In: IEEE-RAS International Conference on Humanoid Robots (Humanoids '15), Workshop on Towards Intelligent Social Robots: Current Advances in Cognitive Robotics, Seoul, South Korea (2015)

Towards Emerging Multimodal Cognitive Representations from Neural Self-Organization

German I. Parisi, Cornelius Weber and Stefan Wermter
Knowledge Technology Institute, Department of Informatics
University of Hamburg, Germany
{parisi,weber,wermter}@informatik.uni-hamburg.de
http://www.informatik.uni-hamburg.de/WTM/

***Abstract*—The integration of multisensory information plays a crucial role in autonomous robotics. In this work, we investigate how robust multimodal representations can naturally develop in a self-organized manner from co-occurring multisensory inputs. We propose a hierarchical learning architecture with growing self-organizing neural networks for learning human actions from audiovisual inputs. Associative links between unimodal representations are incrementally learned by a semi-supervised algorithm with bidirectional connectivity that takes into account inherent spatiotemporal dynamics of the input. Experiments on a dataset of 10 full-body actions show that our architecture is able to learn action-word mappings without the need of segmenting training samples for ground-truth labelling. Instead, multimodal representations of actions are obtained using the co-activation of action features from video sequences and labels from automatic speech recognition. Promising experimental results encourage the extension of our architecture in several directions.**

Keywords*—Human action recognition, multimodal integration, self-organizing networks.

I. Introduction

The ability to integrate information from different modalities for an efficient interaction with the environment is a fundamental feature of the brain. As humans, our daily perceptual experience is modulated by an array of sensors that convey different types of modalities such as vision, sound, touch, and movement [1]. Similarly, the integration of modalities conveyed by multiple sensors has been a paramount ingredient of autonomous robots. In this context, multisensory inputs must be represented and integrated in an appropriate way such that it results in a reliable cognitive experience aimed to trigger adequate behavioral responses. Multimodal cognitive representations have been shown to improve robustness in the context of action recognition and action-driven perception, learning by imitation, socially-aware agents, and natural human-robot interaction (HRI) [2].

An extensive number of computational models has been proposed that aimed to integrate audiovisual input (e.g. [3][4]). These approaches used unsupervised learning for generalizing visual properties of the environment (e.g. objects) and linking these representations with linguistic labels. However, action verbs do not label actions in the same way that nouns label objects [5]. 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. In fact, action classification has been shown to be particularly challenging since it involves the processing of a huge amount of visual information to learn inherent spatiotemporal dependencies in the data. To tackle this issue, learning-based mechanisms have been typically used for generalizing a set of labelled training action samples and then predicting the labels of unseen samples (e.g. [15][16]). However, most of the well-established methods learn actions with a batch learning scheme, i.e. assuming that all the training samples are available at the training phase. An additional common assumption is that training samples, generally presented as a sequence of frames from a video, are well segmented so that ground-truth labels can be univocally assigned. Therefore, it is usually the case that raw data collected by sensors must undergo an intensive pre-processing pipeline before training a model. Such pre-processing stages are mainly performed manually, thereby hindering the automatic, continuous learning of actions from live video streams. Intuitively, this is not the case in nature.

Words for actions and events appear to be among children's earliest vocabulary [6]. A central question in the field of developmental learning has been how children first attach verbs to their referents. During their development, children have at their disposal a wide range of perceptual, social, and linguistic cues that they can use to attach a novel label to a novel referent [7]. Referential ambiguity of verbs could then be solved by children assuming that words map onto the action with most perceptual saliency in their environment. Recent experiments have shown that human infants are able to learn action-label mappings using cross-situational statistics, thus in the presence of piece-wise available ground-truth action labels [8]. Furthermore, action labels can be progressively learned and improved from social and linguistic cues so that novel words can be attached to existing visual representations. This hypothesis is supported by many neurophysiological studies evidencing strong links between the areas in the brain governing visual and language processing, and suggesting high levels of functional interaction of these areas during action learning and recognition [9].

