

Simulating Actions with the Associative Self-Organizing Map

Miriam Buonamente¹, Haris Dindo¹, and Magnus Johnsson²

¹ RoboticsLab, DICGIM, University of Palermo,
Viale delle Scienze, Ed. 6, 90128 Palermo, Italy
{[miriam.buonamente](mailto:miriam.buonamente@unipa.it),[haris.dindo](mailto:haris.dindo@unipa.it)}@unipa.it
<http://www.unipa.it>

² Lund University Cognitive Science,
Lundagård, 222 22 Lund, Sweden
magnus@magnusjohnsson.se
<http://www.magnusjohnsson.se>

Abstract. We present a system that can learn to represent actions as well as to internally simulate the likely continuation of their initial parts. The method we propose is based on the Associative Self Organizing Map (A-SOM), a variant of the Self Organizing Map. By emulating the way the human brain is thought to perform pattern recognition tasks, the A-SOM learns to associate its activity with different inputs over time, where inputs are observations of other's actions. Once the A-SOM has learnt to recognize actions, it uses this learning to predict the continuation of an observed initial movement of an agent, in this way reading its intentions. We evaluate the system's ability to simulate actions in an experiment with good results, and we provide a discussion about its generalization ability. The presented research is part of a bigger project aiming at endowing an agent with the ability to internally represent action patterns and to use these to recognize and simulate others behaviour.

Keywords: Associative Self-Organizing Map, Neural Network, Action Recognition, Internal Simulation, Intention Understanding

1 Introduction

Robots are on the verge of becoming a part of the human society. The aim is to augment human capabilities with automated and cooperative robotic devices to have a more convenient and safe life. Robotic agents could be applied in several fields such as the general assistance with everyday tasks for elderly and handicapped enabling them to live independent and comfortable lives like people without disabilities. To deal with such desire and demand, natural and intuitive interfaces, which allow inexperienced users to employ their robots easily and safely, have to be implemented.

Efficient cooperation between humans and robots requires continuous and complex intention recognition; agents have to understand and predict human

intentions and motion. In our daily interactions, we depend on the ability to understand the intent of others, which allows us to read other’s mind. In a simple dance, two persons coordinate their steps and their movements by predicting subliminally the intentions of each other. In the same way in multi-agents environments, two or more agents that cooperate (or compete) to perform a certain task have to mutually understand their intentions.

Intention recognition can be defined as the problem of inferring an agent’s intention through the observation of its actions. This problem has been faced in several fields of human-robot collaboration [1]. In robotics, intention recognition has been addressed in many contexts like social interaction [2] and learning by imitation [3] [4] [5].

Intention recognition requires a wide range of evaluative processes including, among others, the decoding of biological motion and the ability to recognize tasks. This decoding is presumably based on the internal simulation [6] of other peoples behaviour within our own nervous system. The visual perception of motion is a particularly crucial source of sensory input. It is essential to be able to pick out the motion to predict the actions of other individuals. Johansson’s experiment [7] showed that humans, just by observing points of lights, were able to perceive and understand movements. By looking at biological motion, such as Johansson’s walkers, humans attribute mental states such as intentions and desires to the observed movements. Recent neurobiological studies [8] corroborate Johansson’s experiment by arguing that the human brain can perceive actions by observing only the human body poses, called postures, during action execution. Thus, actions can be described as sequences of consecutive human body poses, in terms of human body silhouettes [9] [10] [11]. Many neuroscientists believe that the ability to understand the intentions of other people just by observing them depends on the so-called mirror-neuron system in the brain [12], which comes into play not only when an action is performed, but also when a similar action is observed. It is believed that this mechanism is based on the internal simulation of the observed action and the estimation of the actor’s intentions on the basis of a representation of ones own intentions [13].

