Neural representations of perceived bodily actions using a categorical frame of reference

Tjeerd Jellema a,*, David I. Perrett b

a Department of Psychology, University of Hull, Cottingham Road, HU6 7RX Hull, UK
b School of Psychology, University of St. Andrews, Fife KY16 9JU, UK

Abstract

In object-centered, or categorical, visual representations of an agent’s actions, the spatial positions of (parts of) the body action are defined with respect to the principal axis of the agent, rather than in relation to the observer. Some cells in the superior temporal sulcus (STS) of the macaque monkey have been reported to use such a categorical frame of reference to code for bodily postures and actions, but their small numbers prevented a detailed investigation. Here we report for the first time that anterior sites in the STS contain a relatively large number of cells that use an object-centered frame of reference to code for animate objects and their actions. We further show that these cells are selectively responsive to much more restricted sections of an action trajectory than previously thought, and that they generalize over highly unusual, hitherto untested, orientations (such as upside-down walking). Quite remarkably, some cells coded for the position of a body part not only with respect to the object’s body structure, but also with respect to objects in the immediate surrounding, such as the supporting substrate the agent was standing on. We illustrate these properties using three cell populations: one responding to rotations of the upper body with respect to the lower body, one responding to forward or backward walking actions, and one responding to flexions of the knees. We discuss a possible role for the object-centered STS cell populations in representing the abstract object structure of flexible, animate objects, which could enable their recognition even when presented in highly unusual poses. We further discuss their possible role in representing goal-directed or intentional actions, and in imitation.

© 2006 Elsevier Ltd. All rights reserved.

Keywords: Object-centered coding; Viewer-centered coding; Macaque monkey; Biological motion; Object recognition; Superior temporal sulcus

1. Introduction

Perceived bodily actions can, in principle, be described within two different coordinate systems: within a viewer-centered or an object-centered coordinate system (e.g. Perrett et al., 1989, 1991). In a viewer-centered system, the view and direction of motion or articulation of the object are defined relative to the observer. In an object-centered description (Hasselmo, Rolls, Baylis, & Nalwa, 1989; Marr & Nishihara, 1978; Perrett et al., 1989; Perrett, Harries, Mistlin, & Chitty, 1990), the principal axis of the object, or another part of the object, is taken as a reference point to define the action. Object-centered descriptions therefore remain constant across different vantage points of the observer.

The neural implementation of these coordinate systems has been the subject of intense research. In particular, the viewer-centered descriptions have received a lot of attention since the discovery of cells in the banks of the superior temporal sulcus (STS; the upper bank to area Tp, Seltzer & Pandya, 1978) of the macaque monkey that explicitly coded for animate objects in a viewer-centered frame of reference (Perrett et al., 1989, 1991). The STS has repeatedly been implicated in the visual analysis of animate objects and their actions (e.g. Bruce, Desimone, & Gross, 1981; Hasselmo, Rolls, & Baylis, 1989; Oram & Perrett, 1994, 1996; Perrett, Rolls, & Caan, 1982). For example, a typical STS cell may respond to an agent walking forward to the right of the observer (i.e. from the observer’s perspective the right profile is visible and motion is to the right), but not to the agent walking backward to the right (left profile view and motion to the left). These cells thus require a specific combination of body view and motion direction, defined from the observer’s perspective. Although these cells do not generalize across changes in perspective view, they usually generalize very well across changes in illumination and size.
In contrast, cells making use of an object-centered system have received relatively little attention over the years compared to the viewer-centered cells, despite their spectacular response characteristics. The main reason for this is probably that they are so rare. Just a few studies specifically focused on the object-centered cells (e.g. Booth & Rolls, 1998; Hasselmo et al., 1989; Logothetis, Pauls, & Poggio, 1995; Perrett et al., 1989, 1991). In some of our studies the object-centered cells were briefly discussed (e.g. Jellema & Perrett, 2002; Jellema, Baker, Oram, & Perrett, 2002; Jellema, Maassen, & Perrett, 2004), but were not investigated in a comprehensive manner. Previous reports suggested that about 5% of STS cells that respond to an action or static posture do so in an object-centered manner, while the vast majority are viewer-centered (e.g. Oram & Perrett, 1996).

Actions are, however, more than simple movements, in that they typically involve goal-directedness, meaning that the action is functionally related to aspects of the environment. This allows for a third way of describing actions: goal-centred descriptions (Perrett et al., 1989, 1990). In these descriptions the intention, goal, or result of the action is used as reference point, rather than the exact nature of the action. The three different descriptions can be thought of as successive levels in a neural hierarchy: object-centered descriptions could result from pooling the outputs of viewer-centered cells, while goal-centred cells could result from pooling specific combinations of object-centered cells with cells coding for the position of the goal (Perrett et al., 1989). Goal-centred cells should also generalize their sensitivity to different actions that have the same goal. Each of these descriptions may be optimally suited for a particular behavioural situation. Viewer-centered descriptions seem most appropriate for interactions with objects under visual guidance (visuomotor behaviour), as e.g. when predator and prey need to interpret each others’ movements relative to themselves in order to catch/escape from the other. However, for object and scene recognition object-centered descriptions seem better suited (Milner & Goodale, 1995), e.g. to support the concept of object constancy despite greatly varying retinal projections. This functional distinction between viewer- and object-centered coding is reminiscent of the functional distinction made by Kosslyn and others between categorical and coordinate representation (Kosslyn, 1987; Kosslyn & Koenig, 1995). Goal-centred descriptions serve a higher level of perception, involving sensitivity to causality and the ability to predict behavioural outcomes. One can thus envisage that any perceived action may be described within a number of different, coexisting, coordinate systems, for different purposes.