In this work, we investigate how associative links between unimodal representations can naturally emerge from the co-occurrence of audiovisual stimuli. We show that it is possible to progressively learn congruent multimodal representations of human actions with neural self-organization using a special type of hierarchical connectivity. For this purpose, we extended our recently proposed neural architecture for the self-organizing integration of action cues [16] with an associative learning layer where action-word mappings emerge

In: IEEE-RAS International Conference on Humanoid Robots (Humanoids '15), Workshop on Towards Intelligent Social Robots: Current Advances in Cognitive Robotics, Seoul, South Korea (2015)

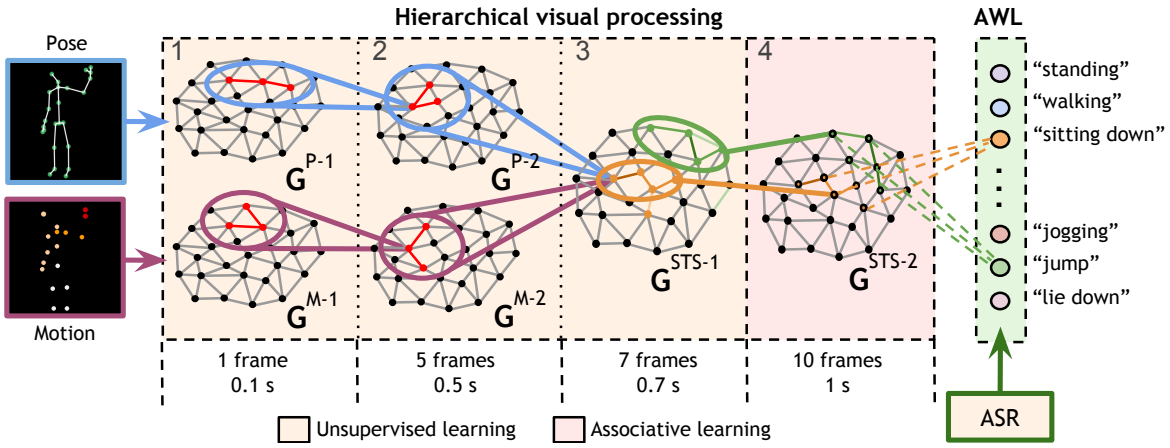


Fig. 1. Diagram of our learning architecture with GWR networks and the number of frames (and seconds) required for hierarchical processing - Layers 1-3: parallel spatiotemporal clustering of visual features and self-organizing pose-motion integration (STS-1). Layer 4: Self-organization of STS-1 representations and associative learning for linking visual representations in STS-2 to the action words layer (AWL) obtained with automatic speech recognition (ASR).

from co-occurring audiovisual inputs using Hebbian-like learning [10]. We implement experience-dependent plasticity with the use of an incremental self-organizing network that employs neurobiologically-motivated habituation for stable learning [11]. The proposed architecture is novel in two main aspects: First, our learning mechanism does not require 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 online fashion with a semi-supervised learning scheme. Second, we propose a type of bidirectional inter-layer connectivity that takes into account the spatiotemporal dynamics of sequences so that symbolic labels are linked to temporally-ordered representations in the visual domain.

In Section II, we describe our hierarchical architecture with incremental self-organizing networks and hierarchical connectivity for multimodal integration. In Section III, we present our conducted experiments and compare our results with other approaches on a dataset of 10 actions using pose-motion cues as visual features and labels obtained from automatic speech recognition. In Section IV, we discuss on-going research efforts for the extension of our model in several directions.

II. PROPOSED METHOD

Our learning architecture consists of 4 hierarchically arranged layers and a symbolic layer of action words (Fig. 1). Layers 1 and 2 consist of a two-stream hierarchy for the processing of pose and motion features. One pathway processes body pose features while the other processes motion flow. The subsequent integration of pose-motion cues is carried out in Layer 4 (or STS-1) to provide movement dynamics in the joint feature space. The motivation underlying hierarchical learning is to obtain progressively specialized neurons coding spatiotemporal dependencies of the input, consistent with the assumption that the recognition of actions must be selective for temporal order. This is achieved by using trajectories of neuron activations from a network for the training of a higher-level network. A detailed description of Layers 1, 2, and 3 is provided by Parisi *et al.* [16].

