Functional Differentiation of Macaque Visual Temporal Cortical Neurons Using a Parametric Action Space

Neurons in the rostral superior temporal sulcus (STS) are responsive to displays of body movements. We employed a parametric action space to determine how similarities among actions are represented by visual temporal neurons and how form and motion information contributes to their responses. The stimulus space consisted of a stick-plus-point-light figure performing arm actions and their blends. Multidimensional scaling showed that the responses of temporal neurons represented the ordinal similarity between these actions. Further tests distinguished neurons responding equally strongly to static presentations and to actions (“snapshot” neurons), from those responding much less strongly to static presentations, but responding well when motion was present (“motion” neurons). The “motion” neurons were predominantly found in the upper bank/fundus of the STS, and “snapshot” neurons in the lower bank of the STS and inferior temporal convexity. Most “motion” neurons showed strong response modulation during the course of an action, thus responding to action kinematics. “Motion” neurons displayed a greater average selectivity for these simple arm actions than did “snapshot” neurons. We suggest that the “motion” neurons code for visual kinematics, whereas the “snapshot” neurons code for form/posture, and that both can contribute to action recognition, in agreement with computational models of action recognition.

Keywords: biological motion, inferotemporal cortex, monkey, movement, single cell, superior temporal sulcus

Introduction

Single cell studies in macaque monkeys have found that neurons of the visual temporal cortex, particularly those of the rostral superior temporal sulcus (STS), respond to displays of body movements (Bruce et al. 1981; Perrett et al. 1985, 1989, 1990; Oram and Perrett 1994, 1996; Jellema and Perrett 2003a, 2003b, 2006; Jellema et al. 2004; Keysers and Perrett 2004; Barraclough et al. 2005). These studies suggested that some STS neurons respond selectively to different classes of perceived body movements, such as grasping, as opposed to locomotion.

In the present study we employed dynamic images of biological motion, allowing a quantitative analysis of the response selectivities of temporal neurons to such stimuli. In order to assess selectivity for visual actions, we employed a parametric action space whose design is similar to previous parametric static shape spaces that have been used to examine shape selectivities of ventral stream visual neurons (Op de Beeck et al. 2001; Pasupathy and Conner 2001; Freedman et al. 2003; De Baene et al. 2007). The action space was created by blending 3 different transitive arm actions: knocking, lifting, and throwing. The 3-way blends differed in the relative amounts to which each of the 3 basic actions contributed to the blend, producing systematic and smooth transitions between the 3 basic action stimuli. Because we were interested in the coding of visual actions, we reduced pattern information to its basics by using stick-plus-point-light figures instead of real actors. The stimuli were derived from actual motion-capture data obtained using real human actors. We used stick-plus-point-light figures instead of full-body figures because the latter contains additional texture, shading and form information, which might have complicated determining which stimulus properties the neuron was really selective for. On the other hand, we did not reduce the stimulus further to point-light displays (Johansson 1973), because we reasoned that stick-plus-point-light figures would produce stronger neural responses than point-light displays only.

Computational and psychophysical work suggests that stimulus similarity is an important factor in determining perceptual categorization (Nosofsky 1984; Ashby and Perrin 1988; Op de Beeck et al. 2001; Palmeri and Gauthier 2004). Thus, one would expect that if STS neurons contribute to action categorization, then their response to different actions would depend on similarities between those actions. Previous work showed that inferior temporal (IT) neurons rather faithfully represent the ordinal similarity relations between static shapes (Op de Beeck et al. 2001; De Baene et al. 2007). In the present study we asked whether this also holds true for dynamic images of actions. Actions are a more complex visual stimulus than static shapes, because visual actions inherently consist of spatio-temporal changes of form. It is thus an open question whether single neurons can code for similarities between actions and how this coding might evolve over the course of the action. Human psychophysical research employing the same stimuli that we have employed here has shown that the perceptual representation of similarities between action blends fits the parametric stimulus configuration rather well (Pollick et al. 2007), indicating that our stimulus set is well-suited for investigating the neural coding of action similarity.

A second set of questions that we examined relates to the relative contributions of form and motion information to the coding of actions. Because different actions usually differ in both form and motion, action coding can, in principle, be based on either form or motion information, or some combination of both. The potential contributions of form and motion information has been nicely shown by the computational model of action recognition proposed by Giese and Poggio (2003): they postulated an action-processing stream based on motion analysis, functioning in tandem with a parallel stream based on form information, present in the static snapshots of the action.
sequence. In order to determine the relative contributions of form versus motion to the neural responses to our action stimuli, we compared responses to the action movies and to static presentations of snapshots taken from those movies. In addition, we systematically reduced the stimulus (Tanaka 1996) by methodically removing limbs of the human figure until only the moving arm or the end-effector, that is, the wristpoint, was left. This allowed us to assess the degree to which the responses to the actions are driven by body configuration. The actions we employed can be well-characterized quantitatively because they are restricted to a single limb and most of the information concerning the action is present in the end-effector itself (Pollick et al. 2007). This allowed us to correlate the single cell responses to differences in form and motion parameters within the parameterized stimuli. Finally, we examined whether the neurons would respond to rigid translation of the representative static snapshots. Biological motion is essentially nonrigid and thus, by comparing responses to rigid translation and nonrigid motion, we could ascertain whether the neurons’ responses are specific for biological motion.

Bruce et al. (1981) and Perrett’s group (Oram and Perrett 1994, 1996; Jellema and Perrett 2003a, 2003b, 2006; Jellema et al. 2004; Barralough et al. 2005) observed responses to perceived actions in the upper bank of the rostral STS (superior temporal polysensory [STP]). This region contains neurons with large receptive fields (RFs) that respond better to moving than to static stimuli (Bruce et al. 1981; Baylis et al. 1987). STP neurons can be selective for motion direction (Bruce et al. 1981; Baylis et al. 1987; Oram et al. 1993) and complex motion patterns (e.g., optic flow: Anderson and Siegel 1999; structure from motion: Anderson and Siegel 2005) and thus is a candidate region for processing dynamic images of actions. Responses to perceived hand actions (Perrett et al. 1989) and locomotion (Barralough et al. 2005, 2006) have also been observed in the lower bank of the rostral STS, which is part of IT. Recent functional magnetic resonance imaging (fMRI) work in awake monkeys has reported activation for a hand grasping an object, compared with static presentations of snapshots of the action, in both banks of the STS (Nelissen et al. 2006). In light of this previous work, we searched for responsive neurons in the lower and upper banks of the STS, as well as in the lateral convexity of IT, and compared the responses properties of these temporal regions.

Materials and Methods

Subjects and Surgery

Two male rhesus monkeys (Macaca mulatta, monkey BR and LU) served as subjects. Before conducting the experiments, aseptic surgery was performed under isoflurane anesthesia to attach a head fixation post to the skull and to stereotactically implant a plastic recording chamber (Crist Instruments). The implantation of the recording chambers was guided using preoperative structural magnetic resonance scans. The recording chambers were positioned dorsal to the rostral temporal cortex, allowing a vertical approach. We recorded from both hemispheres of monkey BR; the recording chamber on this subject’s right hemisphere was positioned at anterior/posterior coordinates which were comparable to those of the other animal (monkey BR right hemisphere: 16 mm anterior, 24 mm lateral; monkey LU: 16 mm anterior, 21 mm lateral), whereas the recording chamber on the opposite hemisphere was located more posteriorly (monkey BR left hemisphere: 10 mm anterior). The anterior-posterior range of the recordings across animals was from 9 to 18 mm anterior to the external auditory meatus. At several points during the course of the recordings, we took anatomical MRI scans (3 Tesla—1-mm resolution), with copper sulfate-filled tubes inserted into the grid at a number of recording positions. By comparing these MRI images with the microdrive depth readings of the white and gray matter transitions and with that of the skull base during the recordings, we estimated the recording positions and assigned the neurons to the lateral convexity of IT, upper bank, fundus or lower bank of the STS. All animal care, experimental and surgical procedures followed national and European guidelines and were approved by the K.U. Leuven Ethical Committee for animal experiments.

The stimuli were presented on a cathode ray tube (CRT) with a frame rate of 60 Hz. Personal computers running custom software controlled every aspect of the stimulus presentation and of the passive fixation task. The monkey was seated in a monkey chair with its head fixed facing the display, whereas the position of one eye was monitored using infrared video-based eye trackers. Initially, we employed ISCAN (120 Hz), whereas in later recordings, we used the EYELINK (1000 Hz) system.

Extracranial single unit recordings were performed using epoxy-like-insulated tungsten microelectrodes (0.5–2 MΩ measured in situ) that we lowered into the brain using a Narishige microdrive. The electrode was advanced through a guiding tube that was fixed in a plastic grid (Crist Instruments) firmly positioned inside the recording chamber. The electrode signal was amplified and filtered using conventional single unit recording equipment. Single units were isolated online using a custom made DSP-based spike discriminator that accepted a spike whenever its level-triggered waveform crossed several criterion boxes positioned at different times and levels on the display. The time stamps of the isolated spikes, stimulus events, and eye position were stored by a personal computer (1-ms resolution), using home made software written in LABVIEW.