Theories of ventral stream function, and the psychological and computational models of object recognition, postulated object-centered (view-point independent) coding of objects as an efficient way of storing object information in memory, because only one such description would be sufficient (Biederman, 1995; Goodale & Milner, 1992; Marr, 1982; Marr & Nishihara, 1978; Wallis & Rolls, 1997). Such coding would enable the neural system to achieve object constancy, which facilitates object recognition. Therefore, it has been somewhat puzzling as to why large proportions of cells in anterior sections of STS that use an object-centered frame of reference. In addition, it shows that object-centered coding is even more specific and precise than previously thought. The collection of data from a relatively large number of cells coding in an object-centered manner also allows us to investigate questions that were beyond the scope of the previous studies. Questions about the response specificity and about the uniformity of the distribution of object-centered cells along the length of the STS. The present study indicates that the object-centered cells can be tuned to highly specific action-segments, and that they are more prevalent at more anterior positions in the STS. We will discuss the functional significance of the proposed accumulation of categorical spatial coding toward the pole of the temporal lobe, and relate the findings to existing ideas about the role of the STS in action understanding and social cognition.

2. Methods

2.1. Subjects and physiological procedures

The experiments were performed on two awake rhesus macaque monkeys (Macaca mulatta, age 4-6 years). A detailed description of the surgical procedures can be found elsewhere (e.g. Oram & Perrett, 1996). Animal care and experimental procedures were performed in accordance with UK Home Office guidelines.

2.2. Recording

Single cell recordings (using standard methods, see Oram & Perrett, 1996) were made while the monkey was seated in a primate chair. Spikes were captured online onto a PC (CED1401plus and Spike2 software, Cambridge Electronic Design, UK). Additionally, spikes were stored on an audio track of a HiFi video tape recorder. The stimulus events (seen from the subject’s perspective) were recorded with a video camera, and stored simultaneously on the video track of the same tape. Eye movements were recorded with a 2nd (infra-red sensitive) camera mounted onto the primate chair. The signals from the two cameras were integrated (Panasonic, VHS video mixer, WAVE7) prior to recording. The signal from the eye camera was also recorded separately on a 2nd video tape recorder, synchronised with a time-code generator and frame counter (VTIC, Horia VG50), for offline analysis of eye position (lateral, Sensorometric Instruments, Germany). Trials in which the eyes were closed were discarded. Analysis of the eye positions indicated no relationship between the pattern of fixation and trial type or response magnitude (cf. Jellema & Perrett, 2003a).

2.3. Stimuli and testing procedures

Visual stimuli were presented either on film projected onto a screen at life size, or live from behind a fast rise-time liquid crystal shutter (aperture 20 cm by 20 cm at a distance of 15 cm). In some cases a mechanical shutter with a larger aperture was used to avoid narrowing the scope of view of the subject. The usual procedure was that initial searching for response specificity was done with real body presentations. That is, the experimenter showed a wide variety of (non-articulated) body postures/actions in front of the subject. After the general response specificity had been determined in this way, more confined real and video stimuli were examined. The video stimuli tended to yield results similar to the real stimuli although usually at slightly reduced spiking frequencies (see Fig. 2 for a direct comparison of responses obtained to real and video presentations of walking apes). The real stimuli were shown at 1.5–4 m distance from the subject. Retinal images of live presented bodies varied from about 67° × 23° (vertically × horizontally, at 1.5 m distance), to 26° × 9° (at 4 m distance). Five to 12 repetitions were tested per stimulus condition in pseudo-random order. Control stimuli included objects of comparable size moved in comparable ways at the same velocity (e.g. a box attached onto the end of a stick, making rotational
motions, served as a control for head and/or upper body rotations). The best control stimuli, however, were formed by other actions and postures, performed by the same human agent, because they resembled the effective actions more closely than any non-living control object could ever do, yet evoked significantly smaller responses.

Once the particular action the cell was tuned to was defined, the frame of reference used by the cell to describe that action was determined by presenting the same action at different perspective views. For all actions the four main views (front, left and right profile and back view) were presented with the principal body axis oriented vertically. These different views clustered the distribution of views preferred by cells (Perrett et al., 1991).

For about half of the cells the perspective views were additionally presented with the principal axis oriented horizontally, or 180° inverted. Two populations of cells were examined in detail: one population responding to upper body rotations and one to walking actions. Testing procedures were as follows. For the cells sensitive to upper body rotations, the total trajectory that the upper body could make was divided into four sub-trajectories, which were presented separately: (1) rotation starting in a bent forward posture and ending in an upright position, (2) rotation starting in the upright position and ending in a bent backward posture, (3) rotation starting in a bent backward posture and ending in the upright position, and (4) rotation starting in the upright position and ending in the bent forward posture. Each of the sub-rotations extended over approximately 45° (see the schematic drawings at the top of Fig. 1). The actions were routinely performed at four different perspective views: front view, left and right profile view, and back view. Eighteen of the 43 cells of this population were further tested with actions 1 and 4 performed with the principal body axis oriented horizontally. These horizontal rotations were presented at two different viewing perspectives: front view (both clockwise and anti-clockwise rotations) and back view. Presentations were performed live, i.e. the vertical rotations were performed by the experimenter while sitting on a rotating chair 1.5 m in front of the subject (see inset to the left of Fig. 1). The actions with the principal axis oriented horizontally were accomplished by the experimenter lying on a table in front of the subject. It is important to realize that as a consequence of changing the viewing perspective, the direction of the rotations changed with respect to the subject.

