Detection and Identification of Rare Audiovisual Cues

Insperata accident magis saepe quam quae speres.
(Things you do not expect happen more often than
things you do expect) Plautus (ca 200(B.C.)

Project no: 027787

DIRAC

Detection and Identification of Rare Audio-visual Cues

Integrated Project
IST - Priority 2

DELIVERABLE NO: D 3.3
Conclusions from the First Neurophysiological STP
Single-Cell Experiments

Date of deliverable: 30.06.2007
Actual submission date: 31.07.2007

Start date of project: 01.01.2006 Duration: 60 months

Organization name of lead contractor for this deliverable: **KU Leuven**

Revision [1]

Project co-funded by the European Commission within the Sixth Framework Program (2002-2006)		
Dissemination Level		
PU	Public	X
PP	Restricted to other program participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

DIRAC
Detection and Identification of Rare Audiovisual Cues

Insperata accident magis saepe quam quae speres.
(Things you do not expect happen more often than
things you do expect) Plautus (ca 200(B.C.)

D3.3 Conclusions from the First Neurophysiological STP Single-Cell Experiments

K.U. Leuven

Abstract:
We report the analysis and conclusions of the first single cell recording study in macaque visual temporal cortex, in particular rostral areas STP and ventral STS and the lateral convexity of IT. The novel feature of our study is the use of a parametric space of simple arm actions. The actions were displayed by stick figures. Our results support an action coding scheme in which the motion of an end-effector is analyzed by neurons that do not respond to static presentations of snapshots but require motion. At the population level, the information contained in the responses of these so called "motion" neurons are sufficient to compute the similarities among novel, unfamiliar actions but the neurons themselves do not represent actions as such since they responded only to segments of an action sequence. Thus, further integration of these responses is needed to obtain a full action code. These "motion" neurons were predominantly located in the dorsal bank and fundus of the STS, while neurons in the more ventral and lateral parts of the visual temporal cortex responded to static snapshots as well as to actions. We found that these so called "snapshot" neurons represent the similarities among the actions to a lesser degree than the motion neurons. Further research using more complex, multi-limb actions as well involving a comparison of the sensitivity of the neural and behavioral responses is underway to understand the contribution of these different neurons to action coding.

Table of Content

1. Introduction

Single cell studies in macaque monkeys have found that neurons of the rostral Superior Temporal Sulcus (STS) respond to visual displays of body movements (Bruce et al., 1981; Oram and Perrett, 1994; Oram and Perrett, 1996; Perrett et al., 1989; Perrett et al., 1985; Perrett et al., 1990; Jellema et al., 2004; Jellema and Perrett, 2003; Jellema and Perrett, 2006; Barraclough et al., 2005; Keysers and Perrett, 2004). Most of these studies employed real-life stimuli, e.g. the experimentator walking in front of the animal, which precludes stimulus control up to the standards of studies in other visual cortical areas. Nonetheless, these studies suggested that some STS neurons responded selectively to different classes of perceived body movements, such as grasping versus locomotion.

In the present study we employed well controlled dynamic visual images of biological motion, allowing a quantitative analysis of the responses of temporal cortical neurons to the dynamic stimuli. Since we were interested in the coding of the dynamics of visual actions, we reduced the form information by using stick figures instead of real actors. However, the motion as well as the bodily patterns were derived from motion capture data of real human actors. We used stick figures instead of full body figures since the latter contains additional texture, shading and form information, which might render it difficult to determine the stimulus properties the neuron is selective for. On the other hand, we did not reduce the stimulus further to point light displays (Johansson, 1973), since we reasoned that stick figures would produce stronger responses than point light displays.

The present work had several novel features which enabled us to address important questions of visual action coding. Firstly, we employed a parametric action space which in design is similar to previous parametric static shape spaces that were used to examine quantitatively the shape selectivity of ventral stream visual neurons (Op de Beeck et al., 2001; Freedman et al., 2003; Pasupathy and Connor, 2001; De Baene et al., 2007). The action space was created by blending 3 different transitive arm actions: knocking, lifting and throwing. The three-way blends differed in the amount to which each of the three actions contributed to the blend, producing systematic and smooth variations between the stimuli.

Computational and psychophysical work suggests that perceptual categorization depends strongly on stimulus similarity (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, their response would depend on the similarity between actions (distance functions; see Edelman, 1999). Previous work showed that inferior temporal (IT) neurons represent rather faithfully the ordinal similarity relations between static shapes (Op de Beeck et al., 2001; DeBaene et al., 2007). In the present study we established whether this also holds for dynamic images of actions. Actions are a more complex visual stimulus than static shapes since visual actions by nature consists of temporal changes in form, and thus it is an open question whether single neurons can code for similarities between actions and how this coding evolves over the course of the action movie. Human psychophysical research that employed the same action space that we employed here showed that the perceptual representation of the similarities between these action blends fits the parametric stimulus configuration rather well (Pollick et al., 2007), indicating that our stimulus set is adequate to study the neural coding of action similarity.

The second set of questions that we examined relates to contribution of form versus motion information to the coding of actions. Since different actions usually differ in both form and motion, action coding can be based on form and/or motion information. The potential contribution of form and motion information is nicely shown in the computational model of action recognition by Giese and Poggio (2003): they postulated an action processing stream that is based on motion analysis and a parallel one that is based on form information present in the snapshots of the action sequence. In order to determine the contribution of form versus

motion information to the neural responses to the action stimuli, we compared the responses to the dynamic action sequences with those to static presentations of representative snapshots from the movies. In addition, we reduced the stimulus (Tanaka, 1996) by systematically removing limbs of the human figure until the moving arm or the end-effector only – i.e. the wrist point -was left.

Unlike in previous single cell studies of action coding, the actions we employed can be well characterized quantitatively since these are restricted to one 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 spatial form differences among the stimuli and to motion parameters such as velocity and acceleration. Finally, we examined whether the neurons would respond to rigid, translation of the representative snapshots or reduced stimuli. Biological motion is essentially non-rigid and thus by comparing the responses to rigid translation and the non-rigid motions of the actions, we could ascertain whether the neurons' responses are specific for non-rigid motion.

Bruce et al. (1981) observed responses to walking humans in the upper bank/fundus of the rostral STS (Superior Temporal Polysensory (STP)). This region contains neurons with large receptive fields 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 it is a candidate region to process dynamic images of actions. Indeed, following Bruce et al. (1981), Perrett and colleagues reported responses to perceived actions in STP (Oram and Perrett, 1996; Oram and Perrett, 1994; Baraclough et al., 2006; Jellema et al., 2004; Baraclough et al., 2004; Jellema et al., 2003; Jellema et al., 2006). However, responses to perceived hand actions (Perrett et al., 1989) and locomotion (Baraclough et al., 2004; Baraclough et al., 2006) have also been observed in the ventral bank of the rostral STS, well outside STP. Also, recent awake monkey fMRI work observed activation to a hand grasping an object, compared to static presentations, in both banks of the STS (Nelissen et al., 2006). Thus we searched for responsive neurons in both ventral and dorsal banks of the STS, as well as the lateral convexity of IT, and compared the responses in the different tests for the two banks.

2. Methods

2.1 Subject and Surgery

Two male rhesus monkeys (*Macaca mulatta*; Monkey L and B) served as subjects. Before conducting the experiments, aseptic surgery under isoflurane anesthesia was performed to attach a head fixation post to the skull and to stereotactically implant a plastic recording chamber (Cryst Instruments). The implantation of the recording chambers was guided using preoperative structural magnetic resonance (MRI) scans of each of the animals. The recording chambers were positioned dorsal to the rostral STS, allowing a vertical approach. We recorded from both hemispheres of monkey B: the recording chamber on its right hemisphere was positioned at coordinates which were comparable to those of the other animal, while the recording chamber on the other hemisphere was located more posteriorly. The explored anterior-posterior range of the recordings across animals was from 7 to 19 mm anterior to the auditory meatus. At several instances during the course of the recordings, we took MRI scans (3 Tesla – 1mm resolution), with a copper sulphate tube inserted in the grid at several recording positions. By comparing these MRI images, depth readings of the white and gray matter transitions and of the skull base during the recordings, with the microdrive readings (referenced to the bottom of the grid), we were able to estimate the recording positions and specifically assign the neurons to the upper bank/fundus STS versus 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.