The animals performed a passive fixation task during the recording sessions. The trial started with the onset of a small square fixation target located in the middle of the display. When the monkey fixated this target for 500 ms, the stimulus was presented together with the fixation target. The monkey was required to fixate the target during the entire duration of the stimulus presentation, that is, 2000 ms for the action movies, and another 200 ms after stimulus offset. If the monkey kept its gaze within a square fixation window (approximate size 1.7°) from the beginning of fixation until the end of the 200-ms poststimulus fixation period (2700 ms in total), the trial was accepted as valid and the monkey obtained a liquid reward.

Stimuli and Tests

Parametric Action Space

All neurons were sought and tested with a set of 21 movies depicting human stick-plus-point-light figures that performed transitive arm actions. The duration of all movies was identical and lasted 2000 ms (120 frames shown at a 60-Hz frame rate). The stick-plus-point-light figures measured approximately 6° in height and 1.5° in width. Initially, the movies were displayed centrally. During the course of the experiments, however, we noted that monkey LU developed a tendency to produce pursuit-like eye movements toward the end of the movie. To ensure stable fixation, we presented the movies in the contralateral visual field at 1.5° eccentricity during the remainder of the recordings in both animals. Recorded neurons in which there was a significant effect of eye movements on neural responses were not analyzed further (see Results). Results from the central and the slightly eccentric positions are pooled in all the population analyses of the present paper.

The 21 movies consisted of 3 real actions, that is, knocking, lifting, and throwing and their 2-way and 3-way blends (Fig. 1). We will refer to these 3 real actions as the “prototypical actions” throughout this paper. The blends were created using the algorithm of Kovar and Gleicher (2003), which preserves the constraints of biomechanical movement. The 3 real actions were selected from a library of 3D motion capture data (Ma et al. 2006). The blending algorithm required a conversion of the data files into BVH (Biovision hierarchical data) format which was done using 3D Studio Max and Lifeforms Software (Dekeyer et al. 2002). Following the creation of blended movements, all movement files (real and blended) were converted from BVH format...
to data files containing the time-varying 3D coordinates of 15 points on the body. To assure that all movement stimuli were of equal duration, the 3D data for each body point were resampled using cubic spline interpolation so that the motion of each point would consist of 120 samples. These 120 samples of each of the 15 body points were then rendered as 120 individual frames measuring 256 × 256 pixels, with these 15 points connected by lines in the form of the human body. The blends were produced in steps of 20%, thus having 3 × 4, 2-way action blends (Fig. 1A outer triangle) and 6, 3-way action blends (Fig. 1A inner triangle). The parametric action space of these 21 actions can be approximated by a 2D triangular configuration with the 3 motion-captured actions lying at the vertices (Fig. 1A). This 2D configuration is a linear approximation of the more complex higher-dimensional parametric configuration that obeys all the constraints of the weight differences between the 21 movies. For our purposes, the triangular approximation is sufficient and will be used throughout this paper for data displays and analyses.

Figure 1B shows undersampled sequences of static snapshots (selected frames taken from the action sequence) of the 3 motion-captured action movies, 5, 2-way blends and 1, 3-way blend. Because the movies were generated from actual human actions that we wished to keep as realistic as possible, there were slight differences in the start snapshots of the 3 actions. Mean speed and arm trajectories differed systematically between actions. The speed of the wristpoint end-effector increased from lifting to throwing and from throwing to knocking. Within an action movie, the speed of the wristpoint varied according to a multimodal function. Similar multimodal speed functions, but reduced in amplitude, were present for the elbow- and shoulderpoints. These variations in the speed within and across actions allowed us to determine the effect of speed on the responses of the neurons (see Results). In each action movie, the wrist moves to the right followed by a movement to the left, back to the starting position. Thus, each action can be divided naturally into 2 parts consisting of a rightward arm motion followed by a leftward, return arm motion. The vertical range of the arm movement differed between the actions, with the lifting and throwing movements containing less upward travel than the knocking movements (Fig. 1B).

Each neuron was tested with all the action movies from the parametric action space, presented in an interleaved, pseudorandom fashion (parametric action space test). At least 4 unaborted trials (median = 7.5) per movie were presented. If the isolation of the neuron was still adequate at the conclusion of this test, the neuron was subsequently tested with one or several of the following tests.

**Static and Translating Snapshots**

For each action we selected 6 static frames that were representative of the variation between the snapshots within an action. Each snapshot was shown for 300 ms, which is sufficiently long to produce responses in visual temporal cortical neurons (Keysers et al. 2001; De Baene et al. 2007).

In addition to static presentations of selected snapshots, we also translated the same 6 snapshots across a part of the display. The velocity (speed and direction) of translation for each snapshot was adjusted to approximate the mean velocity of different phases of the actions. We employed one speed and motion axis for the knocking action prototype, 2 for the throwing action prototype and 3 for the lifting action prototype. The translation was always 2000 ms in duration and thus the translation amplitude depended on the speed. The translation was centered at the same spatial position as that used when presenting the action movies and the static snapshots. We used 2 opposing directions of movement for each motion axis.

The purpose of the static snapshot and translating snapshot tests was to compare the response of a neuron to an action movie with selected static and/or translating snapshots of the same movie. During the course of the experiments, we developed different versions of these tests. In the initial version, after choosing the most- and least-effective actions based on the responses of the neuron in the preceding parametric action space test, static presentations of 6 different snapshots of each of the 2 movies were interleaved with the presentation of the 2 action movies. At least 6 unaborted trials (median = 10) per condition were presented.

In a later version, the test consisted of the effective movie and static and translating snapshot presentations. These were presented in an interleaved fashion with a minimum of 4 unaborted trials per condition (median = 6).
Reduced-Action Configurations

We reduced the complexity of the stick-plus-point-light figure by systematically deleting body parts/limb configurations in successive steps until only the arm (3 points connected by 2 lines) that performed the action remained present. In later recordings, neurons were tested with further simplifications: the wrist- and elbow-point connected by one line (underarm) and the wristpoint only. The reduced-action movies had the same duration and were presented at the same spatial location as the original, nonreduced movie.

The purpose of the reduced-action configuration test was to measure the effect of deleting parts from the stick-plus-point-light figure on the response of the neuron. In the initial version of this test, the most reduced version consisted of the arm only, whereas a later version also presented movies depicting 2 additional reductions (the underarm and the wristpoint only). Reduced-action configuration movies of the most- and least-effective actions and the corresponding unreduced movies were presented interleaved in a pseudorandom fashion with a minimum of 4 unaborted trials (median = 8 for the reduction up to the arm; median = 6 for the reduction up to the wristpoint).

Position

The purpose of the position test was to measure the dependency of the response on the spatial location of the movie. The most effective movie was shown at 17 different positions. In addition to the foveal presentation, 8 positions were located on a small square grid with a spacing of 1.5°, and 8 positions were located on a larger square grid with a 3° spacing. The center of each grid corresponded to the foveal position. All the positions were presented interleaved with at least 4 unaborted trials (median = 5).

Reversed Temporal Order

For each action movie, consisting of 120 frames, the temporal order of the frames was reversed, resulting in movies played backward.

The reversed temporal order test consisted of 4 conditions: the movies of the most- and least-effective actions and versions in which the temporal order of the frames was reversed. These 4 conditions were presented interleaved with a minimum of 5 unaborted trials (median = 10). In this test, the snapshots of the original and temporally reversed movies were identical and thus the original and reversed movies differed only in their kinematics.

Analysis of Neural Activity

Parametric Action Space Test

To test whether the neural responses differed significantly from baseline activity, for each trial we computed the average firing rates for a baseline and a stimulus analysis window. The baseline window started 400 ms before stimulus onset and ended at stimulus onset. The stimulus window started 50 ms after stimulus onset, to allow for the minimum response latency of the neuron, and lasted 2000 ms. To account the empirical fact (see Results) that responses for a given stimulus can vary during the presentation of the action movie and the possibility that the temporal evolution of the response is used by subsequent stages of processing. We quantified the fit between the neural-based ISOMAP solution and the stimulus-based, triangular, action space configuration by computing the sum of the squared errors between the spatial coordinates of corresponding points for the 2 spaces after we Procrustes-rotated the neural-based ISOMAP space toward the stimulus-based parametric action space. We refer to this dissimilarity value as the Procrustes Distance measure.

We tested the statistical significance of the action selectivity by performing a split-plot ANOVA using the 21 actions as a between-trial factor and the responses in 5 successive bins of 400 ms, starting 50 ms after stimulus onset, as a within-trial factor. Because the responses could vary during the action sequence, we used these binned responses instead of an average across the entire 2000 ms. Action selectivity was considered to be statistically significant when there was a significant main effect of the factor action or a significant interaction between the 2 factors. Type 1 error was set at 0.05.

