Available online at www.sciencedirect.com

ScienceDirect

Cognitive Systems Research xxx (2016) xxx–xxx

Cognitive Systems RESEARCH

www.elsevier.com/locate/cogsys

Emergence of multimodal action representations from neural network self-organization

German I. Parisi [a,*], Jun Tani [b], Cornelius Weber [a], Stefan Wermter [a]

[a] Knowledge Technology Group, Department of Informatics, University of Hamburg, Germany
[b] Department of Electrical Engineering, KAIST, Daejeon, Republic of Korea

Received 30 March 2016; received in revised form 21 July 2016; accepted 15 August 2016

Abstract

The integration of multisensory information plays a crucial role in autonomous robotics to forming robust and meaningful representations of the environment. In this work, we investigate how robust multimodal representations can naturally develop in a self-organizing manner from co-occurring multisensory inputs. We propose a hierarchical architecture with growing self-organizing neural networks for learning human actions from audiovisual inputs. The hierarchical processing of visual inputs allows to obtain progressively specialized neurons encoding latent spatiotemporal dynamics of the input, consistent with neurophysiological evidence for increasingly large temporal receptive windows in the human cortex. Associative links to bind unimodal representations are incrementally learned by a semi-supervised algorithm with bidirectional connectivity. Multimodal representations of actions are obtained using the co-activation of action features from video sequences and labels from automatic speech recognition. Experimental results on a dataset of 10 full-body actions show that our system achieves state-of-the-art classification performance without requiring the manual segmentation of training samples, and that congruent visual representations can be retrieved from recognized speech in the absence of visual stimuli. Together, these results show that our hierarchical neural architecture accounts for the development of robust multimodal representations from dynamic audiovisual inputs.

Keywords: Human action recognition; multimodal integration; self-organizing neural networks

1. Introduction

As humans, our daily perceptual experience is modulated by an array of sensors that convey different types of modalities such as visual, auditory, and somatosensory information. The ability to integrate multisensory information is a fundamental feature of the brain that provides a robust perceptual experience for an efficient interaction with the environment (Ernst & Bulthoff, 2004; Stein & Meredith, 1993; Stein, Stanford, & Rowland, 2009). Similarly, computational models for multimodal integration are a paramount ingredient of autonomous robots to forming robust and meaningful representations of perceived events (see Ursino, Cuppini, & Magosso (2014) for a recent review). There are numerous advantages from the crossmodal processing of sensory inputs conveyed by rich and uncertain information streams. For instance, the integration of stimuli from different sources may be used to attenuate noise and remove ambiguities from converging or complementary inputs. Multimodal representations have been shown to improve robustness in the context of action recognition and action-driven perception, human-robot interaction, and sensory-driven motor

* Corresponding author.
E-mail address: parisi@informatik.uni-hamburg.de (G.I. Parisi).

http://dx.doi.org/10.1016/j.cogsys.2016.08.002
1389-0417/

behavior (Bauer, Magg, & Wermter, 2015; Kachouie, Sedighadeli, Khosla, & Chu, 2014; Noda, Arie, Suga, & Ogata, 2014). However, multisensory inputs must be represented and integrated in an appropriate way so that they result in a reliable cognitive experience aimed to trigger adequate behavioral responses. Since real-world events unfold at multiple spatial and temporal scales, artificial neurocognitive architectures should account for the efficient processing and integration of spatiotemporal stimuli with different levels of complexity (Fonlupt, 2003; Hasson, Yang, Vallines, Heeger, & Rubin, 2008; Lerner, Honey, Silbert, & Hasson, 2011; Taylor, Hobbs, Burroni, & Siegelmann, 2015). Consequently, the question of how to acquire, process, and bind multimodal knowledge in learning architectures represents a fundamental issue still to be fully investigated.

A number of computational models have been proposed that aim to effectively integrate multisensory information, in particular audiovisual input. These approaches generally use unsupervised learning for obtaining visual representations of the environment and then link these features to auditory cues. For instance, Vavrečka and Farkaš (2014) presented a connectionist architecture that learns to bind visual properties of objects (spatial location, shape and color) to proper lexical features. These unimodal representations are bound together based on the co-occurrence of audiovisual inputs using a self-organizing neural network. Similarly, Morse, Benitez, Belpaeme, Cangelosi, and Smith (2015) investigated how infants may map a name to an object and how body posture may affect these mappings. The computational model is driven by visual input and learns word–to–object mappings through body posture changes and online speech recognition. Unimodal representations are obtained with neural network self-organization and multimodal representations develop through the activation of unimodal modules via associative connections. The development of associations between co-occurring stimuli for multimodal binding has been strongly supported by neurophysiological evidence (Fiebelkorn, Foxe, & Molholm, 2009).

However, the above-mentioned approaches do not naturally scale up to learn more complex spatiotemporal patterns such as action–word mappings. In fact, action words do not label actions in the same way that nouns label objects (Gentner, 1982). While nouns generally refer to objects that can be perceived as distinct units, action words refer instead to spatiotemporal relations within events that may be performed in many different ways with high spatial and temporal variance. Humans have an astonishing capability to promptly process complex events, exhibiting high tolerance to sensory distortions and temporal variance. The human cortex comprises a hierarchy of spatiotemporal receptive fields for features with increasing complexity of representation (Hasson et al., 2008; Lerner et al., 2011; Taylor et al., 2015), i.e. higher-level areas process information accumulated over larger spatiotemporal receptive windows. Therefore, further work is required to address the learning of multimodal representation of spatiotemporal inputs for obtaining robust action–word mappings.

To tackle the visual recognition of actions, learning-based approaches typically generalize a set of labeled training action samples and then predict the labels of unseen samples by computing their similarity with respect to the learned action templates. In particular, neurobiologically-motivated methods have been shown to recognize actions with high accuracy from video sequences with the use of spatiotemporal hierarchies that functionally resemble the organization of earlier areas of the visual cortex (e.g. Giese & Poggio, 2003; Jung, Hwang, & Tani, 2015; Layher, Giese, & Neumann, 2014; Parisi, Weber, & Wermter, 2015). These methods are trained with a batch learning scheme, and thus assuming that all the training samples and sample labels are available during the training phase. However, an additional strong assumption is that training samples, typically represented as a sequence of feature vectors extracted from video frames, are well segmented so that ground-truth labels can be univocally assigned. Therefore, it is usually the case that raw visual data collected by sensors must undergo an intensive preprocessing pipeline before training a model. These preprocessing stages are mainly performed manually, thereby hindering the automatic, continuous learning of actions from live video. Intuitively, this is not the case in nature.

Words for actions and events appear to be among children's earliest vocabulary (Bloom, 1993). A central question in the field of developmental learning has been how children first attach verbs to their referents. During their development, children have a wide range of perceptual, social, and linguistic cues at their disposal that they can use to attach a novel label to a novel referent (Hirsch-Pasek, Golinkoff, & Hollich, 2000, chapter 6). Referential ambiguity of verbs may then be solved by children assuming that words map onto the most perceptually salient action in their environment. Recent experiments have shown that human infants are able to learn action–word mappings using cross-situational statistics, thus in the presence of sometimes unavailable ground-truth action words (Smith & Yu, 2008). Furthermore, action words can be progressively learned and improved from linguistic and social cues so that novel words can be attached to existing visual representations. This hypothesis is supported by neurophysiological studies evidencing strong links between the cortical areas governing visual and language processing, and suggesting high levels of functional interaction of these areas for the formation of multimodal representations of audiovisual stimuli (Belin, Zatorre, Lafaille, Ahad, & Pike, 2000; Foxe et al., 2000; Belin, Zatorre, & Ahad, 2002; Pulvermüller, 2005; Raij, Uutela, & Hari, 2000).

From a neurobiological perspective, neurons selective to actions in terms of time-varying patterns of body pose and motion features have been found in a wide number of brain structures, such as the superior temporal sulcus (STS), the parietal, the premotor and the motor cortex (Giese & Rizzolatti, 2015). In particular, it has been argued that

Please cite this article in press as: Parisi, G. I., et al. Emergence of multimodal action representations from neural network self-organization. *Cognitive Systems Research* (2016), http://dx.doi.org/10.1016/j.cogsys.2016.08.002

the STS in the mammalian brain may be the basis of an action-encoding network with neurons driven by the perception of dynamic human bodies (Vangeneugden, Pollick, & Vogels, 2009), and that for this purpose it receives converging inputs from earlier visual areas from both the ventral and dorsal pathways (Beauchamp, 2005; Garcia & Grossman, 2008; Thirkettle, Benton, & Scott-Samuel, 2009). Furthermore, neuroimaging studies have shown that the posterior STS shows a greater response for audiovisual stimuli than to unimodal visual or auditory stimuli (Beauchamp, Lee, Argall, & Martin, 2004; Calvert, 2001; Senkowski, Saint-Amour, Hfle, & Foxe, 2011; Wright, Pelphrey, Allison, Mckeown, & Mccarthy, 2003). Thus, the STS area is thought to be an associative learning device for linking different unimodal representations and accounting for the mapping of naturally occurring, highly correlated features such as body pose and motion, the characteristic sound of an action (Barraclough, Xiao, Baker, Oram, & Perrett, 2005; Beauchamp et al., 2004) and linguistic stimuli (Belin et al., 2002; Stevenson & James, 2009; Wright et al., 2003). These findings together suggest that multimodal representations of actions in the brain play an important role for a robust perception of complex action patterns, with the STS representing a multisensory area in the brain network for social cognition (Adolphs, 2003; Allison, Puce, & McCarthy, 2000; Beauchamp, 2005; Beauchamp, Yasar, Frye, & Ro, 2008).