Our long term goal is to endow an agent with the ability to internally represent motion patterns and to use these patterns to recognize and simulate other’s behaviour. The study presented here is part of a bigger project whose first step was to efficiently represent and recognize human actions [14] by using the Associative Self-Organizing Map (A-SOM) [15]. In this paper we want to use the same biologically-inspired model to predict an agent’s intentions by internally simulating the behaviour likely to follow initial movements. As humans do effortlessly, agents have to be able to elicit the likely continuation of the observed action even if an obstacle or other factors obscure their view. Indeed, as we will see below, the A-SOM can remember perceptual sequences by associating the current network activity with its own earlier activity. Due to this ability, the A-SOM could receive an incomplete input pattern and continue to elicit the likely continuation, i.e. to carry out sequence completion of perceptual activity over time.

We have tested the A-SOM on simulation of observed actions on a suitable dataset made of images depicting the only part of the persons body involved in the movement. The images used to create this dataset was taken from the “INRIA 4D repository ³”, a publicly available dataset of movies representing 13 common actions: check watch, cross arms, scratch head, sit down, get up, turn around, walk, wave, punch, kick, point, pick up, and throw (see Fig. 1).

This paper is organized as follows: A short presentation of the A-SOM network is given in section II. Section III presents the method and the experiments for evaluating performance. Conclusions and future works are outlined in section IV.

2 Associative Self-Organizing Map

The A-SOM is an extension of the Self-Organizing Map (SOM) [16] which learns to associate its activity with the activity of other neural networks. It can be considered a SOM with additional (possibly delayed) ancillary input from other networks, Fig. 2.

Ancillary connections can also be used to connect the A-SOM to itself, thus associating its activity with its own earlier activity. This makes the A-SOM able to remember and to complete perceptual sequences over time. Many simulations prove that the A-SOM, once receiving some initial input, can continue to elicit the likely following activity in the nearest future even though no further input is received [17] [18].

The A-SOM consists of an $I \times J$ grid of neurons with a fixed number of neurons and a fixed topology. Each neuron n_{ij} is associated with $r + 1$ weight vectors $w_{ij}^a \in R^n$ and $w_{ij}^1 \in R^{m_1}$, $w_{ij}^2 \in R^{m_2}$, \dots , $w_{ij}^r \in R^{m_r}$. All the elements of all the weight vectors are initialized by real numbers randomly selected from a uniform distribution between 0 and 1, after which all the weight vectors are normalized, i.e. turned into unit vectors.

At time t each neuron n_{ij} receives $r + 1$ input vectors $x^a(t) \in R^n$ and $x^1(t - d_1) \in R^{m_1}$, $x^2(t - d_2) \in R^{m_2}$, \dots , $x^r(t - d_r) \in R^{m_r}$ where d_p is the time delay for input vector x^p , $p = 1, 2, \dots, r$.

The main net input s_{ij} is calculated using the standard cosine metric

$$s_{ij}(t) = \frac{x^a(t) \cdot w_{ij}^a(t)}{\|x^a(t)\| \|w_{ij}^a(t)\|}, \quad (1)$$

The activity in the neuron n_{ij} is given by

$$y_{ij} = [y_{ij}^a(t) + y_{ij}^1(t) + y_{ij}^2(t) + \dots + y_{ij}^r(t)] / (r + 1) \quad (2)$$

where the main activity y_{ij}^a is calculated by using the softmax function [19]

³ The repository is available at <http://4drepository.inrialpes.fr>. It offers several movies representing sequences of actions. Each video is captured from 5 different cameras. For the experiments in this paper we chose the movie “Julien1” with the frontal camera view “cam0”.