The cells sensitive to walking actions were tested with real walking actions in the four basic directions (toward, away from, to the left and to the right of the subject). Walking was performed both forward (i.e. compatible walking, in which the direction of motion matches the direction that the body points) and backward (i.e. incompatible walking, in which the direction that the body points

Fig. 1. Responses of a cell which used an object-centered frame of reference to code for upper body rotation over a restricted section of the rotation trajectory. (A) Mean responses (+1 S.E.) are shown. The cell responded vigorously to the sight of the experimenter rotating the upper body forward, starting at the vertical upright position and ending at a 45° bent forward posture (actions 1 and 4, thin and thick black arrows, when seen from any of four perspectives (front, back, right and left profile views), with the principal body axis oriented upright (a–d, action 4 > actions 1, 2 and 3, P < 0.0001). Actions 1, 2 and 3 did not excite this cell from any of the perspectives. The response selectivity for action 4 persisted when the principal body axis was oriented horizontally (e–g, action 4 > action 1, P < 0.0001). The drawings at the left of the graph show a plan view of the subject and the experimenter (drawn to scale, 1.5 m distance subject–agent) for the front view. (B) Interaction of head and body articulations. The cell’s responses to the upper body rotations could be modulated by simultaneous rotation of the head (indicated by small open arrows). The data represent the averaged responses over the four vertical perspective views (vertical body orientation; left, right, front and back views). The large response to action 4 was significantly reduced when the head was simultaneously turned upward (P < 0.002, Newman–Keuls), whereas the small response to the ineffective action 2 was significantly increased when the head was simultaneously turned downward (P < 0.004). The drawings at the top show the agent at left profile view; the figure in interrupted outline represents the starting position, the solid outline represents the final position.
is opposite to the direction of the motion). In addition, videos of walking agents were presented upside-down, giving the impression of walking along the ceiling. The films were projected at life size onto a screen.

2.4. Data analysis

Spikes from multiple single neurons were routinely discriminated on- or off-line using template matching (Spike2, Cambridge Electronic Design, UK). Spike counts were performed during presentation of the stimuli, usually at 1 s intervals. Cell responses were analysed using ANOVAs, Newman–Keuls post hoc testing and t-tests (significance level at \( P < 0.05 \)).

2.5. Cell localization

A detailed description of the cell localization procedure can be found elsewhere (Jellema, Baker, Wicker, & Perrett, 2000). At completion of each experiment, frontal and lateral X-ray photographs were taken with the electrode still in place, to locate the electrode and the recorded cells with respect to specific bone landmarks. After the final experiment, an electrolytic microlesion was produced at the site of recording. The subject was then sedated and given a lethal dose of anaesthetic. After transcardial perfusion the brain was removed, coronal sections (25 μm) were cut, photographed and stained. The X-ray photographs were aligned with the histological sections to determine the cell locations (accuracy =1 mm). Histological reconstructions of cells responding in a viewer- or object-centered manner to walking or bending actions are shown in Fig. 4 for one subject. All cells were localized in the upper bank of the STS, between 9 and 21 mm anterior to the inter-aural line.

3. Results

3.1. Object-centered coding for the sight of articulating body parts

A total of 272 visually responsive STS cells were tested in 2 subjects for their tuning to bodily articulations, a wide range of bodily actions, and motions of non-living control objects. Of these 272, 208 cells responded significantly more to 1 specific bodily action compared to any of the other actions and the control stimuli. The remaining 64 cells responded less specifically to a variety of motions and were discarded. Specific actions that the cells were tuned to included limb articulations such as arm reaching or leg extension, head rotations, upper body rotations and bowing. In such actions the rest of the body would remain static, but for other cells the effective action involved motion of the body as a whole in addition to articulations of body parts, as e.g. in walking and jumping.

The present study focuses in particular on two object-centered cell populations: 38 cells responsive to forward or backward rotations of the upper body with respect to the lower body, and 43 cells responsive to walking actions. These cells were singled out because they were found relatively frequently, which allowed for extensive testing and manipulation.

3.2. Object-centered coding for bending actions of the upper body

A population of 38 cells in STSAs (the anterior part of STS) were sensitive to either a forward or a backward rotation of the upper body with respect to the lower body (decreasing and increasing the angle between the chest and the upper leg, respectively). Cells have previously been described in STSa that responded specifically to rotational movements of objects such as limbs, in which one end of the object is fixed while the other end rotates about the fixed end (Perrett, Harries, Benson, & Chitty, 1990). The cells responding to the upper body bending forward or backward across a restricted trajectory as described here, may be of this form.

The vast majority of the cells of this population (32 out of 38, 84%) responded to a bending action of the torso irrespective of the perspective view (with principal body axis oriented vertically). Of the 18 cells tested with horizontal rotations, 15 cells maintained their selectivity to the particular action that was effective in the vertical rotations, while only three cells stopped responding to the previously effective action.

Six of the 38 cells tested did not use an object-centered frame of reference, they showed preference for several but not all of the perspective views. Three of them responded to a specific rotational action at all perspective views except the front view, one cell responded to all perspective views except the back view, and the two other cells responded to a particular bending action at two perspective views but not to the other two views. The responsiveness of these six cells could be envisaged to reflect a transition from a purely viewer-centered frame of reference (just one perceptive view) to a truly object-centered one.

The 32 cells using the object-centered frame of reference typically responded to just one of the four sub-actions mentioned above at each of the perspective views, with the two actions directed away from the vertically axis (i.e. actions 2 and 4) most often responding to. Twenty-two cells responded to just action 2 (no response to actions 1, 3 and 4), four cells responded to just action 4 (no response to actions 1, 2 and 3). Some of the cells that responded to action 1 also responded somewhat to action 2 (four cells), and some responding to action 3 also responded somewhat to action 4 (two cells). Note that actions 1 and 2 are consecutive actions in the same direction covering the full 90° rotation, as are actions 3 and 4.

