Lessons from the Primate Visual System

Guy A. Orban[1,2]

[1] Departments of Neuroscience KU Leuven
[2] University of Parma

Abstract. The primate visual system can perform an astonishing array of tasks as reflected by the correspondingly large portion of the cerebral cortex devoted to analyzing retinal signals. Although a potential source of inspiration for computer vision, with a few exceptions, progress has been slow in this field. Principal obstacles are the lack of any exhaustive list of what vision achieves in humans and the restricting of areas of investigation to a few topics such as motion, object categories and the control of a few actions such as reaching or saccades. Here I will review how we integrated several experimental techniques to address a question that arose from interactions with computer vision scientists more than fifteen years ago: the extraction of 3D surfaces. This goal is achieved by a new type of higher-order visual neuron: the gradient-selective neurons. Neurons selective for speed gradients were initially discovered in motion processing areas, such as MT/V5, MSTd and FST, located in the monkey superior temporal sulcus (STS). Subsequently, neurons selective for disparity gradients were discovered in shape processing areas, such as TEs and AIP. By combining these single-cell studies with fMRI in human and awake monkey, we were able to localize similar neurons to human cortical areas. In the second part I address my present interest in understanding the visual signals related to the actions of conspecifics, which is perhaps the ultimate challenge of motion processing, but which receives surprisingly little attention in vision. The understanding of observed actions exemplifies my statement that to be useful visual signals have to leave the visual system, as signals related to biological motion in the STS are indeed relayed to parietal regions involved in the control of diverse actions to be understood as actions.

1 The Primate Visual System

The visual system of the primate occupies a substantial portion of the cortical surface. Cortical surface averages 978 cm^2 per hemisphere in humans [1], compared to 105 cm^2 in monkeys [2]. Thus the macaque cortex is just under ten times smaller than human cortex, while this ratio is 1:1000 in the mouse, making this species less attractive as experimental model. In the macaque about 60% of cortical surface is visual in nature and about 40 visual cortical areas have been identified [3]. In humans, 30% of cortex is thought to be visual, which in absolute size is still greater than monkey visual cortex. Since for homologous areas the ratio of surface area between macaque and human ranges from 2 to 6 fold, one can expect slightly more visual cortical areas in humans than monkeys, perhaps 50 or so. Nineteen areas have been identified so far in occipital cortex by retinotopic mapping [4].

A. Fusiello et al. (Eds.): ECCV 2012 Ws/Demos, Part I, LNCS 7583, pp. 462–467, 2012.

Although it is clear to everybody that the human visual system is able to perform an astonishing array of different tasks, it is difficult to provide an exhaustive list. Therefore I have stressed [5] the importance of studying the visual system from the backside, the level at which connections with other parts of the brain are made. At this level the visual processing is completed and it becomes possible to define the input-output relationships for a given visual task. Since many high-level visual cortical areas have connections outside the visual system, there are multiple exit points from the system, and not necessarily at the same hierarchical level. Very generally, the tasks performed by the visual system have been segregated into recognition/categorization/discrimination and motor-control tasks, which are performed by the ventral, occipito-temporal and dorsal, occipito-parietal streams respectively [6].

There is some evidence that the dorsal stream itself consists of two substreams, depending on whether visual information enters through MT/V5 or V6 [7]. In a similar vein I suggest that the ventral stream also includes multiple substreams, related respectively to the processing of the scene presenting to the observer, the objects and the conspecifics in the scene. This view removes one of conundrums of present visual neuroscience: how to reconcile what are called *category specific* processing with *general object* processing. One category specific set of regions are the place areas, most notable the parahippocampal place area in ventral occipito-temporal cortex. I propose that this area plus afferents, represents the scene processing ventral substream. The other category specific regions are the body and face areas, which in monkeys are located in the upper and lower bank of the rostral and middle STS. This corresponds to the conspecific ventral substream, leaving the remaining infero-temporal cortex as the general object ventral substream. These substreams project to the nearby hippocampal formation, and to the prefrontal cortex.

The anatomical and functional complexity of the visual system would seemingly render any investigation of this system extremely difficult. Recently, however, much progress has been achieved by combining single-cell studies in the macaque brain with parallel functional imaging of both the animal model and human, an approach pioneered by our group in Leuven [8, 9]. Indeed this strategy allows one to establish links between single-cell and fMRI studies which are complementary, in the same species and then use the fMRI comparisons to move between species. Using this strategy we were able to address a question that arose year ago from discussions with Olivier Faugeras and other computer vision scientists in the Insight EU projects.