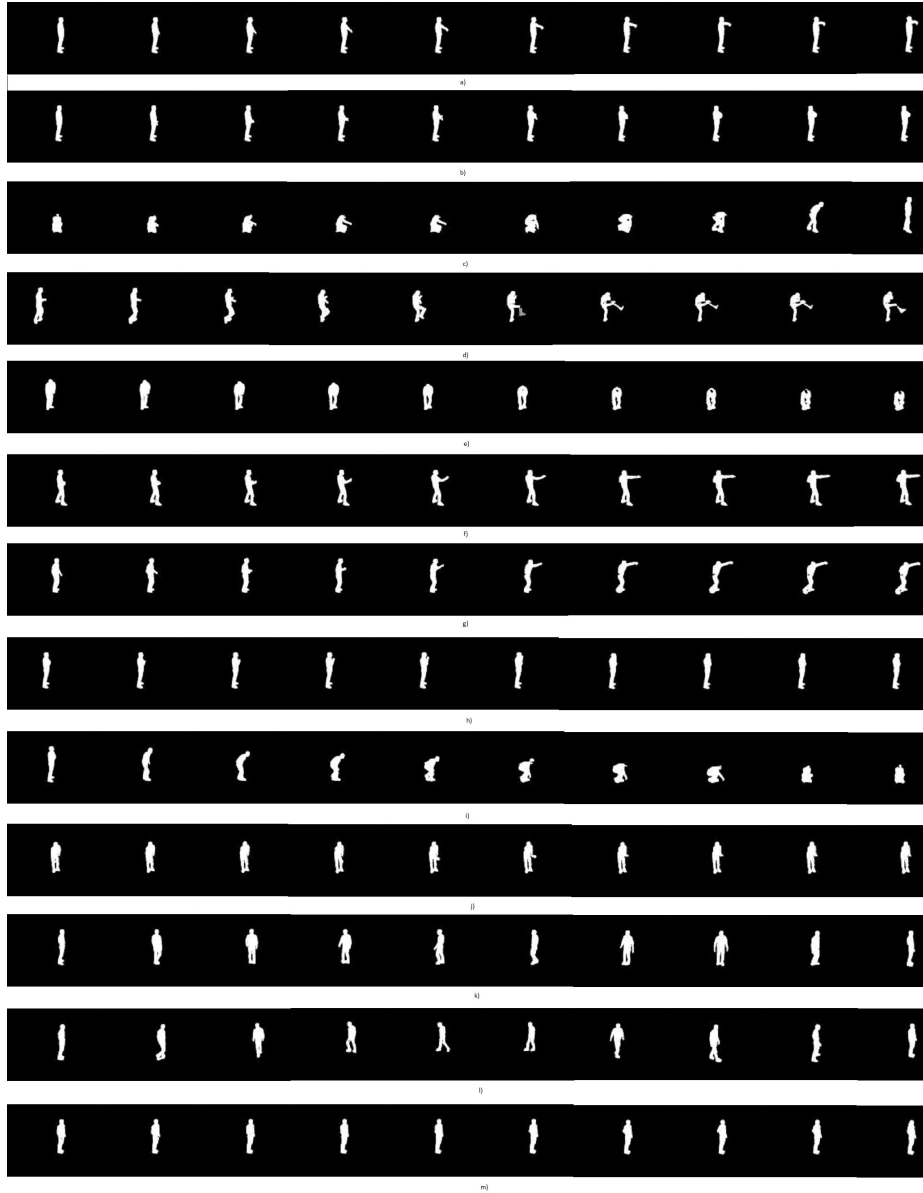


Fig. 1. Prototypical postures of 13 different actions in our dataset: check watch, cross arms, get up, kick, pick up, point, punch, scratch head, sit down, throw, turn around, walk, wave hand.

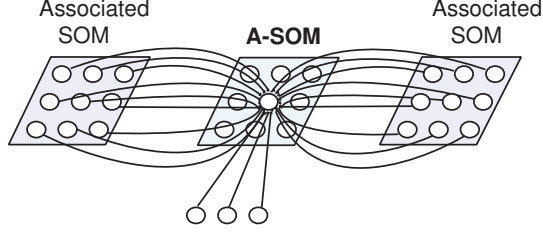


Fig. 2. An A-SOM network connected with two other SOM networks. They provide the ancillary input to the main A-SOM (see the main text for more details).

$$y_{ij}^a(t) = \frac{(s_{ij}(t))^m}{\max_{ij}(s_{ij}(t))^m} \quad (3)$$

where m is the softmax exponent.