From a neurobiological perspective, a large number of studies has shown that the superior temporal sulcus (STS) in the mammalian brain is the basis of an action-encoding network with neurons that are not only driven by the perception of dynamic human bodies, but also by audiovisual integration [13]. Therefore, the STS area is thought to be an associative learning device for linking different unimodal representations, accounting for the mapping of naturally occurring, highly correlated features such as shape, motion, and characteristic sound [14]. In our proposed architecture, we implement an associative learning network in Layer 4 (or STS-2) where action-word mappings are progressively learned from co-occurring audiovisual inputs using a self-organizing connectivity scheme.

A. Self-Organizing Hierarchical Learning

Our model consists of hierarchically-arranged Growing When Required (GWR) networks [11] 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. The activity of a neuron is computed as a function of the distance between the input and its weight vector. During the training, the network dynamically changes its topological structure to better match the input space following competitive Hebbian learning [10].

Different from other incremental models of self-organization, GWR-based learning takes into account the number of times that a neuron has fired so that neurons that have fired frequently are trained less. The network implements a habituation counter $\eta(t) \in [0, 1]$ to express how frequently a neuron s has fired based on a simplified model of how the efficacy of an habituating synapse reduces over time [12]. The habituation counter is given by

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

where $\eta(s_t)$ is the size of the firing rate for neuron s_t, η_0 is the resting value, $S(t)$ is the stimulus strength, and τ, α are constants that control the behaviour of the curve. A neuron

n is considered to be well trained when $\eta(n)$ is greater than a firing threshold η_T. This is in favour of training existing neurons before creating new ones. New nodes can be created any time if the activity of well-trained neurons is smaller than an activity threshold a_T. 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.

Hierarchical learning is carried out by training a higher-level network with neuron activation trajectories from a lower level network. These trajectories are obtained by computing the best-matching neuron of the input sequence with respect to the trained network with N neurons, so that a set of trajectories of length q is given by

$$\Omega^q(\mathbf{x}_i) = \{\mathbf{w}_{b(\mathbf{x}_i)}, \mathbf{w}_{b(\mathbf{x}_{i-1})}, ..., \mathbf{w}_{b(\mathbf{x}_{i-q+1})}\} \quad (2)$$

with $b(\mathbf{x}_i) = \arg\min_{j \in N} \|\mathbf{x}_i - \mathbf{w}_j\|$.

The STS-1 layer integrates pose-motion features by training the network with vectors of the form

$$\Psi = \{\Omega^q(\mathbf{X}), \Omega^q(\mathbf{Y})\}, \quad (3)$$

where $\mathbf{X}$ and $\mathbf{Y}$ are the activation trajectories from the pose and motion pathways respectively. After STS-1 training is completed, each neuron will encode a sequence-selective prototype action segment.

B. GWR-based Associative Learning

For the higher layer STS-2, we extended the standard GWR algorithm with: 1) asymmetric neural connectivity based on Hebbian learning, and 2) semi-supervised labelling functions so that prototype neurons can be attached to symbolic labels during training. The detailed learning procedure for the creation and update of existing neurons is illustrated by Algorithm 1.

Local lateral connectivity in self-organizing networks is responsible for the correct formation of the topological map. We enhanced standard neuron connectivity by taking into account inherent temporal relations of the input, so that connections between neurons that are consecutively activated are strengthened. For this purpose, we define a connection strength function ρ that increases between activated neurons b_{t-1} and b_t at time $t-1$ and t respectively (Algorithm 1, Steps 6c and 7b). This type of connectivity scheme is asymmetric in the sense that $\rho(b_{t-1}, b_t)$ increases while $\rho(b_t, b_{t-1})$ remains unchanged, thereby fostering temporally-ordered representations of actions from neuron activation trajectories.

We extend the unsupervised GWR for semi-supervised learning so that action labels will be attached to prototype neurons during the training phase in an online fashion (Algorithm 1, Steps 6d and 7c). We implement a mechanism for label propagation that takes into account how well trained neurons are before propagating labels to their neighbours. For this purpose, we define two labelling functions: one for when a new neuron is created, and the other for when the neuron is updated. Provided that b_t is the index of the best-matching neuron and that ξ_t is the label of $\mathbf{x}_t$, and that we denote a missing label with -1, when a new neuron r_t is created, its label $\lambda(r_t)$ is assigned according to:

$$\gamma^{new}(b_t, \xi_t) = \begin{cases} \xi_t & \xi_t \neq -1 \\ \lambda(b_t) & \text{otherwise} \end{cases} \quad (4)$$

