In: *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN'20),* pp. 591-596, Bruges, Belgium, October 2-4, 2020.

Self-Organizing Kernel-based Convolutional Echo State Network for Human Actions Recognition

Gin Chong Lee[1], Chu Kiong Loo[2], Wei Shiung Liew[2], and Stefan Wermter[3] *

1- Multimedia University - Faculty of Engineering and Technology
Jalan Ayer Keroh Lama, 75450 Melaka. - Malaysia

2- University of Malaya - Faculty of Computer Science and Information Technology
50603 Lembah Pantai, Kuala Lumpur. - Malaysia

3- University of Hamburg - Department of Informatics
Vogt-Koelln-Str.30, 22527 Hamburg. - Germany

Abstract. We propose a deterministic initialization of the Echo State Network reservoirs to ensure that the activation of its internal echo state representations reflects similar topological qualities of the input signal which should lead to a self-organizing reservoir. Human actions encoded as a multivariate time series signal are clustered before using the clustered nodes and interconnectivity matrices for initializing the S-ConvESN reservoirs. The capability of S-ConvESN is evaluated using several 3D-skeleton-based action recognition datasets.

1 Introduction

Current research in human action recognition (HAR) focuses on the challenge for efficient and effective modeling the temporal features of human actions in 3-dimensional space. Echo state networks (ESNs) are one suitable method for encoding the temporal context due to its short-term memory property. The random assignment of the ESN's input and reservoir weights reduces the computational complexity compared to backpropagation but also increases instability and variance in generalization [1]. Using self-organizing kernel networks in the formation of ESN reservoirs ensures that the activation of its internal echo state representations reflects similar topological qualities of the input signal, acting as a feature map which should lead to a self-organizing reservoir [2]. Inspired by the notion that input-dependent self-organization is decisive for the cortex to adjust the neurons according to the distribution of the inputs [3], the potential of unsupervised self-organizing learning seems to be one of the most encouraging and the most biologically plausible.

This work proposes an approach to implement a self-organizing kernel network in performing deterministic initialization of the input weights and recurrent hidden weights in the ESN stage. This paper is organized as follows: Section 2 briefly discusses the implementation of self-organizing kernel networks, while

*This research was supported by the Georg Forster Research Fellowship for Experienced Researchers from Alexander von Humboldt-Stiftung/Foundation, and the ONRG International Fund (IF017-2018).

Section 3 reports the results of a benchmarking experiment using several 3D-skeleton-based HAR datasets, as well as the effects of manipulating the SOM hyperparameters. Concluding remarks and future perspectives are drawn in Section 4.

2 Self-Organizing Kernel-based Convolutional Echo State Networks (S-ConvESN)

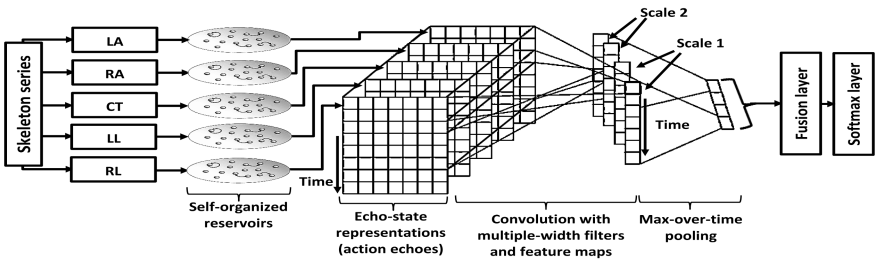


Fig. 1: Implementation of self-organizing kernel-based reservoirs in multi-step, multi-channel convolutional Echo State Network with 5 channels, 2 filters, and 2 time scales.

The general process of the S-ConvESN is shown in Figure 1. Each channel corresponds to the joint coordinate trajectories for a single body part. Clustering and convolution were performed separately for each channel.