The degree of stimulus selectivity of a neuron was quantified by computing the depth of selectivity index (DOS; Rainer and Miller 2000):

$$\text{DOS} = \left( \frac{1}{n-1} \sum_{i=1}^{n} R_i / R_{\text{max}} \right)$$

where $n =$ number of actions (21), $R_i =$ mean firing rate to the $i$th action and $R_{\text{max}} =$ max$|R_i|$. The DOS can vary between 0 (equal response to all actions) and 1 (when there is a response for only one action). The mean firing rate was computed for the long analysis window extending from 50- to 2050-ms post-stimulus onset. The DOS index is a measure of the breadth of action tuning.

In addition, we computed the ANOVA based Omega Square Modulation index, $\omega^2$ (Kirk 1968) which estimates the proportion of variance of the neural response due to stimulus variations. The Omega Square Modulation index takes into account response differences among the 21 movies as well as trial-by-trial variability and was computed for the long analysis window extending from 50- to 2050-ms post-stimulus onset, as well as for the shorter, successive 250-ms time periods, starting at 50 ms and ending at 2050 ms. The Omega Square Modulation index is useful for relating the degree of response modulation by the action stimuli, to the Procrustes Distance measures for the successive 250-ms time periods. An overall low degree of response modulation by the action stimuli will result in relatively low Procrustes distances, because the neurons do not distinguish the actions particularly well.

We correlated the instantaneous firing rate of each neuron to the velocity of the wristpoint end-effector, which contained the most motion, using multiple regression. For this analysis we estimated the instantaneous firing rate of the neuron after convolving the response of the neuron with a Gaussian kernel having a standard deviation of 25 ms. Then, the smoothed instantaneous firing rates, $R$, were fitted by a quadratic polynomial: $R = ax + bx^2 + cy + dy^2 + exy + f$, with $x$ and $y$ being the horizontal and vertical coordinates, respectively, of the endpoint of the instantaneous velocity vectors. Velocity and response were binned at a frame rate of 60 Hz, using several time delays for the neural response, ranging from 1 to 15 frames, that is, up to 250 ms. The regressions relating response to velocity were performed across the 21 movies.
Velocity combines speed and direction. To determine whether the response is related to speed per se, we fitted the instantaneous firing rates, $R$, by a quadratic polynomial: $R = ax + bx^2 + c$ with $x$ being the speed of the wristpoint. These response-speed regressions were performed across the 21 movies for the different time delays. To determine whether the response is related to the direction of the wristpoint motion, we computed the linear-circular correlation coefficient (Mardia and Jupp 2000) between the instantaneous firing rates and the instantaneous motion directions for speeds greater than 1°/s. Note that the linear-circular correlation coefficient varies between 0 and 1. These correlations between response and direction were computed across the 21 movies for the same time delays as were used for the velocity and speed regression analysis (see above).

Responses to Static Snapshots compared with Actions
Because the responses of the neurons varied during the course of the action movie, averaging a response across the 2000 ms of its duration can underestimate the response to specific parts of the action. To avoid any such underestimation of the neural response to the action sequence, we used peak firing rate instead of average firing rate as the response measure when comparing the responses of the action and snapshot presentations. The peak firing rate was computed, after convolving the response with a Gaussian kernel, using a standard deviation of 25 ms, in an interval of 50-2050 ms for the action, and 50-350 ms for the static snapshots presentations. Net peak firing rates were calculated by subtracting the average firing rate, obtained in the baseline analysis window for all conditions, from the peak firing rate observed during the stimulus presentation.

To quantitatively compare the responses to static snapshot presentations with responses to the corresponding action (presented in an interleaved fashion in the same test), we computed the following Action Index $= (P - max Ps)/(P + max Ps)$, with $P$ being the net peak firing rate for the action, and max $Ps$ the maximum net peak firing rate for the 6 static snapshots. Taking the average instead of the maximum of the 6 net peak firing rates would have underestimated the snapshot response in neurons that showed different, and thus selective, responses to the individual snapshots.

If the response to the action movie were to depend on the response to the static snapshots, then one would expect a positive correlation between the selectivity for static presentations of individual snapshots and the modulation of the response during the action. Therefore, we compared the degree of selectivity for the static presentations of the snapshots to the response modulation observed during the action sequence.

The degree of selectivity within the most effective action, as tested in the snapshot test, was calculated for the convolved responses in an analysis window extending from 200 to 2050 ms after stimulus onset using a Within-action Modulation index $= (max Pa – min Pa)/(max Pa + min Pa)$, with $max Pa$ the maximum gross peak firing rate of the action and $min Pa$ the minimum gross peak firing rate of the action. We did not take into account the first 200 ms, because it may reflect transient responses related to stimulus onset. Using analysis windows extending from 50 to 2050 ms, however, produced highly similar results ($r = 0.92$ between these indices for the 1850- and 2000-ms long windows). Again, this index was computed for neurons for which the peak response exceeded 10 spikes/s.

The degree of selectivity for the static presentations of the snapshots was quantified by the following index: Snapshot Selectivity index $= (max Ps – min Ps)/(max Ps + min Ps)$, with $max Ps$ the gross peak firing rate for the most effective static snapshot and $min Ps$ the gross peak firing rate for the least-effective static snapshot. This index was computed for the neurons for which the peak response to any of the static presentations of a snapshot exceeded 10 spikes/s. We will report Snapshot Selectivity indices using the 50- to 350-ms analysis window (see above), although Snapshot Selectivity indices computed using responses from 200 to 350 ms produced highly similar results ($r = 0.95$).

Responses to Translating Snapshots compared with Actions
To compare the responses to translating snapshots with those of their corresponding actions (presented in an interleaved fashion in the same test), we computed a Translation index $= (Pa - max Pr)/(Pa + max Pr)$, with $Pa$ being the net peak firing rate (computed after convolution with a Gaussian; see above) to the action and max $Pr$ the maximum net peak firing rate of the 12 translating snapshots, that is, 6 different snapshots, with each shown translating in 2 opposite directions.

To compute this index, we employed an 1850 ms analysis window that started 200 ms after stimulus onset for both the action and translating snapshots, to reduce the inclusion of transient responses caused by stimulus onset, and therefore unrelated to the motion. However, Translation indices computed on the full 2000-ms window produced similar results ($r = 0.91$ between these indices for the 1850- and 2000-ms windows).

To examine whether the responses of the neurons depended on the direction of motion along the dominant axis of the wrist motion, we computed the following index: Best Direction index $= (Pb - Pto)/(Pb + Pto)$, with $Pb$ being the net peak firing rate for the translated snapshot condition that produced the largest response and Pto the net peak firing rate of the translating snapshot with the opposite direction. This Best Direction index was computed only for those neurons that showed a maximum net peak firing rate of at least 10 spikes/s for one of the translating snapshots.

Reduction-Action Configuration Test
To quantitatively compare the responses of the isolated arm movement with its corresponding full-body action configuration, we calculated an Arm Reduction index $= (Pa - Parm)/(Pa + Parm)$, with $Pa$ as the net peak firing rate for the full stimulus configuration and $Parm$ the net peak firing rate for the arm in isolation. The firing rates were computed using the analysis window of 50-2050 ms and using smoothed responses (see above). Similar reduction indices were also computed for the other reduction conditions. As an index of the degree of selectivity, we computed a Between-Action Selectivity index $= (Pab - Pw)/(Pab + Pw)$, with $Pab$ and $Pw$ being the net peak firing rates for the most- (“best”) and least-effective (“worst”) actions, respectively. This index was calculated for the full-body, the arm-only and the wristpoint-only conditions.

Position Test
Because the arm trajectories differed in all the actions, one possible explanation of selective responses for the actions would be that the RfS of the neurons contain one or more “hot spots.” In this case, the temporal response profile in a particular action condition will differ as a function of the spatial location of the stimulus, because the hot spot will be traversed by the arm at different moments in time or not at all. To examine this possibility, the smoothed responses at the different tested spatial positions were aligned to the time of the peak response at the foveal position. Thus, the occurrence of the smoothed peak response at the foveal position was set to zero and this temporal displacement was subtracted from the smoothed response profiles of all other parafoveal positions.

Other analyses of the responses of the neurons in the various tests are described in the relevant Results sections.

Results
We searched for temporal cortical neurons that responded to stick-plus-point-light movies depicting arm actions. We recorded from 240 responsive neurons in 2 animals. Each of these neurons gave a significant response to one or more stimuli of the parametric action space test (ANOVA; $P < 0.05$). Forty-two neurons were eliminated from the data set because, for these neurons, the eye positions during stimulus presentation differed significantly among the 21 actions (ANOVA on eye positions in 400 ms analysis windows during stimulus presentation; $P < 0.025$ in either $x$ or $y$ direction) and the neural responses were significantly related to the differences in eye movements (Kruskal–Wallis test on responses sorted according to horizontal and vertical eye position; $P < 0.05$). The eye position traces for the remaining 198 neurons (monkey LU: 96; monkey BR: 102) were similar for the
different actions: the mean eye positions for the 3 “prototypical” actions (i.e., the stimuli at the corners of the stimulus space) differed by less than 0.1 deg during the entire stimulus presentation (Supplementary Fig. 1).