In this work, we investigate how congruent multimodal representations of actions can naturally emerge from the co-occurrence of audiovisual stimuli. In particular, we propose an approach where associative links between unimodal representations are incrementally learned in a self-organizing manner. For this purpose, we extended our recently proposed spatiotemporal hierarchy for the integration of pose-motion action cues (Parisi et al., 2015) to include an associative network layer where action–word mappings develop from co-occurring audiovisual inputs using asymmetric inter-layer connectivity. Each network layer comprises a self-organizing neural network that employs neurobiologically-motivated Hebbian-like plasticity and habituation for stable incremental learning (Marsland, Shapiro, & Nehmzow, 2002). The proposed architecture is novel in two main aspects: First, our learning mechanism does not require the manual segmentation of training samples. Instead, spatiotemporal generalizations of actions are incrementally obtained and mapped to symbolic labels using the co-activation of audiovisual stimuli. This allows us to train the model in an incremental fashion also in the presence of occasionally unlabeled samples. Second, we let asymmetric inter-layer connectivity emerge taking into account the spatiotemporal dynamics of sequences so that symbolic labels are linked to temporally-ordered representations in the visual domain. This kind of connectivity allows the bidirectional retrieval of audiovisual inputs, i.e. it is possible to retrieve action words from processed visual patterns and, conversely, to activate congruent visualizations of learned actions from recognized action words. We conduct a set of experiments with a dataset of 10 full-body actions, using body pose-motion cues as visual features and action labels obtained from automatic speech recognition. Experimental results show that we achieve state-of-the-art recognition performance without the need to manually segment training samples, and that this performance is not drastically compromised as the number of available labeled samples is decreased.

2. Methods

Our neural architecture consists of a self-organizing hierarchy with four network layers for the unsupervised processing of visual action features and the development of associative connections between learned action representations and symbolic labels. An overall diagram of the architecture is shown in Fig. 1. Network layers 1 and 2 comprise a two-stream hierarchy for the processing and subsequent integration of body pose and motion features, resembling the ventral and the dorsal pathway respectively for the processing of complex motion patterns (Giese & Poggio, 2003). The integration of pose and motion cues is carried out in network layer 3 (or G^{STS}) to provide movement dynamics in the joint feature space (Parisi et al., 2015). Hierarchical learning from contiguous Growing When Required (GWR) networks (Marsland et al., 2002) shapes a functional hierarchy that processes spatiotemporal visual patterns with an increasing level of complexity by using neural activation trajectories from lower-level layers for training higher-level layers. For learning multimodal representation of actions, network layer 4 (or G^{STSm}) implements a self-organizing algorithm where action–word mappings are developed by binding co-occurring audiovisual inputs using bidirectional inter-layer connectivity. For this purpose, we extended the traditional GWR learning algorithm with a mechanism for semi-supervised label propagation and enhanced synaptic connectivity for learning prototype neural activation patterns in the temporal domain. The proposed learning algorithm is referred to as Online Semi-Supervised GWR (OSS-GWR). The self-organizing associative connectivity between G^{STSm} and the Action Words layer (AWL) will yield an incremental formation of congruent action–word mappings for the bidirectional retrieval of audiovisual patterns.

2.1. A self-organizing spatiotemporal hierarchy

Experience-driven development plays a crucial role in the brain (Nelson, 2000), with topographic maps being a common feature of the cortex for processing sensory inputs (Willshaw & von der Malsburg, 1976). Different models of neural self-organization have been proposed to resemble the dynamics of basic biological findings on Hebbian-like learning and map plasticity (e.g., Fritzke, 1995; Kohonen, 1988).

Please cite this article in press as: Parisi, G. I., et al. Emergence of multimodal action representations from neural network self-organization. *Cognitive Systems Research* (2016), http://dx.doi.org/10.1016/j.cogsys.2016.08.002

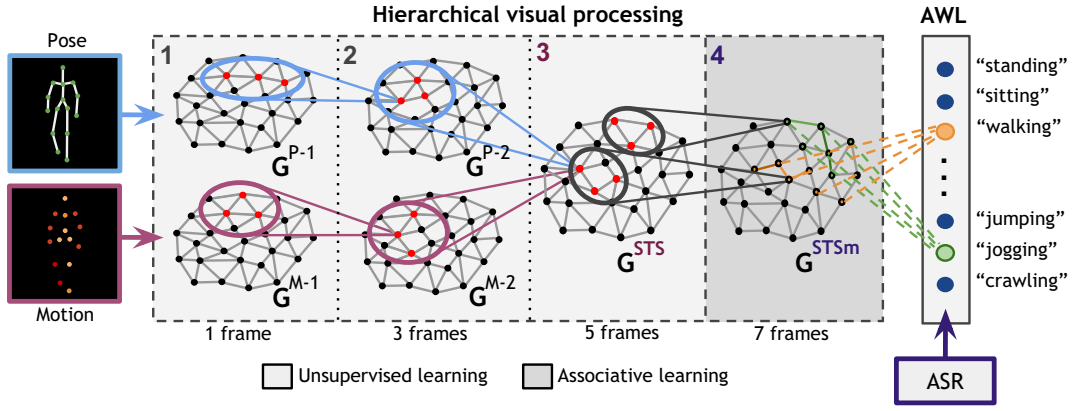

Fig. 1. Diagram of our learning architecture with GWR networks and the number of frames required for hierarchical processing. Layers 1–3: parallel spatiotemporal clustering of visual features and self-organizing pose-motion integration (G^{STS}). Layer 4: Self-organization of G^{STS} representations and associative learning for linking visual representations in G^{STSm} to the action words (AWL) obtained from automatic speech recognition (ASR).

Our learning model consists of hierarchically-arranged GWR networks (Marsland et al., 2002) that obtain progressively generalized representations of sensory inputs and learn inherent spatiotemporal dependencies. The GWR network is composed of a set of neurons with their associated weight vectors linked by a set of edges. During the training, the network dynamically changes its topological structure to better match the input space following competitive Hebbian learning (Martinetz, 1993). Different from other incremental models of self-organization that create new neurons at a fixed growth rate (e.g. Fritzke, 1995, 1997), GWR-based learning creates new nodes whenever the activity of trained neurons is smaller than a given threshold. The amount of network activation at time t is computed as a function of the distance between the current input $\mathbf{x}(t)$ and its best-matching neuron $\mathbf{w}_b$:

$$a(t) = \exp(-\|\mathbf{x}(t) - \mathbf{w}_b\|). \quad (1)$$

Additionally, the training algorithm considers the number of times that a neuron has fired so that recently created neurons are properly trained before creating new ones. For this purpose, the network implements a firing counter $\eta \in [0, 1]$ to express how frequently a neuron has fired based on a simplified model of how the efficacy of an habituating synapse reduces over time (Stanley, 1976). The firing counter is given by

$$\eta(t) = \eta_0 - \frac{S(t)}{\alpha} \cdot (1 - \exp(-\alpha_t/\tau)), \quad (2)$$

where η_0 is the resting value, $S(t)$ is the stimulus strength, and τ and α are constants that control the behavior of the curve.

The use of an activation threshold and firing counters to modulate the growth of the network leads to create a larger number of neurons at early stages of the training and then tune the weights of existing neurons through subsequent training iterations (epochs). This behavior is particularly convenient for incremental learning scenarios since neurons will be created to promptly distribute in the input space, thereby yielding fast convergence through iterative fine-tuning of the topological map. The GWR algorithm will then iterate over the training set until a given stop criterion is met, e.g. a maximum network size or a maximum number of iterations. The learning procedure for GWR is illustrated by Algorithm 1 (except for steps 3, 7.c, 8.c, 9, and 10 that are implemented by the OSS-GWR only).

Algorithm 1. OSS-GWR - In layers 1, 2, and 3 of our architecture, we use GWR learning, while in layer 4 (G^{STSm}) we use OSS-GWR.

1: Create two random neurons with weights $\mathbf{w}_1$ and $\mathbf{w}_2$.
2: Initialize an empty set of spatial connections $E = \varnothing$.
3: [OSS-GWR only] Initialize an empty set of temporal connections $P = \varnothing$ and a set of label–to–action references $V = \varnothing$.
4: At each iteration t, generate an input sample $\mathbf{x}(t)$ with sample label ξ
5: Select the best and the second-best matching neuron such that:
$b = \arg\min_{n \in A} \|\mathbf{x}(t) - \mathbf{w}_n\|$,
$s = \arg\min_{n \in A/\{b\}} \|\mathbf{x}(t) - \mathbf{w}_n\|$.
6: Create a connection $E = E \cup \{(b, s)\}$ if it does not exist and set its age to 0.
7: If $(\exp(-\|\mathbf{x}(t) - \mathbf{w}_b\|) < a_T)$ and $(\eta_b < f_T)$ then:
a: Add a new node r $(A = A \cup \{r\})$ with $\mathbf{w}_r = 0.5 \cdot (\mathbf{x}(t) + \mathbf{w}_b), \eta_r = 1$,
b: Update edges: $E = E \cup \{(r, b), (r, s)\}$ and $E = E/\{(b, s)\}$,
c: [OSS-GWR only] Initialize neuron label (Eq. 4): $\lambda(r) = \gamma^{\text{new}}(b, \xi)$.
8: If no new neuron is added:
a: Update the best-matching neuron $\mathbf{w}_b$ and its neighbors i:
$\Delta\mathbf{w}_b = \epsilon_b \cdot \eta_b \cdot (\mathbf{x}(t) - \mathbf{w}_b)$,
$\Delta\mathbf{w}_i = \epsilon_n \cdot \eta_i \cdot (\mathbf{x}(t) - \mathbf{w}_i)$,
with the learning rates $0 < \epsilon_n < \epsilon_b < 1$.