2 A Largely Solved Problem: Extraction of 3D Shape

3D shape and depth

Some years ago Faugeras and his team were developing techniques for measuring depth in the view fields of robot cameras, producing dense maps of 3D coordinates covering the scene. This seemed an overwhelming quantity of data and I suggested that the brain might use a more synthetic solution and represent the 3D surfaces directly, rather than keeping track of all the precise depth values. Basically, I was suggesting that the brain was building representations using the first and second orders of

depth, which specify tilted planes and curved surfaces. It took me almost 15 years to demonstrate this process for most depth cues, except shading, and to extend this knowledge to humans using the strategy described above. This long journey is summarized here, but a full account can be found in [9]. It however makes two important points: 1) neuroscience can indeed find solutions to 'hard' problems that are useful to computer vision scientists and 2) biological vision and computer vision apparently progress at different time scales, with visual neuroscience being slower. It is fair to say, however, that with more money and a team fully devoted to this objective we could have advanced much faster.

Gradient-selective neurons

Gradient-selective neurons were discovered in the nineties at the end of the Insight projects. Xiao et al [10] showed that MT/V5 neurons were selective for linear speed gradients representing planes tilted in depth and relied upon their antagonistic surround for this selectivity, thus providing a mechanistic explanation. Subsequently, similar neurons were found in MSTd [11], and invariance for mean speed demonstrated. The most complete description was provided in [12], comparing first and second-order speed-gradient selective neurons in MT/V5 and FST. Many of the second-order neurons in FST were selective for saddle-shaped surfaces, an intriguing finding. Given that many joints are in fact saddle-shaped, at least from certain view points, this suggests a manner in which shape signals can be injected into the motion stream to extract action-related signals, which combine motion and shape. Similar gradient-selective neurons were also described for texture and disparity gradients [9].

Using parallel functional imaging to extend knowledge to human brain

The second leg of the journey involves demonstrating in the model system, here the macaque, that the presence of gradient-selective neurons can be captured by fMRI responses in a given paradigm. As paradigm for investigating 3D structure from motion we introduced the comparison between viewing randomly connected lies rotating in depth and those same lines translating in the fronto-parallel plane [13]. It turns out that in the monkey this contrast activates only a few higher-order visual areas: notably MT/V5 and FST [8]. In fact gradient-selective neurons in FST respond differentially to these two stimuli [12] thus validating the paradigm. Recently, the human homologues of MT/V5 and some of its satellites have been mapped retinotopically [4]. Human MT/V5 and to some degree putative FST (pFST) are also activated by the comparison of rotating vs translating random lines, suggesting that they house speed-gradient selective neurons, as do their monkey counterparts.

3 Action Understanding and Processing of Actor Characteristics

The three stages of observed action processing

The best known example of processing action-observation signals is that of observing grasping in the monkey. Single-cell studies have shown that ventral premotor neurons

in F5 [14] and parietal neurons in PFG [15] and AIP are involved in both the planning of a given grasping action to be executed by the subject and in the observation of that same grasping action performed by another, whether a monkey or human. AIP and PFG are known to project to F5 [16]. Combining fMRI with tracer studies have shown that AIP receives input from a region in the rostral lower bank of the STS, processing observed grasping, while PFG receives input from a grasping-processing region in the upper bank of STS, near STPm [17]. These regions receive from more caudal parts of the STS. We propose that action processing originates in MT/V5, where local motion is processed, and its satellites such as FST, which may provide shape information based on 3D SFM, see above. The second step is LST [18] that overlaps with the middle body patch [19], where imaging experiments suggest that biological motion is extracted. Finally, these signals are then relayed forward along the STS. The exact subdivisions and their connections are unknown, but they include the areas projecting to PFG and AIP. Thus the presumed sequence of processing steps is as follows: MT/V5 which receives directly from V1, FST, LST, rostral STS, AIP or PFG, and F5.

This processing of action observation signals in these three stages, including an occipito-temporal, parietal and premotor stage is now believed to be a general feature of the action observation networks in human and non-human primates. Of course this raises the question about homologies between areas in the two species, but progress has been made in this direction, at the occipito-temporal [19] and parietal levels [20].