2.2 Apparatus

The stimuli were presented on a CRT with a frame rate of 60 Hz. The stimulus presentation and task was controlled by a PC running home made software. The monkey was seated in a monkey chair with its head fixed facing the display. The position of one eye was measured using infrared video-based eye trackers. Initially we employed ISCAN (120Hz) while in later recordings, we used the EYELINK (1000Hz) system.

Extracellular single unit recordings were recorded using FHC epoxylite insulated tungsten microelectrodes (0.7 – 2 MOhm measured in situ) that we lowered into the brain using a Narishige microdrive. The electrode was lowered through a guiding tube that was fixed in a plastic grid (Cryst Instruments) which was positioned in the recording chamber. The electrode signal was amplified and filtered using conventional single unit recording equipment. Single units were isolated on line using a custom made DSP based spike discriminator that accepted a spike when its waveform crossed several boxes that differed in time and level. The time stamps of the isolated spikes, stimulus events, and eye position were stored by a PC (1 ms resolution), using home made software written in LABVIEW.

2.3 Task

The animals performed a passive fixation task during the recordings. The trial started with the onset of a small square fixation target that was located in the middle of the display. When the monkey fixated this target for at least 500 ms, the stimulus was presented together with the fixation target. After presentation of the stimulus the monkey was required to fixate another 200 ms the target. If the monkey kept its gaze within a square fixation window (approximate size 1.8 deg) from the beginning of fixation until the end of the 200 ms post-stimulus fixation period, the trial was accepted as valid and the monkey obtained a liquid reward.

2.4 Stimuli

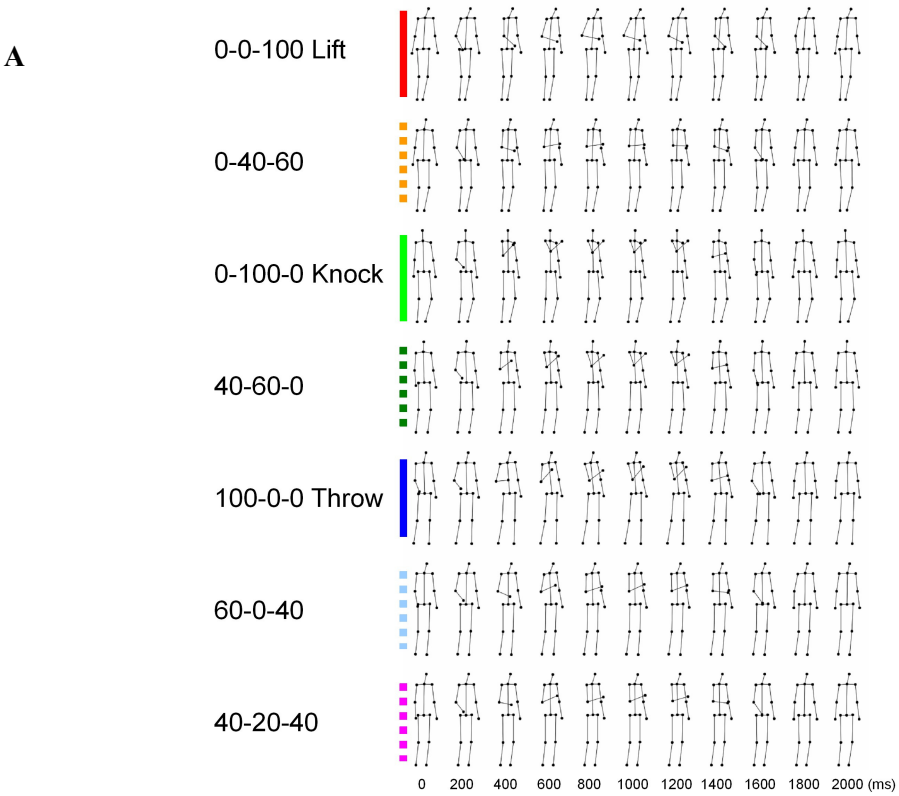
2.4.1 Basis Stimulus Set

All neurons were searched and tested with a set of 21 movies depicting human stick figures that performed arm actions. The duration of all movies was approximately 2 s (120 frames shown at a 60 Hz frame rate). The stick figures measured approximately 6 deg in height and 1.5 deg in width. Initially the movies were displayed centrally. During the course of the recordings we noted that monkey L developed a tendency to produce pursuit-like eye movements at the end of the movies. To ensure stable fixation, we presented the movies in the contralateral field at 1.5 deg eccentricities during the remainder of the recordings in both animals. Results from the central and (slight) eccentric positions are pooled in all the population analyses of the present report.
The movies consisted of 3 real actions, knocking, throwing and lifting, and their three way blends. The blends were created using the algorithm of Kovar and Gleicher (2003), which preserves biomechanical movement constraints.

Figure 1A shows undersampled sequences of snapshots of the 3 real action movies and 4 blends . The blends were produced in steps of 20%, thus having 3 x 4 “two-action” blends (outer triangle in Figure 1B) and 6 “three-action” blends (inner triangle in Figure 1B). The parametric action space of these 21 actions can be approximated by triangular 2D configuration with the 3 real actions being the endpoints (Figure 1B). This 2D configuration is a non-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 triangle approximation is sufficient and will be used throughout the paper for data displays and analyses.

Since the movies were generated from real human actions and we wished to keep these as realistic as possible, there were slight differences in the start posture between the actions (see Figure 1A). However, mean speed and arm trajectories differed systematically between the differed actions. Figure 2 shows that the speed of the end-effector (wrist point) increases from lift to throw and from throw to knock. Also, these speed profiles clearly show that within an action movie the speeds vary with a multimodal distribution. Similar multimodal speed distributions, but much reduced in amplitude, are present for the elbow and the shoulder point. These variations in speed within and across actions were used 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 two parts consisting of rightward arm motion followed by a leftward, return arm motion. The vertical ranges of the arm movement differs between the actions, with the lift and throw movements going less upward than the knock movements (see Figure 1A).



B.

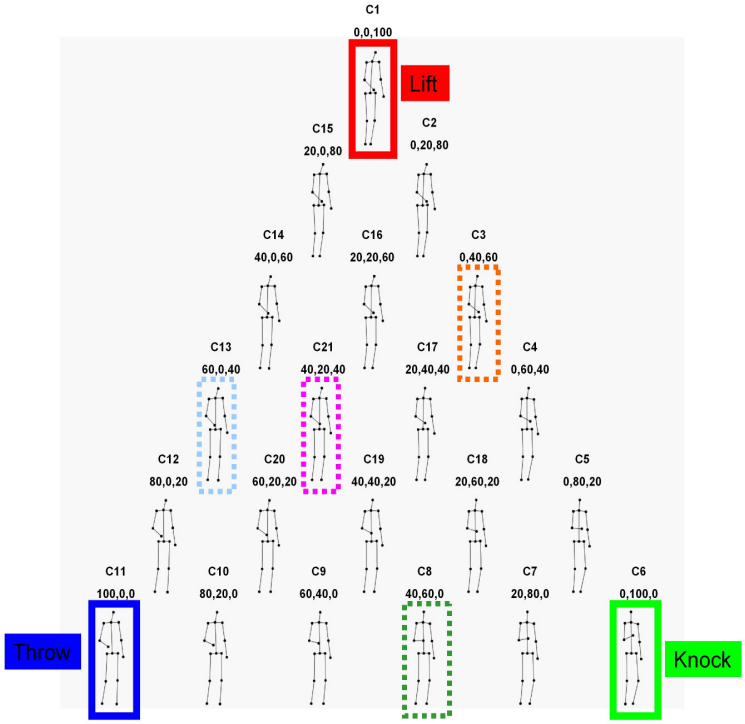


Figure 1. Basis stimulus set. A. Snapshots downsampled to one every 12 frames of 7 action movies. Action coordinates are indicated to the left (%Throw, %Knock ,%Lift) B. Triangular stimulus space with the coordinates of the 21 movies (%Throw, %Knock, %Lift and condition numbers). One chosen snapshot (occurring at approximately 300 ms in the movie) is shown for each movie. Colored boxes indicate the actions shown in A.