The ancillary activity $y_{ij}^p(t)$, $p=1,2,\dots,r$ is calculated by again using the standard cosine metric

$$y_{ij}^p(t) = \frac{x^p(t - d_p) \cdot w_{ij}^p(t)}{\|x^p(t - d_p)\| \|w_{ij}^p(t)\|}. \quad (4)$$

The neuron c with the strongest main activation is selected:

$$c = \operatorname{argmax}_{ij} y_{ij}(t) \quad (5)$$

The weights w_{ijk}^a are adapted by

$$w_{ijk}^a(t+1) = w_{ijk}^a(t) + \alpha(t) G_{ijc}(t) [x_k^a(t) - w_{ijk}^a(t)] \quad (6)$$

where $0 \leq \alpha(t) \leq 1$ is the adaptation strength with $\alpha(t) \rightarrow 0$ when $t \rightarrow \infty$.

The neighbourhood function $G_{ijc}(t) = e^{-\frac{\|r_c - r_{ij}\|}{2\sigma^2(t)}}$ is a Gaussian function decreasing with time, and $r_c \in R^2$ and $r_{ij} \in R^2$ are location vectors of neurons c and n_{ij} respectively.

The weights w_{ijl}^p , $p=1,2,\dots,r$, are adapted by

$$w_{ijl}^p(t+1) = w_{ijl}^p(t) + \beta x_l^p(t - d_p) [y_{ij}^a(t) - y_{ij}^p(t)] \quad (7)$$

where β is the adaptation strength.

All weights $w_{ijk}^a(t)$ and $w_{ijl}^p(t)$ are normalized after each adaptation.

In this paper the ancillary input vector x^1 is the activity of the A-SOM from the previous iteration rearranged into a vector with the time delay $d_1 = 1$.

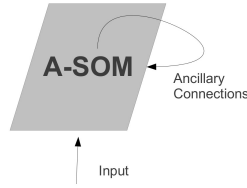


Fig. 3. The model consisting of an A-SOM with time-delayed ancillary connections connected to itself.

3 Experiment

We want to evaluate if the bio-inspired model, introduced and tested for the action recognition task in [14], Fig. 3, is also able to simulate the continuation of the initial part of an action. To this end, we tested the simulation capabilities of the A-SOM. The experiments scope is to verify if the network is able to receive an incomplete input pattern and continue to elicit the likely continuation of recognized actions. Actions, defined as single motion patterns performed by a single human [20], are described as sequences of body postures.

The dataset of actions is the same as we used for the recognition experiment in [14]. It consists of more than 700 postural images representing 13 different actions. Since we want the agent to be able to simulate one action at a time, we split the original movie into 13 different movies: one movie for each action (see Fig. 1). Each frame is preprocessed to reduce the noise and to improve its quality and the posture vectors are extracted (see section 3.1 below). The posture vectors are used to create the training set required to train the A-SOM. Our final training set is composed of about 20000 samples where every sample is a posture vector.

The created input is used to train the A-SOM network. The training lasted for about 90000 iterations. The generated weight file is used to execute tests. The implementation of all code for the experiments presented in this paper was done in C++ using the neural modelling framework Ikaros [21]. The following sections detail the preprocessing phase as well as the results obtained.

3.1 Preprocessing phase

To reduce the computational load and to improve the performance, movies should have the same duration and images should depict the only part of the body involved in the movement. By reducing the numbers of images for each movie to 10, we have a good compromise to have seamless and fluid actions, guaranteeing the quality of the movie. As Fig. 4 shows, the reduction of the number of images, depicting the “walk action” movie, does not affect the quality of the action reproduction.