Fig. 1 illustrates a typical cell of this population, which coded in an object-centered manner for the sight of action 4 (forward articulation of the upper body), irrespective of perspective view (Fig. 1a–d). A two-way repeated measures ANOVA with view (front, left, right and back) and motion (actions 1, 2, 3 and 4) as within-subject repeated measures variables showed a significant main effect for motion \( F(3,15) = 95, P < 0.0001 \), reflecting the preference this cell had for action 4, no significant main effect for view \( F(3,15) = 1.8, P = 0.18 \), reflecting the generalization across viewpoints, and no significant significant interaction effect \( F(9,45) = 24.2, P = 0.08 \). The preference for action 4 was maintained when the principal body axis was oriented horizontally (Fig. 1e–g). At the horizontal orientation only actions 1 and 4 were presented, and were seen only from the front and back views. A three-way repeated measures ANOVA with type of motion (actions 1 and 4), orientation (upright and horizontal) and view (front and back) as within-subject variables, confirmed the maintained preference for action 4 across orientation, indicated by a significant main effect for motion \( F(1,5) = 117, P < 0.0001 \). There were no significant main effects for view \( F(1,5) = 9.2, P = 0.51 \) or orientation \( F(1,5) = 1.0, P = 0.36 \).
reflecting the generalization over view and orientation. All inter-
actions were non-significant ($P > 0.15$).

The direction of motion per se did not determine the observed
responses: action 4 seen from the front view was directed toward
the observer, while the same action 4 seen from the back view
was directed away from the observer (while at profile views the
motion direction was to the left and right of the observer), yet
action 4 was effective from any of these views. Thus the action
was defined not with respect to the observer, but with respect to
a reference point located in the object (the moving body) itself,
which is the hallmark of object-centered coding.

To further pinpoint the exact nature of the action being coded,
we assessed whether the rotation of the whole upper body was
strictly necessary, or whether similar rotations of just the head
could produce similar responses. Twenty-three cells responding
to upper body rotations in an object-centered manner were tested
with rotations of just the head in the four main (vertical) perspec-
tives. The majority of cells (16/23, 70%) responded significantly
less to just head rotations compared to upper body rotations. The
head rotations, however, did usually produce responses signif-
icantly above control levels. Of the remaining seven cells, six
cells responded equally well to both types of head and body
rotation, one cell responded better to head than body rotations.
This suggests that the cell population as a whole represented
rotations of the upper body with respect to the lower body.

This is not to say that head direction was not an important
factor in controlling cell responses, for cell responses could be
modulated by head direction (Fig. 1B, left). Six cells were tested
with actions 2 and 4 while the head was simultaneously rotated
horizontally as much as possible. Thus, during forward torso
rotation from vertical (action 4), the initially forward and down-
ward looking head was seen to rotate upward (see schematic
drawing at top of Fig. 1B); during the backward rotation from
vertical (action 2), the initially forward and upward looking head
was seen to rotate downward. This manipulation had a dramatic
effect on the responses as shown in Fig. 1B. The large response
to forward bending (action 3) was significantly reduced when,
simultaneously, the head was turned upward, while the very
weak response to backward bending was significantly increased
by simultaneously turning the head downward. Similar results
were obtained for each perspective view (found in all six cells).
This finding allowed us to further narrow down the response
selectivity: it not only needed the upper body to move toward
the lower body, it had to be done in such a way that the head was
facing the lower body.

An object-centered frame of reference would in many cases
be an efficient way to code for the direction of attention of an
individual. For example, attention directed upward towards the
sky can equally well be typified by an upward directed head seen
from any perspective view (although the face view may be most
informative because of the visibility of the eyes). However, the
target of attention does not necessarily have to be located in the
agent’s environment, the target can also be a part of the agent’s
own body. Could the cells of this population have been coding for
the direction of attention, as conveyed by the direction the head
pointed in, either toward/away from the ground, or, alternatively,
toward/away from the lower body? The findings are compatible
with the notion that the cells were coding for attention directed
increasingly toward/away from the lower body. The findings are
not compatible with attention directed at the ground or the sky,
since the cells maintained their response selectivity during the
horizontally oriented trunk rotations. The finding that backward
rotation of the head (so that head and body pointed in opposite
directions) could enhance the weak response to an ineffective
action and, conversely, could weaken the strong response to an
effective action (Fig. 1B) is also in agreement with this idea.

3.3. Object-centered coding for walking actions

Cell populations have previously been described in STS that
specifically respond to whole body movements, such as walk-
ing, rather than body part articulations (Jellema & Perrett, 2002;
Oram & Perrett, 1994, 1996). Many of these cells are respon-
sive to a particular combination of a view and a direction of
movement, and thus code in a viewer-centered coordinate sys-
tem. About one-third of cells that respond selectively to either
forward or backward walking maintain this response selectivity
when the real walking agent is replaced by Johansson point-light
walking motions (in which the form of the body is defined only
by light patches attached to points of limb articulation) (Oram &
Perrett, 1994). This indicates that the majority (2/3) of the cells
requires real form information. The viewer-centered cells may
respond to the experimenter walking forwards to the right, but
not to the experimenter walking forwards to the left, nor to the
experimenter walking backwards to the left or right. However,
in the present study at more anterior sites in STS, we report a
relatively large proportion of cells responding to walking actions
using an object-centered frame of reference.

