SELF-ORGANIZING RESERVOIR NETWORK FOR ACTION RECOGNITION

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

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

[2,3]Faculty of Computer Science and Information Technology Universiti Malaya, 50603 Kuala Lumpur, Malaysia

Email: gclee@mmu.edu.my[1], ckloo.um@um.edu.my[2*](corresponding author), liew.wei.shiung@gmail.com[3]

DOI: https://doi.org/10.22452/mjcs.vol35no3.4

ABSTRACT

Current research in human action recognition (HAR) focuses on efficient and effective modelling of 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. However, the random initialization of the ESN's input and reservoir weights may increase instability and variance in generalization. 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, a Self-Organizing Reservoir Network (SORN) is developed based on Adaptive Resonance Theory (ART) and Instantaneous Topological Mapping (ITM) as the clustering process to cater deterministic initialization of the ESN reservoirs in a Convolutional Echo State Network (ConvESN) and yield a Self-Organizing Convolutional Echo State Network (SO-ConvESN). SORN ensures that the activation of ESN's internal echo state representations reflects similar topological qualities of the input signal which should yield a self-organizing reservoir. In the context of HAR task, human actions encoded as a multivariate time series signals are clustered into clustered node centroids and interconnectivity matrices by SORN for initializing the SO-ConvESN reservoirs. By using several publicly available 3D-skeleton-based action recognition datasets, the impact of vigilance threshold and reservoir perturbation of SORN in performing clustering, the SORN reservoir dynamics and the capability of SO-ConvESN on HAR task have been empirically evaluated and analyzed to produce competitive experimental results.

Keywords: Echo State Networks, Action Recognition, Self-Organizing Networks, Deep Neural Networks

1.0 INTRODUCTION

In contemporary society, assistive technologies may improve the quality of life of the motor disabled people community [1] and elderly population [2], from physical rehabilitation to the remote monitoring of health and safety conditions. Various research and development works have been conducted to deal with the variability of these technologies crucial to real world applications, particularly, to build an artificial intelligence-based machine that can correctly understand the intentions of humans in order to assist humans better [3]-[13].

With the advancement and growth of off-the-shelf solutions and affordable devices such as the Microsoft Kinect or the ASUS Xtion Pro, which can reliably acquire depth data in addition to the regular RGB data, human action recognition (HAR) has emerged and attracted the interest of many research areas such as computer vision, machine learning and pattern recognition, to build practical and reliable solutions in detecting, analyzing and recognizing human actions in videos [14]. Successful applications of HAR which can be noticeable in both commercial products and researches include visual surveillance [15][16], human computer interaction [17][18], physical rehabilitation [19] and autonomous driving vehicle [20].

In this work, we focus on HAR based on 3D-skeleton joints extracted from depth maps. The reasons for this are threefold: (1) skeleton joints are not affected under changing light environments [21] (2) skeleton joints have compact representation of human body which may diminish computational costs [22], and (3) it is privacy preserving to represent a person by using skeleton joints [23]. Recent research in skeleton-based HAR focuses on the challenge for efficiently and effectively modelling the temporal features of human actions in 3-dimensional space. One type of reservoir computer that is known as Echo State Networks (ESNs) [24] has emerged as one suitable method for encoding the temporal context owing to its simplicity and short-term memory property. The random assignment of the ESN's input and reservoir weights reduces the computational complexity compared to backpropagation through time. Convolutional Echo State Network (ConvESN) [25] has been introduced as a unified architecture that link the research areas of both reservoir computing and convolutional deep learning for HAR tasks.

Applying Convolutional Neural Network (CNN) to replace the linear regression in ESN made ConvESN to be able to understand complex action echo states. Despite the fact that promising recognition performance are accomplished using these approaches, the random initialization of the ESN's input and reservoir weights may increase instability and variance in generalization [26]. In other words, the performance of ESN may differ even if identical set of hyperparameters are employed to repeat the same task [27].

To address this problem, in this work, we expand upon ConvESN's body of work and contribute a novel self-organizing reservoir design for ESN stage through an approach called Self-Organizing Reservoir Network (SORN). Our approach is 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 [28], the potential of unsupervised self-organizing learning seems to be one of the most encouraging and the most biologically plausible approach for designing self-organizing reservoir. Considering that ESN is suitable for modelling temporal features of human actions, using self-organizing learning in the formation of its 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 deterministic and self-organizing reservoir [29].

SORN is proposed based on the notions of Adaptive Resonance Theory (ART) [30] and Instantaneous Topological Mapping (ITM) [31]. Human actions encoded as a multivariate time series signals are first clustered using SORN. SORN can automatically learn the topological qualities of the input signal and generate clustered node centroids and centroid interconnectivity maps. These feature maps are then used for deterministic initialization of the input weights and recurrent hidden weights in the ESN stage of ConvESN. The resulting novel implementation is known as Self-Organizing Convolutional Echo State Network (SO-ConvESN). Similar to ITM, SORN consists of a clustered topology of nodes from the unsupervised learning of training dataset. Each node representing a sufficiently dissimilar training sample or archetype, while sufficiently similar training samples are 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. The number of nodes in SORN is not defined a priori. Thanks to the features of ART, SORN is having similar network architecture to handle plasticity and stability dilemma by using best-matching node selection, vigilance test, and node learning.

In addition, depending on the iteration index, Adam [32] optimizer's learning rate in ConvESN monotonically decreases may cause the training traps in local minima. In turn, model may be very sensitive to the initial learning rate selection. SO-ConvESN tackles this problem by adopting a state-of-the-art method, known as Cyclical Learning Rate (CLR) [33], into Adam optimizer during the CNN parameters optimization. SO-ConvESN implements CLR to oscillate the learning rate between upper and lower bounds to break the training out of local minima and saddle points.

Our main contributions can be summarized as follows: 1) We propose a self-organizing reservoir network for clustering of input samples which is well-suited to generate node centroids and interconnectivity maps compatible for ESN reservoir. 2) Integrating the SORN into ConvESN in the formation of deterministically initialized ESN reservoirs to ensure 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. 3) Characterizing the dynamics of self-organizing reservoir for stability and echo state property. 4) Adopting CLR into optimization of convolutional stage parameters to oscillate the learning rate between upper and lower bounds in order to break out of saddle points and local minima during training. 5) We have set up experiments to investigate the dynamics of reservoirs generated by SORN, feasibility and performance of SO-ConvESN in 3D-skeleton based human action recognition task using several publicly available 3D-skeleton-based action recognition datasets. In addition, several non-linear transformations have been implemented as fusion layer during classification stage solely for performance assessment.

The remainder of the paper is divided as follows: The related works on HAR 3D skeleton-based using ESN-based approaches are firstly reviewed. We then describe the details of SORN for generating self-organizing reservoirs and SO-ConvESN for HAR tasks. Next, the simulation experiments are presented and the results are analyzed based on several publicly available benchmarking datasets. Lastly, concluding remarks are presented.

2.0 RELATED WORKS

With the growth of reasonably priced depth cameras, such as the Microsoft Kinect or the ASUS Xtion Pro, HAR based on skeleton joints has emerged and attracted the interest of many research areas. Human actions are often considered as temporal series of human body movements which may encompasses several body parts

simultaneously. One of the efforts focuses on the challenge for efficiently and effectively modelling the temporal features of human actions in 3-dimensional space. Existing research on HAR using 3-dimensional trajectories of human skeleton joints mainly exploits machine learning approaches such as Support Vector Machines [34]-[36], Multilayer Perceptrons [36], Dynamic Time Warping [37]-[39], Hidden Markov Models [38], and Decision Trees [35][36]. Nevertheless, these approaches occasionally abandon temporal features that may carry the information over the duration of intervals between activities. This motivates the study of HAR approach that uses memory mechanism such as the implementation in Recurrent Neural Networks (RNN) [40]. RNN is widely known to be suffering from the effects of exploding or vanishing gradients due to training of network architectures with many layers through gradient propagation. To overcome the error-prone gradient computations, Long Short-Term Memory (LSTM) [41] networks have been introduced by integrating gating mechanisms into an RNN architecture. LSTM has been adopted to skeleton based action recognition to further improve the learning of temporal context of skeleton sequences.

An alternative paradigm to the traditional RNN training is under the term "Reservoir Computing" (RC). It has become popular and showed high performance in time-series prediction. A special RC implementation is called ESNs [24]. 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 reduce the computational complexity compared to backpropagation through time. Nonetheless, the literature about the applications of ESN-based approach for HAR tasks by using 3D skeleton joints is very limited. Indicative studies are those of Claudio et. al. [42], Luiza et. al. [27] and Q. Ma et. al. [25]. Claudio et. al. [42] introduced bidirectional Leaky Integrator ESNs (LI-ESNs) to address the issue of learning with temporal 3D body joints through direct processing without additional feature extraction steps. This bidirectional approach split the reservoir into two portions to yield better state representation of input. This approach requires each input sequence is entirely available during the encoding process. Moreover, LI-ESNs take advantage of training-free input weight in reservoir whereby random initialization is implemented. Whereas, recurrent weight is described by a permutation matrix. On the other hand, recent studies are putting efforts on including supplementary contextual cues with 3D body joints with the aim to improve performance and consistency of the HAR. Luiza et. al. [27] proposed such an approach to take into consideration the interaction of an action being performed with objects being manipulated. ESN learning integrates 3D joint coordinates and information of object in order to handle ambiguities during the activity. It is considered as multi-label classification task since the number and type of objects may change between assorted actions. Object label may influence the reservoir internal representation. Similar to Claudio's approach [42], input weight and recurrent weight in reservoir are not input-dependent and are randomly initialized. Performance of ESN may differ even if same set of parameters are employed to repeat the same HAR task. This may increase the performance inconsistency of the ESN model.