Consecutive images were subtracted to depict the only part of the body involved in the action, focusing in this way the attention on the movement exclusively. This operation further reduced the number of frames for each movie



Fig. 4. The walk action movie created with a reduced number of images.



Fig. 5. a) The sequence of images depicting the check watch action; b) The sequence of images obtained by subtracting consecutive images of the check watch action.

to 9, without affecting the quality of the video. As can be seen in Fig. 5, in the “walk action” only the arm is involved in the movement.

To further improve the system’s performance, we need to produce binary images of fixed and small size. By using a fixed boundary box, including the part of the body performing the action, we cut out the images eliminating anything not involved in the movement. In this way, we simulate an attentive process in which the human eye observes and follows the salient parts of the action only. To have smaller representations the binary images depicting the actions were shrunk to 30×30 matrices. Finally, the obtained matrix representations were vectorized to produce 9 posture vectors $p \in R^D$, where $D = 900$, for each action. These posture vectors are used as input to the A-SOM.

3.2 Action Simulation

The objective was to verify whether the A-SOM is able to internally simulate the likely continuation of initial actions. Thus, we fed the trained A-SOM with incomplete input patterns and expected it to continue to elicit activity patterns corresponding to the remaining part of the action. The action recognition task has been already tested in [14] with good results. The system we set up was the same as the one used in [14] and consists of one A-SOM connected to itself with time delayed ancillary connections. To evaluate the A-SOM, 13 sequences each containing 9 posture vectors were constructed as explained above. Each of these sequences represents an action. The posture vectors represent the binary images that form the videos and depict only the part of the human body involved in the action, see Fig.6

We fed the A-SOM with one sequence at a time, reducing the number of posture vectors at the end of the sequence each time and replacing them with null vectors (representing no input). In this way, we created the incomplete input