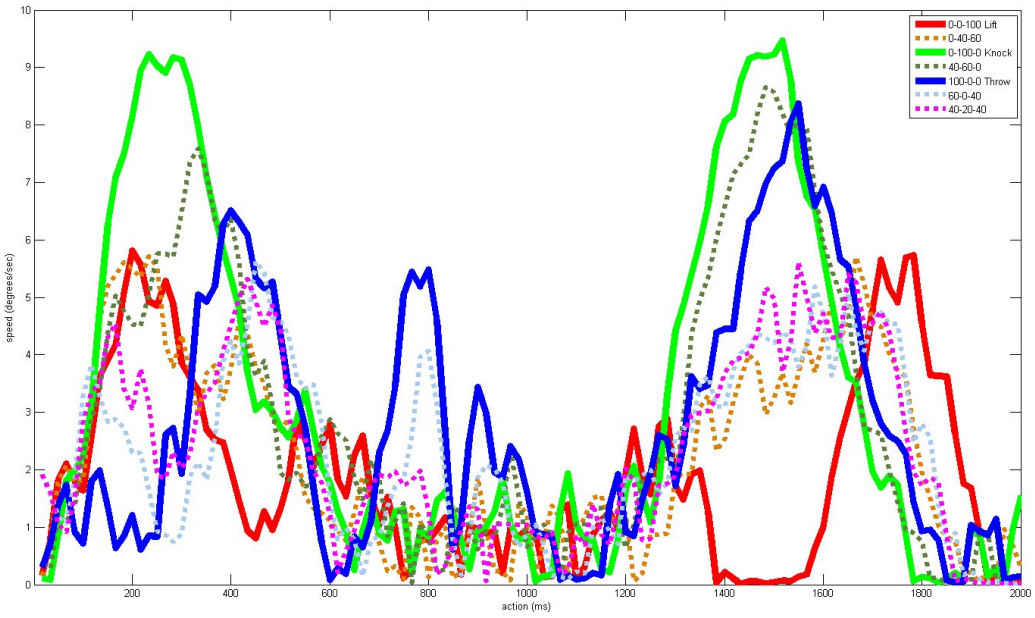


Figure 2. Speed of the end-effector (wrist point) during the action for the same 7 actions as depicted in Figure 1A. Same color code of the actions as in Figure 1.

2.4.2 Static Snapshots

For each movie we selected 6 snapshots that were representative of the variations between the different postures of an action. Each snapshot was shown for 300 ms, which is sufficiently long to produce responses in inferior temporal neurons (De Baene et al., 2007).

2.4.3 Translating Snapshots

In addition to static presentations of selected snapshots we also translated the same 6 snapshots across the screen. The velocity of translation of each snapshot was equated to that of mean velocity of the wrist dot at the moment the snapshot occurred in the movie sequence. The translation duration was always 2 s and thus the translation amplitude depended on the speed. The translation was centered at the same spatial position as used when presenting the movies and the static snapshots. Two opposing movement directions were employed so that we could assess the direction selectivity of the neuron.

2.4.4 Reduced Stimuli

We reduced the complexity of the line figure by systematically deleting body parts/limbs of the figure in successive steps until only the arm (3 dots connected by 2 lines) that performed the action remained present (Figure 3). Later in the recordings, some neurons were tested with further reductions: the wrist and elbow dot connected by one line and the wrist dot only. The reduced action movies had the same duration and were presented at the same spatial location as the original, non-reduced movie.

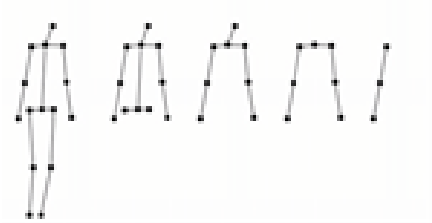


Figure 3. Steps in the stimulus reduction illustrated for a snapshot of one action. Note that the actual reduced stimuli were movies of the action and not static presentations of snapshots. Also, some neurons were tested with 2 further reductions: two points connected by a line (underarm) and the wrist point only.

2.5 Test

2.5.1 Basis Stimulus Set Test.

The responses of each neuron to the basis stimulus set were measured. The movies were presented in an interleaved, pseudorandom fashion. When the isolation of the neuron was still adequate, the neuron was subsequently tested in one or several of the following tests.

2.5.2 Snapshot and Translation Test

The purpose of these tests was to compare the response of the neuron to an action movie and static and/or translating snapshots of the movie. During the course of recording we developed different versions of this test. In the initial version, after choosing the most and least effective action, based on the responses of the neuron in the preceding basis stimulus set test, static presentation of 6 different snapshots of each of the 2 movies were interleaved with the presentation of the 2 actions movies. In a later version, the test consisted of the effective

movie, static snapshot presentations of that movie and translating snapshot presentations. All conditions were presented in an interleaved fashion.

2.5.3 Reduction Test

The purpose of these tests was to measure the effect of deletion of parts of the stick figure – the non-informative parts regarding the motion related to the action (see Figure 3) – on the responses of the neuron. Reduced action movies of the most and least effective movie and the latter two movies were presented interleaved in a pseudorandom fashion. In the initial version of this test, the most reduced version consisted of the arm only, while in a later version, movies depicting two further stimulus reductions (the underarm and the wrist point only) were also presented.

2.5.4 Position Test

The purpose of these tests was to measure the dependency on the spatial location of the movie on the response. We ran several position tests. In one version, the most and least effective action movies were shown at 9 positions that were located on a rectangular grid with a spacing of 5.3. deg. In a second version, the most effective movie was shown at 17 different positions with a spacing of 1.5 deg. The different locations were presented interleaved. The center of the grids corresponded to the foveal position.

2.6 Data Analysis

2.6.1 Basis Stimulus Set Test

To test whether the neural responses differed significantly from baseline activity, we computed for each trial the average firing rate in a baseline and 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 response latency of the neuron – and lasted 2000 ms. The significance of a stimulus related response was tested by a split-plot ANOVA (Kirk, 1968) with baseline versus stimulus activity as a repeated measure within-trial factor and stimulus condition (21 actions) as an across-trial factor. Responses were considered to be statistically significant when there was a significant main effect of the baseline-stimulus activity factor or a significant interaction between the two factors. Type 1 error was set at 0.05.

To determine whether a population of neurons represented the parametric similarities between the stimuli, we computed for each pair of stimuli their Euclidean distance based on the neural responses to these stimuli. As neural response we took the normalized response (maximum response of the neuron equal to 1). The response-based Euclidean distance between two stimuli was computed by subtracting for each neuron the normalized response to these stimuli, summing the squared response differences across neurons, and the taking the square root of this sum. The matrix of all pairwise Euclidean distances was then subjected to the non-linear multi-dimensional scaling (MDS) method ISOMAP (Tenenbaum et al., 2000). The latter represents in a low-dimensional space the geodesic distances between the responses to the stimuli. We favored this non-linear dimensionality reduction technique instead of the classical MDS since it captures better the distances along non-linear surfaces. The ISOMAP algorithm has one free parameter, k, and all the results reported in the present paper were obtained with $k = 4$.

The ISOMAP analysis was performed on the mean normalized firing rates computed for the whole stimulus duration as well as for firing rates computed for shorter successive 250 ms long periods. Also, in order to determine how well a neuronal population can represent the

stimulus space when taking into account changes in firing rate during the evolution of the action sequence, we performed an ISOMAP analysis using Euclidean distances computed on firing rates in successive 50 ms long intervals. For this analysis, the pairwise Euclidean distances were computed as the square root of the squared sum of the response differences for N neurons x 40 time intervals. Thus, this analysis takes into account the empirical fact (see Results) that responses for a given stimuli can vary dramatically 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 (the Procrustes Distance measure) between the ISOMAP derived 2D stimulus representation and the parametric triangular stimulus configuration by computing the sum of squared errors between the spatial coordinates of corresponding stimuli for the two spaces after we Procrustes rotated the ISOMAP derived space towards the parametric space.