Existing approaches in HAR may also severely depend on feature extraction by heuristic handcrafted mechanism, which could be inadequate due to human domain knowledge [43][44]. Particularly, it is decisive to extract multiscale temporal features of 3D skeleton sequences to find dynamical locally similar patterns for performance improvement. With the recent advancement of deep learning that exhibits its superior performance of automatic high-level feature extraction in HAR [45], ConvESN [25] has been introduced as a unified architecture that link the research areas of both Convolutional Neural Network (CNN) and RC for HAR. ConvESN considers the 3D skeleton sequences as multivariate time series and learns the dynamics and multiscale features from echo states. ConvESN could also be viewed as using CNN as a complementary of the ESN-based approach in handling 3D skeleton joints where CNN learns spatial features whereas ESN-based approaches may just lack spatial information. In ConvESN, input weight and recurrent weight in reservoir are also randomly initialized. Besides, ConvESN implemented Adam [32] optimization for the CNN parameters which could be sensitive to the initial learning rate selection. Poor selection in initial learning rate may not be adequate to leave local minima or saddle point during weights learning. Instead, adopting CLR [33] during CNN training by oscillating the learning rate between upper and lower limit could lead to reduced thorough learning rate tuning experiments.

The aforementioned ESN-based approaches apply random assignment of its input and reservoir weights with the aim in reducing the computational complexity. Unfortunately, randomly initialized weights for network training increases instability and variance in generalization [26]. Palangi et.al. [46] has demonstrated that it is desirable to learning these weights by adapting them to the data. Learning one or both of the input and recurrent weights in an ESN shows better performance in classification problem. This has motivated us to propose and investigate a novel approach to implement deterministic and self-organizing reservoir in ESN. 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 [28], the potential of unsupervised self-organizing learning seems to be one of the most encouraging and the most biologically plausible.

Recently, Boccato et al. [47] implemented unsupervised learning strategies to initialized the canonical ESN's reservoir. This approach applied Self-Organizing Map (SOM) [48] and Growing Neural Gas (GNG) [49] for the adaptation of the reservoir input weight for stability. SOM and GNG are essential algorithms of topological self-organizing clustering. SOM is characterized by a predefined and fixed topological map while GNG has extended the adaptive abilities of SOM by inserting new nodes in regular intervals. Both of these self-organizing algorithms heavily rely on the assumption that the training data are statistically independent which makes them difficult in handling trajectory data, such as human action skeleton joints. To overcome this limitation, ITM [31] has been introduced which is well suited for fast adaptation of topological map formation for training data with strong correlation. Moreover, SOM and GNG algorithms suffer from plasticity-stability dilemma [50]. To circumvent the plasticity-stability dilemma, ART [30] has been proposed as a self-organizing clustering algorithm which is motivated by the learning process in the human brain. Combining the benefits of ART and ITM and bridging these algorithms with ESN's reservoir design in handling HAR task seems to be motivating. In addition, ESN remains a black-box algorithm. The existing ESN-based approaches often lack explainability consideration. Hence, this work proposes the SORN to perform deterministic initialization of the ESN reservoir weights and extracts explanatory information to characterizes the dynamics of the reservoir to ensure stability and satisfaction of echo state property (ESP) [24].

3.0 MATERIALS AND METHODS

SO-ConvESN is characterized based on a SORN for clustering the body action sequences to generate self-organizing reservoirs and a ConvESN for multiscale feature extraction and classification. The overview architecture of SO-ConvESN with three filters, for ease of visualization, is illustrated in Fig. 1. The scope of this work focuses on HAR task based on 3D-skeleton single-view and single-person-based scenario. SO-ConvESN divides each skeleton series into five channels correspond to 3D-skeleton joint coordinate trajectories of five single body parts; left arm (LA), right arm (RA), central trunk (CT), left leg (LL) and right leg (RL), respectively. Self-organizing clustering is performed by SORN separately for each channel and the corresponding reservoir is created to obtain action echo states. To explain our proposed approach, SORN for learning the node centroids and the interconnectivity maps is firstly presented. Then, the SO-ConvESN for human action recognition is introduced.

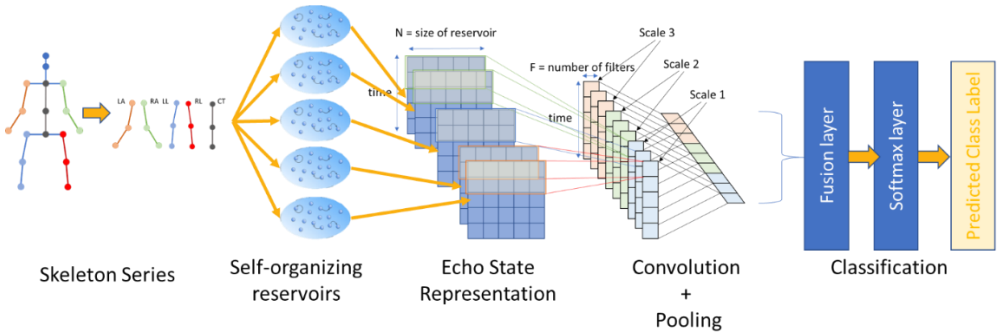


Fig. 1: Overview architecture of SO-ConvESN. Self-organizing reservoirs are generated by SORN and implemented in multi-scale, multi-channel ConvESN with 3 time-scales, 3 filters and 5 channels for human action recognition

3.1 Self-Organizing Reservoir Network (SORN): Learning the Node Centroids and Interconnectivity Maps

This section presents a self-organizing network for reservoir design which we call it SORN. It is specifically developed for the learning of node centroids and interconnectivity maps from skeleton training data. The generated maps preserve the topological properties of the input space at a significantly-reduced dimensionality and are used for deterministic initialization of ESN reservoirs. SORN has an ART-based network architecture to handle plasticity and stability dilemma and is having a topology construction procedure similar to ITM. It consists of a clustered topology of nodes from the unsupervised training of a dataset. The number of nodes in SORN is not defined a priori. 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.

Firstly, we extract five body parts from human skeleton data: left arm (LA), right arm (RA), central trunk (CT), left leg (LL) and right leg (RL). SORN performs learning on skeleton series of each part over time and obtain five corresponding self-organizing reservoirs. Fig. 2 illustrates the clustering of skeleton data of five body-part channels into five corresponding self-organizing reservoirs. The learning algorithm of SORN consists of four parts, namely

best-matching node selection, node matching using vigilance test and node learning which are similar as ART network and a topology construction procedure which is based on ITM.

Assuming $u(t)$ represents the joint coordinates of a single body part at a single time instance t. According to the stochastic resonance theory, adding noise prior to clustering would speed up convergence in a centroid-based clustering algorithm [51]. Hence, the noisy signal is defined as follows

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

where $\eta = [0,1]$ controls the magnitude of the noise, I is an identity matrix having same dimension as $u(t)$ and the noisy signal $z(t)$ would have progressively less noise over time.

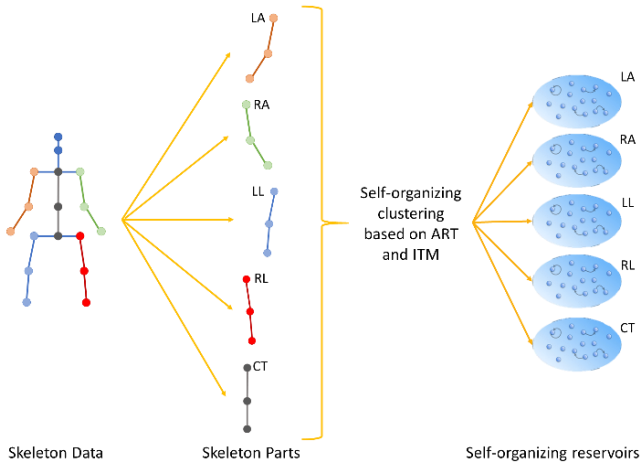


Fig. 2: Self-organizing reservoirs. SORN generates 5 self-organizing reservoirs from the corresponding LA, RA, LL, RL, and CT channels skeleton data.

SORN starts out with best matching node selection. Once the training sample $z(t)$ at instant t is input to the network, two nodes which has the similar state as $z(t)$ are searched: the best matching node b and the second-best matching node s. They are nominated based on a metric which is calculated by Equation 2, Equation 3, and Equation 4 as follows

$$k_l(z(t), w_j) = \|z(t) - w_j\|^2 \tag{2}$$

$$b = \underset{j \in J}{argmin}[k_l(z(t), C)] \tag{3}$$

$$s = \underset{j \in J, j \neq b}{argmin}[k_l(z(t), C)] \tag{4}$$

k_l is the similarity measurement metric between the $z(t)$ and a node j calculated using a Euclidean distance, w_j is the weights for node j, b is the index of the best matching node, s is the index of the second-best matching node, and C is the self-organizing node centroids with J nodes. The second-best matching node is selected for ITM-based topology construction process. If the network is empty, the sample $z(t)$ becomes a new node as

$$K \leftarrow K + 1 \tag{5}$$

$$w_K = z(t) \tag{6}$$

After best-matching node has been determined, node matching using vigilance test is performed to examine whether the sample $z(t)$ is within the vigilance region of node b. It evaluates whether to add a new node or update nodes by using Hebbian rules which is defined as in Equation 7.

$$k_l(z(t), w_b) \leq V \tag{7}$$

where V is the vigilance threshold. In case of the vigilance test is not satisfied, next candidate of best matching node will be searched to satisfy the condition as in Equation 7. In the case where node matching using vigilance test failed with all the existing J nodes, a new node will be added into the network defined as in Equation 5 and Equation 6. Otherwise, the condition in Equation 7 is fulfilled, node learning is performed. State of the best-matching node will be updated as follows

$$w_b = w_b + \epsilon_b(z(t) - w_b) \tag{8}$$

In SORN, the topology construction process is based on the edge adaptation of ITM and least-recently-used node pruning policy. Nodes that have similar information are linked and represented by topological connections. Once node matching using vigilance test occurs and second-best matching node also fulfils the matching condition as in Equation 7, incrementing edges between the best-matching node b and second-best matching node s, $\Delta E(b, s) = 1$, through learning, construct a sparse interconnectivity matrix, E. An edge connecting node b and node s is created if it does not previously exist. Similar to ITM, SORN does not need any edge aging to construct the topological map. For each n-th member of N best-matching node b's neighborhood nodes, check if node s lies inside the vigilance region through node n and node b. If it is, edge connecting node n and node b is removed. Pruning is conducted based on least-recently-used policy at every λ learning cycle. A node k is pruned if it does not have any edges.