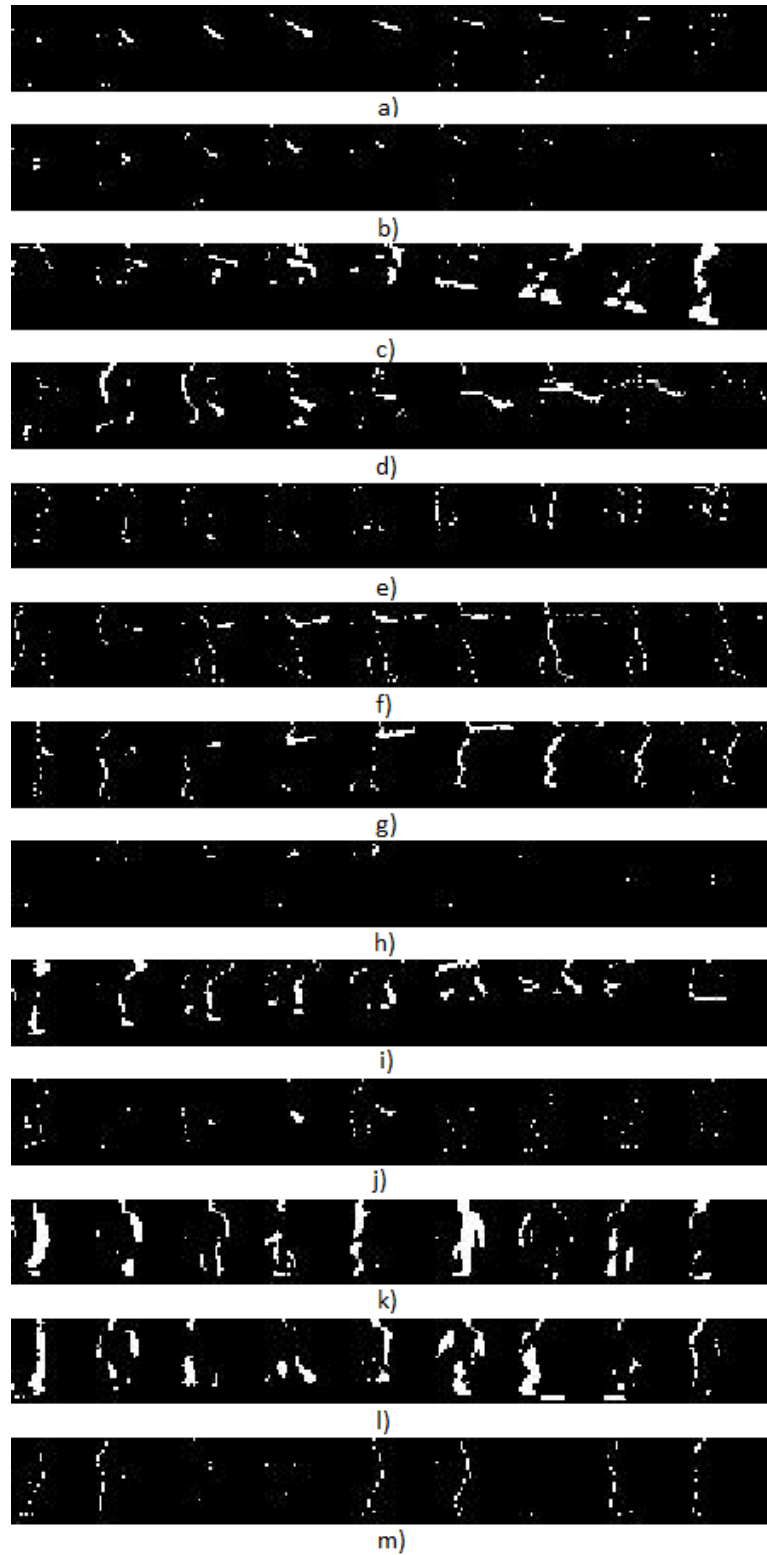


Fig. 6. The parts of the human body involved in the movement of each action. Each sequence was obtained by subtracting consecutive images in each movie. The actions are: a) check watch; b) cross arm; c) get up; d) kick; e) pick up; f) point; g) punch; h) scratch head; i) sit down; j) throw; k) turn around; l) walk; m) wave hand.

that the A-SOM has to complete. The conducted experiment consisted of several tests. The first one was made by using the sequences consisting of all the 9 frames with the aim to record the coordinates of the activity centres generated by the A-SOM and to use these values as reference values for the further iterations. Subsequent tests had the sequences with one frame less (replaced by a null vector representing no input) each time and the A-SOM had the task to complete the frame sequence by eliciting activity corresponding to the activity representing the remaining part of the sequence. The last test included only the sequences made of one frame (followed by 8 null vectors representing no input).

The centres of activity generated by the A-SOM at each iteration were collected in tables, and colour coding was used to indicate the ability (or the inability) of the A-SOM to predict the action continuation. The dark green colour indicates that the A-SOM predicted the right centres of activity; the light green indicates that the A-SOM predicted a value close to the expected centre of activity and the red one indicates that the A-SOM could not predict the right value, see Fig.7. The ability to predict varies with the type of action. For actions like “sit down” and “punch”, A-SOM needed 8 images to predict the rest of the sequence; whereas for the “walk” action, A-SOM needed only 4 images to complete the sequence. In general the system needed between 4 and 9 inputs to internally simulate the rest of the actions. This is a reasonable result, since even humans cannot be expected to be able to predict the intended action of another agent without a reasonable amount of initial information. For example, looking at the initial part of an action like “punch”, we can hardly say what the person is going to do. It could be “punch” or “point”; we need more frames to exactly determine the performed action. In the same way, looking at a person starting to walk, we cannot say in advance if the person would walk or turn around or even kick because the initial postures are all similar to one another.