Forty-three cells were found to be sensitive to whole body
walking actions. Of these, 16 cells used an object-centered frame
of reference, and 27 a viewer-centered frame. This is again a high
percentage of object-centered coding (16/43, 37%) compared
to previous studies (which typically found 5% of cells coding
object-centered). Eight of the 16 object-centered cells responded
to forward walking, the other 8 responded to backward walking,
at each of the 4 main perspective views (toward, to the left and
right, and away from the subject). None of the cells responded
to both forward and backward walking. Seven cells respond-
ing to walking were further tested with upside-down walking
actions. All seven cells maintained their selectivity for walking
compatibility in the upside-down condition. The relatively large
proportion of cells responding to backward walking may seem
surprising given how unusual backward walking is in human
societies. However, in non-human primate societies it is much
more common, as is true for actions carried out upside-down.

One of these seven cells is illustrated in Fig. 2. This cell
responded vigorously to backward walking, and hardly to for-
ward walking (left half of Fig. 2). In the upside-down condition
(right half of Fig. 2), the discrimination between backward
and forward walking was maintained, be it somewhat less pro-
nounced (especially in the direction toward the subject). Note
that neither body view, nor direction of motion, could explain the
responses, since these were identical for backward and forward
walking, yet only backward walking excited the cell. A three-
way repeated measures ANOVA, with as within-subject factors type of walking (forward or backward), orientation (upright or upside-down) and perspective view (front, left profile, right profile, back view), showed a significant main effect for type of walking ($F(1,3) = 895, P < 0.001$), reflecting the preference of this cell for backward walking. There was no main effect for orientation of walking ($F(1,3) = 0.24, P = 0.66$), indicating that the cell did not discriminate between upright and upside-down backward walking, and thus indicating that gravity in the stimulus scene did not affect the responses of the cell. There was no main effect for view ($F(3,9) = 0.56, P = 0.65$) indicating the cell did not discriminate between the four perspective views of the agent. The two-way interaction of type of walking x orientation was significant ($F(1,3) = 712, P < 0.001$), reflecting the less rigorous discrimination between forward and backward walking in the upside-down orientation compared to the upright orientation. The two-way interaction of orientation x view was also significant ($F(3,9) = 12.7, P = 0.001$), reflecting the reduced response to the back view in the upside-down orientation. All other interactions were non-significant ($P > 0.25$).

3.4. Object-centered coding for other actions

In addition to the bending and walking actions, a variety of other actions were coded for in object-centered manner, however, cells specifically tuned to the sight of these other actions were much less frequently found. The actions included rotations of the head over the shoulder, arm extensions and knee flexions/extensions. Fig. 3 gives an example of responses of a cell of the latter category, some of which responded exclusively to flexing of the knees, others exclusively to extending them. The cell responded when the experimenter, standing in front of the subject at 2 m distance, flexed the knees, while keeping the upper body upright (as illustrated in schematic drawings at the top of Fig. 3 a–d). This action caused a vigorous response when seen from each of the four main perspectives (a–d). The opposite action, i.e. straightening the knees resulted in a much reduced response compared to the responses in (a)–(d). The orientation of the main body axis is horizontal (achieved by the experimenter lying on a table). (g) The feet never lost contact with the wall throughout the bending and straightening cycles (see text for details). In this condition the cell maintained the selectivity for knee bending. (h) As (g) except without the feet touching the wall (minimally 50 cm distance between feet and wall). In this condition the cell did not respond at all. Up, upward motion by means of straightening of the knees; Down, downward motion by means of flexing of the knees.
ing in the straight upright posture, hardly produced a response. A two-way repeated measures ANOVA with type of motion (flexing and straightening) and perspective view (front, left and right profile and back view) as within-subject variables, showed a significant main effect for type of motion \((F(1,4) = 678, P < 0.0001)\), reflecting the preference for the flexing action, but no significant main effect for view \((F(3,12) = 1.4, P = 0.28)\), reflecting the lack of preference for a particular view, and no interaction effect between motion and view \((F(3,12) = 52.4, P = 0.08)\).

The simplest explanation for the cell’s responsiveness would of course be that it responded to an object moving downward. Therefore, a number of additional manipulations were performed. Downward (and upward) motion of a box attached to the end of a stick (Fig. 3e) did not produce any responses different from spontaneous level. Thus, downward motion per se could not explain the responses. One could argue that since STS cells are typically tuned to the actions of articulating animate objects, the box did not work because it was not an animate object. Therefore, we presented an agent jumping from a 40 cm high elevation while keeping the knees straight. This constituted a lowering of the body without knee flexing, and produced a very much reduced response (as well as sore knees for the agent; Fig. 3f) compared to the body lowering through knee flexing (a–d). From this it followed that just the downward motion of the experimenter was also unable to explain the responses of this cell. We next manipulated the orientation of the main body axis, to find out whether the cell coded truly object-centered for knee flexion, and whether gravity played a role (e.g. did the cell respond only to knee flexion that reduced the distance between the upper body and the earth?). Knee flexion with the principal body axis oriented horizontally was achieved by the experimenter lying on a mobile table, at 2 m distance from the subject. Two different versions were presented. In the first, the feet made contact with the wall of the testing room throughout the flexing and straightening cycles. In this scenario, knee extension was achieved by exerting force onto the wall, thereby pushing the mobile table (with the experimenter on top) away from the wall. To achieve knee bending, a second experimenter (who remained out of sight) pushed the table towards the wall. In the second version, the feet never touched the wall (50 cm minimal distance between feet and wall). The experimenter simply flexed and straightened the knees in the air just above the surface of the table. Interestingly, only knee bending with the feet maintaining contact with the wall produced a response (Fig. 3g), while knee flexing without the feet making contact with the wall produced no response at all. This allowed us to further define the response specificity. Apparently, the cell responded to the upper body moving toward the substrate the feet made contact with (floor or wall) while the motion toward the substrate had to involve a flexing of the knees. Gravity may, however, still have played a role because the response to horizontal knee bending was significantly smaller than to its vertical counterpart \((P < 0.001, t\text{-test}, \text{comparing the responses to the front views in a and g})\).