On conclusion of the learning, SORN extracts the node centroids C and the interconnectivity maps E which are compatible for ESN reservoir. During the deterministic initialization of the reservoir parameters, clustered node centroid weights C with J nodes are rescaled by the input scaling parameter $-I_S$ to I_S as in Equation 9.

$$c_j = I_S.\left(2\left[\frac{c_j - min(C)}{max(C) - min(C)}\right] - 1\right) \tag{9}$$

for $j \in J$ and are used to initialize input weights $W^{in} = C$ while the interconnectivity matrix E is rescaled by the spectral radius parameter S_R and is infused into recurrent hidden weights W^{res} as in Equation 10.

$$W^{res} = S_R \frac{E}{\lambda_{max}(E)} \tag{10}$$

where $\lambda_{max}(E)$ is the eigenvalue of interconnectivity matrix E with the largest value, to yield a self-organizing reservoir with N neurons.

3.2 Characterize the Dynamics of Self-Organizing Reservoir Using Recurrence Quantification Analysis (RQA) For Stability

Similar to typical ESN, the configurations of input scaling parameter I_S and spectral radius parameter S_R play crucial roles to ensure the stability of the self-organizing reservoir generated by SORN. Additionally, S_R must be kept below 1 to ensure the echo state property (ESP) [24]. To ensure stability and ESP in SORN, we exploit the recurrence quantification analysis (RQA) technique proposed by Bianchi [52] to characterize the dynamics of self-organizing reservoir for manual tuning of I_S and S_R.

In order to adopt RQA in the learning of SORN, recurrent plots (RPs) first need to be constructed from the echo state representation (ESR). Consider a D-dimensional joint coordinates of a single body part at a single time instance t represented as $u(t)$ and an initial echo-state $x(0) \in R^N$ in the self-organizing reservoir, the update equation for the system is given as

$$x(t+1) = f(W^{res}x(t) + W^{in}u(t+1)) \tag{11}$$

where W^{in} and W^{res} are initialized using the clustered node centroid weights C and the interconnectivity matrix E respectively. It is noted that both clustered node centroid weights C and interconnectivity matrix E are the key components learned by SORN. SORN executes deterministic initialization of ESN reservoirs that is crucial to ensure

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. In other words, parameters in these reservoirs are not randomly fixed by ESN but deterministically initialized by SORN. Similar to ConvESN, the ESR is generated by projecting the time-series input into the self-organizing reservoirs according to Equation 11 and it is denoted by N-by-T dimensional matrix as

$$\boldsymbol{X} = (x(0), \dots, x(T-1))^T \tag{12}$$

Each of the ESR state $x(t)$ can be considered as a multivariate time series with N state variables generated according to Equation 11. A RP is created by a $T \times T$ binary matrix $\boldsymbol{R}$ with its element R_{ij} defined as

$$R_{ij} = \Theta\big(\tau_{RP} - d(x[i], x[j])\big) \tag{13}$$

for $1 \leq i, j \leq T$, where T is the length of the ESR state $x(t)$, $d(\cdot,\cdot)$ is a dissimilarity measure, $\Theta(\cdot)$ is the Heaviside function, and $\tau_{RP} > 0$ is a user-defined threshold for identifying recurrences. We refer the reader to [52] for further details description. Fig. 3 summarizes the process of generating RP from ESR.

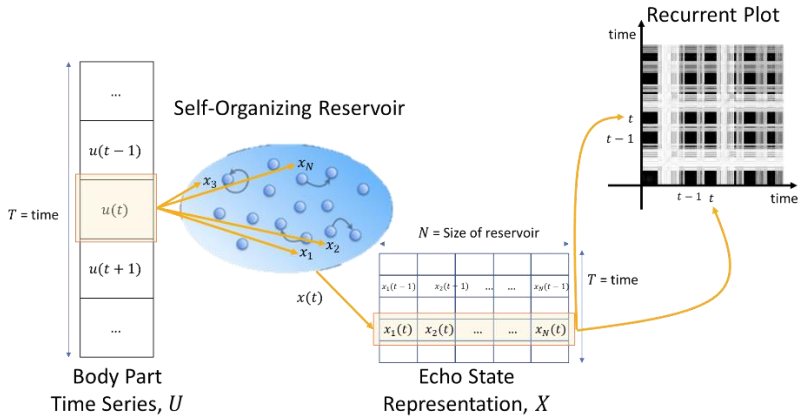


Fig. 3: Generating RP from ESR. $u(t)$ is fed into self-organizing reservoir with N neurons to generate echo state $x(t)$. Recurrent plot is constructed from echo state representation based on Bianchi's RQA techniques [52].

Based on the generated binary matrix $\boldsymbol{R}$, Bianchi suggested a novel stability criterion for reservoirs of ESN which is known as maximum diagonal line length, L_{MAX}. It is a RQA complexity measure based on the diagonal lines in RPs of ESR to quantify the reservoir stability.

To reinforce the configuration of I_S and S_R, we further express and analyze the interpretability of the dynamic of neuron activations in reservoirs. By using RPs in conjunction with RQA measures, we visualize the ESRs generated by self-organizing reservoir in order to characterize their laminarity, time dependence and chaoticity [52]. In HAR, human actions are encoded as a multivariate time series signal. We hypothesize that the generated ESR should exhibit laminarity and time dependence dynamics in RPs and RQA measures.

A reservoir presents laminar phase if echo state changes very slowly over a number of adjacent time steps. It can be identified by the presence of large black rectangles in RPs and characterized by high value of laminarity, $LAM \in [0,1]$. Meanwhile time dependence can be identified by non-uniformly distributed RPs and relies on the RQA measure of the determinism level, $DET \in [0,1]$ to quantify its amount and it would yield a value close to zero when time dependence does not exist. In term of chaoticity, it can be identified by the presence of short and erratic diagonal lines in RPs and characterized by low value of recurrence rate (RR).

By observing the laminarity, time dependence and chaoticity in the ESRs visualization at different values of I_S and S_R, optimal and stable configuration of a reservoir could be formulated. The generated stable self-organizing reservoirs are later infused into the ConvESN [25] to yield a SO-ConvESN for HAR task.

3.3 Self-Organizing Convolutional Echo State Network (SO-ConvESN)

In this section, we exploit the concept of SORN to initialize the reservoir in an ESN variant. Particularly, integrating novel SORN reservoir design into ConvESN yields SO-ConvESN. The general architecture of the SO-ConvESN for human action recognition task consists of four stages: the self-organizing reservoirs, echo state representation, convolutional-pooling, and classification.

As described in the previous section, self-organizing reservoir projects joint coordinates $u(t)$ into $x(t)$ according to the update process as in Equation 11 to yield ESR denoted by Equation 12. Self-organizing reservoir is deterministically initialized via two key components learned by SORN which are the clustered node centroid weights C and the interconnectivity matrix E respectively.

Human action recognition needs to model the temporal features of human actions by maintaining multiscale feature and time invariance. Multiscale features are derived from X matrix, the ESR, using multiple filters widths and feature maps. Multiscale temporal shift-invariance is maintained using max-over-time pooling. Assuming $w_{kj} \in R^{(K\times N)}$ denotes the j-th filter with k-width, the convolution result with w_{kj} is given as:

$$c_{kj} = (c_0, c_1, \dots, T-k+1{:}\,T)^T \tag{14}$$

$$c_m = f\left(\sum_i \left(w_{kj}.\, z^i_{m:m+k-1}\right)\right) \tag{15}$$

where $m = [0, 1, 2, \dots, T-k+1]$ is the index of the sliding window, z^i_m is the temporal window, f is the nonlinear activation function and $\cdot$ denotes a dot-product operation. Max-over-time pooling is used in the pooling layer to obtain the extracted features, combined based on relevance as shown in Fig. 1.

In the final classification stage, all pooled features passed through a fully connected layer followed by softmax layer. Softmax function defined outputs as the conditional distribution $P(C_s|u)$ over action labels, where C_s denotes the s-th class of actions.

Following the direction of ConvESN, SO-ConvESN uses CNN to decode complex ESRs in order to understand the action echo states generated from self-organizing reservoirs. In this work, we additionally implement several non-linear transformations: Residual Network (ResNet), Fully Convolutional Neural Network (FCN) and Multilayer Perceptron (MLP) [53] in replacing the classification stage (as shown in Fig. 1) of SO-ConvESN. Pooled features are input into these non-linear transformations, as depicted in Fig. 4, for performance assessment in view of searching the optimal fusion method. This work follows the hyperparameters settings of the non-linear transformations in [53].

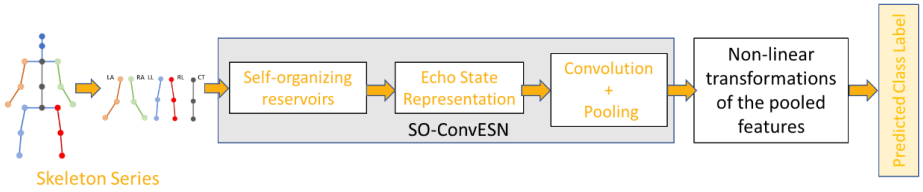


Fig. 4: Non-linear transformation is implemented as alternative fusion method in the classification stage of SO-ConvESN.

In self-organizing reservoirs, input scaling I_S and spectral radius S_R are fixed via the RQA technique as highlighted in previous section. Whereas in CNN stage, rectified linear unit (ReLU) [53] is used as non-linear activation function due to its efficiency in terms of computation and its fast convergence. Adam optimizer and cross entropy error function are used during CNN training.