In monkey LU we explored 13 guiding tube positions between 11 and 18 mm anterior and 17 and 23 mm lateral. Responsive units were found for 11 of these guiding tube positions covering the full extent of the range explored. Of the 96 neurons in monkey LU, 60 were localized in the upper bank and fundus of the STS, 17 in the lower bank and 19 in the lateral convexity of IT. In monkey BR we searched for responsive neurons using 29 guiding tube positions, between 9 and 16 mm anterior and 18 and 24 mm lateral. Responsive units were found for 19 of these positions. At 9 mm posterior, a patch of neurons \((N = 18)\) was found in the fundus of the STS that showed strong motion-related responses. We will label this patch of STS neurons as STPm (superior temporal polysensory middle; Nelissen et al 2006) to distinguish these neurons from the more anterior ones recorded in both animals. At more anterior sites in monkey BR, 21, 30, and 33 responsive neurons were found in the upper bank/fundus, lower bank and lateral convexity, respectively. Unless otherwise noted we will label the neurons of the upper bank and fundus of the STS, anterior to STPm, as upper bank/fundus STS neurons.

The different regions were sampled unevenly with repeated recordings at locations showing responses to the actions. It was clear during the recordings that neurons responding to these dynamic stick-plus-point-light figures were organized in patches, because no responsive neurons were observed at several guiding tube positions in the lower or the upper bank of the rostral STS, despite repeated penetrations.

**Neural Responses in the Parametric Action Space**

Figure 2 displays the responses of 3 representative neurons to the 21 action movies within the parametric action space. These 3 neurons manifested 2 sorts of response modulation: a systematic modulation of the responses within the action space, that is, a selectivity for action movies, and a modulation of the responses within the actions itself, that is, for segments of an action movie. The neuron in Figure 2A was recorded in the STPm region of monkey BR. It responded selectively to the second part of the action movies, when the arm was moving downward, and mainly for the \textit{throwing} and \textit{knocking} actions and their blends. The neuron in Figure 2B, recorded in the upper bank/fundus of the STS of monkey BR, showed a multimodal response pattern for \textit{lifting} and \textit{throwing} actions. The neuron shown in Figure 2C, recorded in the lower bank of the STS, exhibited strong selective responses to particular segments of only a small number of actions.

We tested the statistical significance of the action selectivity by performing a split-plot ANOVA using as factors the 21 actions, and the responses in successive bins of 400 ms starting 50 ms after stimulus onset. Because the responses could vary during the action sequence, we used these binned responses instead of an average across the entire 2000 ms. A large majority of responsive neurons \((157/198; 79\%)\) showed action selectivity, either by having a main effect of the factor action or an interaction between the 2 factors.

Typically, the neurons showed a systematic tuning in the action space. Their response was best for a particular action and declined with increasing distance between this preferred action and another given action. This sort of tuning in the action space may allow these neurons to represent the similarities among the action movies. In order to determine how a population of neurons represent the parametric action...
space we employed a nonlinear MDS technique, that is, ISOMAP (Tenenbaum et al. 2000), examining the interstimulus response differences for all stimulus pairs (see Methods). In this first analysis, the average firing rates were computed for the whole 2000 ms duration of the action, ignoring variations in the responses during the course of the action. Figure 3B shows the 2D configuration of the action space based on the normalized responses of all 198 responsive neurons. The Scree plot showed that this 2D representation was an optimal low-dimensional solution, and provided an excellent fit to the data, explaining 94% of the variance (Fig. 3C). The roughly triangular neural representation reflected to a large extent the original parametric, triangular stimulus configuration (Fig. 3A), thus the neural space preserved the stimulus rank along the 5, 2-way blend lines and along the 6, 3-way blend actions. The 1D solution provided a poorer fit (70% explained variance) indicating that the responses of the neurons do not depend merely on variations of a 1D stimulus parameter, such as mean speed, for example. We thus conclude that this unselected population of visual temporal cortical neurons rather faithfully represent the ordinal similarity relationships among dynamic action stimuli.

Comparing Neural Responses for Actions and Static Snapshots

Figure 4 shows the responses of 3 representative neurons to an effective action and to static presentations of snapshots from that same action movie. The neuron in Figure 4A, which is the same as that in Figure 2C, was recorded in the lower bank of the STS. It responded strongly to the static presentations of a snapshot: the response to the most effective snapshot was at least as high as the peak response to the action sequence. Furthermore, the neuron responded selectively to the different snapshots, showing an increased response when one arm is bent and crosses the other arm (“Napoleon” posture). Thus this neuron displayed form selectivity. The neuron in Figure 4B, recorded in the lower bank of the STS, also responded equally well to the action movies and static presentations of its snapshots. However, in contrast to the neuron in Figure 4A, this neuron responded similarly to the different snapshots. The neuron in Figure 4C, recorded in the upper bank/fundus of the STS, displayed little or no response to the static presentations of the snapshots, although it responded very well to the action movie. Thus, the latter neuron needed motion in order to respond.

The examples shown in Figure 4 illustrate the observed range of responses to the static snapshot presentations: from no response to responses equaling those of the action movies. For each of the neurons tested with static snapshots (Table 1), we computed an Action index (see Methods). An Action index of 0 indicates that the neuron responded equally well to static snapshots and the action movie, whereas an index of 1 indicates no response to the static snapshots. The bulk of the neurons in the lower bank (mean Action index = -0.06; N = 26; SD = 0.13) and lateral convexity (mean Action index = -0.02; N = 33; SD = 0.27) responded equally well to the static snapshots and to the action, whereas the majority of the upper bank/fundus STS neurons (mean Action index = 0.42; N = 41; SD = 0.36) and all tested STPm neurons (mean Action index = 0.82; N = 15; SD = 0.19) responded much more strongly to the action than to the static snapshots. The differences between the Action indices of these regions were highly significant (one-way ANOVA; P < 0.0001). Post hoc Bonferroni-corrected t-tests shows that the mean Action index of the STPm neurons differed significantly from those of the other 3 regions (all Ps < 0.00002) and that the mean Action index of the upper bank/fundus STS differed from that of the lower bank STS and lateral convexity (P < 0.000001), whereas the indices of the lower bank STS and lateral convexity were not significantly different (Fig. 5). These results suggest a marked functional specialization: most upper bank/fundus STS neurons, including STPm, require motion, whereas most lower bank and lateral convexity IT neurons respond to both dynamic and static stick figures.

Effect of Action Configuration Reductions on the Neural Responses

Figure 6 shows the responses of 2 representative neurons to an effective action and its reduced-action configurations. The lower bank STS neuron in Figure 6A, which is the same as that in Figure 4A and Figure 2C, still responded when the legs were omitted but stopped responding when the trunk was also removed and only the moving arm remained visible (Fig. 6A: reduced-action configurations). This demonstrates the form
selectivity of this neuron and suggests that the crossing of the 2 arms is a critical feature for this neuron. On the other hand, the neuron of Figure 6C, recorded in the upper bank/fundus of the STS, still responded well to the motion of the arm alone and with an even higher firing rate for the motion trajectory of the end-effector only.

The effect of stimulus reduction was quantified for each of the tested neurons (Table 1) by computing an Arm Reduction index (see Methods). An Arm Reduction index of zero indicates that the neuron responds equally as well to the arm alone as it did to the full body, whereas an Arm Reduction index of 1 indicates no response to the isolated arm action. The distributions of the Arm Reduction index differed among the 4 regions (one-way ANOVA; \( P < 0.0005 \)). The neurons in the lower bank of the STS (mean Arm Reduction index = 0.20; \( N = 22; \ SD = 0.34 \)) and the lateral convexity of IT (mean Arm Reduction index = 0.18; \( N = 32; \ SD = 0.26 \)) showed, on average, responses that were more strongly attenuated by limiting the stimulus to the arm alone than did the STPm neurons (mean Arm Reduction index = -0.07; \( N = 11; \ SD = 0.06 \)) and upper bank/fundus of the STS neurons (mean Arm Reduction index = -0.02; \( N = 46; \ SD = 0.21; \) all \( P < 0.03 \)), although such intraregional differences were less pronounced than for the snapshot test (Fig. 5). Indeed, several neurons in the lower bank STS and the lateral convexity also responded well to the isolated arm action.

We also computed similar contrast indices for the other stimulus-reduction conditions. The great majority of the neurons responded similarly to full-body and body-without-legs configurations (mean contrast indices of the 4 regions ranging between -0.007 and 0.02). When a similar contrast index was computed for the neurons tested with the wristpoint only, results were more pronounced than with the arm only (Fig. 5). The mean Point Reduction indices for the lower bank STS (mean Point Reduction index = 0.60; \( N = 9; \))
of the most effective action differed by less than a factor of 1.5 (Arm Reduction index or Point Reduction index < 0.20). The latter avoided noisy indices arising from weak responses. We selected those neurons whose responses differed by at least a factor of 1.5 between the most and least-effective action conditions for the full-body or the arm-only (Between-Action Selectivity index > 0.20). For the neurons tested with the arm-only, 37 neurons fulfilled these 2 criteria. For these neurons, the average Between-Action Selectivity index was significantly larger (paired t-test; P < 0.0001) for the full-body (mean Between-Action Selectivity index = 0.41; SD = 0.16) than for the arm-only displays (mean Between-Action Selectivity index = 0.16; SD = 0.22). Also, for the 19 neurons that fulfilled these criteria and were tested with the wristpoint-only, the mean Between-Action Selectivity index was significantly smaller (paired t-test; P < 0.002) for the wristpoint-only (mean Between-Action Selectivity index = 0.15; SD = 0.19) than for the full-body displays (mean Between-Action Selectivity index = 0.36; SD = 0.17). This was due to a relative increase in response to the least-effective action for the wristpoint-only condition compared with the full-body conditions. Thus, although these neurons responded well to the wristpoint only, they provided less information about the arm trajectory than when the full-body or even just the shoulders and arms were displayed, implying that wristpoint trajectory did not fully determine their response selectivity.

