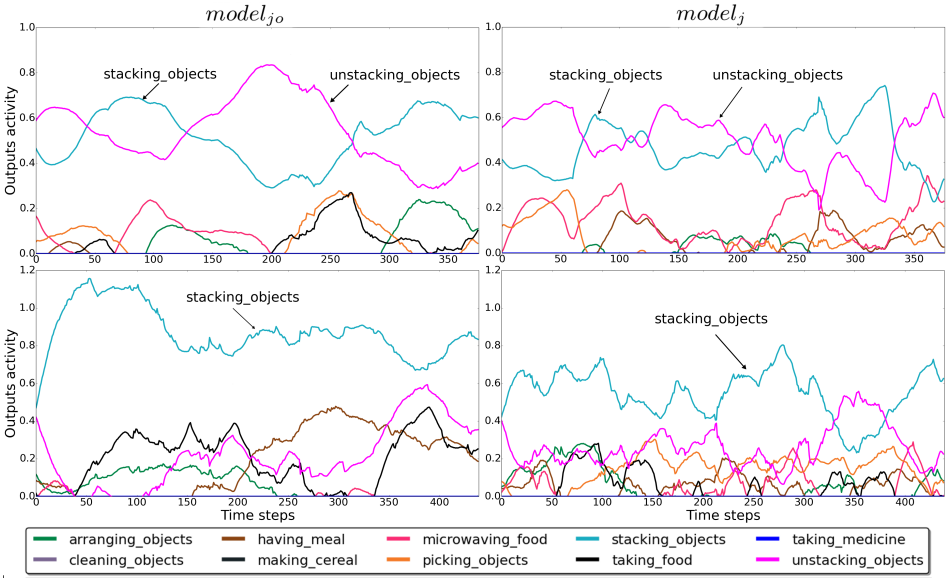


Fig. 2: Output units activations of the two ESN models during testing on an unseen subject, $model_{jo}$ (left column) and $model_j$ (right column). The ground truth label in all cases is *stacking objects*. First row: the models interchange *stacking objects* with *unstacking objects*, two activities involving the same objects and similar body motions. Second row: the models classify correctly the sequence from a different fold.

time step.

The set of global parameters influencing the dynamics of the ESN, namely sparsity and spectral radius of W^{res}, sparsity and scaling of W^{in}, the leaking rate and the ridge regularization parameter were chosen through a Bayesian optimization search using the *hyperopt* python toolbox [13]. We fixed the number of neurons to $30, 50, 100, 300$ and optimized the other parameters directly by maximizing the accuracy in activity recognition. The accuracy was computed using the 4-fold cross validation found in the literature for the CAD-120 dataset [2, 6, 5], i.e. the ESN model was trained on videos of 3 subjects and tested on an *unseen* subject. This type of cross-validation is quite challenging since different subjects perform the same action in a different manner. Since even with the same set of parameters, the performance of ESN fluctuates due to random initialization of weights W^{in} and W^{res}, we averaged the accuracy over 30 trials run with different random reservoir initializations.

4 Results

Extensive parameters search with *hyperopt* gave us two different sets of optimal global parameters for the two models presented in Section 3. For the $model_{jo}$,

In: European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), pp. 465-470, Bruges, Belgium (2016)

where the vector of object labels is fully connected to the reservoir neurons (Figure 1a) we had size of reservoir of 300 neurons, leaking rate $\alpha = 0.02$, spectral radius of the reservoir matrix 0.3 and ridge regularization parameter $\beta = 5.2 * 10^{-10}$. For the $model_j$, where the vector of object labels is connected only to the output units (Figure 1b), we had size of reservoir of 300 neurons, leaking rate $\alpha = 0.06$, spectral radius 1.2 and ridge parameter $\beta = 7.2 * 10^{-5}$. The spectral radius greater than one is not violating the echo state property in our case. When using a leaky-ESN (i.e. $\alpha < 1$), the *effective* spectral radius, which should be kept smaller than one, is different than the spectral radius of W^{res} [7]. The low leaking rate is to be expected in long data sequences as is the case of CAD-120 dataset, where the length of each sequence varies from ≈ 150 to ≈ 900 frames per video. Across different categories of activity, averaging readouts of output units over the whole sequence frames gave better results than the other strategies that we tested (described in Section 3). Examples of outputs generated by the two models are given in Figure 2 for comparison. We inspected that the activities interchanged by both models in more than half of the misclassifications were the ones including the same category of objects and similar body motions, e.g. *stacking objects* and *unstacking objects*. In fact in these activities the subject repeats the same sequence of atomic actions: reaching, moving and placing objects. The classification accuracy of the ESN models using