2.1 Self-Organizing Kernel-Based Network

Self-organizing maps (SOM) [4] consist of a clustered topology of nodes from unsupervised training of a dataset, each node representing a sufficiently dissimilar training sample or archetype, while sufficiently similar training samples were often represented by one node or a cluster of nodes. The generated maps preserve the topological properties of the input space at a significantly-reduced dimensionality. According to the stochastic resonance theory, adding noise prior to clustering would speed up convergence in a centroid-based clustering algorithm [5]. Assuming $x(t)$ represents the joint coordinates at a single time instance t, clustering was conducted as follows:

$$z(t) = x(t) + \frac{\eta}{t^2} I \tag{1}$$

$$\mathrm{CIM}(z(t), w_j, \sigma_j) = [\kappa_{\sigma_j}(0) - \kappa_{\sigma_j}(z(t) - w_j)]^{\frac{1}{2}} \tag{2}$$

$$b = \arg\min_{j \in J} [\mathrm{CIM}(z(t), C, S)] \tag{3}$$

where $\eta = [0, 1]$ controls the magnitude of the noise, and the noisy signal $z(t)$ would have progressively less noise over time, CIM is the correntropy-induced

metric between the $z(t)$ and node j calculated using a Gaussian kernel function κ_{σ_j} [6], σ_j is the kernel bandwidth, S is the vector of individual kernel bandwidths, w_j is the weights for node j, b is the index of the best-matching node, and C is the self-organized centroids with J nodes. Nodes are added and updated using Hebbian rules:

$$\begin{cases} w_m = w_m + \epsilon_m(x(t) - w_m) & \text{if CIM}(z(t), w_b, \sigma_b) <= V \\ K \leftarrow K + 1, w_K = x(t), & \text{else} \end{cases} \tag{4}$$

where m represents the indices of the best-matching node and its topological neighbors, V is the vigilance threshold, and ϵ_m is the learning rate where the best-matching node is updated much faster than its neigbors. A sparse interconnectivity matrix is constructed by incrementing edges between the best-matching node b_1 and second-best matching node b_2, $\Delta E(b_1, b_2) = 1$. On conclusion of the clustering, the node centroids C and the interconnectivity matrix E were extracted to be used for initializing the reservoir of the ConvESN [7].

2.2 Convolutional Echo State Network

The architecture of the S-ConvESN consists of three layers; the input weight layer W^{in}, the reservoir layer W^{res}, and the output weight layer W^{out}. Given a time-series input $u = (u(0), ..., u(T-1))$, an initial state $x(0) \in \mathbb{R}^N$ in the reservoir, and an output series $y = (y(0), ..., y(T-1))$, the update equation for the system is given as:

$$x(t+1) \quad = \quad f(W^{res}x(t) + W^{in}u(t+1)) \tag{5}$$

where W^{in} is initialized using the clustered centroid weights C rescaled to the input scaling parameter $I_s = 0.1$, while W^{res} is initialized using the interconnectivity matrix E rescaled to $[-0.5, 0.5]$ and multiplied by the Spectral Radius parameter $S_r = 0.99$ to observe the echo state property [8].

Multiscale temporal invariance is maintained using max-over-time pooling, and multiscale features are derived from echo-state representations (ESR) using multiple filters widths and feature maps. Assuming $w_{kj} \in \mathbb{R}^{k \times N}$ denotes the j-th filter with k-width, the convolution result with w_{kj} is given as:

$$c_{kj} \quad = \quad (c_0, c_1, ..., c_{T-k+1:T})^T \tag{6}$$

$$c_m \quad = \quad f(\sum_i \alpha^i_{kj} \cdot (w_{kj} * z^i_{m:m+k-1}) + b) \tag{7}$$

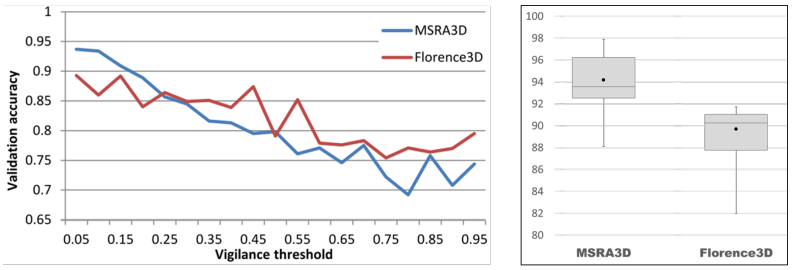
where $m = [1, 2, ..., T-k+1]$ is the index of the sliding window, z^i_m is the temporal window, f is the nonlinear activation function, α^i_{kj} is the connective weight between the i-th channel reservoir and the j-th filter with k-width, and $*$ denotes a dot-product operation.