The degree of stimulus selectivity of a neuron was quantified by computing the omega square index (Kirk, 1968) which estimates the proportion of variance due to stimulus variations. The omega square index takes into account differences in response among the 21 movies as well as trial-to-trial variability. The omega square selectivity index was computed using the responses for the whole stimulus duration as well as for the shorter 250 ms long successive periods during the stimulus presentation.

We correlated the instantaneous firing rate of each neuron to the instantaneous stimulus speed. For this analysis we estimated the instantaneous firing rate of the neuron by convolving the response of the neuron with a Gaussian with s.d. = 25 ms. The responses were correlated with the speed of the end-effector point – both binned at the frame rate of 60 Hz – using several time delays between the speed and neural response measures. These correlation analyses were performed across the 21 movies. Also, we performed the analyses separate for the two halves of the action sequences, i.e. for the two different movement directions (to the right versus to the left). This was done (1) to prevent the underestimation of the speed-response correlation for direction selective neurons, and (2) to determine whether the response-speed correlations differed between the two halves of the action sequences, i.e. between directions.

2.6.2 Responses to Static and Translating Snapshots Compared to Actions

Since the responses of the neurons varied during the course of the action movie, averaging the responses across the 2000 ms action duration can underestimate the response to a part of the action. To avoid this underestimation of the neuron's response to the action sequence, we used peak firing rate instead of average firing rate as response measure when comparing the response to the actions and the snapshot presentations. The peak firing rate was computed after convolving the response of the neuron using a Gaussian with an s.d. of 25 ms. In all these analyses we employed net peak firing rates, which were calculated by subtracting the average firing rate obtained in the baseline analysis window from the peak firing rate observed during the stimulus presentation.

To compare quantitatively the responses to static presentations of the snapshots and to the corresponding action (presented in an interleaved fashion in the same test), we computed the following Snapshot index:

Snapshot index = (peak firing rate action – maximum peak firing rate snapshots)/ (peak firing rate action + maximum peak firing rate snapshots).

We took the maximum instead of the average of the peak firing rate to the different snapshots since taking the average of the responses to the different snapshots would have underestimated the snapshot response in neurons that showed different, selective response to the individual snapshots.

The degree of selectivity for the static presentations of the snapshots was quantified by the following index:

Snapshot selectivity index = (peak net firing rate for most effective snapshot – peak net firing rate for least effective snapshot)/ (peak firing rate for most effective snapshot + peak net firing rate for least effective snapshot).

This index was computed for those neurons for which the peak net response to any of the static presentations of a snapshot exceeded 10 spikes/sec.

To compare quantitatively the responses to translations of the snapshots and to the corresponding action (presented in an interleaved fashion in the same test), we computed a Translation index:

Translation index = (peak firing rate action – maximum peak firing rate translations)/ (peak firing rate action + maximum peak firing rate translations).

We took the maximum of the peak firing rates to each of the two directions of translation of the different snapshots. Also, to prevent considering responses related to stimulus onset instead of to translation, we determined in this analysis the peak firing rate using an analysis window that started 200 ms after stimulus onset and ended 50 ms after stimulus offset. This 1850 ms long analysis window was used to compute the peak firing rate for the translating snapshot conditions as well as for the action movie condition of the same test.
The translating snapshots were presented in each of two directions that differed by 180 deg. This allowed us to determine whether the neuron was responding in a direction selective way to the translating snapshots. We computed a direction selectivity index using the peak firing rate in the 1850 ms analysis window. This index was computed only for those neurons that showed a maximum peak net firing rate of at least 10 spikes/s in at least one of the translation conditions. The index measures the degree of direction selectivity at the translation condition that produced the strongest response:

Best direction index = (Peak firing rate at best direction – Peak firing rate at opponent direction) / (Peak firing rate at best direction + Peak firing rate at opponent direction).

2.6.3 Reduction Test

To compare quantitatively the response to the full "body" action and the arm only movie, we computed the arm reduction index for the effective action:

Arm reduction index = (Peak firing rate full body – Peak firing rate arm) / (peak firing rate full body + Peak firing rate arm).

We computed this index using peak net firing rates for the full 2000 ms analysis window. Similar indices were computed for the other reductions (see Results).

2.6.4 Position Test

Again, all analyses were done using peak net firing rates. Since the arm trajectories differ between the different actions, one possible explanation of selective responses to the different movies would be that the receptive field of the neurons contains one or more hot spots. In such a scenario, the temporal response profile in a particular action condition will differ as a function of the spatial location of the stimulus, since the hot spot will be traversed by the arm at different moments in time or not at all. To examine this, we compared the time with respect to stimulus onset of the net peak firing rate for the most effective action presented at different spatial locations.

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

3. Results

We searched for temporal cortical neurons that responded to movies of a stick figure that performed simple arm actions. We will report here on the responses of 240 temporal cortical neurons recorded in two monkeys. Each of these neurons responded significantly to one or more stimuli of the basis set (ANOVA; $p < 0.05$). We explored the rostral STS and the lateral convexity of IT. In monkey L we explored 13 guiding tube positions ranging between 11 and 19 mm anterior. Responsive units were found for 11 of these guiding tube positions, ranging between 11 and 19 mm anterior. Of these responsive neurons, 81 were localized in the upper bank and fundus of the STS, 26 in the lower bank and 23 in the more ventral lateral convexity. In monkey B we searched for responsive neurons using 29 guiding tube positions, ranging between 7 and 18 mm anterior. Responsive units were found for 19 of these guiding tube positions. At 7 mm posterior, a patch of neurons ($N = 20$) was found in the medial part of the lower bank/fundus of the STS that showed strong motion related responses and these neurons will be treated as a separate population. Since it is possible that these neurons are located in the motion area LST defined by Nelissen et al. (2006) in a monkey fMRI study, we will refer to this patch of neurons as "putative LST" neurons. More anteriorly, 23, 30 and 37 responsive neurons were found in the upper bank, lower bank and lateral convexity, respectively. The different regions were sampled unevenly with repeated recordings at locations showing responses to the actions. It was clear from the recordings that neurons responding to these dynamic stick figures were organized in patches, since at several guiding tube positions no responsive neurons were observed in the ventral or the dorsal bank of the rostral STS, despite repeated penetrations.

3.1 Responses to the Basis Stimulus Set: Examples of Single Neurons

Figure 4 shows the responses of 4 representative neurons to the 21 action movies of the basis stimulus set. The neuron of Figure 4A was recorded in the "putative LST" region of monkey B. It responded selectively to the second part of the action movies, when the arm was moving downward. Thus, this neuron responded only to a particular temporal segment of the action instead of to the whole action sequence. In addition, the neuron responded much stronger to some actions than to others and its response decreased with increasing distance from the most effective action.

The neuron of Figure 4B was recorded in the dorsal bank of the STS of monkey L. Its response was strongly modulated during the course of the action, responding much stronger during particular action segments than during others. It was selective for particular actions, with the response decreasing with increasing distance from the most effective action. Another neuron recorded in the dorsal bank of the STS showing tuning in the action space is shown in Figure 4C. Figure 4D shows a neuron recorded in the ventral bank of the STS that showed strong selective response to a particular segment of a small number of the actions.

