Extended Hopfield Network for Sequence Learning: Application to Gesture Recognition

André Maurer, Micha Hersch and Aude G. Billard

Ecole Polytechnique Fédérale de Lausanne (EPFL)
Swiss Federal Institute of Technology Lausanne
Autonomous Systems Laboratory
CH-1015 Lausanne, Switzerland Email: aude.billard@epfl.ch

Abstract. In this paper, we extend the Hopfield Associative Memory for storing multiple sequences of varying duration. We apply the model for learning, recognizing and encoding a set of human gestures. We measure systematically the performance of the model against noise.

1 Introduction

The work we present here is part of a research agenda, that aims at modeling the neural correlates of human ability to learn new motions through the observation and replication of other's motions [1, 2]. In this paper, we investigate the use of a biologically plausible mechanism for recognizing, classifying and reproducing gestures.

Associative memories based on Hebbian learning, such as the *Hopfield network*, are interesting candidates to model the propensy of biological systems to encode and learn complex sequences of motion [3]. The Hopfield network is known predominantly for its ability to code static patterns. However, recent work extended the Hopfield model to encode a time series of patterns [4]. In the present work, we extend this model to encode *several* sequences of patterns in the same model. While the capacity of such RNN models have been studied at length in simulation [5, 6], there has been yet little work demonstrating their application to the storage of real data sequences. Here, we validate the model for encoding human gestures and measure the performance of the model in the face of a large amount of noise.

Fundamental features of human ability to imitate new motions are a) the ability to robustly recognize gestures from partially occluded demonstrations (this is tightly linked to our ability to predict the dynamics of the motion from observing only the onset of the motion); and b) to store and reproduce a generalized version of the motion, that encapsulates only the key features of the motion. We show that the model can successfully reproduce these two key features.

2 Experimental Set-up and Model

Figure 1 shows a schematic of the data flow across the complete architecture. Input to the system consists of the kinematic data of human motion. The data is

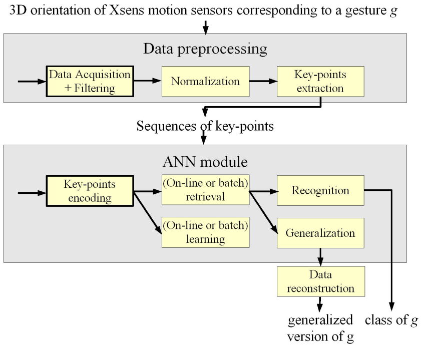

Fig. 1. Schematic of the data flow across the complete architecture.

first preprocessed to smooth and normalize the trajectories, as well as to reduce the dimensionality of the dataset to a subset of keypoints. The time series of keypoints is encoded and classified in a set of Artificial Neural Networks (ANNs). Training of the ANNs results in the storage of a generalized form of each of the demonstrated gestures. The system outputs either the class of the gesture g or the generalized form of the gesture corresponding to the class g.

Data acquisition and preprocessing: Data consist of 45 gestures, composed of the 4 angular trajectories of the arm (shoulder abduction-adduction, flexion-extension and humeral rotation, and elbow flexion-extension) of 8 demonstrators during 5 repetitions of drawing the stylized letters A, B, C, D, E, (see figure 2).

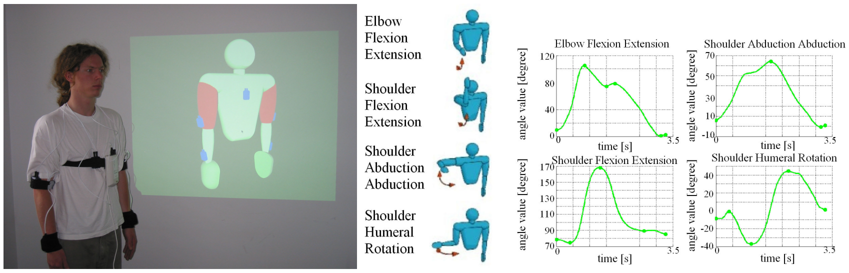

Fig. 2. The demonstrator's motions are recorded by a set of Xsens motion captors, attached to the torso, upper and lower arms (left). The information is then used to reconstruct the trajectories of 4 joint angles of the arm (middle). (Right) Each subplot corresponds to the trajectory of one of the 4 joint angles. Circles represent the key-points used for training the model .