Purpose of processing in the STS: actor characteristics

The preceding overview of the processing of visual action-observation signals indicates that the first occipito-temporal stage comprises multiple sublevels of processing, raising questions regarding the function of its more anterior components. If one compares monkey STS with the presumed homologous regions in humans [19], it is clear that these levels are considerably expanded in humans. Here I propose that the output of these more rostral STS fields processing biological motion, together with the static body and static and dynamic face signals, represent the identity and state of the actor. They are thus part of the conspecific ventral substream defined in point 1. The *state of the actor* refers to his physiological, mental and emotional state. Physiological state refers to age, gender, vigor, fertility, health and disabilities, features typically used in a clinical evaluation by a physician. This state is largely provided by static visual signals such as body size, face configuration, texture, and color of the skin [20]. Mental state includes the intention with respect to the observer that can be signaled by gaze direction, or rationality of the actor, which can be inferred from the way the action is performed given constraints imposed by the environment or the load borne by the subject [22]. The emotional state refers to the overall mood and emotion, indicated by motion kinematics [23] as well as affect relative to the observer as witnessed by vitality signs [24]. Some of this emotional information can be amplified by multimodal combinations of auditory and visual signals: think of the sound created by somebody walking, and how this used in movies to create an atmosphere or tension. We propose that representing the actor and his state is the primary function of monkey STS and its homologous regions in humans. Interestingly neighboring auditory regions in humans have been shown to represent the speaker [25], complementing visual information concerning the conspecifics in our environment. The visual

processing in STS provides a fine-grained analysis of the visual features of observed actions. Signals from this analysis are also dispatched to the posterior parietal cortex (PPC), and sent to the various areas involved in planning the actions in the observers' repertoire, as we have shown for grasping.

Processing of observed actions in parietal cortex
There is general agreement that posterior parietal cortex is involved in the sensorimotor transformations underlying the planning of various actions [26]. Because different actions such as locomotion, grasping, and reaching, require different types of sensory information, we propose that actions are planned in parallel in various subcomponents of PPC. Recent adaptation studies [27] are consistent with the notion that the PPC is organized according to the type of action planned and not the effector used to perform the action, as frequently proposed [28]. If we additionally hypothesize that the visual signals from the STS relating to action observation project to the portion of the PPC involved in planning that same action (generalized mirror principle), one predicts that the PPC regions concerned with action observation are also organized according to the action type. Evidence in this direction has begun to accumulate [29].

Such an organization would indicate that human cortex has discovered a short-cut for solving the intractable visual problem of recognizing the actions of others. This may be an important lesson for computer vision. Interestingly, it is presently thought that the PPC only houses the general plan for various categories of actions, assuming that the typical effector will be used [29]. Thus to design an artificial system that 'understands' human actions, it must be possible to emulate this strategy by designing a computer vision system that plans actions with human effectors, and then mapping the visual signals onto those plans. If the plans are to be used in a robot, anthropomorphic or not, the next stage can transform these plans by mapping them onto the robot effectors, thus mimicking a presumed function of human premotor cortex, namely adapting the general plan to the specific effectors used in the action.

References

1. Van Essen, D.C., Glasser, M.F., Dierker, D.L., Harwell, J., Coalson, T.: Parcellations and Hemispheric Asymmetries of Human Cerebral Cortex Analyzed on Surface-Based Atlases. Cereb. Cortex. e-Pub. (2011)
2. Van Essen, D.C., Glasser, M.F., Dierker, D.L., Harwell, J.: Cortical Parcellations of the Macaque Monkey Analyzed on Surface-Based Atlases. Cereb. Cortex. e-Pub. (2011)
3. Felleman, D.J., Van Essen, D.C.: Distributed hierarchical processing in the primate cerebral cortex. Cereb. Cortex 1, 1–47 (1991)
4. Kolster, H., Peeters, R., Orban, G.A.: The retinotopic organization of the human middle temporal area MT/V5 and its cortical neighbors. J. Neurosci. 30, 9801–9820 (2010)
5. La vision, mission du cerveau. Orban GA, College de France/Fayard, Paris (2007)
6. Goodale, M.A., Milner, A.D.: Separate visual pathways for perception and action. Trends Neurosci. 15, 20–25 (1992)
7. Rizzolatti, G., Matelli, M.: Two different streams form the dorsal visual system: anatomy and functions. Exp. Brain Res. 153, 146–157 (2003)
8. Vanduffel, W., Fize, D., Mandeville, J.B., Nelissen, K., Van Hecke, P., Rosen, B.R., Tootell, R.B., Orban, G.A.: Visual motion processing investigated using contrast agent-enhanced fMRI in awake behaving monkeys. Neuron. 32, 565–577 (2001)