Algorithm	Accuracy%	Precision%	Recall%
Koppula et al. [2]	80.6 ± 1.1	81.8 ± 2.2	80.0 ± 1.2
Koppula et al. [6]	83.1 ± 3	87.0 ± 3.6	82.7 ± 3.1
Rybok et al. [5]	78.2	*	*
ESN $model_{jo}$ (average)	81.5 ± 6	81.5 ± 7.3	80.9 ± 6.2
ESN $model_{jo}$ (best)	$\mathbf{88.7 \pm 3.6}$	$\mathbf{90.3 \pm 3.3}$	$\mathbf{88.3 \pm 3.7}$
ESN $model_j$ (average)	80.0 ± 5.7	79.7 ± 9	79.4 ± 5.9
ESN $model_j$ (best)	87.1 ± 0.1	90.3 ± 1.4	86.7 ± 0.0

Table 1: Performance results of different methods on the CAD-120 dataset not using ground-truth temporal segmentation.

the best global parameters, averaged over 30 random reservoir initializations, as well as their best runs with one specific reservoir instance are reported in Table 1. The latter outperform other approaches. The standard deviation of the average performance has been calculated over all 30 trials and averaged across four folds used in CAD-120 dataset. The standard deviation of the best runs has been calculated across the four folds of one trial. A direct comparison of the results in Table 1 needs some caution though. The approaches mainly differ in two points: (i) we use directly ground-truth object labels, while other approaches use visual features extracted from object tracking and detection, and (ii) the recognition of high-level activities in the other approaches depends on the successful recognition of shorter sequences of atomic actions called subactvities, while in our approach only high-level activity labels are used. In fact we assume that by using the information about objects position in the scene and sub-activity labels, the performance of our ESN models would further increase.

In: European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), pp. 465-470, Bruges, Belgium (2016)

5 Conclusions and future work

Our study has shown a successful application of ESN models for the task of complex human activity recognition, where combination of articulated body motion and manipulated objects is required. The accuracy of the implemented models was evaluated through a challenging benchmark dataset of RGB-D videos comprising long sequences of high-level activities. The recognition rates obtained so far are comparable with the best state of the art and motivate further experiments which can potentially enhance the results, e.g. by using teacher signals for smaller atomic actions and using a deeper architecture which takes care of learning skeletal data sequences in order to remove noise caused by tracking errors. The work presented in this paper provides us with a stepping stone towards real-time activity recognition, which is a crucial task for several applications including human-robot interaction.

References

[1] R. Poppe. A survey on vision-based human action recognition. *Image and vision computing*, 28(6):976-990, 2010.

[2] H. S. Koppula et al. Learning human activities and object affordances from rgb-d videos. *The Intl. Journal of Robotics Research*, 32(8), 951-970, 2013.

[3] N. Hu et al. Learning latent structure for activity recognition. In proc. of *IEEE Intl. Conf. on Robotics and Automation* (ICRA 2014), pp. 1048-1053, 2014.

[4] A.Gupta et al. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(10), 1775-1789, 2009.

[5] L. Rybok et al. "Important stuff, everywhere!" Activity recognition with salient proto-objects as context. In proc. of *IEEE Winter Conference on Applications of Computer Vision* (WACV), pp. 646-651, 2014.

[6] H. Koppula, and A. Saxena. Learning spatio-temporal structure from rgb-d videos for human activity detection and anticipation. In proc. of the 30^{th} *Intl. Conf. on Machine Learning* (ICML-2013), pp. 792-800, 2013.

[7] H. Jaeger et al. Optimization and applications of echo state networks with leaky-integrator neurons. *Neural Networks*, 20(3), 335-352. 2007.

[8] H. Jaeger. The "echo state" approach to analysing and training recurrent neural networks-with an erratum note. Bonn, Germany: German National Research Center for Information Technology GMD Technical Report 148 (2001): 34.

[9] S. Scherer et al. Real-time emotion recognition from speech using echo state networks. In *Artificial neural networks in pattern recognition* pp. 205-216, 2008.

[10] X. Hinaut and S. Wermter. An Incremental Approach to Language Acquisition: Thematic Role Assignment with Echo State Networks. In proc. of *Artificial Neural Networks and Machine Learning* (ICANN 2014), pp 33-40, 2014.

[11] X. Hinaut et al. A recurrent neural network for multiple language acquisition: Starting with English and French. In *Proc. of the NIPS 2015 workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches*. Montreal, Canada, 2015.

[12] J. Han et al. Enhanced computer vision with microsoft kinect sensor: A review. *IEEE Transactions on Cybernetics*, 43(5):1318-1334, 2013.

[13] J. Bergstra et al. Hyperopt: A Python library for optimizing the hyperparameters of machine learning algorithms. In proc. of the 12^{th} *Python in Science Conference* pp. 13-20, 2013.