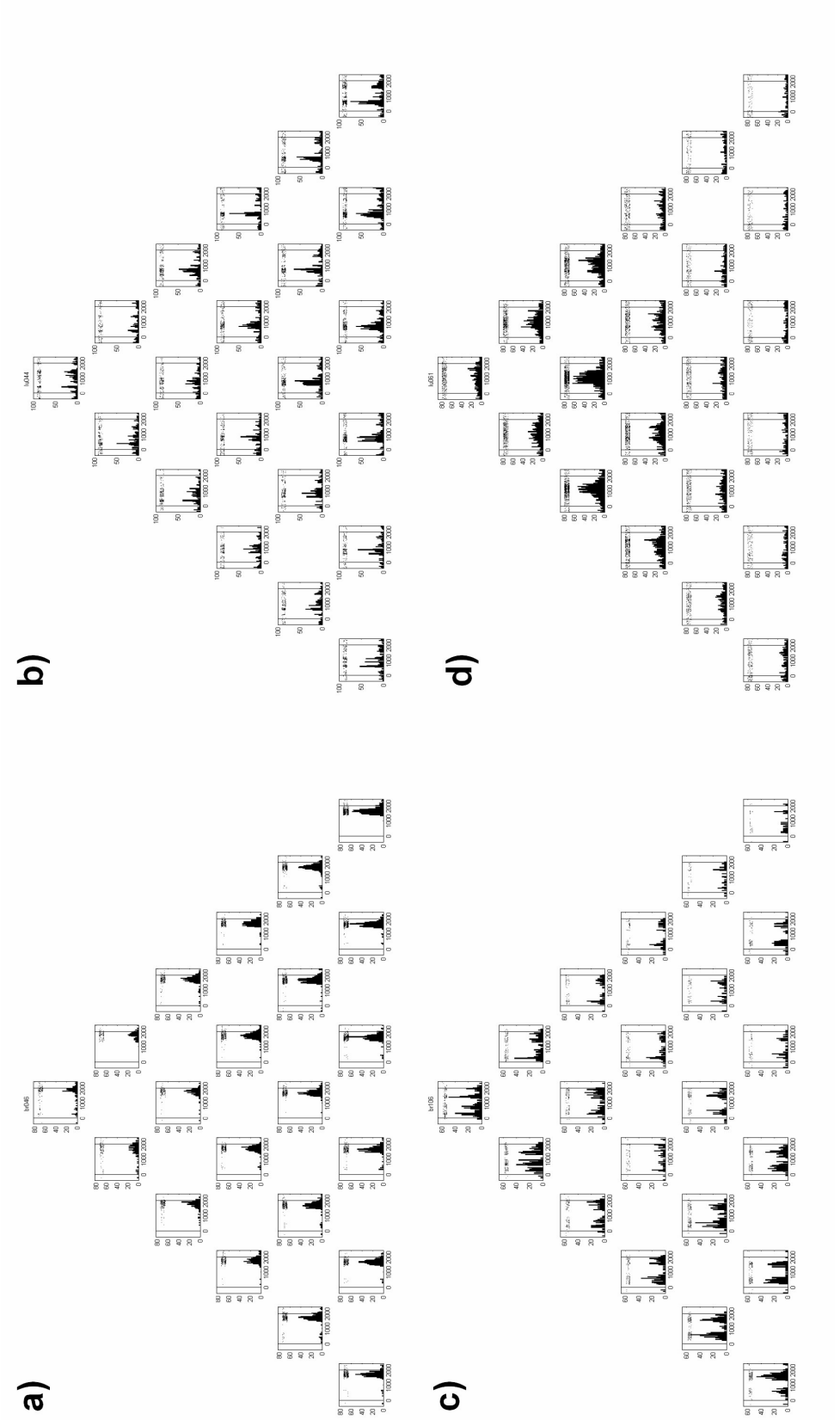


Figure 4. *Responses of 4 temporal cortical neurons to the 21 action movies of the parametric space. The stimuli are ordered as in Figure 1B (Top: Lift; Bottom left: Throw; Bottom right: Knock). A: Putative LST neuron; B and C: dorsal STS neuron; D: ventral STS neuron.*

3.2 Responses to the Basis Stimulus Set: Population Analysis

The single cell examples of Figure 4 are representative for our population of recorded neurons in (1) that they responded to a segment of the action movie instead of to the whole action and (2) that when showing tuning in the action space, their response decreases with increasing distance from the effective action in the action space. If the average responses of the population of neurons show tuning in the action space, then one would expect that as a population they would be able to represent the similarities between the actions. However, this neural action space can be a deformed with respect to the parametric action space, reflecting the intrinsic sensitivity of the population of neurons to differences among particular actions.
In order to determine how the population of 240 neurons represent the parametric action space we performed a non-linear MDS, ISOMAP, on the response differences for all stimulus pairs (see Methods). The firing rates were computed for the whole 2 s stimulus duration, ignoring the variations of the responses during the course of the action. The Scree plot showed that a 2D solution was the optimal low dimensional one and provided an excellent fit to the data, explaining 95% of the variance. Figure 5B shows the 2D configuration of the stimulus space based on the normalized responses of the 240 neurons. The roughly triangular neural stimulus space (Figure 5B) reflects to some degree the parametric, triangular stimulus configuration (Figure 5A): the neural space preserves the stimulus rank along the 3 sides of the triangle (the two-way bends) and also the ordinal relationships among the 3 way blends are preserved. The triangular configuration however is deformed: the neurons distinguish on average more the lift from the other 2 real actions than the knock from the throw. The 1D solution provided a bad fit (normalized response: 77% explained variance) indicating that the responses of the neurons do not merely depend on variations of a one-dimensional stimulus parameter, such as for example mean speed. Thus we conclude that this – unselected - population of visual temporal cortical neurons represent rather faithfully the ordinal similarity relationships among dynamic action stimuli.

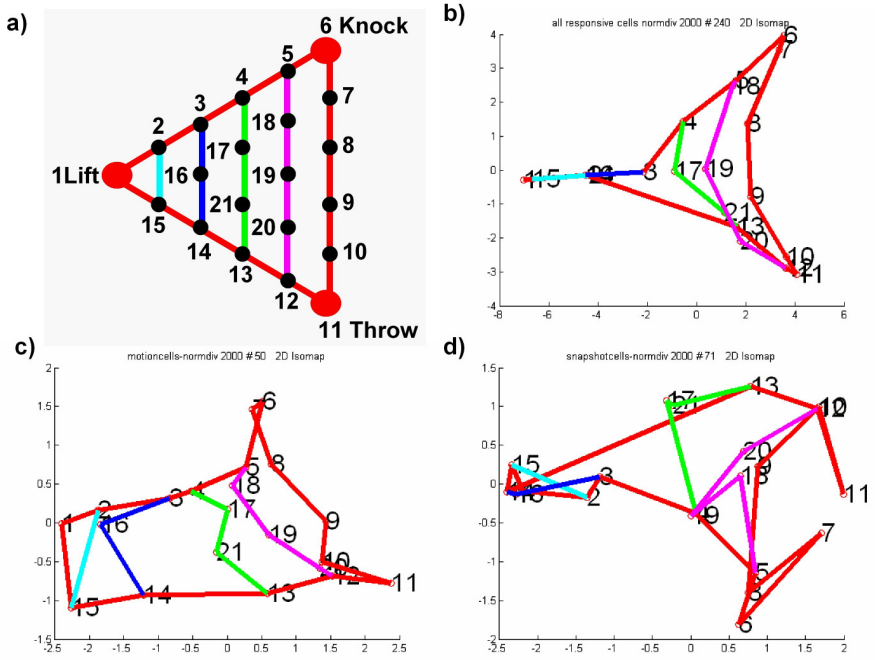


Figure 5. *ISOMAP solutions based on the responses of all neurons (B), the motion neurons (C) and the snapshot neurons (D). The parametric configuration is shown in A using the same color codes as for B-D.*

The distribution of the most effective action was significantly different from a uniform distribution (Chi Square test; $p < 0.00001$) and showed 3 peaks that corresponded to the real actions (Figure 6). Thus, 44 % of the neurons showed the greatest response to one of the 3 real actions, which is much larger than the expected 14%. The same preference for the real actions was seen for the responses averaged across neurons and this for both unnormalized and normalized responses (Figure 6; repeated measure one way ANOVA on unnormalized net responses: $P < 0.0005$; on normalized responses: $P < 0.005$). Overall, the mean response was greatest for the Lift action. Note that this preference for the real actions, the extremities of the parametric space, cannot be explained by simple stimulus parameters such as mean speed or mean position of end-effector since the latter vary monotonically between the real actions, unlike the response preference.