The results obtained through this experiment allowed us to speculate about the ability of the A-SOM to generalize. The generalization is the network’s ability to recognize inputs it has never seen before. Our idea is that if the A-SOM is able to recognize images as similar by generating close or equal centres of activity, then it will also be able to recognize an image it has never encountered before if this is similar to a known image. We checked if similar images had the same centres of activity and if similar centres of activity corresponded to similar images. The results show that the A-SOM generated very close or equal values for very similar images, see Fig.8. Actions like “turn around”, “walk” and “get up” present some frames very similar to each other and for such frames the A-SOM generates the same centres of activity. This ability is validated through the selection of some centres of activity and the verification that they correspond to similar images. “Check watch”, “get up”, “point” and “kick” actions include in their sequences frames depicting the movement of the arm that can be attributed to all of them. For these frames the A-SOM elicits the same centre of activity, see Fig. 9. The results presented here support the belief that our system is also able to generalize.

MOVIE1: Check watch										MOVIE2: Cross arms										MOVIE3: Get up												
Number of images used for the simulation										Number of images used for the simulation										Number of images used for the simulation												
9	8	7	6	5	4	3	2	1		9	8	7	6	5	4	3	2	1		9	8	7	6	5	4	3	2	1				
Image1	2	6	2	6	2	6	2	6	2	6	2	6	2	6	2	6	2	6	2	6	2	6	2	6	2	6	2	6	2	6		
Image2	14	5	14	5	14	5	14	5	14	5	14	5	14	5	14	5	14	5	14	5	14	5	14	5	14	5	14	5	14	5	14	
Image3	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	
Image4	6	0	6	0	6	0	6	0	6	0	6	0	6	0	6	0	6	0	6	0	6	0	6	0	6	0	6	0	6	0	6	
Image5	0	6	0	6	0	6	0	6	0	6	0	6	0	6	0	6	0	6	0	6	0	6	0	6	0	6	0	6	0	6	0	6
Image6	0	3	0	3	0	3	0	3	0	3	0	3	0	3	0	3	0	3	0	3	0	3	0	3	0	3	0	3	0	3	0	3
Image7	14	0	14	0	14	0	14	0	14	0	14	0	14	0	14	0	14	0	14	0	14	0	14	0	14	0	14	0	14	0	14	
Image8	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	
Image9	14	6	14	0	14	0	14	0	14	0	14	0	14	0	14	0	14	0	14	0	14	0	14	0	14	0	14	0	14	0	14	

Fig. 7. Simulation results: The tables show the ability of the A-SOM to continue the likely continuation of an observed behaviour. Dark green colour indicates that the A-SOM is able to simulate, light green colour indicates that the A-SOM predicts a value very close to the expected one, and red colour indicates that the A-SOM predicts the wrong value. The system needs between 4 and 9 inputs to internally simulate the rest of the sequence.

		Similar Images																
Check watch	Cross arms	Winners	Point	Punch	Winners	Scrath head	Cross arms	Winners	Sit down	Get up	Winners	Turn around	Get up	Winners	Turn around	Walk	Winners	
		14,5	0,0			0,0	3,10		14,6	14,14			7,3	9,14			14,2	8,12
		1,0	0,0			6,6	6,0				8,7	9,14			5,24	5,14		
		0,3	2,8			0,3	13,0											

Fig. 8. Similar images have similar centres of activity. The A-SOM elicits similar or equal centres of activity for images that are similar.













Same winner values					
Winners	Check watch	Cross arms	Point	Sit down	Walk
14 14					
14 0					
14 6					

Fig. 9. Images with the same centres of activity (winners). The frames present similar features which lead the A-SOM to elicit the same centre of activity.

4 Conclusion

In this paper, we proposed a new method for internally simulating behaviours of observed agents. The experiment presented here is part of a bigger project whose scope is to develop a cognitive system endowed with the ability to read other's intentions. The method is based on the A-SOM, a novel variant of the SOM, whose ability of recognition and classification has already been tested in [14]. In our experiment, we connected the A-SOM to itself with time delayed ancillary connections and the system was trained and tested with a set of images depicting the part of the body performing the movement. The results presented here show that the A-SOM can receive some initial sensory input and internally simulate the rest of the action without any further input.