Algorithm 1 Semi-supervised Associative GWR

1: Create two random neurons with weights $\mathbf{w}_1$ and $\mathbf{w}_2$
2: Initialize an empty set of connections $E = \emptyset$.
3: At each iteration t, generate an input sample $\mathbf{x}_t$
4: For each neuron n, select the-best matching node and the second-best such that:
$b_t = \arg\min_{n \in A} \|\mathbf{x}_t - \mathbf{w}_n\|$
$s_t = \arg\min_{n \in A/\{b_t\}} \|\mathbf{x}_t - \mathbf{w}_n\|$
5: Create a connection if it does not exist
5a: $E = E \cup \{(b_t, s_t)\}$ and set age of E_{b_t,s_t} to 0.
6: If $(exp(-\|\mathbf{x}_t - \mathbf{w}_{b_t}\|) < a_T)$ and $(\eta(b_t) < f_T)$ then:
6a: Add a new neuron r_t between b_t and s_t with $\mathbf{w}_{r_t} = \kappa \cdot (\mathbf{w}_{s_t} + \mathbf{x}_t)$
6b: Create edges and remove old edge: $E = E \cup \{(r_t, b_t), (r_t, s_t)\}$ and $E = E/\{(b_t, s_t)\}$
6c: Connection strengths: $\rho_{(b_{t-1},r_t)} = 1, \rho_{(b_{t-1},b_t)} = 0$
6d: Initialize label: $\lambda(r_t) = \gamma^{new}(b_t, \xi_t)$
7: Else, i.e. no new neuron is added, update $\mathbf{w}_{b_t}$ and its neighbours i:
7a: $\Delta\mathbf{w}_{b_t} = \epsilon_b \cdot \eta(b_t) \cdot (\mathbf{x}_t - \mathbf{w}_{b_t})$ and $\Delta\mathbf{w}_{i_t} = \epsilon_n \cdot \eta(i) \cdot (\mathbf{x}_t - \mathbf{w}_{i_t})$,
with $0 < \epsilon_n < \epsilon_b < 1$
7b: Increase connection strength $\rho_{(b_{t-1},b_t)}$
7c: Update label: $\lambda(b_t) = \gamma^{update}(b_t, s_t, \xi_t)$
7d: Increment the age of all edges connected to b_t.
8: Reduce the firing counters η according to Eq. 1.
9: Remove all edges with ages larger than a_{max} and remove neurons without edges.
10: If the stop criterion is not met, go to step 3.

Provided that s_t is the index of the second best-matching neuron, the update labelling function for $\lambda(b_t)$ is defined as:

$$\gamma^{update}(b_t, s_t, \xi_t) = \begin{cases} \xi_t & \xi_t \neq -1 \\ \lambda(s_t) & (\xi_t = -1) \wedge (\eta(s_t) \geq \eta_T) \\ \lambda(b_t) & \text{otherwise} \end{cases} \quad (5)$$

This mechanism results in the correct propagation of labels so that labels attach to neurons based on the co-occurrence of audiovisual inputs, thereby avoiding the need of manual segmentation for ground-truth labelling.

C. Action-Word Mappings

During the learning in STS-2, unsupervised visual representations of actions are linked to symbolic action labels $\lambda_j \in L$, with L being the set of j possible words. Action words will then have a one-to-many relation with STS-2 neurons, i.e. neurons can be attached to only one label in L. It is possible that neurons change label during the learning phase based on the self-organizing process of label propagation. For clarity, we now refer to the symbolic connectivity layer of words as the "action words" layer (AWL).