For the number of kernels, it defines the number of convolution filters that does the convolution of ESR vectors. Convolution layers extract the multi-scale features from ESR vectors. The number of kernels may have impact on

the extraction of these features. Therefore, it is crucial to determine the proper number of filters during this process. This work conducted a preliminary experiment to configure the number of kernels. In addition, we implement a new method called CLR proposed by Leslie Smith [33] for setting the learning rate of the CNN in order to eliminate the need to empirically determine the optimal value of the learning rate in Adam optimizer. This work sets, prior training, the boundary values for the learning rate to cyclically vary within a limit. Complementary to adaptive learning rate, CLR method does not require any extra computation at all and it is computationally simpler. Moreover, CLR can be combined with adaptive learning rate. For this reason, this work adopts CLR into the Adam optimization of the convolution kernels weights during CNN training.

4.0 RESULTS AND DISCUSSION

With the aim of analyzing the dynamics and capability of the self-organizing reservoir generated by SORN and demonstrating the recognition performance of the SO-ConvESN, this section firstly presents a series of simulation experiments and analyses using two publicly available skeleton-based action recognition datasets: MSR-Action 3D (MSRA3D) [55] and Florence3D-Action (Florence3D) [56]. Each of the datasets comprises different set of actions and gestures. Simulation experiments show that our proposed approach achieves competitive performance with respect to the state-of-the-art approaches. Next, we deployed SO-ConvESN in the application for elderly rehabilitation exercise via AHA3D [57] dataset to demonstrate the feasibility of our model in applied implementation. Finally, different non-linear transformation methods were experimentally evaluated for performance comparison in order to search for better fusion method during classification in SO-ConvESN.

4.1 Evaluation Datasets

MSRA3D is one of the most popular benchmark datasets for HAR. It includes 20 skeleton joints for 20 different activities: high-arm wave, horizontal arm wave, hammer, hand catch, forward punch, high throw, draw X, draw tick, draw circle, hand clap, two-hand wave, side boxing, bend, forward kick, side kick, jogging, tennis swing, tennis serve, golf swing, and pick-up and throw. Kinect-like sensor was used to capture these activities at 15 fps performed, for 2 or 3 times, by 10 different subjects, resulting a total of 567 sequences with 23797 skeleton frames. Some researches [58][59] discarded 10 of them with excessive noise, the skeletons are either missing or corrupted. However, in this work, complete original dataset is used to evaluate the noise handling capability of the proposed approach.

Florence3D includes 15 skeleton joints for 9 different activities: wave, drink from a bottle, answer phone, clap, tight lace, sit down, stand up, read watch, and bow. Kinect sensor was used to capture these activities performed, for 2 or 3 times, by 10 different subjects, resulting a total of 215 sequences. The main challenges in this dataset are the high intraclass variation where same action is performed with both left and right hands and the existence of very similar actions such as drink from a bottle and answer phone.

The AHA3D includes 21 skeleton joints for 4 different standard fitness exercises: 30-second chair stand, 8-feet up and go, 2-minute step test and unipedal stance. Kinect sensor was used to capture these exercises performed by 21 different subjects, resulting a total of 79 different 3D skeletal videos with 171753 skeleton frames. Each video contains between 1 to 3 runs of the same exercise being performed. The main challenges in this dataset are the exercises being performed by 21 young and elderly subjects, 11 were young and 10 were elderly. 5 of the subjects were male, and 16 were female.

4.2 Implementation Details

For SORN reservoir, input scaling, I_S and spectral radius, S_R were set to 0.1 and 0.99 respectively, and reservoir sizes ranged from 30 to 100. We have empirically shown the analysis of SORN reservoir dynamics to express the interpretability in setting the value of I_S and S_R in previous section using RQA techniques. For the CNN stage of SO-ConvESN, multiple sliding window widths were chosen to be 2, 3 and 4 for multiscale feature extraction and optimal number of kernels for each width were manually explored by setting the scales from 16 to 256. The CLR implemented triangular learning rate policy due to its simplicity [33]. Its base limit and upper limit were set at 0.001 and 0.006 respectively, following the guideline from Bengio [60].

Pre-processing of raw skeletal sequence in all datasets were first performed before evaluating SO-ConvESN. The given raw skeleton joints in the dataset were not in a normalized coordinate system, origin of each skeleton is

different. For each skeleton sequence, average of the hip center, left, and right joints was first computed, then the origin of each frame was normalized to this point. Normalized skeleton sequences may contain different levels of smoothness, hence, Savitzky-Golay smoothing filter [61] was applied to smoothen the joint trajectories. Lastly, for various length trajectories, skeleton sequences were padded with zeros up to the maximum length value.

To facilitate comparison with state-of-the-art results, training and testing protocols were applied as follows. MSRA3D used the standard validation protocols [55], three training and validation sets were created with half of the subjects used for training and the other half for validation. For Florence3D, following the protocol applied in [56], ten-fold cross-validation method was used for training and validation. For AHA3D, following the training and testing protocol [57], 79 skeletal videos were split as 39 videos for training, 20 for validation and 20 for testing. All performance metrics were calculated by averaging over 100 runs. In each of the experiments, we use accuracy as assessment metric to evaluate the performance of the action recognition approach. Higher the accuracy implies better performance.

4.3 Effects of Vigilance Threshold and Reservoir Perturbation in SORN

This section presents the analysis on the impact of vigilance threshold and reservoir perturbation of SORN in performing clustering for HAR task. Two benchmark datasets were used: MSRA3D and Florence3D. Vigilance thresholds were tested for a range from 0.05 to 0.95. In addition to benchmarking for different vigilance thresholds, reservoir perturbation was also considered. Clustering was conducted for different initial noise distribution scales, $\eta = [0;\ 0.1;\ 0.01;\ 0.001]$.

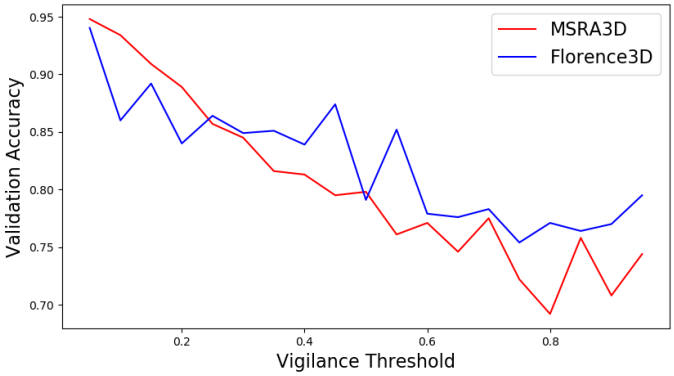


Fig.5: Validation accuracy of the SO-ConvESN in response to clustered node centroids with varying vigilance thresholds. Red solid line and blue solid line represents the validation result of training the model using MSRA3D dataset and Florence3D dataset respectively. Validation accuracy in both models decreases as soon as vigilance threshold increases to higher values

The optimal clustering configuration for HAR is obtained by setting the vigilance threshold to a low value (see Fig. 5). In both models, setting the vigilance threshold to 0.05 resulted in the peak validation accuracy compared to other vigilance thresholds. Comparing different magnitudes of reservoir perturbation, SO-ConvESN shows 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.

This suggests that the feature map requires high-granularity clusters generated by SORN to represent a comprehensive set of unique joint coordinates. It can be achieved by setting the vigilance threshold to a low value. On the other hand, adding controllable user-defined noise prior to clustering according to Equation 1 and set it to 0.1 could rise the performance of SORN.

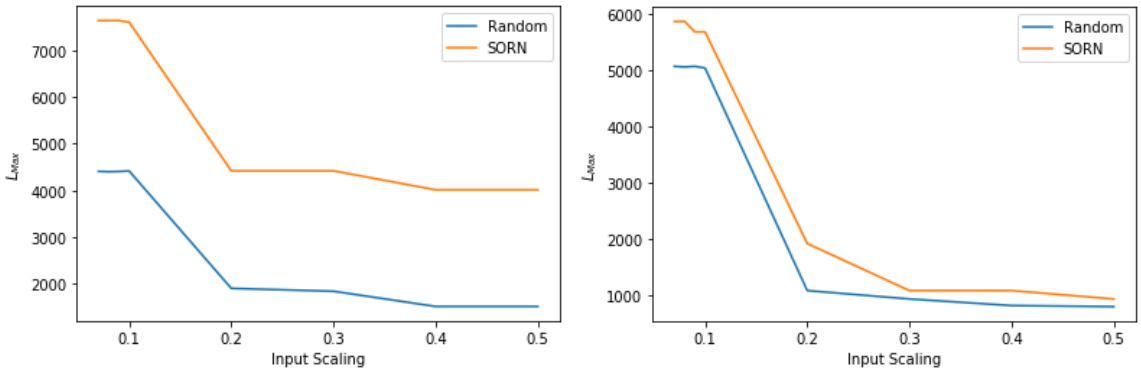
4.4 Characterize Dynamics of Self-Organizing Reservoir Using Recurrent Quantification Analysis

In this section, we execute the RQA techniques to visualize and characterize the dynamics of self-organizing reservoir generated by SORN at different configurations of two hyperparameters: input scaling, I_S and spectral radius, S_R, and hence obtain the optimal settings for stability. We also compare the RQA results of self-organizing reservoir with randomly initialized reservoir for the interpretability of the dynamic of neuron activations in

reservoirs. Two benchmark datasets, namely MSRA3D and Florence3D, were used for this analysis. For ease of visualization, we only considered the RPs generated from ESRs of left arm trajectories based on Equation 13 for RQA, as similar findings have been discovered for each of the five channels corresponding to five body parts.