Note that the above analysis underestimates the effect of stimulus reduction on the selectivity for the whole population of neurons because we included only neurons that responded well in the reduced conditions. Indeed, if the whole population of neurons was considered, the reduction had a strong effect on selectivity. The mean Between-Action Selectivity index was 0.03 (N = 111; SD = 0.26) for the arm-only conditions, compared with 0.18 (N = 111; SD = 0.23; paired t-test; P < 0.0001) for the full-body displays and 0.01 (N = 68; SD = 0.22) for wristpoint-only compared with 0.20 (N = 68; SD = 0.22; paired t-test; P < 0.0001) for the full-body displays presented in the same test.

Comparing Neural Responses between Actions and Translating Snapshots

Figure 7 shows the responses of 3 neurons to translating snapshot presentations. The neuron of Figure 7A is an upper bank/fundus STS neuron that responded well to the motion present in the action but failed to respond to translating snapshots of that same action. Thus, rigid motion of the line configuration was not sufficient to drive this neuron. The 2 other neurons of Figure 7, recorded from the upper bank/fundus STS (Fig. 7B) and from the STPm region (Fig. 7C) did respond well to the translating snapshots with the neuron of Figure 7C, showing strong direction selectivity. The distributions of the Translation indices for all tested neurons (Table 1) differed among the 4 regions (Fig. 5): an ANOVA showed a significant effect of region (P < 0.05). Post hoc Bonferroni-corrected t-tests showed that this was due to STPm neurons having significantly larger Translation indices (mean Translation index = 0.27; N = 8; SD = 0.11) compared with neurons recorded in the lateral convexity (mean Translation index = -0.04; N = 19; SD = 0.18; Bonferroni-corrected t-test; P < 0.02). The upper bank/fundus STS (mean Translation index = 0.07; N = 27; SD = 0.32) contained some neurons that responded much less strongly to the translating snapshots (for example the

SD = 0.31) and lateral convexity (mean Point Reduction index = 0.31; N = 20; SD = 0.26) were larger (Bonferroni-corrected t-tests; Ps < 0.006) than the value obtained for the upper bank/fundus STS neurons (mean Point Reduction index = 0.03; N = 39; SD = 0.35). The difference between Point Reduction indices for the lower bank STS and lateral convexity was not statistically significant and only a few of the neurons tested in these regions responded as well to the single wristpoint action as they did to the full-body stimulus. These results also showed that many STS neurons, especially those in the upper bank, respond as well or better to the motion of the isolated arm, and to the motion of the end-effector wristpoint in isolation, as they do to the whole body.

The Reduction indices were computed on the responses to the most effective action and thus do not inform us about possible changes in selectivity across different actions when the stimulus is simplified. To address this question, we computed a Between-Action Selectivity index (see Methods) that compares the net peak firing rates for the 2 actions that were used in the reduced-action configuration test. We computed such indices for the full-body, arm-only and wristpoint-only conditions of those neurons for which the response to the full-body and reduced-stimulus configurations

Figure 6. Neural responses to full-body and reduced-action configurations compared. Peristimulus time histograms with raster plots for 2 example neurons recorded with the reduced-action configuration test. (A) Systematic reductions of the stimulus configuration from the full-body action configuration down to the presentation of the wristpoint-only. Only the first frame of action C1 (kicking prototype) is shown. Note that the actual test contained dynamic sequences of these reduced configurations. (B) Neuron recorded in the lower bank of the STS, corresponding to Figure 2C, tested with the reduced-action configurations of C5 down to the arm-only presentations. A ‘Napoleon’ posture with the right arm wristpoint covering the left arm elbow point was necessary to drive this neuron. (Arm Reduction index = 0.71). (C) Neuron recorded in the fundus/upper bank of the STS, tested with the reduced-action configurations of C21 down to the presentation of the wristpoint dynamic sequence, which drove the response even more strongly than the full action configuration (Arm Reduction index = -0.19; Point Reduction index = -0.32). Time course (in ms) and response strength (firing rate in spikes/s) are indicated on the action sequence scale. Vertical lines mark stimulus onset and offset. Stimulus durations were 2000 ms.

Comparing Neural Responses between Actions and Translating Snapshots

Figure 7 shows the responses of 3 neurons to translating snapshot presentations. The neuron of Figure 7A is an upper bank/fundus STS neuron that responded well to the motion present in the action but failed to respond to translating snapshots of that same action. Thus, rigid motion of the line configuration was not sufficient to drive this neuron. The 2 other neurons of Figure 7, recorded from the upper bank/fundus STS (Fig. 7B) and from the STPm region (Fig. 7C) did respond well to the translating snapshots with the neuron of Figure 7C, showing strong direction selectivity. The distributions of the Translation indices for all tested neurons (Table 1) differed among the 4 regions (Fig. 5): an ANOVA showed a significant effect of region (P < 0.05). Post hoc Bonferroni-corrected t-tests showed that this was due to STPm neurons having significantly larger Translation indices (mean Translation index = 0.27; N = 8; SD = 0.11) compared with neurons recorded in the lateral convexity (mean Translation index = -0.04; N = 19; SD = 0.18; Bonferroni-corrected t-test; P < 0.02). The upper bank/fundus STS (mean Translation index = 0.07; N = 27; SD = 0.32) contained some neurons that responded much less strongly to the translating snapshots (for example the
neuron in Fig. 7A) than to the action movie but most of the neurons that responded to the action movies also responded to translation. The lower bank STS neurons responded equally well to the translating snapshots and the action (mean Translation index = 0.18 that provided the highest performance for the neurons tested with static snapshots). These neurons were recorded predominantly in the upper bank/fundus of the STS and in the lateral convexity of IT.

To assess direction selectivity along the dominant axes of the wristpoint motion, we computed a Best Direction index (see Methods) for the neurons responding with at least 10 spikes/s in any of the directions of the translating snapshots. The Direction indices were, on average, low, indicating relatively weak direction selectivity along these motion axes. The differences between regions were statistical significant (one-way ANOVA; \( P < 0.01 \), Fig. 5) with a significantly larger mean Best Direction index for the STPm region (mean Best Direction index = 0.43; \( N = 8; SD = 0.27 \); Bonferroni-corrected \( t \) tests; \( P < 0.05 \)) than for the other regions (means Best Direction indices = 0.24, 0.19, 0.18; \( N = 27, 12 \) and 19; SD = 0.19, 0.13 and 0.15 for upper bank/fundus STS, lower bank STS, lateral convexity of IT, respectively).

**Functional Differentiation of Neurons: “Snapshot” and “Motion” Neurons Defined**

To determine whether we could distinguish between different groups of neurons based on their response properties, we correlated the Action, Arm Reduction, Translation and Best Direction index values for each pair of neurons. These correlations were computed for the 49 neurons for which values were obtained for all 4 indices. Then, the \( 49 \times 49 \) matrix containing the pairwise 1-Pearson \( r \) distance values was subjected to cluster analysis (Ward’s method). As shown in Figure 8A, 2 distinct clusters of neurons were obtained with one cluster (18/49 neurons) containing only upper bank/fundus STS and STPm neurons, whereas the other cluster (31/49) consisted of all of the lower bank STS and lateral convexity neurons, in addition to 40% of the upper bank/fundus STS neurons.

This cluster analysis suggests that the 4 indices differentiate—although not perfectly—lower bank STS/lateral convexity neurons from upper bank STS/STPm neurons. To examine which index yielded the best separation of the regions, we determined with Receiver Operator Characteristic (ROC) analysis how well one can classify a neuron as belonging to the upper bank STS/STPm versus lower bank STS/lateral convexity given its index value. This analysis revealed that the Action index produced the best separation of the regions (classification performance—area under the ROC curve = 0.90; \( N = 115 \)), followed by the Arm Reduction index (area = 0.75; \( N = 111 \)). The Translation (area = 0.66; \( N = 66 \)) and Best Direction indices (area = 0.60; \( N = 66 \) produced rather poor separations of the regions (chance performance = 0.50).