The responses cannot be explained by a shortening of the body because bending of the knees in the air (body axis horizontally) shortened the body similarly to the vertical actions, but did not evoke a response. Moreover, shortening of the body with the principal axis oriented vertically by means of rotating...
the upper body forward towards the lower body also failed to evoke responses (data not shown).

An interesting aspect of these cells was that they combined sensitivity to the relative positions of body parts with sensitivity for environmental cues, i.e. the substrate with which the feet made contact. Feet and substrate had to be seen to make contact otherwise the cell would not respond. This is reminiscent of STS cells described in earlier reports responding for example to the sight of a hand stroking a piece of fur, but not to the same action performed a few centimeters above the fur (Perrett et al., 1989). The combination of object-centered coding for an action and sensitivity to the object the action is directed at, confers a special quality to the cell. It enables the cell to code for an action in a goal-centred frame of reference (cf. Perrett et al., 1991).

The cells sensitive to walking actions (Section 3.3) were not tested with walking in which the feet did not make contact with a substrate (which might be achieved by presenting films in which the background is digitally removed). Therefore we cannot assess the necessity of a substrate for the walking-sensitive cells. For the cells sensitive to upper body rotations (Section 3.2) we know that the feet were not required to make contact with a substrate, since in the rotations where the principal body axis was oriented horizontally the feet were moving freely in the air, yet the responses were not reduced in magnitude compared to the actions where the feet made floor contact.

3.5. Anterior–posterior gradients for object- and viewer-centered coding in the STS

The relative high incidence of cells in the STS coding in an object-centered manner for actions found in the current study, compared to previous studies, may be related to the more anterior position of the recording tracks in the current study. The data in both subjects showed a trend towards an increase in the number of object-centered coding cells at more anterior positions, and a trend towards an increase in the number of viewer-centered cells at more posterior positions. These data, however, have to be taken cautiously, since very few studies exploring the neural representations of actions looked this far anterior in the STS. Fig. 4 shows the locations of the cells responding to walking and upper body rotations described in this report within consecutive coronal sections of the STS for one subject. In addition to the tendency for object-centered cells to be more prevalent at anterior sites, a striking feature is that all of the cells were located in the upper bank of the STS (the recording tracks always sampled from both the upper and lower banks of the STS). With the accuracy of 1–2 mm in cell localization, we cannot with certainty exclude the possibility that some of them were located in the lower bank, but the tendency for these action-sensitive cells to be in the upper bank is clear.

4. Discussion

4.1. Object-centered coding in the STS

The cells described in this report coded for the relative positions of specific parts of an (animate) object occupied with respect to either other parts of the object, the object’s principal axis, or cues in the environment, rather than with respect to the observer. The frame of reference used by these cells therefore is of the object-centered, or categorical, type, which typically describes the spatial positions of objects (or parts of objects) in terms of their relative position to other objects (which may be parts of the same object) (Kosslyn, 1987).

The relatively high incidence of object-centered coding in the anterior part of the STS, as reported in the current study, prompts several questions. First of all, why did previous studies in the STS not find them? These earlier studies suggested viewer-centered coding to be by far the preferred way of coding for STS cells. For example, the incidence of cells using an object-centered frame to code for walking never exceeded 5% of the total number of walking sensitive cells (Oram & Perrett, 1996). We propose that object-centered coding is more prevalent at anterior locations in the STS. The main argument for this is that current AP coordinates ranged from 9 to 21 mm anterior to the inter-aural line, while about 60% of cells coding for bodily actions did so in an object-centered way, whereas AP coordinates in the previous reports ranged from 0 to 10 mm anterior to the inter-aural line, with typically 5% of cells coding object-centered (Perrett et al., 1991; Wachsmuth et al., 1994). Furthermore, within the 9–21 mm AP range of the current study, we found a trend towards more object-centered coding at anterior locations, although this did not reach significance. This issue requires further investigation on more subjects, also encompassing cells responding to other types of actions and postures, before firm conclusions can be drawn.

Postulating object-centered coding in the anterior STS suggests that the anterior STS occupies a higher position in the hierarchical flow of information (of form and motion) in the temporal lobe than the posterior STS, since object-centered coding is thought to be the culmination of a pooling process of viewer-centered representations (Perrett et al., 1991). The exact flow of information through the temporal lobe is, however, still being debated. It has been suggested that the feed forward processing hierarchy of form information consists of an initial sweep within the inferior temporal lobe towards the pole, which then doubles back in posterior direction through STS, while motion information would feed forward from area MT (medial temporal) via posterior STS to anterior STS. In an alternative view, the initial flow in the inferior temporal lobe projects at multiple stages onto different areas of the STS (Wallis & Rolls, 1997). Two lines of inquiry seem to suggest the anterior STS sites represent a later stage in the processing chain than the posterior sites. (1) Latencies of cells in the STS tend to be the longest at anterior sites (Baylis, Rolls, & Leonard, 1985; Perrett, Hietanen, Oram, & Benson, 1992). (2) The anterior part of the STS, not its posterior part, provides the main output projections onto e.g. the orbitofrontal cortex (Barbas, 1993).

4.2. What could be the functional role of categorical object-centered coding in the anterior STS?

We will discuss here three potential functions these cells may contribute to: (1) the recognition of non-rigid, animate, objects,
(2) the forming of goal-centered descriptions of actions, and (3) the ability to imitate others’ behaviour.