Firstly, we analyzed the variation of reservoir stability as indicated by L_{MAX} against different configurations of input scaling, I_S. Knowing that S_R must be fixed at value less than 1 to follow echo state property [24], we first selected and set S_R at a boundary value of 0.99 and swept through the I_S values from 0.07 to 0.5 to observe the corresponding variation of L_{MAX}.

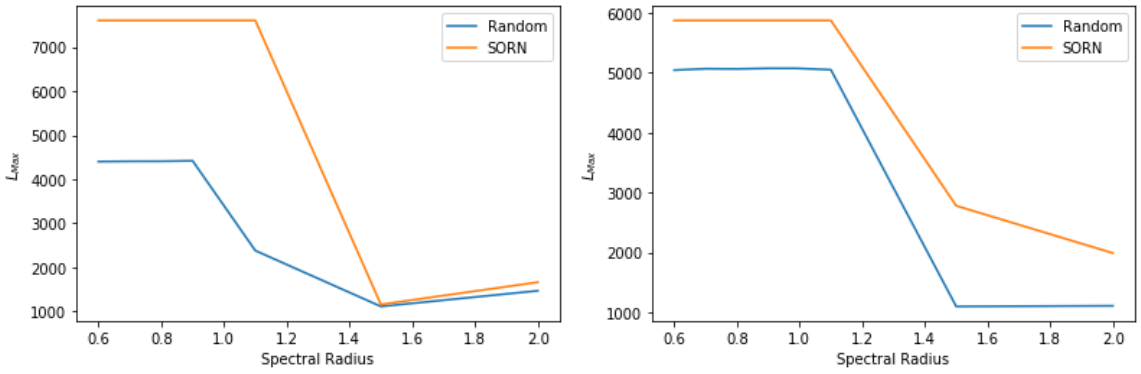
Based on the results produced by MSRA3D dataset and Florence3D dataset, as shown in Fig. 6a and Fig. 6b respectively, L_{MAX} remains the highest value at around $I_S = 0.1$ in both randomly initialized reservoir and self-organizing reservoir. This may suggest that setting I_S to 0.1 could produce more stable reservoir. It is crucial to highlight that setting any value of I_S ranges from 0.07 to 0.5, self-organizing reservoir at all times shows higher L_{MAX} as compared to randomly initialized reservoir.



(a) (b)

Fig.6: Stability of reservoirs based on RQA metric, L_{MAX} in response to different configuration of input scaling, I_S using two datasets. (a) MSRA3D dataset. (b) Florence3D dataset. Orange line represents the L_{MAX} changes of self-organizing reservoir generated by SORN meanwhile blue line indicates the result using random reservoir. Self-organizing reservoir shows higher stability in both models as compared to random reservoir. Self-organization ensures activation of ESN's internal ESR reflects similar topological qualities of the input action signal and hence create more stable ESR

Next, we further analyzed the variation of L_{MAX} against different configurations of spectral radius, S_R. Taking into account that setting I_S less than or equal to 0.1 produces more stable reservoir, we then fixed I_S at 0.1, swept through S_R values from 0.6 to 2.0 and observed the corresponding changes of L_{MAX}.



(a) (b)

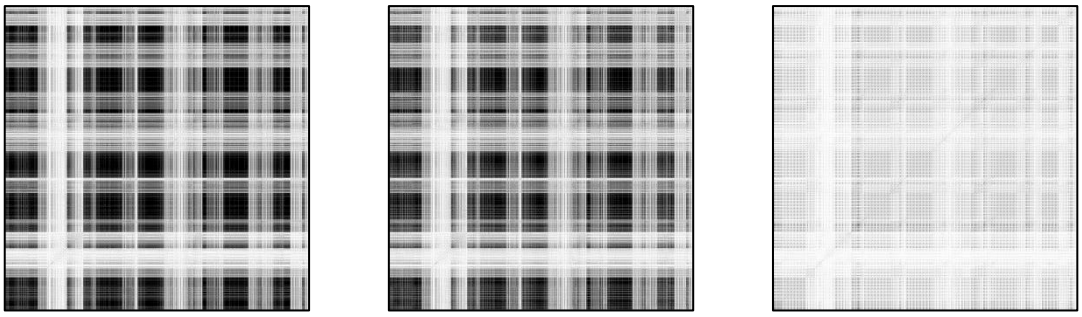
Fig.7: Stability of reservoirs based on RQA metric, L_{MAX} in response to different configurations of spectral radius, S_R using two datasets. (a) MSRA3D dataset. (b) Florence3D dataset. Orange line represents the L_{MAX} changes of self-organizing reservoir generated by SORN meanwhile blue line indicates the result using random reservoir. Self-organizing reservoir shows higher stability in both configurations as compared to random reservoir. The results

justify the importance of self-organization to ensure the activation of ESN's internal ESR reflects similar topological qualities of the input action signal in order to generate stable ESR.

From both Fig. 7a and Fig. 7b, it could be noticed that L_{MAX} remains at highest value around $S_R = 1.0$. This suggest that setting S_R close to 1.0 could produce more stable reservoir. However, in order to follow echo state property [24] and at the same time preserve reservoir stability, S_R could be set to 0.99. Similar to L_{MAX} in response to different configuration of I_S, when arbitrarily setting the value of S_R, self-organizing reservoir generated by SORN continually shows higher L_{MAX} as compared to randomly initialized reservoir which could imply that SORN produced reservoir that offers better stability.

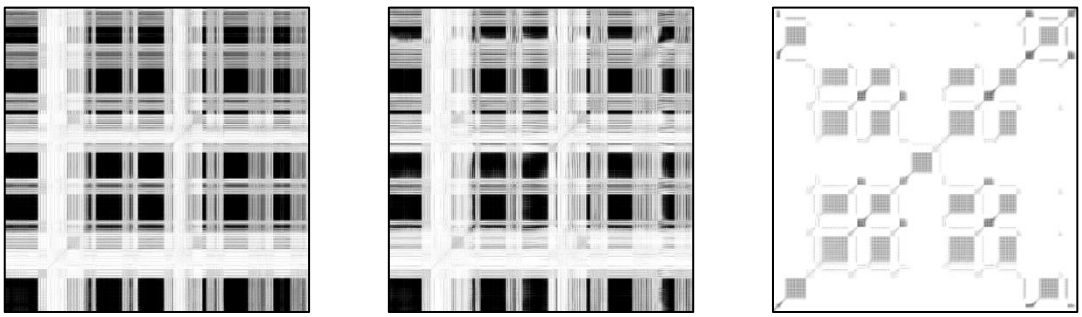
The experimental results of L_{MAX} in response to different configurations of input scaling, I_S and spectral radius, S_R respectively may imply that deterministic initialization of reservoir weights to generate self-organizing reservoir could ensure better reservoir stability. In other words, self-organization is crucial to ensure the activation of ESN's internal ESR reflects similar topological qualities of the input action signal and hence create more stable ESR. In addition, RQA stability metric known as L_{MAX} has been empirically implemented for manually tuning of the input scaling, I_S and spectral radius, S_R to ensure stability and ESP during reservoir design.

After empirically determined optimal $I_S = 0.1$ and $S_R = 0.99$ for stable reservoir that obey ESP, subsequently, we used RPs in conjunction with few more RQA measures to visualize and quantify the dynamics of ESRs generated by self-organizing reservoirs. We characterized their laminarity, time dependence and chaoticity to express the interpretability of the dynamic of neuron activations in reservoirs. We also compared the results with randomly initialized reservoir. To ease the visualization and interpretability of RPs, we fixed $I_S = 0.1$ and visualized the RPs of ESR for three different configurations of S_R at 0.6, 0.99, and 2.0. We intentionally included S_R at 2.0 merely to visualize the dynamics of ESRs when violation of ESP occurs.



(a) (b) (c)

Fig. 8: Recurrence plots of self-organizing reservoir neuron activations. (a) Setting S_R at 0.6: $L_{MAX} = 5874$, $LAM = 0.999976$, $DET = 0.999999$, $RR = 0.998830$. (b) Setting S_R at 0.99: $L_{MAX} = 5874$, $LAM = 0.999980$, $DET = 0.999999$, $RR = 0.998964$. (c) Setting S_R at 2.0: $L_{MAX} = 1990$, $LAM = 0.950442$, $DET = 0.904688$, $RR = 0.682407$. Both LAM and DET are closer to 1 and RR is at higher value when setting S_R at 0.6 and 0.99 which exhibits significant laminarity and time dependence dynamics but less chaoticity as compared to setting S_R at 2.0



(a) (b) (c)

Fig. 9: Recurrence plots of randomly initialized reservoir neuron activations. (a) Setting S_R at 0.6: $L_{MAX} = 5045$, $LAM = 0.999421$, $DET = 0.999893$, $RR = 0.995116$. (b) Setting S_R at 0.99: $L_{MAX} = 5075$, $LAM = 0.998944$, $DET = 0.999626$, $RR = 0.992221$. (c) Setting S_R at 2.0: $L_{MAX} = 1109$, $LAM = 0.921424$, $DET = 0.876809$, $RR = 0.268180$. Chaoticity is noticeable when setting S_R at 2.0.

Observing the RPs of self-organizing reservoir with S_R set at 0.6 (see Fig. 8a) and 0.99 (see Fig. 8b) and the ones of randomly initialized reservoir (see Fig. 9a and Fig. 9b respectively), echo state changes very slowly over a number of adjacent time steps, i.e., the presence of large black rectangles in RPs, indicates the presence of laminar phase. On the other hand, non-uniformly distributed RPs exhibits time dependence. In addition, measured LAM, DET and RR are close to 1 which characterizes that the reservoirs present significant laminar phase, significant time dependence and less chaoticity respectively. These dynamics reflect that both self-organizing reservoir and randomly initialized reservoir have encoded the multivariate human action time series signal with S_R set at 0.6 or 0.99. However, self-organizing reservoir exhibits higher measured LAM, DET and RR as compared to randomly initialized reservoir which may imply that self-organizing reservoir poses improved laminar phase and time dependence dynamics as well as lesser chaoticity. Particularly, action series consists of frames that has similar short-term histories. Projecting the series onto the reservoir has generated similar echo states which has been visualized by the large black rectangles in RPs. The finding of higher measured LAM proves that the satisfactory of ESP is higher in self-organizing reservoir which is important to ensure better stability in ESN.

