available at www.sciencedirect.com

www.elsevier.com/locate/brainres

BRAIN RESEARCH

Research Report

A connectionist architecture for view-independent grip-aperture computation

Roberto Prevete[a,*], Giovanni Tessitore[a], Matteo Santoro[b], Ezio Catanzariti[a]

[a]Department of Physical Sciences, University of Naples Federico II, Naples, Italy
[b]DISI, University of Genoa, Genoa, Italy

ARTICLE INFO

Article history:
Accepted 28 April 2008
Available online 10 May 2008

Keywords:
Mirror neuron
Grip aperture
View independence
Computational model
Reach-to-grasp

ABSTRACT

This paper addresses the problem of extracting view-invariant visual features for the recognition of object-directed *actions* and introduces a computational model of how these visual features are processed in the brain. In particular, in the test-bed setting of reach-to-grasp actions, *grip aperture* is identified as a good candidate for inclusion into a parsimonious set of hand high-level features describing overall hand movement during reach-to-grasp actions. The computational model NeGOI (neural network architecture for measuring grip aperture in an observer-independent way) for extracting grip aperture in a view-independent fashion was developed on the basis of functional hypotheses about cortical areas that are involved in visual processing. An assumption built into NeGOI is that grip aperture can be measured from the superposition of a small number of prototypical hand shapes corresponding to predefined grip-aperture sizes. The key idea underlying the NeGOI model is to introduce view-independent units (*VIP* units) that are selective for prototypical hand shapes, and to integrate the output of VIP units in order to compute grip aperture. The distinguishing traits of the NEGOI architecture are discussed together with results of tests concerning its view-independence and grip-aperture recognition properties. The overall functional organization of NEGOI model is shown to be coherent with current functional models of the ventral visual stream, up to and including temporal area STS. Finally, the functional role of the NeGOI model is examined from the perspective of a biologically plausible architecture which provides a parsimonious set of high-level and view-independent visual features as input to mirror systems.

1. Introduction

Over the last few years, there has been a keen interest in developing computational models for *action recognition*. A chief motivation for this growing interest stems from applied research in computer science, where efficient algorithms for action recognition are being developed for large numbers of potential applications in, e.g., robotics, intelligent surveillance systems, and sign language recognition (Kruger, 2007). These investigations notably concern mechanisms that enable robotic agents to acquire new behaviours by observing and generalizing the behaviour of other agents (Dillmann, 2005; Kruger, 2007), and feed-back mechanisms that enable agents to control their own behaviour (Desmurget and Grafton, 2000). Another chief motivation for this growing interest stems from computational neuroscience. Our everyday experience shows

[*] *Corresponding author.* Department of Physical Sciences, Complesso Universitario di M. S. Angelo, Via Cintia — 80126, Italy.
E-mail address: prevete@na.infn.it (R. Prevete).

0006-8993/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.brainres.2008.04.076

that human beings are capable of interacting through perception and recognition of actions. Computational models of brain regions involved in the process of action recognition are being developed which provide insights into the mechanisms supporting these high-level functionalities of many primate brains (e.g, see Oztop and Arbib, 2002; Oztop et al., 2006). These computational neuroscience modelling efforts provide the broad context for the work presented here, which concerns the problem of extracting visual features that are needed to recognize object-directed actions in a view-independent fashion. Specifically, we present a biologically plausible computational model of processes which enable one to extract *grip aperture* (Jeannerod, 1984), that is, the aperture between index finger and thumb, *in a view-independent fashion* and in the context of *object-directed actions* (typically, an object-directed action is a sequence of prehension movements relating body effectors, such as hand or mouth, to three-dimensional objects to be grasped or manipulated, such as food morsels, paper-clips, and mugs).