Each trajectory is smoothed using a 1D local Gaussian filter of size 7. From those trajectories, we extract a set of P key-points $\{\theta_p^a, t_p^a\}$ (a=1..4, p=1..P).

A key-point is either the first or last element of the trajectory or an inflexion point (zero velocity). Such a segmentation aims at extracting the correlations between the different joint trajectories. The duration of the whole trajectory is normalized so that two gestures belonging to the same class but performed at different speed are encoded into similar sequences.

Pattern encoding: Each element of the input sequence $\{t^a_p, \theta^a_p\}$ ($a = 1..4$, $p = 1..P$; P being the length of the sequence) is encoded in a 2D matrix $\tilde{x}$ of real values, as follows:

$$(t^{\tilde{a}}_{\tilde{i}}, \theta^{\tilde{a}}_{\tilde{i}}) \rightarrow (\tilde{x}_{u,v})$$

where u=1..M and $v = 1..N$. In order to preserve the notion of neigbourhood across inputs we encode a pair $(t^{\tilde{a}}_{\tilde{p}}, \theta^{\tilde{a}}_{\tilde{p}})$ using a 2D gaussian distribution function f, centered on $\boldsymbol{\mu} = (\mu_t, \mu_\theta)^T$ with standard deviation $\boldsymbol{\sigma} = (\sigma_t, \sigma_\theta)^T$:

$$f(x_u, x_v) = e^{\frac{-\frac{1}{2}(x_u-\mu_t)^2}{\sigma_t^2}} e^{\frac{-\frac{1}{2}(x_v-\mu_\theta)^2}{\sigma_\theta^2}}$$

2.1 ANN module

The general topology of the network is presented in Figure 3. Inputs to the network are sequences of key-points. The sequences are stored in a series of Hopfield networks linked to one another through the matrix of weights W. Each sequence is then classified according to a set of classes $c = 1, .., C$ and C, represented by a set of neurons y_c.

The activity of each neuron, for each angle a, and for each time step t=1..P, $x^a_{u,v}(t)$, as well as the weights $w^a_{u,v,u',v'}(t)$ storing the correlation between the neuronal activities $x^a_{u,v}(t)$ and $x^a_{u,v}(t+1)$ are normalized and bounded in [0..1].

Learning process The learning rule for updating elements of W is a modification of the one presented in [4], to allow storage of several sequences rather than a single one, as well as to allow a non-overlapping encoding with $x^a_{u,v} \in [0..1]$, as opposed to $x^a_{u,v} = \pm 1$.

$$w^a_{u,v,u',v'}(t) = \sum_s x^{s,a}_{u,v}(t) x^{s,a}_{u',v'}(t+1), \quad t = 1, .., P-1 \tag{1}$$

where s indicates the training sequence.

When learning a gesture s belonging to a class $\tilde{c}$, we set the output neurons $y_c = 0 \;\; \forall c \neq \tilde{c}$ and $y_{\tilde{c}} = 1$. Updating the elements of the recognition matrix J is done according to:

$$\mathit{\Delta} J^a_{u,v,c}(t) = x^{s,a}_{u,v}(t) y^{s,a}_c(t) \tag{2}$$

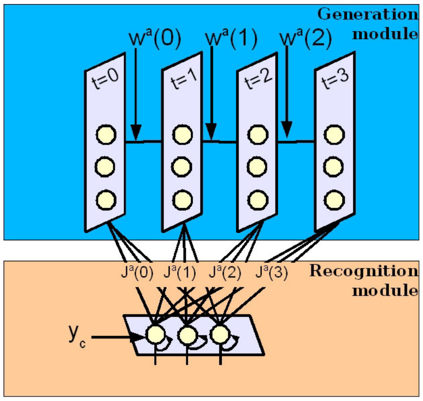

Fig. 3. Network topology. The weight matrix W (generation module) connects all neurons from one layer to all neurons in the next layer. Each output neuron y_c corresponds to a class c of gestures. The weights of the matrix J (recognition module) are set so that the output neuron $y_{\tilde{c}}$ is maximal when a sequence of class $\tilde{c}$ is generated. Each angular trajectory a (a=1..4) is encoded in a separate network.

Retrieval process: In order to retrieve the generalized form of the sequence associated with a given class, we activate one of the y_c neurons and then reactivate the neurons in each layer of the extended Hopfield in sequence for P-1 time steps. That is, we update each neuron $x^a_{u,v}(t+1)$ according to:

$$x^a_{u,v}(t+1) = \sum_{u'} \sum_{v'} w^a_{u,v,u',v'}(t) \cdot x^a_{u',v'}(t), \quad t = 1, .., P-1 \tag{3}$$

Figure 4 shows the history of the network state after the retrieval of a sequence of four elements.

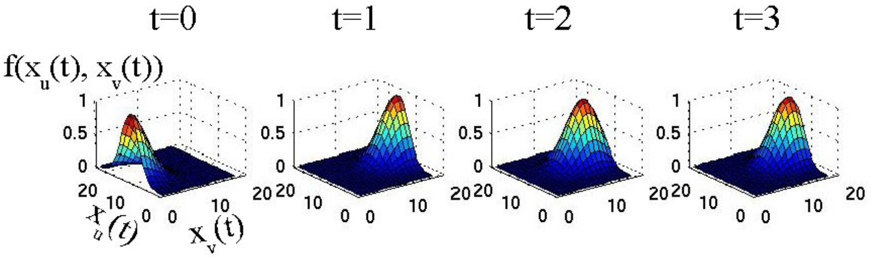

Fig. 4. State of a 4-layer extended Hopfield network while retrieving a sequence of four elements.

During recognition of a gesture, we proceed conversely by activating a subset of the first layers of the extended Hopfield network. Recognition of the class to which the gesture belongs is done by reactivating the output neurons y_c according to:

$$y_c(t+1) = y_c(t) + \sum_a \sum_u \sum_v J^a_{u,v,c}(t) \tag{4}$$

3 Results

We evaluated the performance of the network to classify and regenerate our set of 45 gestures (stylized drawings of the letters A to E). Further, in order to evaluate the network's capacity against a large amount of noise, we generated a *synthetic dataset* of 250 gestures, by adding gaussian noise on one of the gestures belonging to the *real dataset*. Each dataset was divided equally into a *training set* and a *testing set*. Synthetic data were generated by displacing each key-point according to a gaussian distribution function, centered on the original key-point and with a given standard deviation $\boldsymbol{\sigma}^D = (\sigma^D_t, \sigma^D_\theta)^T$, see Figure 5. For each value of σ^D, we generated 10 different gestures. We measured a standard deviation (noise) on the *real dataset* of $\boldsymbol{\sigma}^D = (0.88, 22.23)^T$.

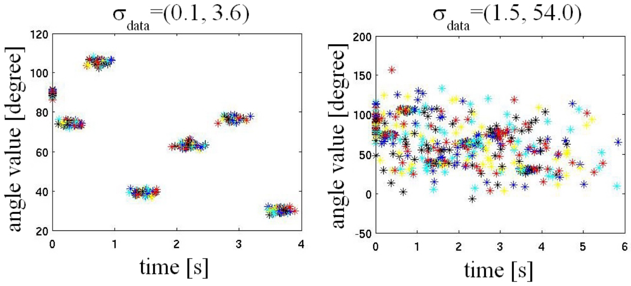

Fig. 5. Sequence of key-points (t, θ) when the noise is generated with $\sigma^D_t = 0.1$ and $\sigma^D_\theta = 3.6$. Clusters do not overlap. Middle: key-points (t, θ) with $\sigma^D_t = 1.5$ and $\sigma^D_\theta = 54.0$. The overlap between clusters is large.