Next, we attempt to violate the ESP by setting S_R at 2.0 in both self-organizing reservoir and randomly initialized reservoir in order to visualize and analyze the dynamics of the reservoirs. The presence of large black rectangles in the RP of self-organizing reservoir remains visible (see Fig. 8c). While the existence of large black rectangles is not significant visible in the RP of randomly initialized reservoir (see Fig. 9c). Moreover, some short and erratic diagonal lines are present which implied that it poses certain amount of chaotic dynamics. This finding could demonstrate visually the impact of violating the echo state property in reservoirs. Particularly, self-organizing reservoir seems to exhibit greater tolerance to chaotic during the violence of ESP whereby the laminarity phase remain visible.

In terms of RQA measures, the respective L_{MAX}, LAM, DET, and RR of both self-organizing reservoir and randomly initialized reservoir generated at $S_R = 2.0$ are low. This indicates that both reservoirs are less stable, reduced in laminarity and in time independence dynamics, and more chaotic during the violation of ESP. Comparing the RQA measures between self-organizing reservoir and randomly initialized reservoir, RR at 0.682407 and L_{MAX} at 1990 could indicate that self-organizing reservoir exhibits less chaotic dynamics and more stable reservoir respectively than randomly initialized reservoir with RR at 0.268180 and L_{MAX} at 1109.

The results show that self-organizing reservoir generated by SORN is more stable than randomly initialized reservoir. Self-organizing reservoir maintains more significant laminarity phase and time dependence whereby it poses higher value (closer to 1) of LAM and DET respectively. It even does not exhibit high chaotic dynamics when echo state property is not followed. In other word, using SORN for deterministic initialization of the ESN's input and reservoir weights may increase stability and improve variance of ESN in encoding the temporal context of human actions in 3-dimensional space. Self-organization ensures that the neuron activations reflect the characteristics of the action time series which may due to the consideration of input distribution in self-organizing reservoir generated by SORN. It could be one of the biologically plausible self-organizing reservoir design approaches. Additionally, by observing the laminarity, time dependence and chaoticity in the ESRs visualization at different values of I_S and S_R, stable configuration of a reservoir could be formulated. The generated stable self-organizing reservoirs are later infused into the ConvESN [25] to yield a SO-ConvESN for HAR task.

4.5 Effect of Implementing CLR in SO-ConvESN

In this section, we analyzed the impact of implementing CLR technique in CNN stage for performing multiscale convolutional process in SO-ConvESN. The results of the experiments using MSRA3D dataset and Florence3D dataset are reported in Table 2.

Table 2: Comparison of recognition accuracy between SO-CONVESN with CLR and SO-CONVESN without CLR.

Methods	Dataset	Average (%)
SO-ConvESN without CLR	MSRA3D	94.21
	Florence3D	89.70
SO-ConvESN with CLR	MSRA3D	**94.80**
	Florence3D	**94.03**

Comparing SO-ConvESN with CLR and SO-ConvESN without CLR, we showed that using state-of-the-art CLR technique during training can help in finding appropriate learning rates for an improved performance. This suggests

that the SO-ConvESN requires CLR to oscillates the learning rate between upper and lower bounds to break out of saddle points and local minima during training. It is interesting to see that the HAR performance of SO-ConvESN with CLR in MSRA3D experiment showed only 0.6% improvement whereas in Florence3D experiment exhibited significant improvement of 3.27% as compared to SO-ConvESN without CLR. It might due to SORN is based on ART model which is a potentially noise-sensitive model. Since MSRA3D dataset consists of training samples with excessive noise, the skeletons are either missing or corrupted and we used the complete original dataset in which noise could be clustered as a new node or update into a node or cluster of nodes during SORN learning. This suggests that the clustered node centroids and feature map generated by SORN are sensitive to the noisy joint coordinates, particularly in MSRA3D experiment, retaining the noisy data and using them during training could be challenging for SO-ConvESN to get away of saddle points and local minima during training even with the adoption of CLR. In general, taking CLR into convolution stage may boost the performance of SO-ConvESN. For the other experiments in the next sections, SO-ConvESN with CLR will be deployed and hereinafter we refer it as SO-ConvESN.

4.6 Comparison with Existing Approaches

In this section, the performance of SO-ConvESN is compared with several existing HAR methods. Table 3 and Table 4 show the state-of-the-art recognition accuracy of cross-subject test and cross validation on MSRA3D dataset and Florence3D dataset respectively.

Table 3: Recognition accuracy on cross-subject test of MSRA3D dataset.

Approaches	Average (%)
Covariance [62]	88.10
Skeletons Lie group [63]	92.40
DHMM+SL [64]	92.91
SO-ConvESN (Our approach)	94.80
Gram matrices rep. [58]	96.90
ConvESN [25]	**97.88**

Table 4: Recognition accuracy on 10-fold cross validation Florence3D-Action dataset.

Approaches	Average (%)
Multi-Part Bag-of-Poses [56]	82.00
Skeletons Lie group [63]	90.88
ConvESN [25]	91.72
SO-ConvESN (Our approach)	**94.03**

For Florence3D, SO-ConvESN achieved 94.03% overall accuracy. There was some confusion between the actions for “wave”, “drink from a bottle”, “answer phone”, and “read watch”, presumably due to all of them having a characteristic arm movement towards the head. The “read watch” action was often misclassified as “answer phone”. For the MSRA3D, SO-ConvESN exhibited 94.80% overall accuracy, performing poorly for “two-hand wave” and “draw circle” actions. The “draw circle” action was often mistaken for the “side boxing” action.

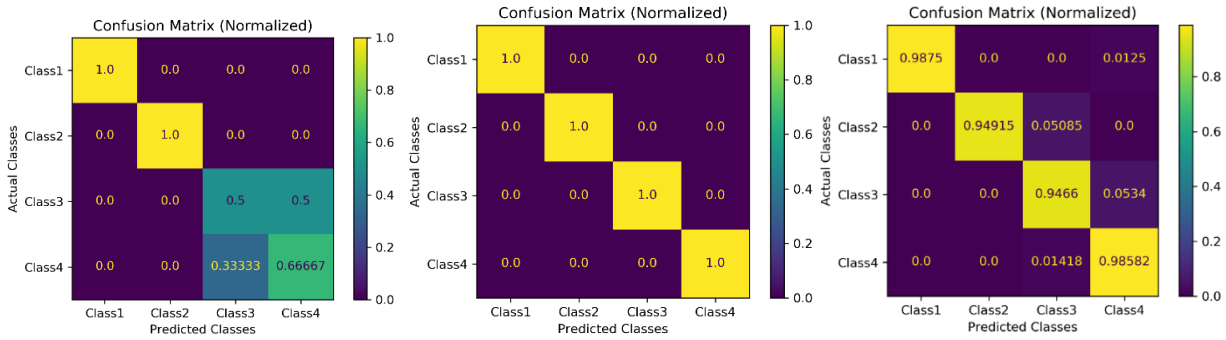
For MSRA3D, our approach achieved lower accuracy particularly when it is compared to ConvESN. Adding user-defined controllable noise prior to clustering according to Equation 1 could rise the performance of SORN. However, SORN seems to fail in handling uncontrollable noise such as the noisy data appear in MSRA3D data. Using noisy data in MSRA3D dataset during training could be challenging for SO-ConvESN because SORN is based on ART model which could be a model potentially sensitive to noise. In addition, the high number of action classes in MSRA3D seems to be the limit that SORN can handle due to class proliferation, which may reduce the generalization capability of SO-ConvESN. Nevertheless, our method produced stable reservoir. Repeated experiments using same hyperparameters configuration could still produce consistent results, as SO-ConvESN takes deterministic initialization of reservoir into account instead of randomly initialized reservoir in ConvESN.

On the other hand, it is interesting to note that our approach outperformed ConvESN in Florence3D. It could be the reason that SORN is good in clustering dataset problem with low noise and lower number of classes. Experimental results on HAR task show that self-organizing reservoir is competitive with state-of-the-art approaches. 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. The proposed reservoir design method is biologically feasible.

4.7 Application of SO-ConvESN on Physical Fitness Exercise of Elderly

An empirical application has been conducted on AHA3D dataset to verify the potential and feasibility of the SO-ConvESN in the recognition of physical fitness exercise for elderly. For a single experimental run, 39 videos were randomly chosen from 79 skeletal videos as training set for SORN learning. Next, the learned node centroids and interconnectivity matrix were used to initialize the recurrent weight and input weight in SO-ConvESN. The same 39 videos were used to train SO-ConvESN and together with randomly selected independent 20 videos for validation. After the SO-ConvESN had been trained, it was deployed for physical fitness exercise recognition using testing set consists of the other 20 videos. The evaluation process had been repeated for 100 runs to recognize four classes: Class 1 is unipedal stance, Class 2 is 8-feet up and go, Class 3 is 30-second chair stand and Class 4 is 2-minute step. All performance metrics were calculated by averaging over 100 runs.



(a) (b) (c)

Fig.10: The confusion matrices of AHA3D. (a) Worst-case with testing accuracy at 80%. (b) Best-case with testing accuracy at 100%. (c) Overall testing accuracy at 96.30%. Class 1 denotes unipedal stance, Class 2 denotes 8-feet up and go, Class 3 denotes 30-second chair stand and Class 4 denotes 2-minute step.