The development of connections between STS-2 and AWL depends upon the co-activation of audiovisual inputs. More specifically, the connection between a STS-2 neuron and its symbolic label in AWL will be strengthened if the neuron is activated within a time window in which also the label is activated by an audio signal. In the case that no audio stimulus occurs during the creation or adaptation of a STS-2 neuron,

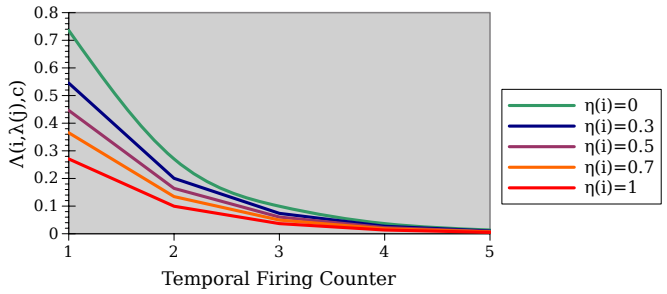


Fig. 2. Temporal strength function $\Lambda(i, \lambda_j, t) = 2 \cdot [exp(\eta(i) + c(\lambda_j, t))]^{-1}$ for different firing rates (y-axis) and sequence counters (x-axis). It can be seen how greater values are given to well-trained neurons activated at the beginning of the sequence.

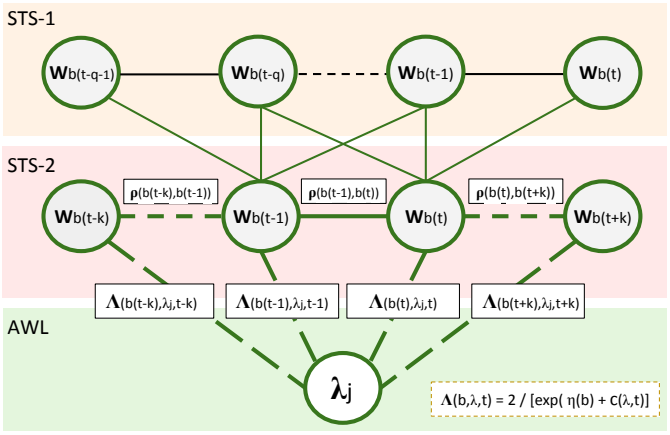


Fig. 3. Inter-layer connectivity scheme: Neurons in the STS-2 layer result from the hierarchical learning of STS-1 activation trajectories. STS-2 neurons use recurrent connection strength ρ to preserve temporal relations of the input. Connectivity between STS-2 and AWL emerges taking into account neuron firing rates and the order of activation.

symbolic labels will instead be updated according to our semi-supervised label propagation rules (Eq. 4 and 5). This scheme takes into account the temporal order of activation in a given sequence of consecutively fired neurons. This is in favour of the generation of temporally-ordered trajectories generalizing one prototype action sequence. For a generic labelled neuron i in STS-2 fired at time t, its connection strength with the symbolic label λ_j becomes:

$$\Lambda(i, \lambda_j, t) = 2 \cdot [exp(\eta(i) + c(\lambda_j, t))]^{-1}, \qquad (6)$$

where $c(\lambda_j, t)$ is the sequence counter and $exp(\eta(i) + c(\lambda_j, t))$ expresses the exponential relation between the firing counter of the neuron and its sequential order within the set of neuron activations with the same label. This function yields greater values for connections of well-trained nodes that activate at the beginning of a sequence. The counter $c(\lambda_j, t)$ will increase while $\lambda(b_t) = \lambda_j(t)$ and reset when this condition does not hold. The temporal strength function for different firing rates and sequence counters is depicted in Fig. 2 for a window of 5 neuron activations. A diagram of inter-layer connectivity between STS-1, STS-2, and AWL is shown in Fig. 3.

D. Action Word from Visual Recognition

At recognition time, we classify previously unseen video sequences to match one of the training actions. For this purpose, we define a recognition function $\varphi : \Omega \rightarrow \Lambda$ on the basis of a single-linkage strategy [19] such that each new trajectory sample ω_{new} from STS-1 is labelled with an action word $\lambda_j \in \Lambda$ associated to the STS-2 neuron $\mathbf{w}$ that minimizes the distance to the new sample:

$$\varphi(\omega_{new}) = \arg \min_{\lambda_j} (\arg \min_{\mathbf{w} \in N(\lambda_j)} \|\mathbf{w}_n - \omega_{new}\|). \qquad (7)$$

The hierarchical flow is composed of 4 networks, with each subsequent network neuron encoding a window of 3 neurons from the previous one, with the exception of STS-2, which processes 4-neuron trajectories. Therefore, this classification algorithm returns a new action label every 10 samples (1 second of video operating at 10 frames per second). By applying a temporal sliding window scheme, we get a new action label for each frame.