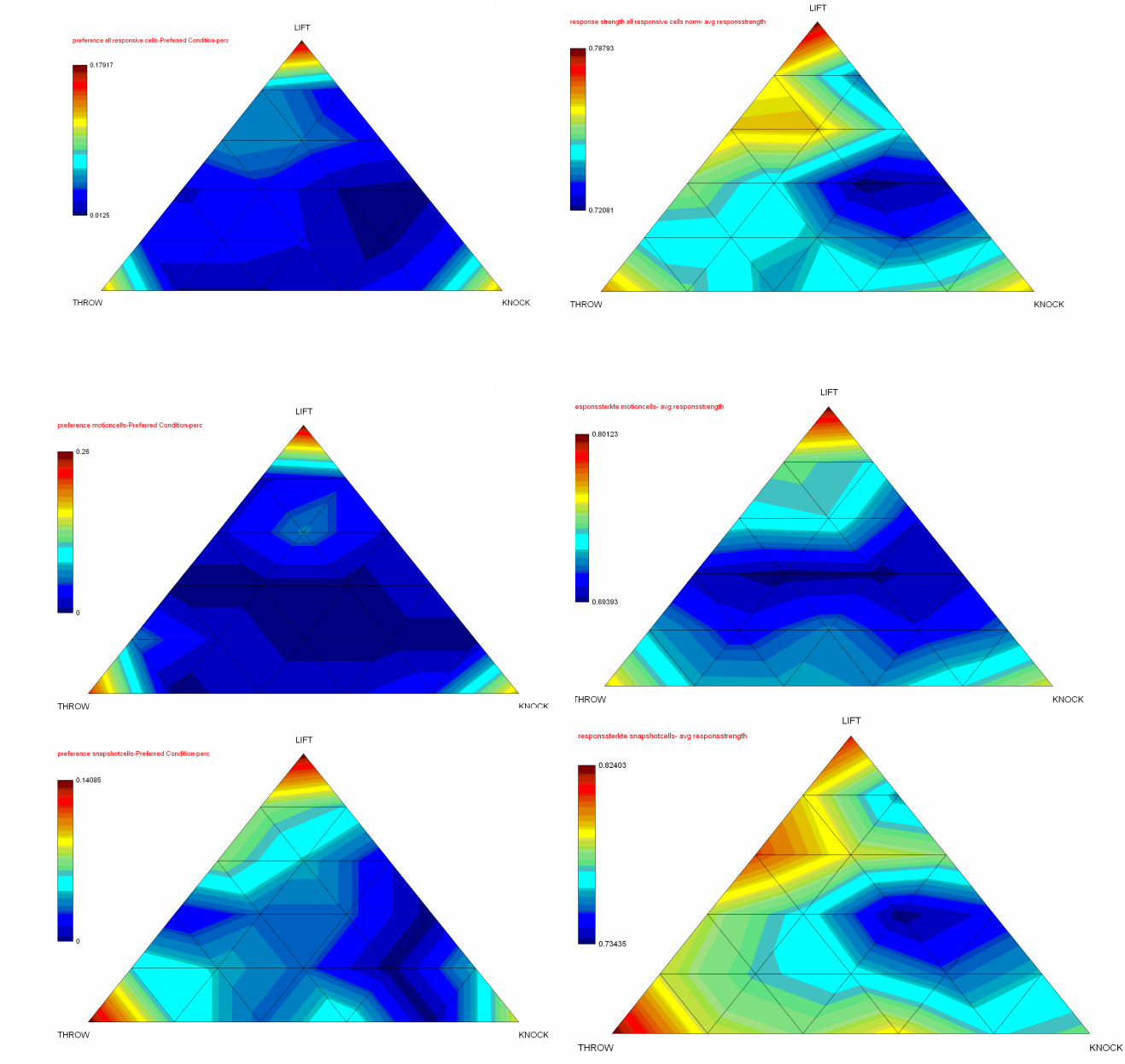


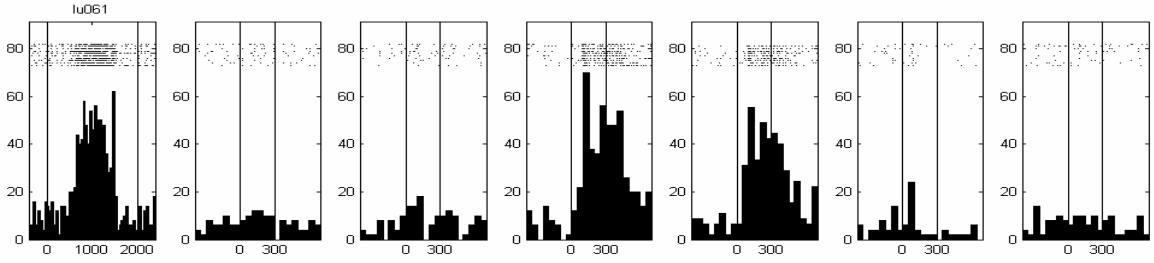
Figure 6. *Distributions of preferred actions (left column) and average normalized responses (left column) for all neurons (top row), motion neurons (middle row) and snapshot neurons (bottom row). Preferred action and average response are indicated by a color code (red: more frequent/ stronger response) after linear interpolation between neighboring actions. Configuration as in Figure 1B.*

3.3 Responses to Static Presentations of Snapshots

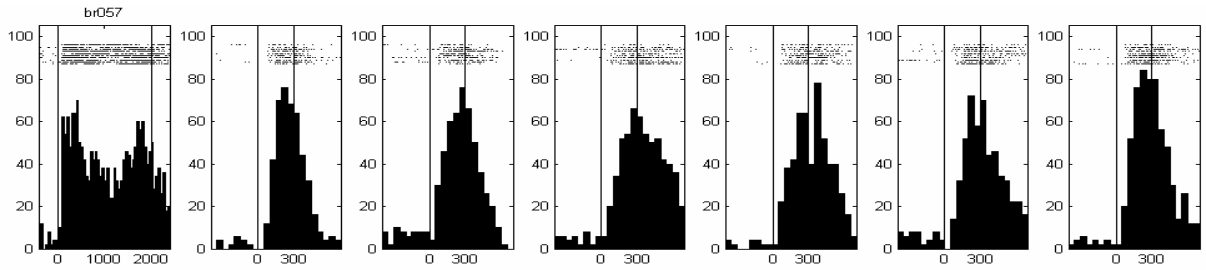
Figure 7 shows the responses of 3 representative neurons to an effective action and static presentations of snapshots of that action movie. The neuron of Figure 7A responds strongly to the static presentations of a snapshot – in fact the response to the most effective snapshot is at least as high as the response to the action sequence. Furthermore the neuron responded selectively for the different snapshots, showing an increased response when one arm is bend and that crosses the other arm. Thus this neuron displayed form selectivity.

The neuron of Figure 7B also responded equally well to the action movies as to static presentations of its snapshots. However, in contrast to the neuron of Figure 7A, this neuron responded similarly to the different snapshots. The neuron of Figure 7C displayed little if no responses to the static presentations of the snapshots, although it responded very well to the action movie. Thus the latter neuron needs motion in order to respond.

