Available online at www.sciencedirect.com

ScienceDirect

Cognitive Systems Research 43 (2017) 208–221

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
Available online 22 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