E. Visual Sequence from Action Word

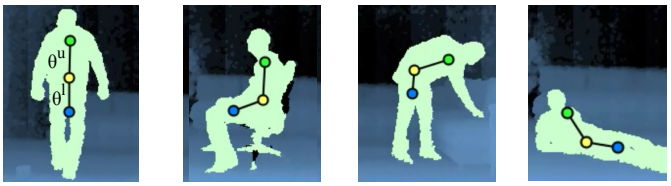
We use the strength function ρ to obtain prototype visual representations of actions from recognized action words. We expect that each action word will activate a trajectory that represents a prototype action sequence in the STS-2 layer. Therefore, after recognizing an action word λ_j from speech, the STS-2 neuron that maximizes Eq. 6 is selected as the first element of a sequence and used to generate temporally-ordered prototype representations of actions by recursive ρ-connectivity. This mechanism can be used in practice to assess how well the model has learned action dynamics and whether it has accounted for linking action words to visual representations.

III. EXPERIMENTS

We now present our experimental set-up and results on a dataset of full-body actions. In contrast to previous training procedures [15][16], for these experiments action samples from sequential frames were not manually segmented. Instead, action labels were recorded from speech so that action-word mappings of training samples resulted from co-occurring audiovisual inputs using our label propagation strategy. To evaluate our system, we compared new obtained results with recently reported results using GWR-based hierarchical processing with manual segmentation for ground-truth labelling [16].

A. Audiovisual Inputs

Our action dataset is composed of 10 full-body actions performed by 13 subjects [15]. 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 640x480, an operation range from 0.8 to 3.5 m at 30 frames per second. The dataset contains the following actions: standing, walking, jogging, sitting, lying down, crawling, pick up, jump, fall down, and stand 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. 4) as described in [15].



a) Walking b) Sitting c) Pick up object d) Lying down

Fig. 4. Representation of full-body movements from our action dataset [15]. We estimate three centroids: C_1 (green), C_2 (yellow) and C_1 (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 recording action labels, we used automatic speech recognition from Google's cloud-based ASR enhanced with domain-dependent post-processing [18]. The post-processor translates each sentence in the 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 sentence. An advantage of this approach is 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[1] 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 seconds during its execution, i.e. for each repetition in the case of jump, fall down, and stand up, and every 2 seconds for cyclic actions (standing, walking, jogging, sitting down, lying down, crawling). This approach has the advantage of assigning labels to continuous video streams without the manual segmentation of visual features.

B. Evaluation

For a fair 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 seconds for each periodic action (standing, walking, jogging, sitting, lying down, crawling) and 30 repetitions for each goal-oriented action (pick up object, jump, fall down, stand up). Both the training and the test sets contained data from all subjects. For GWR learning, we used the following training parameters: insertion threshold $a_T = 0.9$, learning rates $\epsilon_b = 0.3$, and $\epsilon_n = 0.006$, $\kappa = 0.5$, maximum age $a_{max} = 50$, firing counter parameters $\eta_0 = 1$, $\tau_b = 0.3$, $\tau_n = 0.1$, firing threshold $\eta_T = 0.01$. For a more detailed discussion on training parameters, please refer to Parisi *et al.* [16]

Experimental results showed that our new approach performs very well ($93,3\%$ average accuracy) with respect to our previous approach based on manual segmentation (94% average accuracy). The confusion matrix for the 10 actions is shown in Fig. 5 (with the rows of the matrix being the instances of actual actions and columns being the instances