Max-over-time pooling was used in the pooling layer to obtain the extracted features, combined based on relevance as shown in Figure 1. In the final layer,

outputs are defined as the conditional distribution $p(C_s|u)$ over action labels, where C_s denotes the s-th class of actions and $p(C_s|u)$ is the output of the softmax function.

3 Experiment Results

With the aim of analyzing the capability of the self-organizing reservoir, S-ConvESN was benchmarked using two skeleton-based action recognition datasets: **MSR-Action 3D(MSRA3D)** [9] and **Florence3D-Action (Florence3D)** [10]. To facilitate comparison with state of the art results, training and testing protocols were applied as follows. MSRA3D used the standard validation protocols [9], three training and validation sets were created with half of the subjects used for training and the other half for validation. For Florence3D, ten-fold cross-validation method was used for training and validation.



(a) (b)

Fig. 2: (a) Validation accuracy of the S-ConvESN in response to clustered centroids with varying vigilance thresholds. (b) The boxplots show the accuracy distribution for the state of the art results for MSRA3D (left) and Florence3D (right).

Vigilance thresholds were tested for a range $[0.05, 0.95]$. In addition to benchmarking for different vigilance thresholds, reservoir perturbation was also considered. Clustering was conducted for different initial noise distribution scales $\eta = [0, 0.1, 0.01, 0.001]$.

As shown in Figure 2 (a), the optimal clustering configuration was obtained by setting the vigilance threshold to a low value. In both models, setting the vigilance threshold to 0.05 resulted in the peak validation accuracy compared to other vigilance thresholds. This suggests that the feature map requires high-granularity clusters to represent a comprehensive set of unique joint coordinates. Comparing different magnitudes of perturbation, S-ConvESN showed improved accuracy when noise distribution was set to 0.1, while setting the magnitude to 0.01 and 0.001 produced no discernible improvement compared to the noise-less result.

Figure 2(b) shows the distribution of the state of the art recognition accuracy for the Florence3D and MSRA3D datasets. The bubble indicates the overall

MSR-Action 3D	
Approaches	*Ave(%)*
Covariance [11]	88.10
Skeletons Lie group [12]	92.40
DHMM+SL[13]	92.91
S-ConvESN(Our approach)	94.21
Gram matrices rep.[14]	96.90
ConvESN [7]	97.88

Florence3D-Action	
Approaches	*Ave(%)*
Multi-Part Bag-of-Poses [10]	82.00
S-ConvESN(Our approach)	89.70
Skeletons Lie group [12]	90.88
ConvESN [7]	91.72

Table 1: Recognition accuracy on cross-subject test of MSR-Action 3D dataset and on 10-fold cross validation of Florence3D-Action dataset

accuracy of S-ConvESN which stays within interquartile range. The confusion matrices for Florence3D and MSRA3D are depicted in Figure 3.

Table 1 shows the state of the art recognition accuracy for MRS-Action 3D dataset and Florence3D-Action dataset respectively. For the Florence3D, S-ConvESN achieves 89.70% overall accuracy. There is some confusion between the actions for "wave", "drink water", "listen to phone", and "look at watch", presumably due to all of them having a characteristic arm movement towards the head. The "check watch" action was often misclassified as "listen to phone". For the MSRA3D, S-ConvESN exhibits 94.21% overall accuracy, performing poorly for "high arm wave" and "draw X" actions. The "draw X" action was often mistaken for the "draw circle" action.

Experimental results on HAR task show that self-organizing reservoir is competitive with state-of-the-art approaches. The proposed reservoir design method is biologically feasible. By implementing the mechanism inspired by cortex neuron adjustment, self-organizing and deterministic initialization of ESN reservoir ensures topological information of the input signal is to be included into the reservoir.