The significance of this extraction problem for the overall action recognition problem is quite evident. First, the need for acquiring information about the dynamics and the poses of individual body parts is widely acknowledged in action recognition inquiries, and cannot be altogether dismissed even by the more “holistic” approaches to action recognition (Ricquebourg and Bouthemy, 2000; Rittscher et al., 2002). In particular, the ability to extract “relevant” features of body parts from visual input assumes a central role for action recognition. Second, if one takes into account the great input variability due to viewpoint variations, then independence from the viewpoint of both action representation and recognition becomes a key aspect of visual action recognition. This aspect has received relatively scarce attention (Parameswaran and Chellappa, 2006) so far, and the computational modelling of brain mechanisms which enable one to extract visual features of body parts in a view-independent fashion is needed to advance our understanding of action representation and recognition processes. Third, in the light of mirror systems models, *object-directed actions* assume special significance in this context. A population of neurons was discovered in macaque pre-motor area F5 which are active (high spike rate) either when the monkey executes an *object-directed action* or when the monkey observes another individual (a conspecific or a human experimenter) who executes a similar object-directed action. In view of their characteristic activation properties, F5 neurons responding in both action observation and execution conditions are called *mirror neurons* (Rizzolatti and Luppino, 2001; Rizzolatti et al., 1996). According to various computational models (Keysers and Perrett, 2004; Oztop and Arbib, 2002; Oztop et al., 2006; Oztop et al., 2005; Prevete et al., 2005; Tani et al., 2005), mirror activity is triggered by view-independent *effector/object descriptions* which apply to both executed and observed actions. Notably, in (Oztop and Arbib, 2002) effector/object descriptions take the form of *high-level feature vectors*. These vectors, called *hand states*, provide information about various hand features, including grip aperture, object–hand distance, and wrist velocity. Hand states are collections of *view-independent* features: one computes the same hand state in both execution and observation conditions with respect to some given object-directed action; and identical hand states give rise to the same mirror neuron output. Thus, the ability to compute the same effector/object description relies on the existence of a mechanism which enables one to extract, in a view-independent fashion, features of hand/object pairs.

This extraction problem requires a careful abstract formulation and analysis. Indeed, hands are highly complex structures with more than 20 degrees of freedom, and therefore the process of making a detailed hand description available from visual inputs is presumably quite demanding from a computational point of view. But how many different features must be taken into account, and how detailed must a hand description be for the purpose of both recognition and grasping control tasks?

In the context of grasping control, the opportunity of using simpler hand description models is argued for in (Iberall and Fagg, 1996; Iberall et al., 1986) and explored there by means of the notion of virtual finger which enables one to reduce the degrees of freedom, and thereby the complexity of the hand control problem.

In the context of recognition tasks, a simplified control strategy during reach-to-grasp actions is suggested, at the output stage, in the reduced number of hand shapes one can effectively assume and, hence, in the reduced number of features that are needed to achieve a meaningful hand description. Empirical evidence for the use of a reduced set of features for representing hand shapes in the context of reach-to-grasp actions is provided by behavioural findings (Santello et al., 2002; Mason et al., 2001). Notably, in Santello's work a principal component analysis is performed over a series of hand features that are monitored while a subject performs (or mimics) a reach-to-grasp action. The outcome of this analysis shows that the first principal component suffices to account for most hand feature variability, and the first two principal components account for almost every aspect of whole hand feature variability. Apparently, coordinated movements of hand fingers result, during reach-to-grasp action, into a reduced number of physically possible hand shapes. This reduction of the number of hand shapes suggests, in turn, the possibility to describe hand shapes during reach-to-grasp actions by a “parsimonious” set of hand features (notice that a significant step in this direction was made by Oztop and Arbib (2002) by introducing the notion of virtual finger).

A parsimonious set of hand features which is both sufficiently powerful to describe hand shapes during object-directed actions and sufficiently simple to be computed in a view-independent fashion without requiring computationally demanding algorithms has not been identified yet. However, grip aperture appears to be a key element of this set, insofar as it is a crucial variable in the dynamical evolution of certain types of grasping actions. Some have even advanced and supported the hypothesis that grasping actions are basically coded in terms of changes in grip aperture (Castiello, 2005; Jeannerod, 1984). During a reach-to-grasp action, grip aperture initially increases until a maximum value is reached which exceeds object size; then grip aperture gradually decreases until it matches the actual object size; the grip-aperture largest value (maximum grip aperture) is reached within 60–70% of the grasp action duration and is linearly correlated with the size of the object. On the whole, grip aperture is a good