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

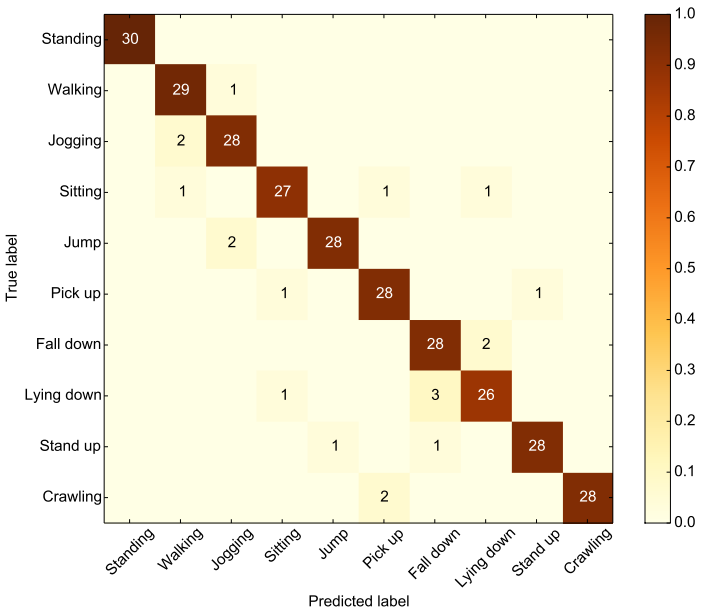


Fig. 5. Confusion matrix for our dataset of 10 actions. The average accuracy is 93,3%.

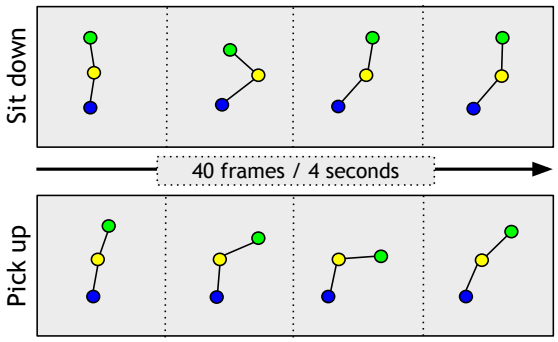


Fig. 6. Example of visual representations in STS-2 that maximize inter-layer connectivity for the actions "Sit down" and "Pick up" generated by speech recognition.

of predicted actions). These promising results encourage to extend our current neural architecture in several directions.

To have a qualitative idea of how well the associative layer has learned action dynamics, we extracted STS-2 neuron trajectories with the first neuron being activated by maximizing the temporal connection strength function (Eq. 6) and the subsequent 4 neurons obtained with ρ-connectivity. The visual representations of the actions "Sit down" and "Pick up" for a time window of 40 frames (4 seconds) are shown in Fig. 6, from which we can argue that the associative layer successfully learns temporally-ordered representations of input sequences.

IV. CONCLUSIONS AND FUTURE WORK

We presented a hierarchical neural architecture for action recognition from audiovisual inputs. In particular, we investigated how associative links between unimodal representations can emerge from the co-occurrence of multimodal stimuli in a self-organized manner. Experimental results on a dataset of 10 full-body actions shows that our learning mechanism

does not require the manual segmentation of training samples for accurate recognition. Instead, generalizations of action sequences are incrementally learned and mapped to symbolic labels using the co-activation of audiovisual inputs. For this purpose, we proposed a type of bidirectional, inter-layer connectivity that takes into account the spatiotemporal dynamics of action samples.

Similar to Vavrečka and Farkaš [4], 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 progressively learned in an unsupervised fashion. Interestingly, our implementation with bidirectional action-to-word connections roughly resemble a phenomenon found in the human brain, i.e. spoken action words elicit receptive fields in the visual area [13]. In other words, visual representations of generalized actions can be activated in the absence of visual inputs, in this case from speech. We have shown that this property can be used in practice to assess how well the model has learned action dynamics.

This work represents an effort towards a more sophisticated model that learns cognitive representations through the self-organizing development of associative links between different modalities. Current research work aims to leverage the proposed neural architecture in several directions. For instance, in the presented model, we assume that labels are provided from speech during the training session for all action samples. We are currently investigating a scenario in which labels are not always provided during training sessions, as it is also the case in nature. Several developmental studies have shown that human infants are able to learn action-label mappings using cross-situational statistics, thus in the presence of not always available ground-truth action labels [8]. Another limitation of our model is the use of domain-dependent ASR. In the future, we plan to avoid this constraint by accounting for learning new lexical features so that the action vocabulary can be dynamically extended during training sessions. For instance, it has been shown that lexical features can be learned using recursive self-organizing architectures [20][21]. Finally, we plan to evaluate our learning architecture with benchmark datasets using a greater number of body features. This is aimed to achieve more complex visual tasks such as the recognition of transitive actions.

ACKNOWLEDGMENT

This work was supported by the DAAD German Academic Exchange Service (Kz:A/13/94748) - Cognitive Assistive Systems Project.