9. Orban, G.: The extraction of 3D shape in the visual system of human and nonhuman primates. Annu. Rev. Neurosci. 34, 361–388 (2011)
10. Xiao, D.K., Marcar, V.L., Raiguel, S.E., Orban, G.A.: Selectivity of macaque MT/V5 neurons for surface orientation in depth specified by motion. Eur. J. Neurosci. 9, 956–964 (1997)
11. Sugihara, H., Murakami, I., Shenoy, K.V., Andersen, R.A., Komatsu, H.: Response of MSTd neurons to simulated 3D orientation of rotating planes. J. Neurophysiol. 87, 273–285 (2002)
12. Mysore, S.G., Vogels, R., Raiguel, S.E., Todd, J.T., Orban, G.A.: The selectivity of neurons in the macaque fundus of the superior temporal area for three-dimensional structure from motion. J. Neurosci. 30, 15491–15508 (2010)
13. Orban, G.A., Sunaert, S., Todd, J.T., Van Hecke, P., Marchal, G.: Human cortical regions involved in extracting depth from motion. Neuron. 24, 929–940 (1999)
14. Gallese, V., Fadiga, L., Fogassi, L., Rizzolatti, G.: Action recognition in the premotor cortex. Brain 119, 593–609 (1996)
15. Fogassi, L., Ferrari, P.F., Gesierich, B., Rozzi, S., Chersi, F., Rizzolatti, G.: Parietal lobe: from action organization to intention understanding. Science 308, 662–667 (2005)
16. Rizzolatti, G., Luppino, G.: The cortical motor system. Neuron. 31, 889–901 (2001)
17. Nelissen, K., Borra, E., Gerbella, M., Rozzi, S., Luppino, G., Vanduffel, W., Rizzolatti, G., Orban, G.A.: Action observation circuits in the macaque monkey cortex. J. Neurosci. 31, 3743–3756 (2011)
18. Nelissen, K., Vanduffel, W., Orban, G.A.: Charting the lower superior temporal region, a new motion-sensitive region in monkey superior temporal sulcus. J. Neurosci. 26, 5929–5947 (2006)
19. Jastorff, J., Popivanov, I.D., Vogels, R., Vanduffel, W., Orban, G.A.: Integration of shape and motion cues in biological motion processing in the monkey STS. Neuroimage 60, 911–921 (2012)
20. Durand, J.B., Peeters, R., Norman, J.F., Todd, J.T., Orban, G.A.: Parietal regions processing visual 3D shape extracted from disparity. Neuroimage 46, 1114–1126 (2009)
21. Stirrat, M., Perrett, D.I.: Valid facial cues to cooperation and trust: male facial width and trustworthiness. Psychol. Sci. 21, 349–354 (2010)
22. Jastorff, J., Clavagnier, S., Gergely, G., Orban, G.A.: Neural mechanisms of understanding rational actions: middle temporal gyrus activation by contextual violation. Cereb. Cortex 21, 318–329 (2011)
23. Pichon, S., Degelder, B., Grèzes, J.: Two different faces of threat. Comparing the neural systems for recognizing fear and anger in dynamic body expressions. J. Neuroimage 47, 1873–1883 (2009)
24. Stern, D.N.: Forms of Vitality: Exploring Dynamic Experience in Psychology and the Arts. Oxford University Press (2010)
25. Campanella, S., Belin, P.: Integrating face and voice in person perception. Trends Cogn. Sci. 11, 535–543 (2007)
26. Andersen, R.A., Buneo, C.A.: Intentional maps in posterior parietal cortex. Annu. Rev. Neurosci. 25, 189–220 (2002)
27. Bruggeman, H., Warren, W.H.: The direction of walking–but not throwing or kicking–is adapted by optic flow. Psychol. Sci. 21, 1006–1013 (2010)
28. Andersen, R.A., Cui, H.: Intention, action planning, and decision making in parietal-frontal circuits. Neuron. 63, 568–583 (2009)
29. Jastorff, J., Begliomini, C., Fabbri-Destro, M., Rizzolatti, G., Orban, G.A.: Coding observed motor acts: different organizational principles in the parietal and premotor cortex of humans. J. Neurophysiol. 104, 128–140 (2010)