Moreover, we verified the ability of the A-SOM to recognize input never encountered before, with encouraging results. In fact, the A-SOM recognizes similar actions by eliciting close or identical centres of activity.

We are currently working on improving the system to increase the recognition and simulation abilities.

Acknowledgements The authors gratefully acknowledge the support from the Linnaeus Centre Thinking in Time: Cognition, Communication, and Learning, financed by the Swedish Research Council, grant no. 349-2007-8695.

References

1. Awais, M., Henrich, D.: Human-robot collaboration by intention recognition using probabilistic state machines. In: *Robotics in Alpe-Adria-Danube Region (RAAD)*, 2010 IEEE 19th International Workshop on Robotics. (2010) 75–80
2. Breazeal, C.: *Designing sociable robots*. the MIT Press (2004)
3. Chella, A., Dindo, H., Infantino, I.: A cognitive framework for imitation learning. *Robotics and Autonomous Systems* **54**(5) (2006) 403–408
4. Chella, A., Dindo, H., Infantino, I.: Imitation learning and anchoring through conceptual spaces. *Applied Artificial Intelligence* **21**(4-5) (2007) 343–359
5. Argall, B.D., Chernova, S., Veloso, M., Browning, B.: A survey of robot learning from demonstration. *Robotics and Autonomous Systems* **57**(5) (2009) 396–483
6. Hesslow, G.: Conscious thought as simulation of behaviour and perception. *Trends in Cognitive Sciences* **6** (2002) 242–247
7. Johansson, G.: Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics* **14**(2) (1973) 201–211
8. Giese, M.A., Poggio, T. *Nat Rev Neurosci* **4**(3) (March 2003) 179–192
9. Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. *IEEE Trans. Pattern Anal. Mach. Intell.* **29**(12) (2007) 2247–2253
10. Iosifidis, A., Tefas, A., Pitas, I.: View-invariant action recognition based on artificial neural networks. *IEEE Trans. Neural Netw. Learning Syst.* **23**(3) (2012) 412–424
11. Gkalelis, N., Tefas, A., Pitas, I.: Combining fuzzy vector quantization with linear discriminant analysis for continuous human movement recognition. *IEEE Transactions on Circuits Systems Video Technology* **18**(11) (2008) 1511–1521
12. Rizzolatti, G., Craighero, L.: The mirror-neuron system. *Annual Review of Neuroscience* **27** (2004) 169–192
13. Goldman, A.I.: *Simulating minds: The philosophy, psychology, and neuroscience of mindreading*. (2) (2006)
14. Buonamente, M., Dindo, H., Johnsson, M.: Recognizing actions with the associative self-organizing map. In: *the proceedings of the XXIV International Conference on Information, Communication and Automation Technologies (ICAT 2013)*. (2013)
15. Johnsson, M., Balkenius, C., Hesslow, G.: Associative self-organizing map. In: *Proceedings of IJCCI*. (2009) 363–370
16. Kohonen, T.: *Self-Organization and Associative Memory*. Springer Verlag (1988)
17. Johnsson, M., Gil, D., Balkenius, C., Hesslow, G.: Supervised architectures for internal simulation of perceptions and actions. In: *Proceedings of BICS*. (2010)
18. Johnsson, M., Mendez, D.G., Hesslow, G., Balkenius, C.: Internal simulation in a bimodal system. In: *Proceedings of SCAI*. (2011) 173–182
19. Bishop, C.M.: *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford (1995)
20. Turaga, P.K., Chellappa, R., Subrahmanian, V.S., Udrea, O.: Machine recognition of human activities: A survey. *IEEE Trans. Circuits Syst. Video Techn.* **18**(11) (2008) 1473–1488
21. Balkenius, C., Morén, J., Johansson, B., Johnsson, M.: Ikaros: Building cognitive models for robots. *Advanced Engineering Informatics* **24**(1) (2010) 40–48