REFERENCES

[1] B.E. Stein, T.R. Stanford, B.A. Rowland. The neural basis of multisensory integration in the midbrain: its organization and maturation. Hear Res 258(1–2):4–15, 2009.

[2] R. Kachouie, S. Sedighadeli, R. Khosla, and M-T Chu. Socially assistive robots in elderly care: a mixed-method systematic literature review. Int. J. Hum. Comput. Interact. 30:369–393, 2014.

[3] A.F. Morse, V.L. Benitez, T. Belpaeme, A. Cangelosi, and L. B. Smith. Posture affects how robots and infants map words to objects. PLoS ONE 10(3): e0116012. doi:10.1371/journal.pone.0116012, 2015.

[4] M. Vavrečka and I. Farkaš. A multimodal connectionist architecture for unsupervised grounding of spatial language. Cogn Comput 6:101–112, 2014.

[5] D. Gentner. Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In S. Kuczaj (Ed.), Language development: Language, thought, and culture 2:301–334, 1982.

[6] L. Bloom. The transition from infancy to language: Acquiring the power of expression. New York: Cambridge University Press, 1993.

[7] K. Hirsh-Pasek, R.M., Golinkoff, and G. Hollich. An emergentist coalition model for word learning. In R. M. Golinkoff et al. (Eds.), Becoming a word learner: A debate on lexical acquisition. New York: Oxford University Press, 2000.

[8] L. Smith and C. Yu. Infants rapidly learn word-referent mappings via cross-situational statistics. Cognition 106(3):1558–1568, 2008.

[9] F. Pulvermueller. Brain mechanisms linking language and action. Nature 6:576–582, 2005.

[10] T. Martinetz. Competitive Hebbian learning rule forms perfectly topology preserving maps. In: ICANN93 (Springer, Helderberg), pp. 427–434, 1993.

[11] S. Marsland, J. Shapiro, and U. Nehmzow. A self-organising network that grows when required. Neural Networks 15:1041–1058, 2002.

[12] J.C. Stanley. Computer simulation of a model of habituation. Nature 261:146–148, 1976.

[13] N.E Barraclough, D. Xiao, C.I. Baker, M.W. Oram, and D.I. Perrett. Integration of visual and auditory information by superior temporal sulcus neurons responsive to the sight of actions. J Cogn Neurosci 17(3):377–91, 2005.

[14] M.S. Beauchamp, K.E. Lee, B.D. Argall, and A. Martin. Integration of auditory and visual information about objects in superior temporal sulcus. Neuron 41(5):809–823, 2004.

[15] G.I. Parisi, C. Weber, and S. Wermter. Human action recognition with hierarchical growing neural gas learning. In Wermter, S., et al., editors. In: International Conference on Artificial Neural Networks (ICANN 2014), pp. 89–96, Hamburg, Germany, 2014.

[16] G.I. Parisi, C. Weber, and S. Wermter. Self-Organizing neural integration of pose-motion features for human action recognition. Frontiers in Neurobotics 9:3, 10.3389/fnbot.2015.00003, 2015.

[17] G.I. Parisi, G. I., F. v. Stosch, S. Magg, and S. Wermter. Learning human motion feedback with neural self-Organization. In: International Joint Conference on Neural Networks (IJCNN), pp. 2973–2978, Killarney, Ireland, 2015.

[18] J. Twiefel, T. Baumann, S. Heinrich, and S. Wermter. Improving Domain-independent cloud-based speech recognition with domain-dependent phonetic post-processing. In: *IEEE Conf. on Artificial Intelligence (AAAI-14)*, pp. 1529–1535, Quebec, Canada, 2014.

[19] O. Beyer, and P. Cimiano. Online labelling strategies for growing neural gas. In: IDEAL11 (Norwich: Springer Berlin Heidelberg), pp. 76–83, 2011.

[20] M. Strickert and B. Hammer. Merge SOM for temporal data. Neurocomputing 64: 39–71, 2005.

[21] A. Andreakis, N. v. Hoyningen-Huene, and M. Beetz. Incremental unsupervised time series analysis using merge growing neural gas. In: Advances in Self-Organizing Maps, pp. 10–18. Springer, 2009.