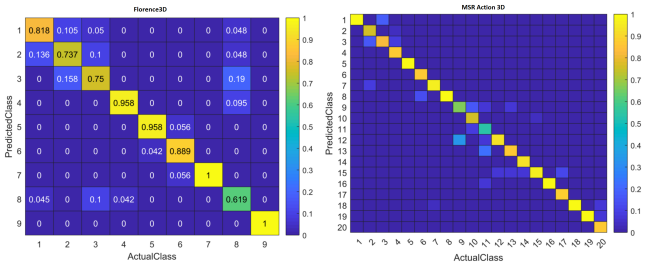


Fig. 3: The confusion matrices for Florence3D and MSR Action 3D datasets.

4 Conclusion

This paper presents a self-organizing kernel-based reservoir design for a convolutional ESN. Deterministic initialization instead of randomly initialization of the

input weight and recurrent hidden weight exhibits successful and feasible application in generating stable reservoir and with proper scaling factor to ensure echo state property. The recognition rates are comparable with the state of the art, and motivate further enhancements on the robustness of the approach such as incremental learning, or by optimizing a number of parameters in the clustering (i.e. node pruning) and reservoir (i.e. weight scaling factors). In addition, we can explore other fusion strategies in the CNN architecture.

References

[1] Q. Wu, E. Fokoue, and D. Kudithipudi, On the statistical challenges of echo state networks and some potential remedies, *arXiv preprint arXiv:1802.07369*, 2018.

[2] L. Boccato, R. Attux, and F. J. Von Zuben, Self-organization and lateral interaction in echo state network reservoirs. *Neurocomputing* 138:297-309, 2014.

[3] C. A. Nelson, Neural plasticity and human development: the role of early experience in sculpting memory systems, *Developmental Science*, 3:2:115-136, 2000.

[4] T. Kohonen, The self-organizing map. *Proceedings of the IEEE* 78:9:1464-1480, 1990.

[5] O. Osoba, and B. Kosko, Noise-enhanced clustering and competitive learning algorithms, *Neural Networks*, 37:132-140, 2013.

[6] N. Masuyama, C. K. Loo, and S. Wermter, A kernel Bayesian adaptive resonance theory with a topological structure, *International Journal of Neural Systems*, 29:5, 2019.

[7] Q. Ma, L. Shen, E. Chen, S. Tian, J. Wang, and G. W. Cottrell, WALKING WALKing walking: Action Recognition from Action Echoes, *proceedings of the International Joint Conference on Artificial Intelligence (IJCAI) 2017*, pages 2457-2463, August 19-25, Melbourne (Australia), 2017.

[8] H. Jaeger, The "echo state" approach to analysing and training recurrent neural networks-with an erratum note. Technical Report, German National Research Center for Information Technology GMD, 148(34), 13, 2001.

[9] W. Li, Z. Zhang, and Z. Liu, Action recognition based on a bag of 3d points. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 9-14, June 13-18, San Francisco (USA), 2010.

[10] L. Seidenari, V. Varano, S. Berretti, A. D. Bimbo, and P. Pala, Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses. In *Conference on Computer Vision and Pattern Recognition Workshops*, pages 479-485, June 23-28, Portland (USA), 2013.

[11] M. E. Hussein, M. Torki, M. A.Gowayyed, and M. El-Saban, Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations. In *International Joint Conference on Artificial Intelligence*, volume 13, pages 2466-2472, 2013.

[12] R. Vemulapalli, F. Arrate, and R. Chellappa, Human action recognition by representing 3d skeletons as points in a lie group. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 588-595, June 24-27, Columbus (USA), 2014.

[13] L. L. Presti, M. L.Cascia, S. Sclaroff, and O, Camps, Hankelet-based dynamical systems modeling for 3d action recognition. In *Image Vision Comput.*, 44(C):29-43, December 2015.

[14] X. Zhang, Y. Wang, M. Gou, M. Sznaier, and O. Camps, Efficient temporal sequence comparison and classification using gram matrix embeddings on a riemannian manifold. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4498-4507, June 26-July 1, Las Vegas (USA), 2016.