4.2.1. The role of object-centered coding in the recognition of animate objects

This report focused on STS cells coding in an object-centered manner for animate postures and actions. Whereas rigid, non-living objects have constituent parts that remain at fixed locations with respect to each other, the flexible, animate objects can adopt a wide range of different poses, due to bending/rotation in their multiple joints. One classical problem in object recognition, which pertains to both types of objects, is that objects project hugely different shapes onto our retinas when seen from different viewpoints, yet all these different shapes have to be recognised as belonging to the same object. However, in the case of animate objects, this problem is further exasperated by the different poses the objects can adopt (cf. Laeng, Shah, & Kosslyn, 1999). There are basically two major accounts that aim to explain object recognition. In one account the representation of an object would contain a series of characteristic perspective views of the object (e.g. Koenderink & Van Doorn, 1979; Logothetis et al., 1995; Tarr, 1995). Novel viewpoints of familiar objects are recognised by transformation, or alignment to the nearest stored view. However, it has been argued that in the case of flexible, animate objects, this account is insufficient, as it cannot deal with the huge variety of unusual poses an animate object can adopt. Several authors have therefore postulated that the animate object’s representation should include an object-centered structural description of qualitative properties of the object (such as head and arms are connected to one end of the trunk, the legs to the other end), in addition to descriptions of the constituent parts separately and their spatial relationships (Laeng, Shah, & Kosslyn, 1999; Laeng, Carlesimo, Caltagirone, Capasso, & Miceli, 2002; Marr, 1982). Such structural descriptions are called categorical, as opposed to coordinate descriptions, which rely on precise spatial positions (Kosslyn, 1987, 1994).

To what extent does our single-cell data from the STS support any of these accounts? The properties of the object-centered cells in the anterior STS may help the forming of the postulated abstract structural object description (cf. Laeng, Shah, & Kosslyn, 1999). Such a description should be viewpoint-independent to enable its abstract nature. It should be noted, however, that categorical spatial descriptions, as defined by Kosslyn (1987, 1994), are not exclusively of the object-centered type, but can also be viewpoint-centered. Both types of descriptions are computed in the STS, and although the viewpoint-centered descriptions are needed to form the latter (Perrett et al., 1991), the viewer-centered ones are likely to serve purposes in their own right as well. They could support the global recognition of animate objects in familiar poses, which does not rely on an abstract structural description. While the above model requires relatively large numbers of object-centered cells, object-recognition through interpolation or transformation of stored canonical views does not seem to require them at all.

How does the postulated involvement of these STS cell populations in animate object recognition using categorical descriptions relate to the activation of parietal circuitry found in studies during categorical processing? One possibility is that animate objects in unusual poses require the observer to make several attentional saccades (overly or covertly) towards conspicuous details of the object in order to identify it, hence the involvement of the parietal supplementary eye fields, whereas objects in familiar poses would not require this scanning process.

The issue of whether the hemispheric lateralization with respect to categorical (left hemisphere) and coordinate (right hemisphere) spatial representations reported in the human brain (Kosslyn, 1987, 1994), is also present in the macaque monkey brain is still largely unknown. Single cell studies in our lab revealing categorical coding were usually conducted in the left hemisphere, but this topic remains to be investigated systematically.

4.2.2. The role of object-centered coding in the forming of goal-centred representations of actions

Though the object-centered descriptions serve purposes in their own right, at the same time they may serve as ‘building blocks’ to help form goal-centered descriptions. The latter are descriptions that combine an action with its goal, such as in reaching for a cup or walking to a door. Thus, the disposition or movements of an agent are specified with respect to a second object or part of the environment (Perrett et al., 1989). For a goal-centered description, the exact perspective view and nature of the action is of secondary importance to the interaction between action and object/goal, because there may be many varieties of a particular action leading to the same goal. It would indeed be far less efficient for a neural system to have to associate the object/goal with every single perspective view of an action. Associating the goal with an object-centered description of an action offers an efficient and economical solution. Cells coding in a goal-centered manner have been identified in the STS (Perrett et al., 1989, 1991), and also in area F5 in the premotor cortex (mirror-neurones) (Di Pellegrino, Fadiga, Fogassi, Gallese, & Rizzolatti, 1992; Keysers & Perrett, 2004).

Goal-centered descriptions often involve sensitivity to causality. Cells sensitive to causality include e.g. those that responded to the sight of a hand manipulating an object, but not to just the hand action without the object present, nor to the (moving) object without the hand action (Perrett et al., 1989). Moreover, physical contact with the object was required: mimicking the action a few centimetres above the object failed to excite the cells. Such sensitivity for spatial relations between the agent and target puts the cells in a position to signal causal relations. The cells reported in the current study that responded to knee flexion provided the feet made contact with a supporting substrate belong to this category. The sensitivity of the cells responding to walking and upper body rotations seemed to be confined to spatial relations within the object, and did not include environmental cues. These cells therefore can provide information on the manner in which an action is performed.

Another aspect of goal-directed descriptions is the intention of the actor. To code for the intention it is irrelevant whether or not the goal is achieved. The forming of goal-centred represen-
tations of actions may be seen as a step toward the realization of representations of others’ intentions. In this view, the proposed accumulation of object-centered coding toward the poles of the temporal lobes in the macaque monkey is especially interesting, since in the human brain the poles of the temporal lobe have specifically been associated with mentalizing abilities, i.e. the attribution of intentions to others (cf. Frith, 2001).

4.2.3. The role of object-centered coding in imitation

When imitating others, attempts to reproduce the goal of their actions can be more efficient than attempts to reproduce the precise movements that are witnessed. After all, the agent performing the action may fail to achieve the intended goal, but the observer who understands the intention or goal of the action can imitate the intended act and attain the intended goal. The generalization over perspective views in the visual coding of actions leading to object-centered, and goal-centered, representations that we described in this report is well suited to support the imitation of behaviour. When an agent demonstrates a particular action, the particular vantage point of the observer is arbitrary. To imitate the action, the same motor pattern must be triggered whatever the perspective view.