Please cite this article in press as: Parisi, G. I., et al. Emergence of multimodal action representations from neural network self-organization. *Cognitive Systems Research* (2016), http://dx.doi.org/10.1016/j.cogsys.2016.08.002

b: Increment the age of all edges connected to b by 1.
c: [OSS-GWR only] Update neuron label (Eq. 5): $\lambda(b) = \gamma^{\text{update}}(b, s, \xi)$.

9: [OSS-GWR only] Create a temporal connection $P_b^{b(t-1)}$ if it does not exist, increase it by a value of 1 and decrease all the others if $P_n^{b(t-1)} > 0$.

10: [OSS-GWR only] Create or update the label–to–action reference $V_b^{\lambda(b)}$ (Eq. 8): $V_b^{\lambda(b)} = \Lambda^{\lambda(b)}(b)$.

11: Reduce the firing counters of the best-matching neuron and its neighbors i: $\Delta\eta_b = \tau_b \cdot \kappa \cdot (1 - \eta_b) - \tau_b$, $\Delta\eta_i = \tau_i \cdot \kappa \cdot (1 - \eta_i) - \tau_i$, with constant τ and κ controlling the curve behavior.

12: Remove all edges with ages larger than μ_{max} and remove neurons without edges.

13: If the stop criterion is not met, repeat from step 4.

The motivation underlying hierarchical learning is to obtain progressively specialized neurons coding spatiotemporal dependencies of the input. This is consistent with neurophysiological evidence supporting increasingly large temporal receptive windows in the mammalian cortex (Hasson et al., 2008; Lerner et al., 2011; Taylor et al., 2015), where neurons in higher areas encode information accumulated over longer timescales. In our architecture, hierarchical learning is carried out by training a higher-level network with neural activation trajectories from a lower-level network. These trajectories are obtained by computing the best-matching neurons for the current input sequence with respect to the trained network with N neurons, so that a set of trajectories of length q is given by

$$\Omega^q(\mathbf{x}(t)) = \{\mathbf{w}_{b(\mathbf{x}(t))}, \mathbf{w}_{b(\mathbf{x}(t-1))}, \ldots, \mathbf{w}_{b(\mathbf{x}(t-q+1))}\}, \quad (3)$$

with $b(\mathbf{x}(t)) = \arg\min_{j \in N} \|\mathbf{x}(t) - \mathbf{w}_j\|$ computing the index of the neuron that minimizes the distance to the current input.

The overall hierarchical flow is illustrated in Fig. 2. The low-level networks G^{P-1} and G^{M-1} learn a set of time-independent primitives that will be used for higher-level representations and should exhibit robust activation regardless of temporal disruptions of the input. The networks G^{P-2} and G^{M-2} process activation trajectories of 3 neurons from the previous layer and the integration of the input is carried out in G^{STS} over activation trajectories of 3 neurons from layer 2. The network layer G^{STS} integrates pose-motion features by training the network with the concatenation of vectors $\Psi = \{\Omega^q(\mathbf{P}) \frown \Omega^q(\mathbf{M})\}$, where $\mathbf{P}$ and $\mathbf{M}$ are the activation trajectories from G^{P-2} and G^{M-2} respectively. Network layer G^{STSm} processes activation trajectories of 3 neurons from G^{STS}, thereby representing visual information over a temporal window of 7 frames. After the training is completed, neurons in G^{STSm} encode sequence-selective prototype action segments, following the assumption that the recognition of actions must be selective for temporal order (Giese & Poggio, 2003; Hasson et al., 2008).

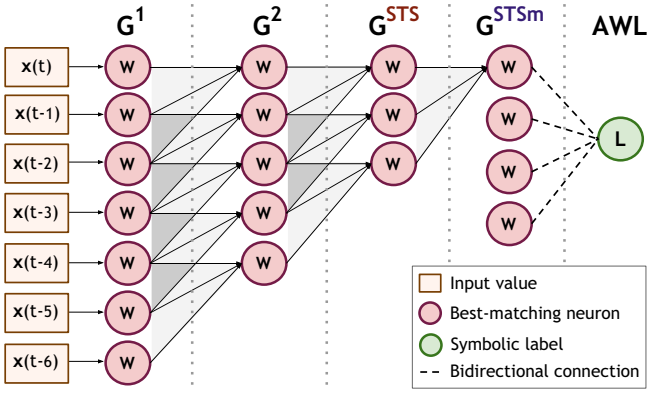

Fig. 2. Hierarchical learning of the last 7 inputs for processing neural activations with a sliding window scheme and asymmetric inter-layer connectivity between G^{STSm} and AWL used for bidirectional retrieval of audiovisual patterns. A neuron in G^{STSm} encodes action segments of 7 inputs. Action labels are predicted from 4 neurons in G^{STSm} (10 inputs), while for each action word in AWL, one onset neuron in G^{STSm} is computed.

2.2. GWR-based associative learning

For the G^{STSm} layer, we extended the standard GWR algorithm with: (1) semi-supervised label propagation functions so that prototype neurons can be attached to symbolic labels also in the absence of labeled samples and (2) enhanced synaptic connectivity in the temporal domain for learning activation patterns of consecutively activated neurons. The detailed learning algorithm for the proposed Online Semi-Supervised GWR (OSS-GWR) is illustrated by Algorithm 1.

2.2.1. Semi-supervised label propagation

For the semi-supervised propagation of labels, we attach labels to neurons by taking into account local connectivity and neural activation patterns. In this way, only labels attached to well-trained neurons are propagated to unlabeled neighbors (Algorithm 1, steps 7.c and 8.c). For this purpose, we defined two labeling functions: γ^{new} for when a new neuron is created, and γ^{update} for when a neuron is updated.

Provided that b is the index of the best-matching neuron of the training sample $\mathbf{x}(t)$ with label ξ and that we denote a missing sample label with the value -1, the label of a new neuron $\lambda(\mathbf{w}_r)$ is assigned according to

$$\gamma^{\text{new}}(b, \xi) = \begin{cases} \xi & \xi \neq -1 \\ \lambda(\mathbf{w}_b) & \text{otherwise} \end{cases} \quad (4)$$

For updating the label of an existing neuron, we also consider whether the current training sample is labeled. If this is not the case, then the best-matching neuron b will take the label of its closest neighbor s, provided that the two neurons have been sufficiently trained as expressed by their firing counters. Given the index of the second-best matching neuron s of $\mathbf{x}(t)$, the update labeling function for $\lambda(\mathbf{w}_b)$ is defined as

Please cite this article in press as: Parisi, G. I., et al. Emergence of multimodal action representations from neural network self-organization. *Cognitive Systems Research* (2016), http://dx.doi.org/10.1016/j.cogsys.2016.08.002

Table 1
Training parameters for the S-GWR and the OSS-GWR used for the classification task of the Iris dataset (results in Fig. 3).

Insertion threshold	$a_T = \{0.35, 0.75\}$
Firing threshold	$f_T = 0.1$
Learning rates	$\epsilon_b = 0.1$, $\epsilon_n = 0.01$
Firing counter	$\tau_b = 0.3$, $\tau_i = 0.1$, $\kappa = 1.05$
Training epochs	20
Labeling threshold (OSS-GWR only)	$\pi_T = 0.5$

$$\gamma^{\text{update}}(\xi, b, s) = \begin{cases} \xi & \xi \neq -1 \\ \lambda(\mathbf{w}_s) & (\xi = -1) \wedge (\pi_s^b \geqslant \pi_T) \\ \lambda(\mathbf{w}_b) & \text{otherwise} \end{cases} \quad (5)$$

$$\pi_j^i = \frac{E_j^i}{1 + \eta_i + \eta_j}, \quad (6)$$

with $E_j^i = 1$ if the neurons i and j are connected and 0 otherwise. Thus, this function yields greater values for interconnected, well-trained neurons, i.e. that have smaller firing counters. The value π_T is used as a threshold to modulate the propagation of a label from s to b.

We evaluated our semi-supervised labeling strategy with a classification task using the Iris benchmark dataset[1] containing 3 classes with 50 four-dimensional samples each. The goal of our experiment was to compare the classification performance of the proposed OSS-GWR with respect to the traditional GWR extended for classification (S-GWR, Parisi et al., 2015) using a decreasing percentage of available labeled samples in the training set. The average accuracy was estimated over 10 runs by removing labels at random positions for each percentage of available labels (from 0% to 100%).

The training algorithm used for this experiment is illustrated by Algorithm 1, excluding steps 3, 9, and 10 which are used in the G^{STSm} layer only, while the training parameters are listed in Table 1. Fig. 3 shows the average recognition accuracy for two different insertion thresholds $a_T = \{0.35, 0.75\}$ used to modulate the number of neurons created by the network, which has also an impact on the classification performance. In a smaller network, a prototype neuron will represent a greater number of samples. Thus, it is more likely that a neuron representing a dense cluster of samples with the same label will be assigned the correct one. It can be seen that the OSS-GWR outperforms S-GWR for the classification task as soon as not all labels are available. Larger deviations from the average accuracy can be observed due to the fact that for each run labels were removed from randomly selected samples and the distribution of missing labels can strongly influence the final outcome, particularly when few samples were labeled. Furthermore, the number of neurons created at each run varied, i.e. ≈ 16 for $a_T = 0.35$ and ≈ 100 for $a_T = 0.75$. This is due to the fact that the weight vectors

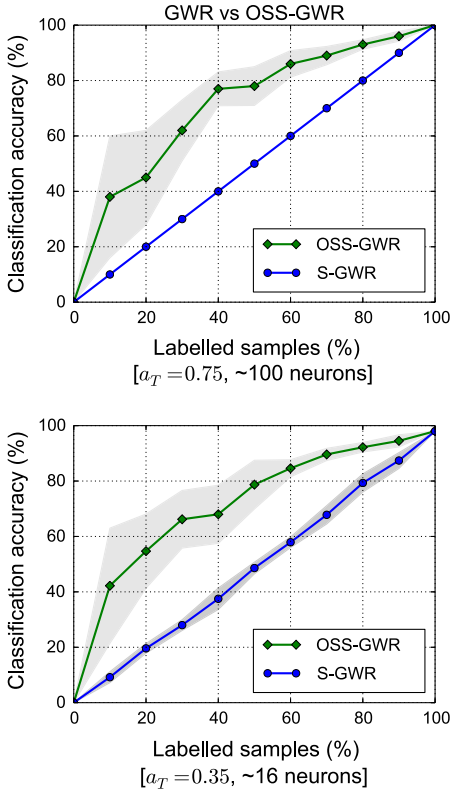

Fig. 3. Average classification accuracy with a decreasing percentage of available labels in the training set for trained S-GWR and OSS-GWR networks with ≈100 neurons (top) and ≈16 neurons (bottom).

for the two neurons initializing the networks were randomly selected from the training samples.