Fig. 10a and Fig. 10b show the normalized confusion matrices of worst-case scenario and best-case scenario of the 100-run experiment respectively. Testing recognition accuracies are 80% for worst-case scenario and 100% for best-case scenario. Fig. 10c depicts the normalized confusion matrix of overall performance averaging over 100 runs. The first row of the matrix shows that out of 480 videos representing a Unipedal Stance, 474 videos were correctly classified, and 6 were wrongly classified as 2-Minute Step. This depicts that the SO-ConvESN classified the Unipedal Stance exercise with 98.75% accuracy. The second row of the matrix shows that out of 826 videos representing an 8-feet up and go, 784 videos were correctly classified, and 42 were wrongly classified as 30-second chair stand. The SO-ConvESN classified the 8-feet up and go with 94.92% accuracy. The third row of the matrix shows that SO-ConvESN performing poorly on 30-second chair stand. Out of 412 videos representing a 30-second chair stand, 390 videos were correctly classified, and 22 were wrongly classified as 2-minute step. The SO-ConvESN classified the 30-second chair stand with 94.66% accuracy. The last row of the matrix shows that out of 282 videos representing a 2-minute step, 278 videos were correctly classified, and 4 were wrongly classified as 30-second chair stand. The SO-ConvESN classified the 2-minute step with 98.58% accuracy. The 30-second chair stand exercise was often mistaken for the 2-minute step exercise. It is noticeable that SO-ConvESN exhibits incorrect recognition only up to one class. It again demonstrates the efficiency of SORN in self-organizing and deterministic initialization reservoir design which considers topological information of the input signal to be included into the reservoir for SO-ConvESN. Experimental results show that SO-ConvESN successfully learned to classify the set of four fitness exercises and that its performance was very promising. The overall testing accuracy of fitness exercise recognized by SO-ConvESN is 96.30%, whereas the baseline approach used in [57] showed an overall accuracy of 91%. This outperformance warrants the potential and feasibility of SO-ConvESN to be applied for deployment in HAR tasks.

4.8 Non-linear Transformations

In order investigate the feasibility of using different non-linear transformation methods in classification stage of SO-ConvESN, in this experiment, we attempt to modify the classification stage of SO-ConvESN by treating pooled

features, which contain multiscale features extracted from ESR, as univariate input time series. Following the direction of Wang et al. [53], three different non-linear transformations were considered. We cascaded self-organizing reservoirs, echo state representation, convolution + pooling with three different non-linear transformation classification methods, namely, MLP, FCN and ResNet to formulate three models respectively. In this experiment, we adapt the configurations from [65]. In first model, MLP was configured to have four fully connected layers, and dropout regularization. Optimization process used AdaDelta optimization, 5000 epochs, learning rate at 1.0. FCN was configured to have 5 layers with 3 convolution layers in second model. Third model configured ResNet to have 11 layers with 9 convolution layers. Both FCN and ResNet used Adam optimization with learning rate set at 0.001 for 2000 epochs and 1500 epochs respectively. In MLP, FCN and ResNet, ReLU was selected as activation function and cross entropy loss function was employed.

We conducted a pairwise comparison to analyze the performance, for each non-linear transformation method against SO-ConvESN with CLR. Particularly, we conducted the Wilcoxon signed rank test on HAR accuracy average rank on the HAR datasets: MSRA3D, Florence3D and AHA3D. The overall accuracy average rank is depicted in critical difference diagram as shown in Fig. 11.

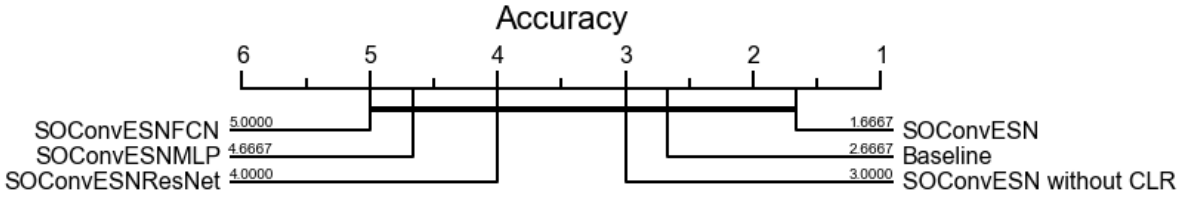


Fig. 11: Critical difference diagram over the accuracy average rank of SO-ConvESN and different non-linear transformation methods on the HAR datasets. SO-ConvESN achieved the highest accuracy average rank.

The horizontal line joins the methods that are not significantly different. The result implies that combining SO-ConvESN with different non-linear transformations under consideration does not show statistically significant difference among them. However, we can see that our proposed approach, SO-ConvESN with CLR, achieved the highest accuracy average rank. The result further demonstrates the effectiveness of our proposed SO-ConvESN and the benefit of the deterministic initialization of reservoir using SORN. Moreover, simple CNN stage with CLR implementation in SO-ConvESN could handle HAR task and exhibited competitive state-of-the-art performance. Using high complexity models such as non-linear transformation, namely MLP, FCN and ResNet as classification methods could be overfit in this HAR task.

5.0 CONCLUSIONS

The random initialization of ESNs input and reservoir weights may increase instability and variance in generalization. Performance of ESN in encoding the temporal context of human actions in 3-dimensional space may be inconsistent even if same set of hyperparameters are employed to repeat the same task. Current progress of this research work presents a SORN which is proposed based on the notions of ART and ITM. SORN is well-suited to generate input-information-driven clustered node centroids and interconnectivity maps that are compatible for self-organizing reservoir design in ConvESN. In the context of HAR task, human actions encoded as a multivariate time series signals are clustered by SORN before using the clustered node centroids and interconnectivity matrices for deterministic initialization of the reservoirs. The resulting network is known as SO-ConvESN. It could be viewed as a novel framework that bridges the area of self-organizing learning and reservoir computing in human action recognition task. We have also adopted CLR into optimization of convolution stage parameters to oscillate the learning rate between upper and lower bounds to break out of saddle points and local minima during training. We have set up experiments to investigate the capability and feasibility of SO-ConvESN in 3D-skeleton based human action recognition task using several publicly available 3D-skeleton-based action recognition datasets. In addition, several non-linear transformations have been implemented as fusion layer during classification stage solely for performance comparison. The impact of vigilance threshold and reservoir perturbation of SORN in performing clustering and the adaptation of CLR in SO-ConvESN were empirically analyzed. The dynamics of self-organizing reservoir were also analyzed. Experimental results show that SO-ConvESN is competitive with state-of-the-art approaches. Deterministic initialization instead of random 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. SO-ConvESN has the potential to be applied in general time series classification

applications. These motivate further enhancements on the robustness of the approach such as incremental learning, or by optimizing a number of hyperparameters in the SORN (i.e., node pruning) and reservoir (i.e., weight scaling factors). Future research may further improve noise handling capability of the SO-ConvESN. Moreover, SORN considers only spatial information during clustering due to the fact that Euclidean distance measurement metric is used during learning process. Temporal learning by using recurrent kernel should be taken into account for betterment.

REFERENCES

[1] Y. J. Chang, S. F. Chen, and J. D. Huang, "A Kinect-based system for physical rehabilitation: A pilot study for young adults with motor disabilities," *Research in Developmental Disabilities*, vol. 32, no. 6, pp. 2566-2570,2011.

[2] P. Rashidi and A. Mihailidis, "A survey on ambient-assisted living tools for older adults," *Journal of Biomedical Health Information.*, vol. 17, no. 3, pp. 579-590,2013.

[3] S. Kwak, B. Han, and J. Han, "Scenario-based video event recognition by constraint flow," in *IEEE Conference Computer Vision Pattern Recognition*,2011, pp. 3345-3352.

[4] L. L. Presti, S. Sclaroff, and A. Rozga, "Joint alignment and modelling of correlated behavior streams," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII) IEEE*, 2013, pp.730-737.

[5] U. Gaur, B. S. Y. Zhu, and A. Roy-Chowdhury, "A string of feature graphs model for recognition of complex activities in natural videos," in *IEEE International Conference Computer Vision*, 2011, pp. 2595-2602.

[6] S. Park and J. Aggarwal, "Recognition of two-person interactions using a hierarchical Bayesian network," in *First International Workshop Video Surveillance*. *ACM*, 2003, pp. 65-76.

[7] I. Junejo, E. Dexter, I. Laptev, and P. Pérez, "View-independent action recognition from temporal self-similarities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 1, pp. 172-185, 2011.

[8] Z. Duric, W. Gray, R. Heishman, F. Li, A. Rosenfeld, M. Schoelles, C. Schunn, and H. Wechsler, "Integrating perceptual and cognitive modelling for adaptive and intelligent human–computer interaction," in *Proceedings of the IEEE*, vol. 90, no. 7. IEEE, 2002, pp. 1272-1289.

[9] A. Thangali, J. P. Nash, S. Sclaroff, and C. Neidle, "Exploiting phonological constraints for handshape inference in asl video," *Proceedings of IEEE Conference Computer Vision Pattern Recognition*, pp. 521-528, 2011.

[10] A. T. Varadaraju, *Exploiting phonological constraints for hand shape recognition in sign language video*. Ph.D. dissertation, Boston Univ., MA, USA, 2013.

[11] H. Cooper and R. Bowden, "Large lexicon detection of sign language," in *Proceedings of International Workshop on Human–Computer Interaction, Springer, Berlin*, Heidelberg, Beijing, P. R. China, 2007, pp. 88-97.

[12] J. M. Rehg, G. D. Abowd, A. Rozga, M. Romero, M. A. Clements, S. Sclaroff, I. Essa, O. Y. Ousley, Y. Li, C. Kim, and et al., "Decoding children's social behavior," in *Proceedings of IEEE Conference Computer Vision and Pattern Recognition*, 2013, pp. 3414-3421.

[13] H. Moon, R. Sharma, and N. Jung, "Method and system for measuring shopper response to products based on behavior and facial expression," *US Patent 8 219 438*, July 10, 2012.

[14] Z. Z. et al., "Deep learning based human action recognition: A survey," in *Proceedings of China Automation Congress*, Oct. 2018, pp. 3780-3785.