Figure 8B shows the relationship between the Action and Arm Reduction indices, for the neurons that were tested with both tests (\( N = 79 \)), for the 4 regions independently. The scatterplot clearly shows that neurons responding more strongly to the action stimuli than to the static snapshot presentations (that is, with large, positive Action indices) also responded more similarly to the arm-only and the full-body action (small Reduction indices). These neurons were recorded predominantly in the upper bank/fundus of the STS including area STPm. On the other hand, neurons that respond similarly to the action and static snapshots can vary in their degree of tolerance to simplification of the action configuration. This type of neuron was found mainly in the lower bank of the STS and in the lateral convexity of IT.

Given that the Action index separated the anatomical regions well (see above ROC analysis and Fig. 8B) and given the theoretical importance of this distinction (i.e., responding to form vs. motion information), we distinguished 2 populations of neurons using an Action index of 0.20 as criterion (striped vertical line in Fig. 8B). This criterion value is close to the Action index of 0.18 that provided the highest performance rate when classifying the anatomical regions (upper bank/fundus STS vs. lower bank STS/lateral IT; all neurons tested with static snapshots (\( N = 115 \)) based on the Action Index. In all further analyses we will use the terms “snapshot” and “motion” neurons to refer to neurons that have an Action index lower or
higher than 0.20, respectively. "Motion" neurons respond to actions but much less so to static snapshot presentations, whereas "snapshot" neurons respond as well to actions as to static snapshot presentations. Figure 9 shows the estimated recording positions and the distribution of the "motion" and "snapshot" neurons; it is clear that the "motion" neurons were recorded mainly in the upper bank/fundus STS regions, whereas the "snapshot" neurons were recorded mainly in the lower bank STS and the IT convexity. The next sections compare the responses of these functionally defined "snapshot" and "motion" neurons.

**Representation of the Parametric Action Space in "Snapshot" and "Motion" Neurons**

To examine whether there is a difference between the representations of the parametric action space in the "snapshot" and "motion" neurons, we applied the ISOMAP technique on the responses of these neurons. The "motion" neurons represented the action space more faithfully than the "snapshot" neurons (Fig. 10A,B). First, the 2D configuration provided a better fit to the data for the "motion" neurons (explained variance = 98%) than for the "snapshot" neurons (explained variance = 83%). In both populations, a 1D solution was suboptimal (explained variances < 70%). Second, the ordinal relationships among the actions are preserved to a greater extent for the "motion" than for the "snapshot" neurons. Third, Procrustes rotation of the neural representation to the action space resulted in a better fit for the "motion" neurons compared with the "snapshot" neurons (Procrustes Distance measures = 0.07 and 0.23 respectively; note that a large value of this dissimilarity measure indicates a worse fit).

It should be remarked that the less faithful representation for the "snapshot" neurons compared with the "motion" neurons does not result from a smaller sample, in fact the opposite was true: the sampled "snapshot" neurons (N = 64) outnumbered the "motion" neurons (N = 51).

The mean maximal firing rates, computed from 50- to 2050-ms poststimulus onset, were similar for "motion" (29 spikes/s; N = 51; SD = 18) and "snapshot" neurons (28 spikes/s; N = 64; SD = 18; Mann-Whitney U test; ns). "Motion" neurons possessed significantly greater action selectivity as measured by the DOS index (see Materials and Methods) than "snapshot" neurons ("motion" neurons: mean DOS = 0.29; N = 51; SD = 0.13; "snapshot" neurons: mean DOS = 0.23; N = 64; SD = 0.12; Mann-Whitney U test; P < 0.005). Also, "motion" neurons were modulated significantly more strongly by the different action stimuli (mean Omega Square Modulation index = 0.19; SD = 0.17) than "snapshot" neurons (mean Omega Square Modulation index = 0.09; SD = 0.15; Mann-Whitney U test;
$P < 0.00001$) which might explain the more faithful neural representation of the action space in the case of the “motion” neurons.

**Neural Responses during the Course of the Action in “Snapshot” and “Motion” Neurons**

Given that profound changes in firing pattern during the course of an action movie (Fig. 2A–C) were present in some cases, we performed a series of analyses that took into account temporal response changes. In a first analysis, we divided the entire action sequence into 8 successive 250-ms segments. The responses were averaged within each of these time segments. In order to take into account response latency, the first segment started 50 ms after stimulus onset and the last segment lasted until 50-ms poststimulus offset. ISOMAP analyses were carried out for all time segments for the full population and for the “snapshot” and “motion” neurons separately. An inspection of the ISOMAP solutions for the full population ($N = 198$), made it clear that the most faithful representations were for the second (300–550 ms), fifth (1050–1300 ms), and seventh (1550–1800 ms) segments, all having Procrustes Distance measures of less than 0.20. For the “motion” neurons the best fit occurred for the second (300–550 ms) and seventh (1550–1800 ms) segment of the action with Procrustes Distance measures smaller than 0.19 (Fig. 11A). Overall, the “snapshot” neurons produced erratic configurations (all Procrustes Distance measures $> 0.36$) with equally good fits for 4 of the 8 segments (Fig. 11B).

These variations in Procrustes Distance measures that were observed during the course of the action correlated significantly with the mean Omega Square modulation indices computed for the responses in each of the 8 segments for both “motion” ($r = -0.76; P < 0.05; N = 8$) and “snapshot” neurons ($r = -0.74; P < 0.05; N = 8$). These significant, negative correlations indicate that the variations seen in both parameters during the course of the action are not mere noise, and that stronger modulation of the neural responses to the various actions coincides with a more faithful representation of action space.

We also computed, for each of the 8 segments, the mean Euclidean distances (averaged across all 210 possible stimulus pairs) for the spatial ($x$, $y$)-position coordinates of the wristpoint. The position differences were computed for segments that started 50 ms earlier (i.e., 0–250 ms, 250–500 ms, etc.) than those used to compute the neural responses. The differences in the spatial position of the wristpoint end-effector across the 8 segments correlated significantly with the Procrustes Distance measures for “snapshot” ($r = -0.80; P < 0.05; N = 8$) but not for “motion” neurons ($r = -0.57; \text{ns}; N = 8$). A similar result was obtained when we computed the mean Euclidean distances of the ($x$, $y$)-position of the full right arm, a combination of the wrist-, elbow-, and shoulderpoints (“snapshot” neurons; $r = -0.89; P < 0.005; N = 8$; “motion” neurons; $r = -0.59; \text{ns}; N = 8$). This suggests that “snapshot” neurons, but less so “motion” neurons, are sensitive to the spatial differences among the stimulus configurations, fitting with the form-driven responses of the former type of neurons.

**Contribution of Kinematics to the Responses of “Snapshot” and “Motion” Neurons**

We correlated the instantaneous firing rate of each neuron to the instantaneous velocity of the wristpoint end-effector using a multiple regression analysis (see Materials and Methods). This was done across the 21 actions for a set of time delays between

---

*Figure 10. Neural representation of the parametric action space in “motion” and “snapshot” neurons. (A and C) Neural representations of the stimulus space obtained after applying ISOMAP (nonlinear MDS) to the mean responses averaged using an analysis window ranging from 50 to 2050 ms: “motion” (A) and “snapshot” (C) neurons. (B and D) Scree plots of the residual variance as a function of the number of dimensions. Same conventions as in Figure 3.*
the motion and instantaneous firing rate. For each neuron we choose that time delay for which the amount of variance in firing rate explained by velocity was the highest. The median neural response variances explained by velocity were 0.35 for the "motion" neurons and 0.17 for the "snapshot" neurons, a highly significant difference (Mann-Whitney U test; \( P < 0.0001 \)). Thus, the response of "motion" neurons correlated more strongly with the velocity of the end-effector than the response of "snapshot" neurons did, which agrees with the motion sensitivity of the "motion" neurons.

Correlations with velocity can be due to correlations with speed or direction or a combination of both parameters. To assess the correlation between direction and instantaneous firing rate we computed the linear-circular correlation coefficients between the direction of the wristpoint and the smoothed instantaneous firing rates of each neuron, binned to the stimulus presentation frame rate of 60 Hz (see Methods and Materials). Note that the linear-circular correlation coefficient can vary between 0 and 1. For each neuron we computed the correlation between direction and firing rate for a set of time delays and choose that time delay for which the correlation was maximal. The median linear-circular correlation was significantly larger for the "motion" than for the "snapshot" neurons (median linear-circular correlations = 0.18 vs. 0.12 respectively; Mann–Whitney U test; \( P < 0.01 \)). Because the mean Best Direction indices were higher for the STPm neurons than for neurons of the other regions, we wondered whether the significantly larger correlation between firing rate and direction for the "motion" compared with the "snapshot" neurons was due to presence of STPm neurons in our sample of "motion neurons." The median linear-circular correlation between direction and firing rate for the 36 "motion" neurons that were recorded anterior to the STPm region was indeed lower and did not differ significantly from the correlations for the "snapshot" neurons (median linear-circular correlations = 0.15 vs. 0.12, respectively; Mann–Whitney U test; ns). This suggests that for the anterior "motion" neurons the contribution of motion direction to the neural responses is small. Note that this population of "motion" neurons ( \( N = 36 \) ) anterior to STPm still represented the parametric action space rather well (Supplementary Fig. 2; Procrustes Distance measure = 0.10).