These results show that the proposed labeling strategy (Eqs. (4)–(6)) yields higher classification performance in the absence of sample labels. The overall approach is said to be online since the algorithm incrementally propagates labels during the training process (Beyer & Cimiano, 2011), in contrast to offline methods where labels are used after the unsupervised training of a network has finished.

2.2.2. Sequence-selective synaptic links

Next, we enhanced standard GWR connectivity by taking into account latent temporal relations of the input, so that connections between neurons that are consecutively activated can be created and incrementally updated. In other words, when the two neurons i and j are activated at time $t-1$ and t respectively, the synaptic link P_j^i between them is strengthened. At each iteration, the link $P_b^{b(t-1)}$ between the best-matching neuron b and the previous winner neuron $b(t-1)$ is increased by a value of 1, while the synaptic links between $b(t-1)$ and the other neurons are decreased if $P_n^{b(t-1)} > 0$ for $n \in A/\{b\}$ (Algorithm 1, step 9). This approach results in the efficient learning of the temporal structure of the input in terms of neural activation trajectories. The highest value of P_n^b will encode the most

[1] http://archive.ics.uci.edu/ml/datasets/Iris.

frequent transition, and thus allowing to estimate a prediction of $b(t+1)$ provided $b(t)$.

Sequence selectivity driven by asymmetric connections has been argued to be a feature of the cortex for neurons encoding optic flow patterns, where an active neuron preactives neurons encoding future patterns, while it inhibits neurons encoding other patterns (Mineiro & Zipser, 1998). This mechanism can be used for iteratively retrieving prototype neurons that encode an action sequence given the onset neuron for that action.

2.3. Action–word mappings

We now describe the asymmetric connectivity between the G^{STSm} layer and the Action Words layer (AWL) that allows the bidirectional retrieval of audiovisual patterns. We will show how it is possible to predict action words from processed visual patterns and, conversely, how to activate congruent visualizations of learned actions from recognized action words.

2.3.1. Action–to–word patterns

During the learning phase, unsupervised visual representations of actions in G^{STSm} are linked to symbolic action labels $\lambda \in L$, with L being the set of possible words. Action words in AWL will then have a one-to-many relation with neurons in G^{STSm}, while neurons can be linked to only one label in L. The development of connections between G^{STSm} and AWL depends on the co-activation of audiovisual inputs. More specifically, the connection between a neuron in G^{STSm} and a symbolic label in AWL will emerge if the neuron is activated within a time window in which also the label is activated by an auditory signal. In the case that no auditory stimulus occurs during the training of neurons in G^{STSm}, the sample label will be given the value -1 to indicate a missing label. Symbolic labels attached to neurons will be updated according to the semi-supervised label propagation rules (Eqs. (4) and (5)).

Given a previously unseen sequence of visual inputs, we want to predict the correct action word by comparing the novel input to prototype action sequences in G^{STS} and then return action labels attached to the best-matching neurons. The hierarchical flow of the visual input is composed of four networks, each of them processing activation trajectories of 3 neurons from the previous layer. Thus, each neuron in G^{STS} represents a prototype sequence encoding 7 consecutive frames (Fig. 2). By applying a temporal sliding window scheme, we get a new action label for each processed frame. To improve the robustness of the label prediction process, we return an action word from 4 neurons consecutively activated in G^{STS} (10 frames). Given a set of 4 labels obtained from the last 4 activated neurons from visual input, we output the statistical mode of the set, i.e. the most frequent label in the set is returned as the predicted action word. If we assume visual input at 10 frames per second, an action word will be predicted for 1 s of video.

2.3.2. Word–to–action patterns

For the development of connectivity patterns from AWL to G^{STSm}, we take into account the temporal order of consecutively activated neurons, yielding the learning of onset neurons in G^{STSm} to be linked with an action label, and from which it is possible to retrieve temporally-ordered prototype sequences for an action word. For a labeled neuron b in G^{STSm} activated at time t, its connection strength with the symbolic label λ becomes:

$$\Lambda^{\lambda}(b) = \frac{1}{2 \cdot \eta_b + c(\lambda, t)}, \tag{7}$$

with $c(\lambda, t)$ being a sequence counter that will increase by 1 when $\lambda(b) = \lambda(b-1)$ and reset to zero when this condition does not hold. Thus, this function expresses the relation between the firing counter η_b of the neuron b and its sequential order within the set of neural activations with the same label, yielding greater connection strengths for well-trained neurons that activate at the beginning of a sequence. The Λ^{λ} function for different neuron firing counters is depicted in Fig. 4 for a temporal window of 6 neural activations.

Word–to–action connectivity patterns are stored in the label–to–action reference V and updated at each training iteration so that $V_b^{\lambda(b)} = \Lambda^{\lambda(b)}(b)$ (Algorithm 1, step 10). The neuron in G^{STSm} with the maximum value of Λ^{λ} can then be selected as the onset neuron of an action label λ representing the first element of a prototype sequence.

We expect that word–to–action connections will develop according to the Λ^{λ} function (Eq. (7)) for each action label. Thus, when an action label $\lambda(j)$ is recognized from speech, the onset neuron in G^{STSm} of that action can be selected as the neuron that maximizes $\Lambda^{\lambda(j)}$, and consequently as the first element of a sequence used to generate prototype visual representations of actions. The index of the onset neuron $\mathbf{w}_v(t)$ for an action label $\lambda(j)$ is defined as:

$$v(t) = \arg\max_n V_n^{\lambda(j)}. \tag{8}$$

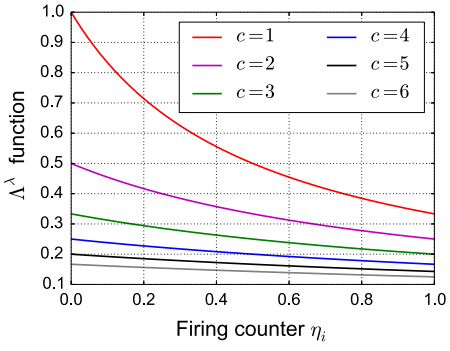

Fig. 4. Values of the Λ^{λ} function (Eq. (7)) for different firing counters η_i and sequence counters $c(\lambda_j, t)$ in the range 1–6 expressing the sequential order of processed samples. It can be seen how greater values are given to neurons activated at the beginning of the sequence, with an increasing response for well-trained neurons (smaller firing counter).

Please cite this article in press as: Parisi, G. I., et al. Emergence of multimodal action representations from neural network self-organization. *Cognitive Systems Research* (2016), http://dx.doi.org/10.1016/j.cogsys.2016.08.002

Next, we can retrieve the next neuron of a prototype action sequence by selecting the maximal temporal synaptic connectivity:

$$v(t+1) = \arg\max_n P_n^{v(t)}, \tag{9}$$

from which we can reconstruct a temporally-ordered sequence of arbitrary length by retrieving the weight vectors for a number of timesteps into the future. For instance, the sequence $(\mathbf{w}_{v_t}, \ldots, \mathbf{w}_{v_{t+3}})$ will generate visual output for a temporal window of 10 frames (1 s). This mechanism can be used in practice to visually assess how well the model has learned action dynamics and whether it has accounted for effectively binding action words to visual representations.

3. Experimental results

We present our experimental set-up and results on a dataset of 10 full-body actions that has been previously used to report recognition performance with manual segmentation for ground-truth labeling (Parisi, Weber, & Wermter, 2014, 2015). For the experiments reported in this paper, instead, action labels were recorded from speech so that action–word mappings of training samples can result from co-occurring audiovisual inputs using unsupervised learning and our strategy for label propagation. To evaluate our system, we compared newly obtained results with recently reported results using hierarchical GWR-based recognition (Parisi et al., 2015). We conducted additional experiments with different percentages of available labeled samples during the training, ranging between 100% (all samples are labeled) and 0%.

3.1. Audiovisual inputs

Our action dataset is composed of 10 full-body actions performed by 13 subjects (Parisi et al., 2014). Videos were captured in a home-like environment with a Kinect sensor installed 1.30 m above the ground. Depth maps were sampled with a VGA resolution of 640 × 480 and an operation range from 0.8 to 3.5 m at 30 frames per second. The dataset contains the following actions: *standing, walking, jogging, picking up, sitting, jumping, falling down, lying down, crawling*, and *standing up*. From the raw depth map sequences, 3D body joints were estimated on the basis of the tracking skeleton model and actions were represented by three body centroids (Fig. 5): C_1 for upper body with respect to the shoulders and the torso; C_2 for middle body with respect to the torso and the hips; and C_3 for lower body with respect to the hips and the knees, with each centroid computed as a point sequence of real-world coordinates $C = (x, y, z)$. To attenuate sensor noise, we used the median value of the last 3 estimated points (yielding action features at 10 frames per second). We then estimated upper and lower orientations θ_u and θ_l given by the slope angles of the line segments $\{C_1, C_2\}$ and $\{C_2, C_3\}$ respectively. As shown in Fig. 5, the values θ_u and θ_l describe the overall body pose according to the orientation of the torso and the legs, which allows to capture significant body features such as the characteristic posture of actions. We computed the body velocity S_i as the difference in pixels of the upper centroid C_1 between two consecutive frames. This centroid was chosen based on the motivation that the orientation of the torso is the most characteristic reference during the execution of a full-body action (Papadopoulos, Axenopoulos, & Daras, 2014).