[15] R. Cichy, A. Khosla, D. Pantazis, A. Torralba, and A. Oliva, "Deep neural networks predict hierarchical spatio-temporal cortical dynamics of human visual object recognition," *arXiv preprint* arXiv:1601.02970, 2016.VOLUME 4, 2016

[16] L. Onofri, P. Soda, M. Pechenizkiy, and G. Iannello, "A survey on using domain and contextual knowledge for human activity recognition in video streams," *Expert Systems With Applications*, vol. 63, no. 1, pp. 97-111, 2016.

[17] L. Presti and M. L. Cascia, "3d skeleton-based human action classification: A survey," *Pattern Recognition*, vol. 53, pp. 130-147, 2016.

[18] G. T. Papadopoulos, A. Axenopoulos, and P. Daras, "Real-time skeleton-tracking-based human action recognition using Kinect data," in *Proceedings of International Conference on Multimedia Modeling, Dublin*, Ireland, 2014, pp. 473—483.

[19] J. D. Huang, "Kinerehab: A Kinect-based system for physical rehabilitation: A pilot study for young adults with motor disabilities," in *Proceedings of 13th International ACM SIGACCESS Conference on Computers and Accessibility, New York*, USA, 2011, pp. 319-320.

[20] Y. Kong and Y. Fu, "Human action recognition and prediction: A survey," *arXiv preprint* arXiv:1806.11230v2, 2018.

[21] S. Gasparrini, E. Cippitelli, E. Gambi, S. Spinsante, and F. Flórez-Revuelta, "Performance analysis of self-organising neural networks tracking algorithms for intake monitoring using Kinect," in *Proceedings of IET International Conference on Technologies for Active and Assisted Living, Kingston*, UK, 2015.

[22] J. H. et al., "Enhanced computer vision with Microsoft Kinect sensor: A review," *IEEE Transactions on Cybernetics*, vol. 43, no. 5, pp. 1318-1334, 2013.

[23] J. R. Padilla-López, A. A. Chaaraoui, F. Gu, and F. Flórez-Revuelta, "Visual privacy by context: proposal and evaluation of a level-based visualisation scheme," *Sensors*, vol. 15, no. 6, pp. 12 959-12 982, 2015.

[24] H. Jaeger, "The "echo state" approach to analysing and training recurrent neural networks-with an erratum note," *Technical Report of German National Research Centre for Information Technology, GMD*, vol. 148, no. 13, p. 172-185, 2001.

[25] Q. Ma, L. Shen, E. Chen, S. Tian, J. Wang, and G. W. Cottrell, "Walking walking walking: Action recognition from action echoes," in *Proceedings of International Joint Conference on Artificial Intelligence, Melbourne*, Australia, 2017, pp. 2457-2463.

[26] Q. Wu, E. Fokoue, and D. Kudithipudi, "On the statistical challenges of echo state networks and some potential remedies," *arXiv preprint* arXiv:1806.11230v2, 2018.

[27] L. Mici, X. Hinaut, and S. Wermter, "Activity recognition with echo state networks using 3D body joints and objects category," in *European Symposium On Artificial Neural Networks, Computational Intelligence and Machine Learning, Bruges*, Belgium, Apr. 2016, pp. 465–470.

[28] C. A. Nelson, "Neural plasticity and human development: the role of early experience in sculpting memory systems," *Developmental Science*, vol. 3, no. 2, pp. 115-136, 2000.

[29] L. Boccato, R. Attux, and F. J. V. Zuben, "Self-organization and lateral interaction in echo state network reservoirs," *Neurocomputing*, vol. 138, pp. 297-309, 2014.

[30] N. S. Grossberg, "Competitive learning: From interactive activation to adaptive resonance," *Cognitive Science*, vol. 11, no. 1, pp. 23-63, 1987.

[31] J. Jockusch and H. Ritter, "An instantaneous topological mapping model for correlated stimuli," in *Proceedings of International Joint Conference on Neural Networks*, vol. 1.Washington, DC, USA, 1999, pp. 529-534.

[32] D. Kingma and J. Lei-Ba, "Adam: a method for stochastic optimization," *arXiv preprint* arXiv:1412.6980, 2015.

[33] L. N. Smith, "Cyclical learning rates for training neural networks," in *IEEE Winter Conference on Applications of Computer Vision, Santa Rosa*, CA, USA, 217, pp. 464-472.

[34] K. Biswas and S. K. Basu, "Gesture recognition using Microsoft Kinect," in *5th IEEE International Conference on Robotics and Automation*, 2011, pp. 100-103.

[35] S. Bhattacharya, B. Czejdo, and N. Perez, "Gesture classification with machine learning using Kinect sensor data," in *3rd International Conference on Emerging Applications of Information Technology*, 2012, pp. 348-351.

[36] O. Patsadu, C. Nukoolkit, and B. Watanapa, "Human gesture recognition using Kinect camera," in *International Joint Conference on Computer Science and Software Engineering*, 2012, pp. 28-32.

[37] M. Dupont and P. F. Marteau, "Coarse-dtw for sparse time series alignment," Douzal-Chouakria, A., Vilar, J., Marteau, P.F. (eds.) *Advanced Analysis and Learning on Temporal Data. Lecture Notes in Computer Science*, vol. 9785, pp. 157-172, 2016.

[38] R. Ibanez, A. Soria, A. Teyseyre, and M. Camp, "Easy gesture recognition for Kinect," *Advances in Engineering Software*, vol. 76, pp. 171-180, 2014.

[39] P. F. Marteau, S. Gibet, and C. Reverdy, "Down-sampling coupled to elastic kernel machines for efficient recognition of isolated gestures," in *Proceedings of the 22nd IEEE Computer Society International Conference on Pattern Recognition*, 2014, pp. 363-368.

[40] G. Lev, G. Sadeh, B. Klein, and L. Wolf, "Rnn fisher vectors for action recognition and image annotation," *arXiv preprint* arXiv:1512.03958v1,2016.

[41] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive neural networks for high performance skeleton-based human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1963-1978, 2019.

[42] C. Gallicchio and A. Micheli, "A reservoir computing approach for human gesture recognition from Kinect data," in *Proceedings of Workshop on Artificial Intelligence for Ambient Assisted Living*, vol. 1803, 2016, pp. 33-42.

[43] Y. Bengio, "Deep learning of representations: looking forward," in *International Conference on Statistical Language and Speech Processing*, *Springer*, 2013, pp. 1-37.

[44] H. F. Nweke, T. Y. Wah, M. Al-garadi, and U. R. Alo, "Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges," *Expert Systems With Applications*., vol. 105, pp. 233-261, 2018.

[45] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *Pattern Recognition Letters*, vol. 119, pp.3-11, 2019.

[46] H. Palangi, L. Deng, and R. Ward, "Learning input and recurrent weight matrices in echo state networks," *arXiv preprint* arXiv:abs/1311.2987, 2013.

[47] Boccato, L., de Faissol Attux, R., & Zuben, F. J. V. (2014). Self-organization and lateral interaction in echo state network reservoirs. Neurocomputing, 138, 297-309.

[48] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biological Cybernetics*, vol. 43, no. 1, pp. 59-69, 1982.

[49] B. Fritzke, "Growing cell structures: A self-organizing network for unsupervised and supervised learning," *Neural Networks*, vol. 7, no. 9, pp. 1441-1460, 1994.

[50] G. A. Carpenter and S. Grossberg, "The ART of adaptive pattern recognition by a self-organizing neural network," *Computer*, vol. 3, pp. 77-88,1988.

[51] O. Osoba and B. Kosko, "Noise-enhanced clustering and competitive learning algorithm," *Neural Networks*, vol. 37, pp. 132-140, 2013.

[52] F. M. Bianchi, L. Livi, and C. Alippi, "Investigating echo-state networks dynamics by means of recurrence analysis," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 2, pp. 427-439, 2018.

[53] Z. Wang, W. Yan, and T. Oates, "Time series classification from scratch with deep neural networks: a strong baseline," in *International Joint Conference on Neural Network*, 2017, pp. 1578-1585.

[54] V. Nair and G. E. Hinto, "Rectified linear units improve restricted Boltz-mann machine," in *Proceedings of 27th International Conference on Machine Learning*, June 2010, pp.807-814.

[55] 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, San Francisco*, USA, 13-18 June 2010, pp. 9-14.

[56] 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 Computer Vision and Pattern Recognition Workshops, Portland*, USA, 2013, pp. 479-485.

[57] J. Antunes, A. Bernardino, A. Smailagic, and D. Siewiorek, "AHA-3D: A labelled dataset for senior fitness exercise recognition and segmentation from 3d skeletal data," in *Vision International Behaviour Understanding Workshop, British Machine Vision Conference, Newcastle upon Tyne*, UK, 2018.

[58] 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, Las Vegas*, USA, 2016, pp. 4498-4507.

[59] E. Cippitelli, S. Gasparrini, E. Gambi, and S. Spinsante, "A human activity recognition system using skeleton data from RGBD sensor," *Computational Intelligence and Neuroscience*, vol. 2016, 2016.

[60] Y. Bengio, "Neural networks: Tricks of the trade, chapter Practical recommendations for gradient-based training of deep architectures," pp. 437-478, 2012.

[61] J. Steinier, Y. Termonia, J. Deltour, and A. Chem, "Smoothing and differentiation of data by simplified least square procedure," *Analytical Chemistry*, vol. 44, no. 11, pp. 1906-1909, 1972.

[62] 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*, vol. 13, 2013, pp.2466-2472.

[63] 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, Columbus*, USA, 2014, pp. 588-595.

[64] L. L. Presti, M. L.Cascia, S. Sclaro, and C. O, "Hankelet-based dynamical systems modeling for 3d action recognition," *Image and Vision Computing*, vol. 44,pp. 29-43, 2015.

[65] H. I. Fawaz, G. Forestier, J. Weber, and et al., "Deep learning for timeseries classification: a review," *Data Mining and Knowledge Discovery*, vol. 33, pp. 917-963, 2019.