A



B



C

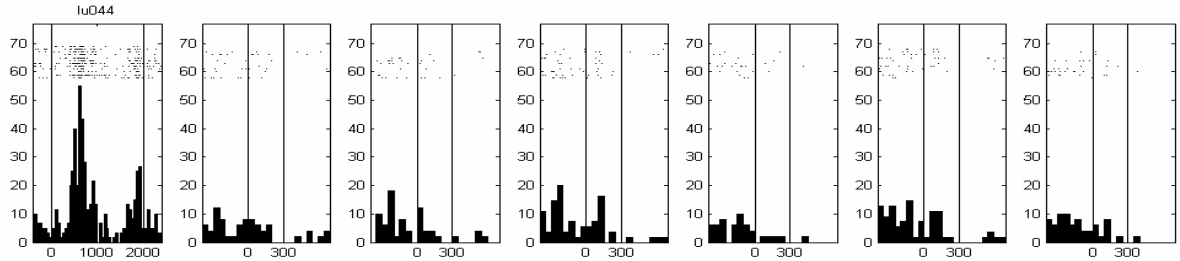


***Figure** 7. Responses of 3 neurons to action movie and static presentations of snapshots compared. For each neuron, the left histogram shows the response to the action movie, while the 6 right columns show the responses to 6 different snapshots. Note that snapshots were presented for 300 ms while the action lasted 2 s.*

The examples of Figure 7 illustrate the variations in responses to the static snapshot presentations: from no responses to responses that equated those to the action movies. For each neuron tested with static snapshots, we computed a *Snapshot index,* which compares the peak firing rate for the action with the maximal peak firing rate for the snapshot presentations (see Methods). A *Snapshot index* of 0 indicates that the neuron responded equally well to the snapshot as to the action movie while an index of 1 indicates no response to the static snapshot. The bulk of the lower bank and lateral convexity neurons responded equally well to the snapshot as to the action movie (median *Snapshot index*: lateral convexity: -0.05; ventral STS: -0.09) while the majority of the upper bank STS neurons (median *Snapshot index*: 0.47) and all putative LST neurons (median *Snapshot index*: 0.84) responded much stronger to the action movie than to the static snapshot presentations. These differences between the *Snapshot indices* of the different regions were highly significant (one way ANOVA; $P < 0.0001$). Post hoc Bonferroni corrected t tests shows that the mean *Snapshot index* of the putative LST differed from those of the other 3 regions (all $Ps < 0.00005$) and that the mean *Snapshot index* of the dorsal STS differed significantly from that of the ventral STS and lateral convexity ($P < 0.000001$), while the ventral STS and lateral convexity indices did not differ significantly. These results suggest a marked functional specialization: upper bank and fundus STS neurons require motion while ventral bank and lateral convexity IT neurons respond to both moving and static stimuli.

3.4 Effect of Stimulus Reductions

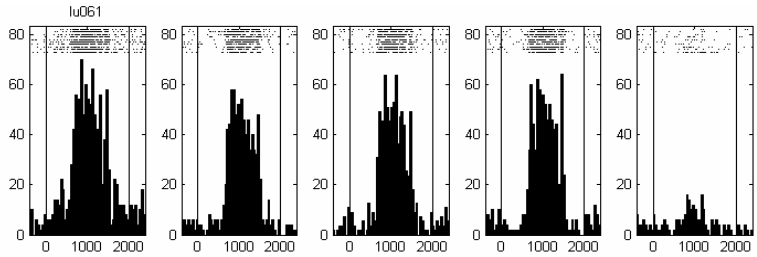
Figure 8 shows the responses of 2 representative neurons to an effective action and reduced versions of the action. The neuron of Figure 8A, which is the same as that of Figure 7A, still responds when the legs are removed but stops responding when the trunk is removed and only the moving arm is visible. This fits the form selectivity of this neuron and suggests that the crossing of the two arms is a critical feature for this neuron. On the other hand, the neuron of Figure 8B still responds well to the motion of the arm alone and even to the motion trajectory of the end-effector (the wrist-dot) only. Thus, as was the case for the responses to the static shapes, there was considerable variation across neurons in the effect of reduction.

The effect of the reduction was quantified by computing an *Arm reduction index.* An *Arm reduction index* of zero indicates that the neuron responds equally well to the arm only as to the full body, while an *Arm reduction index* of 1 indicates no response to the isolated arm action. The distributions of the *Arm reduction index* differed between the 4 regions (One Way ANOVA: $p< 0.0005$). The neurons in the ventral bank of the STS (median *Arm reduction index* = 0.21; N = 26) and the lateral convexity of IT (0 .11; N = 35) showed on average a stronger reduced response for the arm alone than the putative LST (-0.08; N =12) and dorsal bank of the STS neurons (-0.05; N= 56; Bonferroni corrected t tests : all $Ps < 0.05$), although the differences between the regions were less pronounced than for the snapshot test. Indeed, several neurons in IT also responded well to the isolated arm action.

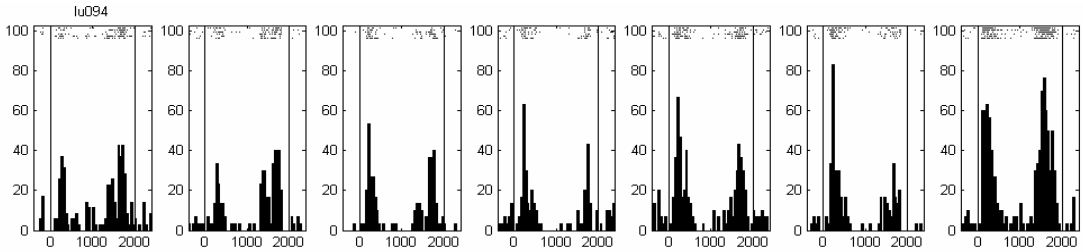
We also computed contrast indices for the other reductions (Figure 3). The large majority of the neurons responded similarly to the full body versus the body without legs configuration (median *Reduction indices* varying between -0.02 and 0). Thus, we found no "body" neurons. The lateral convexity IT neurons showed effects of reduction (median index 0.09) in the next step: removal of the trunk and legs. When computing a similar contrast index for the neurons tested with the wrist point alone, the *Point reduction index,* results were similar to that of the arm alone. The median *Point reduction index* for the lower bank STS (median = 0.71; N = 10) and lateral convexity (median = 0.30; N = 22) were larger (Bonferroni corrected t test; $Ps < 0.005$) than the value for the dorsal STS neurons (-0.03; N = 44), with the difference between the different regions even more pronounced than for the arm only condition. The difference in *Point reduction index* between the ventral STS and lateral convexity was not statistically

significant and only a few neurons in these regions responded as well to the single wrist point action as to the full body stimulus.

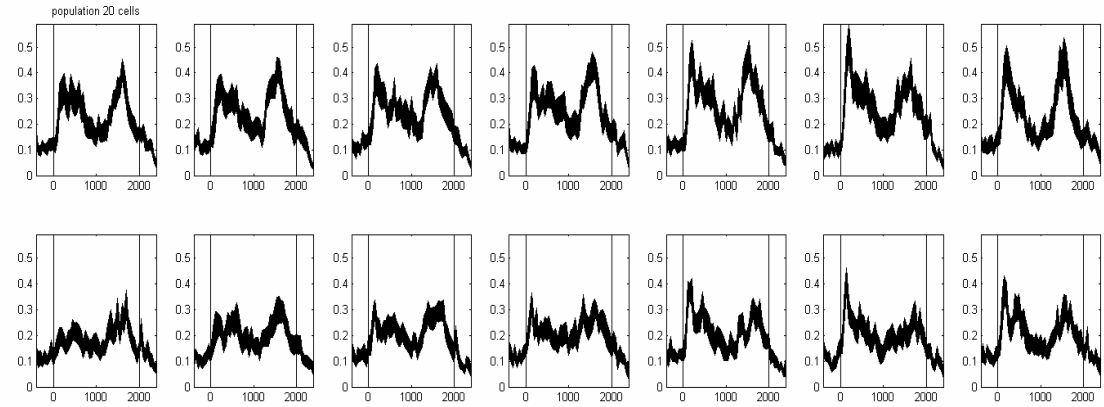
A



B



C



Action ***wrist point***

Figure 8. Responses to the reduced stimuli. A. Responses of a single neuron to the reduced stimuli of Figure 3. The left and right histogram depicts the response to the full body and arm only, respectively. The stimuli in between the full body and wrist point are ordered as in Figure 3. B. Responses of a single neuron to the reduced stimuli including the underarm and wrist point only (6^{th} and 7^{th} histogram, respectively). C. Average responses of 20 action selective neuron that showed an Arm Reduction or Point Reduction index smaller than 0.20 to the reduced stimuli. Same conventions as for B. Top row: effective action; bottom row: less effective action.