For recording action labels, we used automatic speech recognition from Google's cloud-based ASR enhanced with domain-dependent post-processing (Twiefel, Baumann, Heinrich, & Wermter, 2014). The post-processor translates each sentence in a list of candidate sentences returned by the ASR service into a string of phonemes. To exploit the quality of the well-trained acoustic models employed by this service, the ASR hypothesis is converted to a phonemic representation employing a grapheme-to-phoneme converter. The word from a list of in-domain words is then selected as the most likely word candidate. An advantage of this approach are the hard constraints of the results, as each possible result can be mapped to an expected action word. Reported experiments showed that the sentence list approach obtained the best performance for in-domain recognition with respect to other approaches on the TIMIT speech corpus[2] with a sentence-error-rate of 0.521. The audio recordings were performed by speaking the name of the action in a time window of 2 s during its execution, i.e. for each repetition in the case of jumping, picking up, falling down, and standing up, and every 2 s for cyclic actions (standing, walking, jogging, sitting, lying down, crawling). This approach has the advantage of assigning labels to continuous video streams without the manual segmentation of visual features from specific frames.

3.2. Results and evaluation

For a consistent comparison with previous results, we adopted similar feature extraction and evaluation schemes. We divided the data equally into training and test set, i.e., 30 sequences of 10 s for each cyclic action (standing, walking, jogging, sitting, lying down, crawling) and 30 repetitions for each goal-oriented action (jumping, picking up, falling down, standing up). Both the training and the test sets contained data from all subjects.

For the learning in the G^{STSm} layer, we used the following training parameters: insertion threshold $a_T = 0.9$, learning rates $\epsilon_b = 0.3$, $\epsilon_n = 0.006$, firing counter parameters $\tau_b = 0.3$, $\tau_i = 0.1$, $\kappa = 1.05$, maximum age for edges $\mu_{max} = 500$, labeling threshold $\pi_T = 0.5$ (OSS-GWR only). These parameters were empirically found with respect to

[2] TIMIT Acoustic–Phonetic Continuous Speech Corpus: https://catalog.ldc.upenn.edu/LDC93S1.

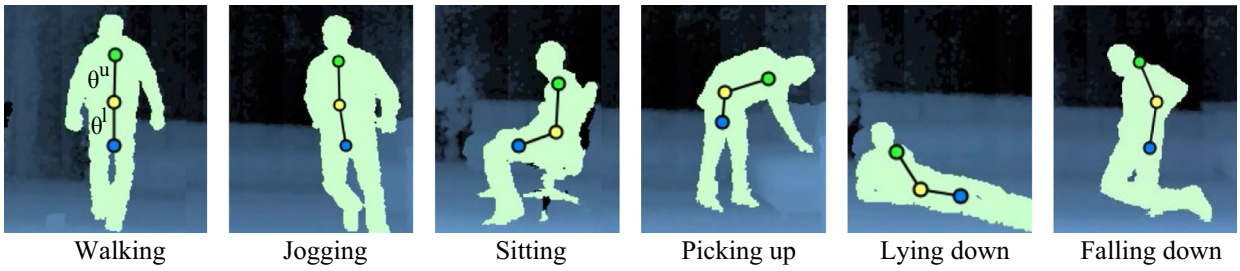

Fig. 5. Representation of full-body movements from our action dataset (Parisi et al., 2014). We estimate three centroids: C_1 (green), C_2 (yellow) and C_3 (blue) for upper, middle and lower body respectively. The segment slopes θ^u and θ^l describe the posture in terms of the overall orientation of the upper and lower body. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

best accuracy in terms of classification performance. Similar to Parisi et al. (2015), each network was trained for 500 epochs over the entire training set. Once a layer was trained, its weights were set fixed and the next higher-level layer was trained. If we consider the 4 network layers of the architecture, it took overall 2000 epochs to obtain a trained neuron in the G^{STSm} network. Layers G^{STSm} and AWL were trained together according to Algorithm 1.

Experimental results showed an average classification accuracy of 93.3%, comparing with the state-of-the-art results of 94% reported by Parisi et al. (2015) that required the manual segmentation of training samples for assigning ground-truth labels. The confusion matrices for the novel OSS-GWR and the S-GWR approaches tested on a set of 10 actions are shown in Figs. 6 and 7 respectively (with the rows of the matrix being the instances of actual actions and columns being the instances of predicted actions). The matrices show that there is a significant similarity on which samples were misclassified, suggesting that misclassification depends more on the visual features than on issues related to the associative learning mechanism via the co-occurrence of audiovisual inputs. For example, actions that are similar with respect to body posture (e.g. walking and jogging, falling down and lying down), tend to be mutually misclassified. The reason for this is that although the

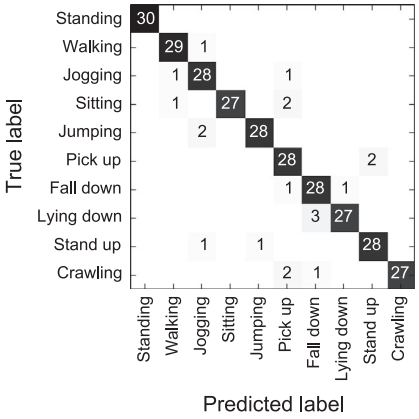

Fig. 6. Confusion matrix for the novel OSS-GWR approach tested on a dataset of actions with an average accuracy of 93.3% (no manual segmentation).

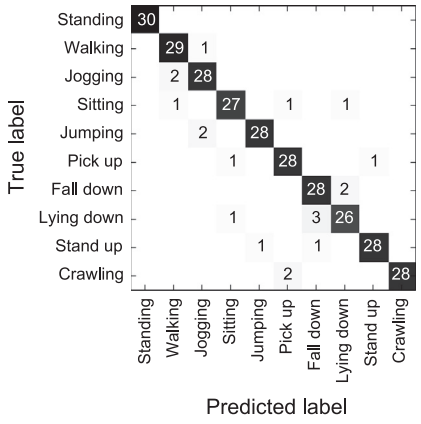

Fig. 7. Confusion matrix for the S-GWR approach (Parisi et al., 2015) tested on a dataset of actions with an average accuracy of 94% (samples manually segmented).

defined features used to learn relevant properties of actions should be sufficient to univocally describe spatiotemporal patterns over different timescales, tracking inaccuracies from the depth sensor may have a negative impact on the extraction of reliable pose-motion cues. While it is possible to embed the detection of sensor noise in low-level networks (Parisi et al., 2015), it is non-trivial to detect inaccurate samples that belong to the feature space, e.g. caused by the (self-)occlusion of body joints. In this case, tracking errors will propagate from low to higher-level layers and lead to the misclassification of samples.

An additional experiment consisted of decreasing the percentage of labeled action samples. Since visual representations are progressively learned without supervision, we expect that the absence of training action labels will not have a catastrophic impact on the correct development of associative connections of audiovisual input (as would be expected for a strictly supervised method). For this purpose, we trained our system with a similar scheme as in the first experiment, but this time we omitted action words from ASR of randomly chosen samples and varied the percentage of available labels from 100% to 0%. Here, with *sample* we do not refer to a single data point (as in the experiment from Section 2.2), but rather to a set of data points represented by the amount of frames for the duration of the audio time window, i.e. 20 frames. The average

Please cite this article in press as: Parisi, G. I., et al. Emergence of multimodal action representations from neural network self-organization. *Cognitive Systems Research* (2016), http://dx.doi.org/10.1016/j.cogsys.2016.08.002

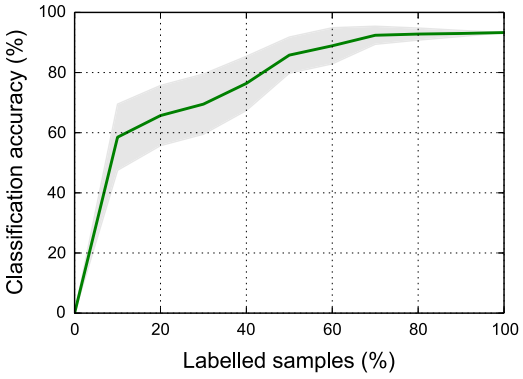

Fig. 8. Average classification accuracy over 10 runs for a decreasing percentage of available action labels on randomly selected training samples.

classification accuracy with different percentages of omitted audio samples for randomly selected samples over 10 runs is displayed in Fig. 8. We can observe that although a decreasing number of available labeled samples during the training phase has a negative impact on the classification performance, this decline is not proportional to the number of omitted action words. As soon as 10% of labeled samples are available during the training, the system shows an accuracy of 58.5%, and accuracy values above 85% can be observed for 50% or more available labeled samples. On the other hand, we found that the timing at which these action words are presented to the AWL layer over the training epochs does have a significant impact on the performance. In fact, best results were obtained if action words are presented when visual representations have reached a certain degree of stability, while associative connections created at early stages of visual development may not be as reliable.

To have an idea of how well the associative layer has learned action dynamics, we generated learned action representations from action words in the absence of visual input. The visualizations were generated from the recognized action words by computing onset neurons in G^{STSm} via the associative connections from AWL (Eq. (8)). For each onset neuron, one-step prediction was made using the temporal connectivity (Eq. (9)) to compute snapshots of 10 frames (1 s of action). The visual representations of the actions *sitting*, *jumping*, and *picking up* for a time window of 1 s are shown in Fig. 9, where we displayed the three body centroids and the motion intensity of the upper-body centroid (black arrow). From these visualizations, we can argue that the associative layer successfully learns temporally-ordered representations of visual input sequences from onset neurons, and therefore that our model accounts for the bidirectional retrieval of audiovisual inputs.