In addition we assessed the contribution of instantaneous speed to the firing rate of each neuron using regression analysis (see Methods and Materials). We related the smoothed instantaneous firing rates of each neuron, binned to the stimulus presentation frame rate of 60 Hz, with the speed of the wristpoint. These regression analyses were performed for a set of time delays between the motion and instantaneous firing rate. For each neuron, we choose that time delay for which the amount of response variance explained by speed was the highest. The median neural response variances explained by the speed were 0.29 for "motion" neurons and 0.10 for "snapshot" neurons, a highly significant difference (Mann–Whitney U test; \( P < 0.0001 \)).

In the above analysis, the correlations between velocity and neural responses were computed across all 21 actions. When changes of the response profile within an action were considered, there were also differences between "motion" and "snapshot" neurons, as is illustrated in Figure 12 for the 7 different actions of Figure 1B. Figure 12 plots the speed profile of the wristpoint end-effector and the average normalized responses of the "motion" and "snapshot" neurons for each of these 7 actions. The average normalized response of "motion" neurons follows the speed profile of the wristpoint more closely than the average normalized response of "snapshot" neurons. Indeed, the median correlations between the instantaneous speed and firing rate, computed for each of the 7 actions, were larger for the "motion" (range of median \( r^2 = 0.35 \) to 0.76) than for the "snapshot" neurons (range of median \( r^2 = 0.02 \) to 0.21; Fig. 12). Thus, the peak responses of the "motion" neurons are not uniformly distributed during the action but favor action segments in which there is motion, that is, when the figure is engaged in an action.

The within- and across-action correlations of speed and response suggests that speed is an important factor in determining the responses of these "motion" neurons. However, in addition to the above reported, albeit weak, correlations between response and direction, analysis of the distribution of preferred actions in the parametric action space indicated that speed is not the only factor determining the neural responses. For the "motion" neurons, 45.1% preferred one of the 3 prototypical actions over the action blends, which is significantly more than the expected 14.3% (Binomial test; \( P < 0.0001 \); Fig. 13B). This bias for the extremities of the parametric action space was also present in the full population of neurons (Fig. 13C) and in the subpopulation of "snapshot" neurons.
neurons (Fig. 13G; 39.4% and 37.5%, respectively, Ps < 0.0001). The distribution of preferred actions is very different from that of the mean speed of the wristpoint (averaged over the whole stimulus duration; Fig. 13D) which varied systematically across the action space, having its maximum at the knocking action prototype and declining monotonically toward the throwing and lifting prototypical actions, and thus suggests that the action tuning does not result from mere speed tuning.

Results obtained in the reversed temporal order test also support the conclusion that speed is not the only factor determining the response of “motion” neurons. In the reversed temporal order test we tested whether neurons could distinguish between the same movies played forward or backward. The example cell shown in Figure 14A, recorded in the upper bank/fundus of the rostral STS, shows that this is indeed the case: the neuron responded more strongly to the reversed ineffective action than to the original ineffective action. In fact, the neuron followed the speed profile more closely for the reversed ineffective movie than when it was played forward. Similar effects of reversal were obtained in the population of 11 “motion” neurons recorded with the reversed temporal order test (Fig. 14B). The nonsignificant correlation between response and speed for the ineffective action when it was displayed as a normal sequence became significant when it was played backwards ($r = 0.15; P = 0.1; N = 11$ vs. $r = 0.81; P < 0.001, N = 11$). The speed-response correlations of the effective actions were significant for the original and reversed movies ($r = 0.69$ and $r = 0.83, P < 0.0001$ for original and reversed temporal order, respectively). Although the purpose of the reversed temporal order test was not to study the effect of speed, the different response profiles in the forward and reversed ineffective conditions, which have the same instantaneous speeds, do indicate that speed is not the sole factor determining the action selectivity.

Effect of Spatial Location of the Actions on the Responses of “Motion” Neurons

One possible explanation of the within-action modulation for “motion” neurons is the presence of a heterogeneity within the neuron’s RF through which the arm moves. If this explanation holds, then the timing of the peak of the response within the action should vary as the spatial position of the action changes because, for different positions, the trajectory of the arm will differ relative to the supposed RF heterogeneity. Figure 15A shows the response of one “motion” neuron in a position test (step size 1.5° in the horizontal and vertical dimensions) for the most effective action. Two effects are noteworthy. Firstly, the response amplitude of the neuron varied with the spatial position of the action stimulus and, secondly, the timings of the peak responses were invariant with regard to spatial position.

We examined position invariance in the timing of the peak response across the population of neurons by aligning, for each neuron, the responses at eccentric positions using the time of the peak firing rate determined at the foveal position. The results for the 13 “motion” neurons for which we had a position test are shown in Figure 15B. Eccentricities tested were 1.5°, 2.1°, 3°, and 4.2°. The population responses showed variation in the height of the peak response with the spatial position of the stimulus, but overall an invariance of the timing of the peak response with respect to spatial position. Note that the arm movement has a maximum amplitude in the x and y direction of about 1.5° and 2.5°, respectively. If the within-action modulations of responses had resulted from a RF heterogeneity or “hot spots” within the RF, then, given the range of the arm trajectories, the peaks should have shifted considerably or even disappeared at the eccentricities tested in Figure 15B. Thus, we conclude that the within-action modulation for “motion” neurons does not merely result from RF heterogeneities.

Relating Static Snapshot Selectivity and Within-Action Modulation for “Snapshot” Neurons

If the initial transient related to stimulus onset is ignored, the average responses of “snapshot” neurons shown in Figure 12 display little overall modulation during the course of the action sequences. However, it should be noted that these are responses averaged across neurons. A neuron-by-neuron visual inspection showed that many “snapshot” neurons do in fact show strong response modulation during the action sequence (e.g., Fig. 2C). The analyses reported above suggested that for such “snapshot” neurons, these within-action modulations correlate poorly with the speed of the wristpoint, a motion that has a maximum amplitude in the x and y direction of about 1.5° and 2.5°, respectively. If the within-action modulations of responses had resulted from a RF heterogeneity or “hot spots” within the RF, then, given the range of the arm trajectories, the peaks should have shifted considerably or even disappeared at the eccentricities tested in Figure 15B. Thus, we conclude that the within-action modulation for “motion” neurons does not merely result from RF heterogeneities.

---

**Figure 12.** Population responses and speed profiles compared. The mean normalized responses (±SEM) to 7 different actions (same actions as in Fig. 1B) for “snapshot” neurons (N = 64, left panel) and “motion” neurons (N = 51, right panel). The x-axis represents stimulus duration (in ms), whereas the y-axis represents the average normalized neural response. The instantaneous speed of the wristpoint for each action sequence is indicated by gray curves and is scaled to the across-action maximum of the average normalized responses for illustrative purposes. Median correlations ($r^2$) between the instantaneous speed and firing rate are indicated for each action and group of neurons.
parameter. Focusing on a form parameter, we examined whether the within-action response modulations observed in some “snapshot” neurons is related to the selectivity for different static snapshots. As expected, there was indeed a significant positive correlation, across “snapshot” neurons, between the Within-action Modulation indices and the Snapshot Selectivity indices ($r = 0.56; P < 0.001; N = 62$; see Methods).

Discussion
The present study examined the coding of visual actions of stick-plus-point-light figures by rostral temporal cortical neurons. A novel feature of the present study was that we employed a stimulus set in which the actions varied parametrically, enabling us to examine whether temporal cortical neurons can represent similarities between actions. Our results show that this is indeed the case, at least with respect to the ordinal relationships between the action stimuli. This suggests that the output of these neurons are useful for action categorization.

Natural actions differ not only in their kinematics but also in form parameters (posture). We examined to what degree rostral temporal cortical neurons respond to the form versus the motion information and found neurons that responded to the action movies but less so to static presentations of snapshots taken from the same movies (“motion neurons”) and other neurons that responded equally well to the action movies and to the static snapshot presentations (“snapshot neurons”). As with many other neural properties (e.g., the distinction between simple and complex RFs in early visual cortex), there was a continuum between the relative degree to which neurons responded to snapshots versus actions. The “motion” neurons were found mainly, but not exclusively, in the fundus and upper bank of the STS, whereas the “snapshot” neurons were found mainly, but not exclusively, in the lower bank of the STS and the lateral convexity of IT, suggesting an anatomical dissociation of this response property. This anatomical dissociation fits previous observations of motion responses in the upper bank of the STS (Bruce et al. 1981; Baylis et al. 1987), but unlike these earlier studies we could show the dissociation using the same stimulus sets in the different areas.