During the learning phase, the system is trained on a set of gestures. During the testing phase, the system is evaluated on its ability to both recognize and regenerate the data.

Recognition Performance: Figure 7, right, shows the recognition rate on the synthetic testing set as an effect of the temporal noise (average over 10 different gestures for each value of σ) with $\sigma^D_t = 3.6\sigma^\theta_D$. The recognition rate τ is given by *the proportion of correctly recognized patterns relative to the total number of patterns.*

We observe that the recognition is perfect for all gestures when the noise is low ($\sigma^D_t \leq 0.25$). However, for high noise ($\sigma^D_t >= 1.0$), the recognition rate decreases importantly.

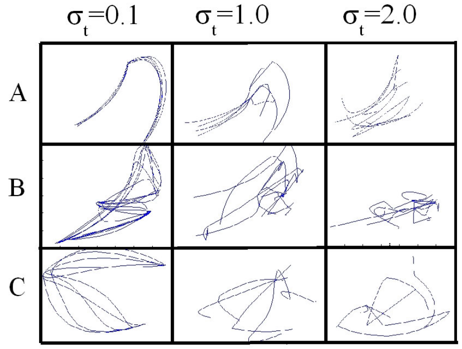

Fig. 6. Distortion of the original gestures with a noise level of $\sigma_t^D = 0.1, 1.0$ and 2 respectively. Sole the gestures on the left are easily recognizable by the human eye.

Data Regeneration: Figure 7, left, shows 3 examples of regenerated gestures, superimposed to a set of 4 training gestures generated with a noise value $\boldsymbol{\sigma}^D = (0.1, 3.6)$. The network generates a generalized form of the gestures that encapsulates the major qualitative features (point of curvature) of the demonstrations.

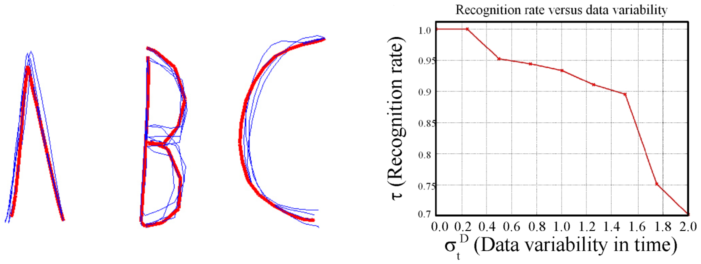

Fig. 7. (Left:) Regenerated gestures (bold line) against a set of 4 examples of demonstrated gestures (thin line) with a noise of $\boldsymbol{\sigma}^D = (0.1, 3.6)^T$. (Right:) Recognition rate as an effect of the noise.

References

1. Arbib, M., Billard, A., Iacoboni, M., Oztop, E.: Mirror neurons, imitation and (synthetic) brain imaging. In: Neural Networks. Volume 13 (8/9). (2000) 975–997
2. Billard, A.: Imitation. In: Handbook of Brain Theory and Neural Networks. Volume 2. MIT Press (2002) 566–569
3. Wang, D.: Temporal pattern processing. The Handbook of Brain Theory and Neural Network **2** (2003) 1163–1167
4. Miyoshi, S., Yanai, H., Okada, M.: Associative memory by recurrent neural networks with delay elements. Neural Networks **17** (2004) 55–63
5. Miyoshi, S., Nakayama, K.: A recurrent neural network with serial delay elements for memorizing limit cycles. In: Proc. of ICANN'95. (1995) 1955–1960
6. Mueller, K.R., Ibens, O.: Sequence storage of asymmetric hopfield networks with delay. In: ICANN'91. (1991) 163–168