4. Discussion

4.1. Summary

We presented a hierarchical neural architecture for learning multimodal action representations from a set of training audiovisual inputs. In particular, we investigated how associative links between unimodal representations can emerge in a self-organizing manner from the co-occurrence of multimodal stimuli. Visual generalizations of action sequences were learned using hierarchically-arranged GWR networks for the processing of inputs with increasingly larger temporal windows. Multimodal action representations in terms of action–word mappings were obtained by incrementally developing bidirectional connections between learned visual representations and action labels from automatic speech recognition. For this purpose, we proposed an associative network with asymmetric inter-layer connectivity that takes into account

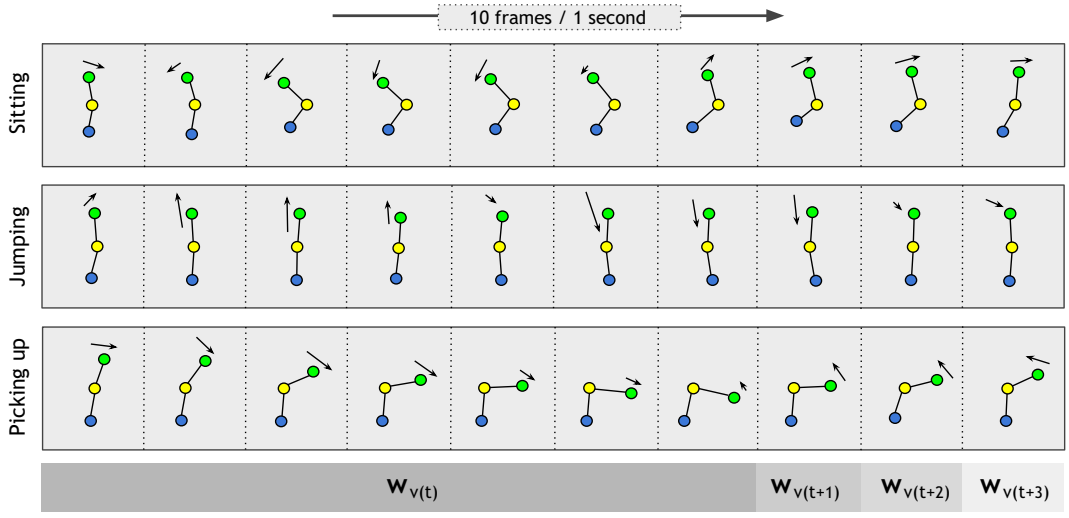

Fig. 9. Example of learned visual representations generated from bidirectional connectivity between G^{STSm} and AWL for the action words *sitting*, *jumping*, and *picking up* obtained from speech recognition. The figure shows the three body centroids and the motion intensity of the upper-body centroid (black arrow) for a time window of 10 frames (1 s) starting from the action onset neuron.

Please cite this article in press as: Parisi, G. I., et al. Emergence of multimodal action representations from neural network self-organization. *Cognitive Systems Research* (2016), http://dx.doi.org/10.1016/j.cogsys.2016.08.002

the spatiotemporal dynamics of action samples and binds co-occurring audiovisual inputs. For this associative layer, we implemented an extended GWR learning algorithm (the OSS-GWR) that accounts for the propagation of action labels in a semi-supervised training scenario and that learns neural activation patterns in the temporal domain through enhanced synaptic connectivity. We conducted experiments with a dataset of 10 full-body actions showing that our system achieves state-of-the-art classification performance without requiring the manual segmentation of training samples. Together, these results show that our neural architecture accounts for the bidirectional retrieval of audiovisual inputs, also in the scenario where a number of action labels is omitted during the training phase.

Interestingly, our implementation of bidirectional action–word connections roughly resembles a phenomenon found in the human brain, i.e. spoken action words elicit receptive fields in the visual area (Barraclough et al., 2005; Miller & Saygin, 2013). In other words, learned visual representations of actions can be activated in the absence of visual inputs, in this case from recognized speech. These visualizations can be generated by computing the onset neuron in the G^{STSm} layer via the developed associative connections to AWL, so that temporally-ordered action snapshots can be obtained from neural activation patterns learned by synaptic connectivity in the temporal domain. We have shown that this property can be used in practice to assess how well the model accounts for learning congruent visual representations of actions from pose-motion features.

4.2. Neurobiologically-motivated multimodal integration

A vast corpus of studies has shown the ability of the brain to integrate multimodal information for providing a coherent perceptual experience (Ernst & Bulthoff, 2004; Stein & Meredith, 1993; Stein et al., 2009). In particular for the integration of audiovisual stimuli, neurophysiological studies have evidenced strong links between the areas in the brain governing visual and language processing for the formation of multimodal perceptual representations (Belin et al., 2000; Belin et al., 2002; Foxe et al., 2000; Pulvermüller, 2005; Raij et al., 2000). However, the question of how to develop artificial models that efficiently process and bind multimodal information has remained an issue to be investigated (Ursino et al., 2014).

The development of associations between co-occurring stimuli for multimodal binding has been strongly supported by neurophysiological evidence (Fiebelkorn et al., 2009; Ursino et al., 2014). Similar to Vavrečka and Farkaš (2014) and Morse et al. (2015), we argue that the co-occurrence of sensory inputs is a sufficient source of information to create robust multimodal representations with the use of associative links between unimodal representations that can be incrementally learned in an unsupervised fashion. However, in contrast to previous models focused on the development of object–word mappings, we focus on the development of associative links between action labels and visual actions, which have high spatial and temporal variance, thereby requiring a processing architecture that accounts for the generalization of inputs at different spatiotemporal scales.

From a neurobiological perspective, neurons selective to actions in terms of complex biological motion have been found in a wide number of brain structures (Giese & Rizzolatti, 2015). For example, in the STS, which is thought to be an associative learning device for linking different unimodal perceptual representations, and consequently crucial for social cognition (Allison et al., 2000; Adolphs, 2003; Beauchamp, 2005; Beauchamp et al., 2008). It has been shown that different regions in the STS are activated by naturally occurring, highly correlated action features, such as pose, motion, the characteristic sound of an action (Barraclough et al., 2005; Beauchamp et al., 2004) and linguistic stimuli (Belin et al., 2002; Stevenson & James, 2009; Wright et al., 2003).

In this paper, we propose a simplified computational model that learns to integrate audiovisual patterns of action sequences. Our model incrementally learns a set of associative connections in a self-organized manner to bind unimodal representations from co-occurring multisensory inputs. Therefore, neurons in the G^{STSm} layer are tuned to multimodal action snapshots in terms of action–word mappings. The focus of our study was the self-organizing development of associative connections between visual and auditory action representations. For audiovisual stimulation, neurons in the posterior STS showed greater response to multimodal stimuli than to unimodal ones, with these multimodal responses being greater than the sum of the single unimodal responses. The modeling of neurobiologically observed principles underlying audiovisual integration in the STS for speech and non-speech stimuli, such as superadditivity (Calvert, Campbell, & Brammer, 2000), spatial and temporal congruence (Bushara, Grafman, & Hallett, 2001; Macaluso, George, Dolan, Spence, & Driver, 2004), and inverse effectiveness (Stevenson & James, 2009), was out of the scope of this paper and will be subject of future research.

Based on the principle of learning associative connections from co-occurring inputs, it is possible to extend the development of associative patterns beyond the audiovisual domain. For instance, several neurophysiological studies have evidenced strong interaction between the visual and motor representations, more specifically including the STS, parietal cortex, and premotor cortex (see Giese & Rizzolatti (2015) for a recent survey), with higher activation of neurons in the motor system for biomechanically-plausible, perceived motion sequences (Miller & Saygin, 2013). From the perspective of our model, we could think of emerging associative connections between auditory, visual, and motor representations in terms of the self-organizing binding of temporally correlated activations. However, while our architecture scales up to a larger number of modalities, it does not account

Please cite this article in press as: Parisi, G. I., et al. Emergence of multimodal action representations from neural network self-organization. *Cognitive Systems Research* (2016), http://dx.doi.org/10.1016/j.cogsys.2016.08.002

for crossmodal learning aspects, e.g. in an embodied robot perception scenario where motor contingencies influence audiovisual mappings (Morse et al., 2015). Consequently, the extension of our model in such a direction would require additional mechanisms for the crossmodal learning of spatiotemporal contingencies built on the basis of modality-specific properties.

4.3. Growing self-organization and hierarchical learning

Motivated by the process of input-driven self-organization exhibited by topographic maps in the cortex (Miikkulainen, Bednar, Choe, & Sirosh, 2005; Nelson, 2000; Willshaw & von der Malsburg, 1976), we proposed a learning model encompassing a hierarchy of Growing When Required (GWR) networks (Marsland et al., 2002). GWR networks have the ability to dynamically change their topological structure through competitive Hebbian learning (Martinetz, 1993) and incrementally match the distribution of the data in input space. Different from other incremental models of self-organization that create new neurons at a fixed growth rate (e.g. Fritzke, 1995, 1997), GWR learning creates new neurons whenever the activity of well-trained neurons is smaller than a given threshold. This mechanism creates a larger number of neurons at early stages of the training and then tune the weights through subsequent training epochs. While the process of neural growth of the GWR algorithm does not resemble biologically plausible mechanisms of neurogenesis (e.g., Eriksson et al., 1998; Gould, 2007; Ming & Song, 2011), it is an efficient learning model exhibiting a computationally convenient trade-off between adaptation to dynamic input and learning convergence. For instance, it has been shown that GWR learning is particularly suitable for novelty detection and cumulative learning in robot scenarios (Marsland, Nehmzow, & Shapiro, 2005).