“Motion” Neurons
Our analyses suggest that the responses of the “motion” neurons depend on the speed of the end-effector. On average, higher speeds produced stronger responses. The correlation between speed and response for the “motion” neurons might not be surprising given that action segments with high speeds are those that contain the majority of the action and thus are the most informative for coding purposes. Indeed, our results show that at least for these simple action stimuli, STS neurons do not code for the action as a whole but for temporal segments of that action. Overall, these temporal segments coincided with those that contain most of the arm movement. However, it should be stressed that speed is certainly not the only determinant of the response of the “motion” neurons, because 1) the neurons represented the action space in a 2D and not a 1D configuration (speed is a 1D parameter), and 2) the distribution of the preferred actions did not fit the distribution of the speeds in the parametric action space. What factors besides speed contribute to the responses of the “motion” neurons? Our correlation analysis showed that direction of wristpoint motion and firing rate were weakly correlated across actions in the “motion neurons.” In addition, most STS neurons anterior to the STPm region were only weakly selective for opposing motion directions along the prominent axes of the wristpoint movement when tested with a translating movement of the snapshots. Both findings suggest
that the contribution of motion direction to the observed action selectivity is not that great in most neurons. The results of the position test showed that the neurons still responded to the same segments of the action at different spatial positions, suggesting that mere spatial RF heterogeneities are not causing the response modulations within and across actions. A third candidate factor is the relative motion of the arm segments. The results from the reduced-action configuration tests showed that the selectivity for the actions decreased when the stimulus was reduced to a single dot (wristpoint), suggesting that the relative motion of the arm points or arm segments indeed contributes to action coding. A fourth potential factor is the temporal context of the stimulus. Recently, Schlack et al. (2007) showed that the responses of MT neurons to accelerating and decelerating stimuli were affected by speed history, resulting from differences in the speed-history-dependent adaptation state between these conditions. If such a dependency upon temporal context is already present at that early stage of visual motion processing, it should not be surprising that similar temporal contextual effects occur at later stages such as the rostral STS (Baker et al. 2001; Jellema and Perrett 2003a). Note that taking motion history into account can contribute to the visual description of the movement in an action, because the movement within the context of different actions can have the same velocity at a particular instant but with a different velocity history.

We propose that the "motion" neurons code for the effector kinematics, allowing a computation of the action by downstream neurons. This proposal is based on the following observations: 1) "motion" neurons respond to segments of the action, 2) they display a tolerance to a reduction of the actor to the effector (arm) only, and 3) their responses partially correlate with motion parameters such as speed. Note that it is unclear how these neurons will respond to more complex actions that are not limited to a single limb: how will these

Figure 14. Selectivity of "motion" neurons to temporally reversed action sequences, that is, actions played backwards. (A) Peristimulus time histograms with raster plots of responses in the reversed temporal order test for one example neuron (lu088) recorded in the upper bank/fundus STS. Gray lines indicate the speed profile of the wristpoint, scaled to the maximum of the response (rows: top = most effective action, bottom = least-effective action; columns: left = original order, right = reversed temporal order). (B) Smoothed averaged normalized responses of the population of "motion" neurons \(N = 11\) tested in the reversed temporal order test and the averaged speed profiles of the motion sequences involved. Same conventions as in (A).

Figure 15. Neural responses at different spatial positions compared. (A) Peristimulus time histograms with raster plots of one "motion" neuron (lu082) recorded in the position test. In this position test, in addition to a foveal presentation, there were 8 positions located on a small grid of 1.5° spacing and 8 positions located on a larger grid of 3° spacing. Time course (in ms) and response strength (firing rate in spikes/s) are indicated on the foveal action sequence. Solid vertical lines mark stimulus onset and offset. (B) The within-action response modulation for "motion" neurons cannot be explained by RF inhomogeneities. The smoothed normalized responses of 13 "motion" neurons at different spatial positions were aligned to the time of the peak response at the foveal position (the same grid spacing holds as in A). The solid vertical line marks the time of the foveal peak response, which was set at 0 ms. The peaks at the parafoveal positions generally occur at times similar to the foveal peak.
neurons respond when several limbs move simultaneously inside their RF (as when viewing a walking person, for instance)?

"Snapshot" Neurons

Recent computational models as well as psychophysical studies have stressed the importance of form information for recognizing actions (Giese and Poggio 2003; Beintema et al. 2006; Lange and Lappe 2006; Lange et al. 2006) and have postulated the existence of units that respond to snapshot of actions, that is, "snapshot" neurons. Indeed, the "snapshot" neurons we found, responded to the form of the actor or parts of the actor during the action and thus such neurons, as a population, can contribute to the action coding by signaling the posture of the actor. Note that most "snapshot" neurons were rather broadly tuned to the different snapshots within a movie, limiting their ability to code for similar actions. Although the possibility cannot be excluded that their selectivity is greater when using full-body displays instead of stick-plus-point-light figures, it does put a limitation on the implementation of pure snapshot-based models of action recognition, such as the template-matching model of Lange and Lappe (2006).

Representation of the Parametric Action Space

We found that the population of "motion" neurons represented the similarities among the action movies more faithfully than the "snapshot" neurons, although the latter outnumbered the former in our neural sample. Thus it is tempting to conclude that the "motion" neurons can contribute more to action coding than the "snapshot" neurons. Given our observation that many neurons respond mainly to subsegments of the action, one would expect that if one takes into account the temporal evolution of the neural response, the coding of the actions will be substantially enhanced. This was indeed the case: when we performed the ISOMAP analysis of a distance matrix based on the concatenation of the responses of the neurons in successive 250 ms analysis windows, the configuration obtained was improved (Fig. 16, "snapshot" neurons: Procrustes Distance Measure = 0.16; "motion" neurons: Procrustes Distance Measure = 0.05). The configuration obtained using the temporal information from the "motion" neurons was excellent, especially when considering the relatively low number of neurons (N = 51) involved.

Note that these neurons do not code for action semantics such as lifting versus knocking. However, as the ISOMAP analysis shows, the information that these "motion" neurons provide, when properly integrated, allows an excellent reconstruction of the similarities between different actions, which can form the basis of the categorization of novel actions (according to their similarities with learned actions).

Tuning for the Extremities of the Parametric Action Space

Interestingly, the average activity of the neurons was larger for the prototypical actions than for the blended actions (Fig. 13). It is tempting to relate this finding to the tuning of static shapes for the extremities of simple shape dimensions (e.g., curvature) as has been observed in IT (Kayaert et al. 2005; De Baene et al. 2007). However, in the present case no such simple dimensions

---

**Figure 16.** Neural representation of the parametric action space in "motion" and "snapshot" neurons computed using 250-ms long segments of the actions. (A and C) Neural representations of the stimulus space obtained after applying ISOMAP (nonlinear MDS) to the mean responses averaged in 8 successive 250-ms long analysis windows: "motion" (A) and "snapshot" (C) neurons. (B and D) Scree plots of the residual variance as a function of the number of dimensions. Same conventions as in Figure 3.
are apparent and this suggests that the tuning for the extremities of a parametric space is induced by stimulus exposure. One possible mechanism that might underlie such tuning for extremities is repetition suppression: the blends are more similar to the other stimuli than are the prototypical actions, which lie on the extremities of the space. One might thus expect stronger similarity-based repetition suppression for the former compared with the latter stimuli. Another possible explanation for the stronger responses to the prototypical actions is that the neurons respond more strongly to natural than to unnatural, that is, the blended, action patterns. However, because the blends themselves appear rather natural, at least to human observers, we find this explanation implausible.

**Contribution of Visual Temporal Cortical Neurons to Visual Analysis of Actions**

Our results support a visual action coding scheme in which the movement of an end-effector is represented by neurons that require motion. At the population level, the information contained in the responses of these “motion” neurons are sufficient to compute the similarities among actions but these neurons do not represent actions as a whole, because they respond only to subsegments of an action sequence. Thus, further integration of these responses is needed to obtain full action categorization. These “motion” neurons are located predominantly in the dorsal bank and fundus of the STS, whereas neurons in the more ventral and lateral parts of the visual temporal cortex respond to static snapshots as well as to actions. These “snapshot” neurons can contribute to action coding by signaling the presence of particular postures within the movement. Our data suggest that the “snapshot” neurons represent the similarities among the action movements to a lesser degree than “motion” neurons. Further research using more complex, multilimb actions and involving a comparison of the sensitivities of the neural and behavioral responses is underway in order to better understand the contribution of these different neurons to action coding.

**Supplementary Material**

Supplementary material can be found at: http://www.ercor.oxfordjournals.org/

**Funding**

Geneeskundige Stichting Koningin Elizabeth, GOA (2005/18), Detection and Identification of Rare Audiovisual Cues (DIRAC, FP6-IST 027787), EF (05/014), IUAP, FWO Flanders, and The Engineering and Physical Science Research Council (UK).

**Notes**

The technical assistance of P. Kayenbergh, G. Meulemans, M. De Paep, W. Depuydt, and S. Verstraeten is gratefully acknowledged. We thank R. Peeters for his help in taking structural magnetic resonance images from our monkeys. Dr Michael Gleicher, Dr Lucas Kovar, and Dr Yingliang Ma helped with the stimulus construction and the blending algorithm. J. V. is a research assistant of the Fund for Scientific Research (FWO) Flanders.

**Address correspondence to Dr Rulfin Vogels, Laboratorium voor Neuro- en Psychofysiologie, K.U. Leuven Medical School, Campus Gasthuisberg, O&N2, Herestraat 49, Bus 1021, 3000 Leuven, Belgium. Email: Rulfin.Vogels@med.kuleuven.be.

**References**


Download BibTeX Entry