Visual information about actions is not only encoded in the macaque’s temporal cortex but also in premotor cortex. Cells in area F5 of the inferior area 6 of the macaque’s premotor cortex respond to the sight of particular actions such as grasping, holding, tearing and manipulating, executed by another monkey or by the experimenter (Di Pellegrino et al., 1992; Gallese, 1996; Gallese & Goldman, 1998; Rizzolatti, Fadiga, Gallese, & Fogassi, 1996). The STS may well provide input to the premotor cortex through projections that relay in parietal cortex (Keysers, 2004). Unlike the temporal cells, the premotor and parietal cells also respond when the subject itself executes an identical action, hence their name ‘mirror-neurones’. Many of these premotor cells seem to use an object-centered or goal-centered frame of reference, rather than a viewer-centered one, and they have been proposed to play a role in imitation and mindreading (Gallese & Goldman, 1998). In the human brain, the sight of real hand movements also activates the premotor cortex and left STS (Grafton, Arbib, Fadiga, & Rizzolatti, 1996; Rizzolatti et al., 1996).

However, there is little behavioural evidence from observational studies or laboratory experiments that monkeys imitate or comprehend the goals of actions of others (Byrne, 1995; Gardiner & Heyes, 1998; Whiten & Ham, 1992, though see Bugnyar & Huber, 1997), in contrast to great apes, who do show evidence for imitating the goals of actions demonstrated visually by humans (Custance, Whiten, & Bard, 1995). Many instances of so-called imitative behaviour made by monkeys may be explained as instances of response facilitation or stimulus enhancement with trial and error learning (Byrne, 1995). Therefore, one could envisage that, in concert with premotor cortex, the object-centered coding in STS supports rudimentary forms of social learning and imitation in monkeys. These rudimentary forms are then the precursors of imitation behaviour as seen in the great apes and humans.

4.3. Categorical spatial relationships within and between objects

In a previous study we described categorical spatial coding in the STS for the location of agents with respect to the relative spatial location in the testing room (Jellema, Maassen, & Perrett, 2004). In contrast, the current study looked at categorical coding for the relative positions of different parts of the same animate object. Although for the knee flexion sensitive cells in the current study the relative position of the object with respect to environmental cues was crucial, the spatial location in the testing room per se was irrelevant. This was true for all cells described here, whereas in the previous study the location in the testing room was a decisive quality. For example, in the previous study some cell populations were found to respond selectively to an agent walking at a position close to the subject, while the same action, same body view and same motion direction, but presented at a different location in the room (e.g. at middle or far locations) did not excite the cells. Other cell populations would respond exclusively to the agent performing the walking action exclusively at the middle or far location in the room. These results can in principle be explained either by an ego-centric frame of reference (specifying the location of the agent with respect to the subject) or by an allocentric frame of reference (specifying the location of the agent with respect to markers in the environment, e.g. the door of the room). As yet, we cannot discriminate between them; to do so would require manipulation of the position/orientation of the object while keeping the agent at fixed locations. Whether ego- or allocentric, the spatial coding in STS seems to rely on the relative, categorical, positions of the (animate) objects, rather than on their precise, coordinate, positions (cf. Koslyn, 1987). Such categorical coding may be achieved by multiplying the spatial code (ego- or allocentric position) specifying where the agent is located by the action code for what the agent is doing (specified in a viewer-, object- or goal-centred way).

4.4. The importance of categorical spatial coding in the STS

Categorical spatial coding in the STS makes a lot of sense when we consider the proposed function of the computations performed in the STS, which is the detection of sequences of actions/postures and their consequences, so as to make predictions about the most likely next event and to determine the goals of these actions (Jellema & Perrett, 2002, 2003a,b, 2005). Such a system would be essential to the detection of intentional behaviour (Saxe, Xiao, Kovacs, Perrett, & Kanwisher, 2004), the prerequisites of which seem to be present in the macaque STS (Jellema et al., 2000). The relative spatial positions, either between parts of one and the same object, or between different objects, rather than their precise, coordinate relationships, are of decisive importance for determining the social relevance of the objects. The STS is generally considered to be one of the pillars of the so-called social brain (Allison, Puce, & McCarthy, 2000; Frith, 2001), i.e. those distributed areas of the brain that specifically deal with the social relevance of stimuli. The spatial
positions of articulated body parts with respect to the body, or spatial relationships of to the observer and/or other agents or objects, provide information about perceived social attitudes such as approach or withdrawal, dominance or submission, and friendliness, cooperativeness or hostility. If intentions in the actions of others can be registered before the actions are completed, then that would give a perceiver an advantage in predicting behaviour in a competitive situation, and in social interactions more generally. The ability to predict behaviour allows for anticipatory responses. Therefore, categorical coding of the spatial relationships within and between animate objects plays an integral part in the analysis of social events in the STS.

There is evidence that object-centered representations also influence the allocation of attention, coming mainly from visual neglect studies. For example, when a patient with spatial neglect ignores certain parts on the left side of an object, then, after rotation of that object around its principal axis so that these parts are now on the right side of space, this patient may continue to ignore these same parts now on the right side of space. Apparent, attention can access multiple frames of reference, of which is object-centered, although object-centered neglect is relatively rare (e.g. Behrmann & Tipper, 1999). On a speculative note, one could envisage that the object-centred-representations in STS feed into the parietal areas controlling eye movements and attention, so as to enable object-centered attention. Reciprocal connections between STS and the inferior parietal lobule exist (e.g. Hari & Perrett, 1991; Rizzolatti & Mattelli, 2003; Seltzer & Pandya, 1994), but this issue remains to be explored.

References