The two parameters modulating the growth rate of the network are the activation threshold and the firing counter threshold. The activation threshold a_T establishes the maximum discrepancy (distance) between the input and its best-matching neuron in the network. For larger values of a_T, the discrepancy expressed by Eq. (1) will be smaller. The firing counter threshold f_T is used to favor the training of recently created neurons before creating new ones. Intuitively, the average discrepancy between the input and the network representation should decrease for a larger number of neurons. On the other hand, there is not such a straightforward relation between the number of neurons and the classification performance. This is because the classification process consists of predicting the label of novel samples by retrieving attached labels to the inputs' best-matching neurons, irrespective of the actual distance between the novel inputs and the selected neurons. Therefore, a convenient value for a_T should be chosen by taking into account the distribution of the input and, in the case of a classification task, the classification performance. For instance, in a scenario with a number of missing labels during the training phase as described in Section 2.2, a better classification performance may be obtained with a smaller number of neurons (Fig. 3).

Our GWR-based hierarchical learning architecture allows to obtain progressively specialized neurons encoding latent spatiotemporal dynamics of the input (Parisi et al., 2015). A hierarchical structure has also the advantage of increased computational efficiency by sharing functionalities of lower levels to obtain representations in higher levels. We implemented hierarchical learning by training a higher-level network with neuron activation trajectories from lower-level representations. After the training is completed, neurons in higher-level layers will encode prototype sequence-selective snapshots of visual input, following the assumption that the recognition of actions must be selective for temporal order (Giese & Poggio, 2003; Hasson et al., 2008). This hierarchical organization is consistent with neurophysiological evidence for increasingly large spatiotemporal receptive windows in the human cortex (Hasson et al., 2008; Lerner et al., 2011; Taylor et al., 2015), where simple features manifest in low-level layers closest to sensory inputs, while increasingly complex representations emerge in deeper layers. Specifically for the visual cortex, Hasson et al. (2008) showed that while early visual areas such as the primary visual cortex (V1) and the motion-sensitive area (MT+) yield higher responses to instantaneous sensory input, high-level areas such as the STS were more affected by information accumulated over longer timescales (~12 s). This kind of hierarchical aggregation is a fundamental organizational principle of cortical networks for dealing with perceptual and cognitive processes that unfold over time (Fonlupt, 2003).

4.4. Future work

Our results encourage the leverage of the proposed architecture in several directions. For instance, so far we have assumed that the training labels provided from speech are correct. On the other hand, several developmental studies have shown that human infants are able to learn action–word mappings also in the presence of missing, ambiguous or sometimes contradictory referents using cross-situational statistics (Smith & Yu, 2008). Thus, it would be interesting to evaluate the robustness of the system if the available labels are sometimes inaccurate or in contradiction with previously learned labels. Furthermore, another limitation of our model is the use of domain-dependent ASR. Although this approach yields the reliable recognition of a set of action words (Twiefel et al., 2014), it has the disadvantage that a specific set of words has to be defined a priori. Therefore, new action words cannot be learned during the training process. We plan to address this constraint by accounting for learning new lexical features so that the action vocabulary can be dynamically extended during training sessions. It has been shown that lexical features can be learned using recursive self-organizing architectures (Strickert & Hammer, 2005), obtaining action

Please cite this article in press as: Parisi, G. I., et al. Emergence of multimodal action representations from neural network self-organization. *Cognitive Systems Research* (2016), http://dx.doi.org/10.1016/j.cogsys.2016.08.002

word representations from a phonemic representation of recognized audio. This extension would comprise a hierarchical stream for processing audio features and, similar to the visual hierarchy, higher-lever representations of speech (words) would be learned from lower-level representations (e.g., phonemes). Such a processing scheme would be in line with neurophysiological evidence supporting the hierarchical processing of aural features in the auditory cortex with increasing temporal receptive windows (Lerner et al., 2011). By considering the aforementioned extensions, the mechanism responsible for developing associative connections should be robust to situations in which action words recognized from speech may not be reliable. Therefore, an additional labeling scheme should be considered that takes into account cross-statistical properties of labels to guarantee a congruent audiovisual mapping.

Finally, our results motivate the extension of our approach for scenarios that require more complex audiovisual inputs, for instance by considering the recognition of transitive actions. This challenging task would require accounting for the learning of action-object relations to be described by more flexible action words, e.g. labeling both the action and the object being used. An interesting question would then be how multiple different modules develop bidirectional connections in order to provide a congruent perceptual experience.

The manual labeling of training sequences is expensive and hinders the automatic, continuous learning of novel information. Thus, research work in the direction of neurocognitive architectures aimed to develop robust multimodal representations from more natural interactions would provide a significant benefit for learning agents in order to trigger proper action-driven behavior in complex environments.

Acknowledgment

This research was partially supported by the DAAD German Academic Exchange Service (Kz:A/13/94748) and the DFG (Deutsche Forschungsgemeinschaft) for the project Cross-modal Learning TRR-169/A5.

References

Adolphs, R. (2003). Cognitive neuroscience of human social behaviour. *Nature Reviews Neuroscience, 4*, 165–178. http://dx.doi.org/10.1038/nrn1056.

Allison, T., Puce, A., & McCarthy, G. (2000). Social perception from visual cues: Role of the STS region. *Trends in Cognitive Sciences, 4*, 267–278. http://dx.doi.org/10.1016/S1364-6613(00)01501-1.

Barraclough, N. E., Xiao, D., Baker, C. I., Oram, M. W., & Perrett, D. I. (2005). Integration of visual and auditory information by superior temporal sulcus neurons responsive to the sight of actions. *Journal of Cognitive Neuroscience, 17*, 377–391. http://dx.doi.org/10.1162/0898929053279586.

Bauer, J., Magg, S., & Wermter, S. (2015). Attention modeled as information in learning multisensory integration. *Neural Networks, 65*, 44–52. http://dx.doi.org/10.1016/j.neunet.2015.01.004.

Beauchamp, M. S. (2005). See me, hear me, touch me: Multisensory integration in lateral occipital-temporal cortex. *Current Opinion in Neurobiology, 15*, 145–153. http://dx.doi.org/10.1016/j.conb.2005.03.011, Cognitive Neuroscience.

Beauchamp, M. S., Lee, K. E., Argall, B. D., & Martin, A. (2004). Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron, 41*, 809–823. http://dx.doi.org/10.1016/S0896-6273(04)00070-4.

Beauchamp, M. S., Yasar, N. E., Frye, R. E., & Ro, T. (2008). Touch, sound and vision in human superior temporal sulcus. *NeuroImage, 41*, 1011–1020. http://dx.doi.org/10.1016/j.neuroimage.2008.03.015.

Belin, P., Zatorre, R. J., & Ahad, P. (2002). Human temporal-lobe response to vocal sounds. *Cognitive Brain Research, 13*, 17–26. http://dx.doi.org/10.1016/S0926-6410(01)00084-2.

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature, 403*, 309–312.

Beyer, O., & Cimiano, P. (2011). Online labelling strategies for growing neural gas. In H. Yin, W. Wang, & V. Rayward-Smith (Eds.), *Proceedings of the international conference on intelligent data engineering and automated learning (IDEAL)* (pp. 76–83). Berlin, Heidelberg: Springer. http://dx.doi.org/10.1007/978-3-642-23878-9_10.

Bloom, L. (1993). *The transition from infancy to language: Acquiring the power of expression*. New York: Cambridge University Press.

Bushara, K., Grafman, J., & Hallett, M. (2001). Neural correlates of auditory-visual stimulus onset asynchrony detection. *Journal of Neuroscience, 21*, 300–304.

Calvert, G. A. (2001). Crossmodal processing in the human brain: Insights from functional neuroimaging studies. *Cerebral Cortex, 11*, 1110–1123. http://dx.doi.org/10.1093/cercor/11.12.1110.

Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology, 10*, 649–657. http://dx.doi.org/10.1016/S0960-9822(00)00513-3.

Eriksson, P. S., Perfilieva, E., Bjork-Eriksson, T., Alborn, A. M., Nordborg, C., Peterson, D. A., & Gage, F. H. (1998). Neurogenesis in the adult human hippocampus. *Nature Medicine, 4*, 1313–1317. http://dx.doi.org/10.1038/3305.

Ernst, M. O., & Bulthoff, H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences, 8*, 162–169. http://dx.doi.org/10.1016/j.tics.2004.02.002.

Fiebelkorn, I. C., Foxe, J. J., & Molholm, S. (2009). Dual mechanisms for the cross-sensory spread of attention: How much do learned associations matter? *Cerebral Cortex, 20*, 109–120. http://dx.doi.org/10.1093/cercor/bhp083.

Fonlupt, P. (2003). Perception and judgement of physical causality involve different brain structures. *Cognitive Brain Research, 17*, 248–254. http://dx.doi.org/10.1016/S0926-6410(03)00112-5.

Foxe, J. J., Morocz, I. A., Murray, M. M., Higgins, B. A., Javitt, D. C., & Schroeder, C. E. (2000). Multisensory auditory–somatosensory interactions in early cortical processing revealed by high-density electrical mapping. *Cognitive Brain Research, 10*, 77–83. http://dx.doi.org/10.1016/S0926-6410(00)00024-0.

Fritzke, B. (1995). A growing neural gas network learns topologies. *Advances in neural information processing systems* (Vol. 7, pp. 625–632). MIT Press.

Fritzke, B. (1997). A self-organizing network that can follow non-stationary distributions. In *ICANN'97: International conference on artificial neural networks* (pp. 613–618). Springer.

Garcia, J. O., & Grossman, E. D. (2008). Necessary but not sufficient: Motion perception is required for perceiving biological motion. *Vision Research, 48*, 1144–1149. http://dx.doi.org/10.1016/j.visres.2008.01.027.

Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. *Language Development: Language, Thought, and Culture, 2*, 301–334.

Giese, M. A., & Poggio, T. (2003). Neural mechanisms for the recognition of biological movements. *Nature Reviews Neuroscience, 4*, 179–192. http://dx.doi.org/10.1038/nrn1057.

Giese, M. A., & Rizzolatti, G. (2015). Neural and computational mechanisms of action processing: Interaction between visual and motor representations. *Neuron, 88*, 167–180. http://dx.doi.org/10.1016/675 j.neuron.2015.09.040.

Gould, E. (2007). How widespread is adult neurogenesis in mammals? *Nature Reviews Neuroscience, 8*, 481–488. http://dx.doi.org/10.1038/nrn2147.

Hasson, U., Yang, E., Vallines, I., Heeger, D. J., & Rubin, N. (2008). A hierarchy of temporal receptive windows in human cortex. *The Journal of Neuroscience, 28*, 2539–2550. http://dx.doi.org/10.1523/jneurosci.5487-07.2008.

Hirsch-Pasek, K., Golinkoff, R., & Hollich, G. (2000). *An emergentist coalition model for word learning: Mapping words to objects is a product of the interaction of multiple cues*. Oxford University Press (pp. 136–165). Oxford University Press.

Jung, M., Hwang, J., & Tani, J. (2015). Self-organization of spatio-temporal hierarchy via learning of dynamic visual image patterns on action sequences. *PLoS One, 10*, e0131214. http://dx.doi.org/10.1371/journal.pone.0131214.

Kachouie, R., Sedighadeli, S., Khosla, R., & Chu, M. (2014). Socially assistive robots in elderly care: A mixed-method systematic literature review. *International Journal of Human-Computer Interaction, 30*, 369–393. http://dx.doi.org/10.1080/10447318.2013.873278.

Kohonen, T. (1988). Self-organized formation of topologically correct feature maps. *Biological Cybernetics, 43*, 59–69.

Layher, G., Giese, M. A., & Neumann, H. (2014). Learning representations of animated motion sequences? A neural model. *Topics in Cognitive Science, 6*, 170–182. http://dx.doi.org/10.1111/tops.12075.

Lerner, Y., Honey, C. J., Silbert, L. J., & Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *The Journal of Neuroscience, 31*, 2906–2915. http://dx.doi.org/10.1523/jneurosci.3684-10.2011.

Macaluso, E., George, N., Dolan, R., Spence, C., & Driver, J. (2004). Spatial and temporal factors during processing of audiovisual speech: A {PET} study. *NeuroImage, 21*, 725–732. http://dx.doi.org/10.1016/j.neuroimage.2003.09.049.

Marsland, S., Nehmzow, U., & Shapiro, J. (2005). On-line novelty detection for autonomous mobile robots. *Robotics and Autonomous Systems, 51*, 191–206.

Marsland, S., Shapiro, J., & Nehmzow, U. (2002). A self-organising network that grows when required. *Neural Networks, 15*, 1041–1058.

Martinetz, T. M. (1993). Competitive Hebbian learning rule forms perfectly topology preserving maps. In *ICANN'93: International conference on artificial neural networks* (pp. 427–434). Amsterdam: Springer.

Miikkulainen, R., Bednar, J. A., Choe, Y., & Sirosh, J. (2005). *Computational maps in the visual cortex*. Springer. http://dx.doi.org/10.1007/0-387-28806-6.

Miller, L. E., & Saygin, A. P. (2013). Individual differences in the perception of biological motion: Links to social cognition and motor imagery. *Cognition, 128*, 140–148. http://dx.doi.org/10.1016/j.cognition.2013.03.013.

Mineiro, P., & Zipser, D. (1998). Analysis of direction selectivity arising from recurrent cortical interactions. *Neural Computation, 10*, 353–371. http://dx.doi.org/10.1162/089976698300017791.

Ming, G. l., & Song, H. (2011). Adult neurogenesis in the mammalian brain: Significant answers and significant questions. *Neuron, 70*, 687–702. http://dx.doi.org/10.1016/j.neuron.2011.05.001, http://dx.doi.org/10.1038/nrn2147.

Morse, A. F., Benitez, V. L., Belpaeme, T., Cangelosi, A., & Smith, L. B. (2015). Posture affects how robots and infants map words to objects. *PLoS One, 10*, e0116012. http://dx.doi.org/10.1371/journal.pone.0116012.

Nelson, C. A. (2000). Neural plasticity and human development: The role of early experience in sculpting memory systems. *Developmental Science, 3*, 115–136.

Noda, K., Arie, H., Suga, Y., & Ogata, T. (2014). Multimodal integration learning of robot behavior using deep neural networks. *Robotics and Autonomous Systems, 62*, 721–736. http://dx.doi.org/10.1016/j.robot.2014.03.003.

Papadopoulos, G. T., Axenopoulos, A., & Daras, P. (2014). Real-time skeleton-tracking-based human action recognition using kinect data. In *MultiMedia modeling* (pp. 473–483). Springer. http://dx.doi.org/10.1007/978-3-319-04114-8_40.

Parisi, G. I., Weber, C., & Wermter, S. (2014). Human action recognition with hierarchical growing neural gas learning. In *Artificial neural networks and machine learning (ICANN)* (pp. 89–96). http://dx.doi.org/10.1007/978-3-319-11179-7_12.

Parisi, G. I., Weber, C., & Wermter, S. (2015). Self-organizing neural integration of pose-motion features for human action recognition. *Frontiers in Neurorobotics, 9*. http://dx.doi.org/10.3389/fnbot.7352015.00003.

Pulvermüller, F. (2005). Brain mechanisms linking language and action. *Nature Reviews Neuroscience, 6*, 576–582. http://dx.doi.org/10.1038/nrn1706.

Raij, T., Uutela, K., & Hari, R. (2000). Audiovisual integration of letters in the human brain. *Neuron, 28*, 617–625. http://dx.doi.org/10.1016/S0896-6273(00)00138-0.

Senkowski, D., Saint-Amour, D., Hfle, M., & Foxe, J. J. (2011). Multisensory interactions in early evoked brain activity follow the principle of inverse effectiveness. *NeuroImage, 56*, 2200–2208. http://dx.doi.org/10.1016/j.neuroimage.2011.03.075.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition, 106*, 1558–1568. http://dx.doi.org/10.1016/j.cognition.2007.06.010.

Stanley, J. C. (1976). Computer simulation of a model of habituation. *Nature, 261*, 146–148. http://dx.doi.org/10.1038/261146a0.

Stein, B. E., & Meredith, M. A. (1993). *The merging of the senses*. Cambridge, MA, US: The MIT Press.

Stein, B. E., Stanford, T. R., & Rowland, B. A. (2009). The neural basis of multisensory integration in the midbrain: Its organization and maturation. *Hearing Research, 258*, 4–15. http://dx.doi.org/10.1016/j.heares.2009.03.012.

Stevenson, R. A., & James, T. W. (2009). Audiovisual integration in human superior temporal sulcus: Inverse effectiveness and the neural processing of speech and object recognition. *NeuroImage, 44*, 1210–1223. http://dx.doi.org/10.1016/j.neuroimage.2008.09.034.

Strickert, M., & Hammer, B. (2005). Merge SOM for temporal data. *Neurocomputing, 64*. http://dx.doi.org/10.1016/j.neucom.2004.11.014.

Taylor, P., Hobbs, J. N., Burroni, J., & Siegelmann, H. T. (2015). The global landscape of cognition: Hierarchical aggregation as an organizational principle of human cortical networks and functions. *Scientific Reports, 5*. http://dx.doi.org/10.1038/srep18112.

Thirkettle, M., Benton, C. P., & Scott-Samuel, N. E. (2009). Contributions of form, motion and task to biological motion perception. *Journal of Vision, 9*, 28. http://dx.doi.org/10.1167/9.3.28.

Twiefel, J., Baumann, T., Heinrich, S., & Wermter, S. (2014). Improving domain-independent cloud-based speech recognition with domain-dependent phonetic post-processing. In *AAAI conference on artificial intelligence* (pp. 1529–1536).

Ursino, M., Cuppini, C., & Magosso, E. (2014). Neurocomputational approaches to modelling multisensory integration in the brain: A review. *Neural Networks, 60*, 141–165. http://dx.doi.org/10.1016/j.neunet.2014.08.003.

Vangeneugden, J., Pollick, F., & Vogels, R. (2009). Neural and computational mechanisms of action processing: Interaction between visual and motor representations. *Cerebral Cortex, 19*, 593–611. http://dx.doi.org/10.1093/cercor/bhn109.

Vavrečka, M., & Farkaš, I. (2014). A multimodal connectionist architecture for unsupervised grounding of spatial language. *Cognitive Computation, 6*, 101–112. http://dx.doi.org/10.1007/s12559-013-9212-5.

Willshaw, D. J., & von der Malsburg, C. (1976). How patterned neural connections can be set up by self-organization. *Proceedings of the Royal Society of London B: Biological Sciences, 194*, 431–445.

Wright, T. M., Pelphrey, K. A., Allison, T., Mckeown, M. J., & Mccarthy, G. (2003). Polysensory interactions along lateral temporal regions evoked by audiovisual speech. *Cerebral Cortex, 13*, 1034–1043. http://dx.doi.org/10.1093/cercor/13.10.1034.

Please cite this article in press as: Parisi, G. I., et al. Emergence of multimodal action representations from neural network self-organization. *Cognitive Systems Research* (2016), http://dx.doi.org/10.1016/j.cogsys.2016.08.002