These results show that many STS neurons, especially those of the upper bank, respond as well and sometimes even better to the motion of the isolated arm and even the end-effector alone than to the whole body. The Reduction indices were computed on the responses to the

selected, effective action and thus do not inform us about possible changes in the selectivity among the action patterns. To assess the latter, we computed a (best-worst)/(best+worst) selectivity index that compares the peak firing rates for the two actions that were used in the reduction test. We computed such selectivity indices for the full body, arm only and wrist point only conditions of those neurons for which the response to the full body and reduced stimulus differed by less than a factor of 1.5 (*Arm Reduction index* or *Point Reduction index* < 0.20). We selected those neurons which showed at least a factor of 1.5 difference in response between the best and worst action conditions for the full body or the reduced stimulus condition (selectivity index > 0.20). For the neurons tested with the arm only, 45 neurons fulfilled these two criteria. For these 45 neurons, the average degree of selectivity was larger for the full body (median selectivity index = 0.32) than for the arm only displays (median = 0.24) but this difference did not reach significance. However, for the 20 neurons that fulfilled these criteria and were tested with wrist point only condition, the action selectivity was significantly smaller (Wilcoxon matched pairs test; $p < 0.005$) for the wrist only displays (median 0.15) than for the full body displays (median = 0.31). This was due to a relative increase of the responses to the worst action (Figure 8C) for the wrist dot only condition compared to the full body conditions. Thus, although these neurons responded well to the wrist point only, they provided less information about the arm trajectory than when the full body (or even the shoulders + arm only –see Figure 8C) was displayed, implying that wrist point only did not fully determine their response selectivity.

3.5 Responses to Translating Snapshots

Figure 9 shows the responses of 3 neurons to the translating snapshots. The neuron of Figure 9A is a neuron that responded well to the motion present in the action but failed to respond when translating snapshots of the actions. Thus, this neuron did not respond to any sort of motion. The two other neurons of Figure 9 did respond well to the translating snapshots with the neuron of Figure 9C showing strong direction selectivity.

The distributions of the *Translation index* differed between the 4 regions: ANOVA showed a significant effect of region ($P< 0.05$) with the putative LST neurons having the largest *Translation index* (median 0.30) compared to the other 3 regions. However, inspection of the distributions of the *Translation index* indicates that the responses to the translating snapshots did not differentiate the regions as well as the responses to static snapshots and even to the reduction. The ventral STS (median index = -0.11) and lateral convexity IT (median = -.02) neurons had overall similar peak firing rates to the translating snapshots and action movies. The dorsal STS contained some neurons that responded worse to the translating snapshots (for example the neuron of Figure 9A than to the action movie but many neurons in this region responded as well to the action as to the translating snapshots (median index = 0.12). The strong responses to the translating snapshots suggest that many of the neurons do not respond only to the non-rigid motion that is typical of biological motion.

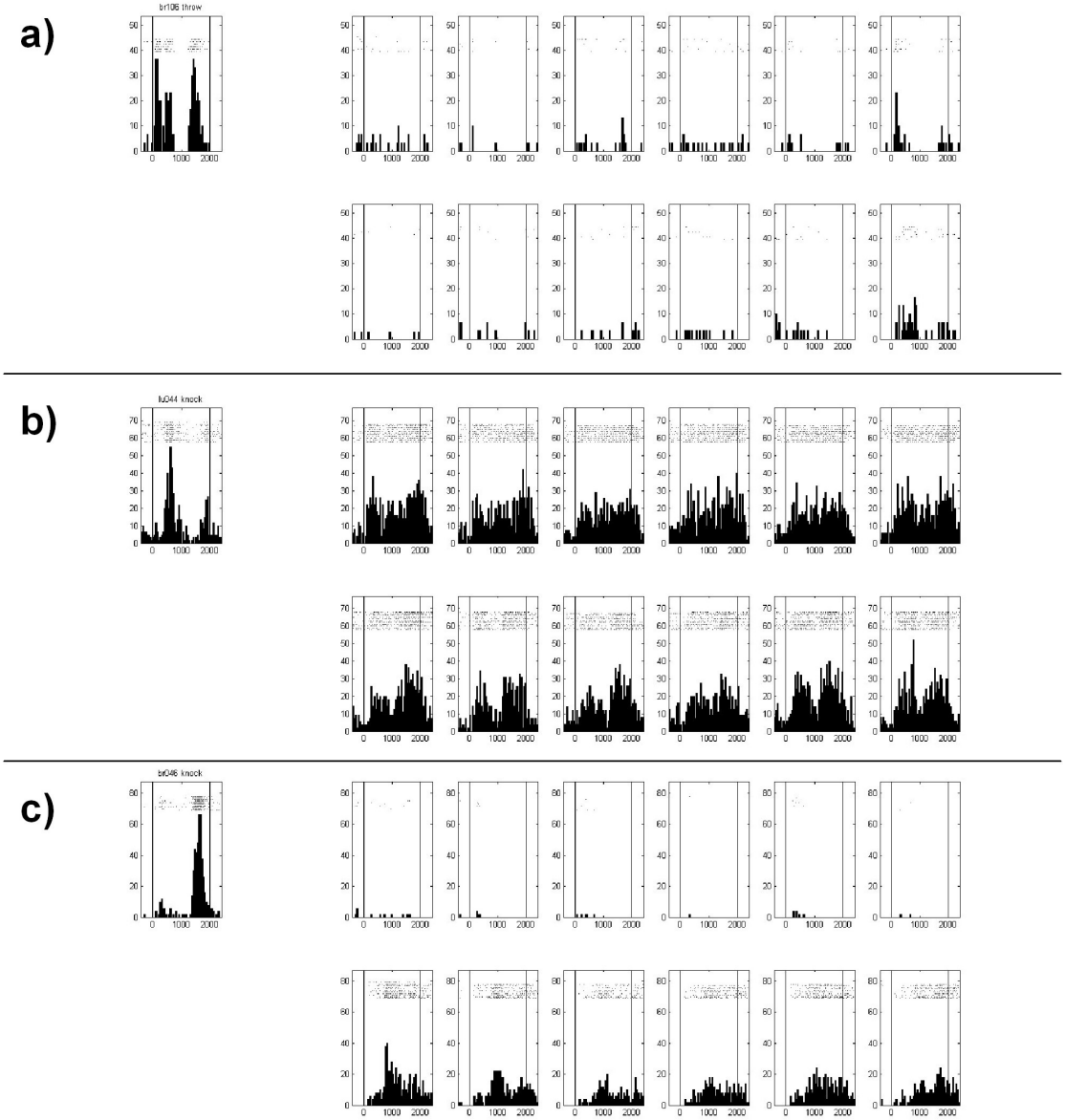


Figure 9. Responses of single neurons to the effective action movie and to translating snapshots compared. The left column shows the responses to the effective action. The upper and lower row of each panel shows the responses to 6 different translating snapshots in 2 different orthogonal directions. A. Single neuron that does not respond to translating snapshots. B. Neuron with strong response to translating snapshots. C. Neuron responding to translating snapshots and showing direction selectivity.

We computed a *Best Direction Index* (see Methods) for the neurons firing with at least 10 spikes/sec in any of the directions of the translating snapshot. The direction indices were on average low, indicating relative weak direction selectivity for the translating snapshot. The difference between the regions did not reach statistical significance (one way ANOVA; $p = 0.053$), although there was a marginal significant trend (Bonferonni t test , $p = 0.059$) for larger *Best Direction Indices* for the putative LST neurons (median = 0.31) compared to the other regions (medians ranging between 0.13 and 0.19).

3.6 Correlation Effects of Static Stimulus Presentation and of Stimulus Reduction

Since the effect of the recorded region appears to be complementary for the *Snapshot* and *Reduction Indices*, we wondered whether these two indices are correlated on a neuron by neuron basis. Figure 10 shows the relationship between the *Snapshot* and *Reduction Indices* for the neurons that were tested with both tests and this separately for the 4 regions. In agreement with our previous analysis, the ventral bank STS and lower convexity neurons cluster together, and thus in the remainder of this report we will consider these neurons as a single population. Figure 10 clearly shows that neurons that responded more strongly to the action stimuli than to the static presentations (large *Snapshot Indices*) responded similarly to the arm only action than to the full body action (small *Reduction Indices*). On the other hand, neurons that respond similarly to the action and static snapshots can vary greatly in their degree of tolerance to stimulus reduction.

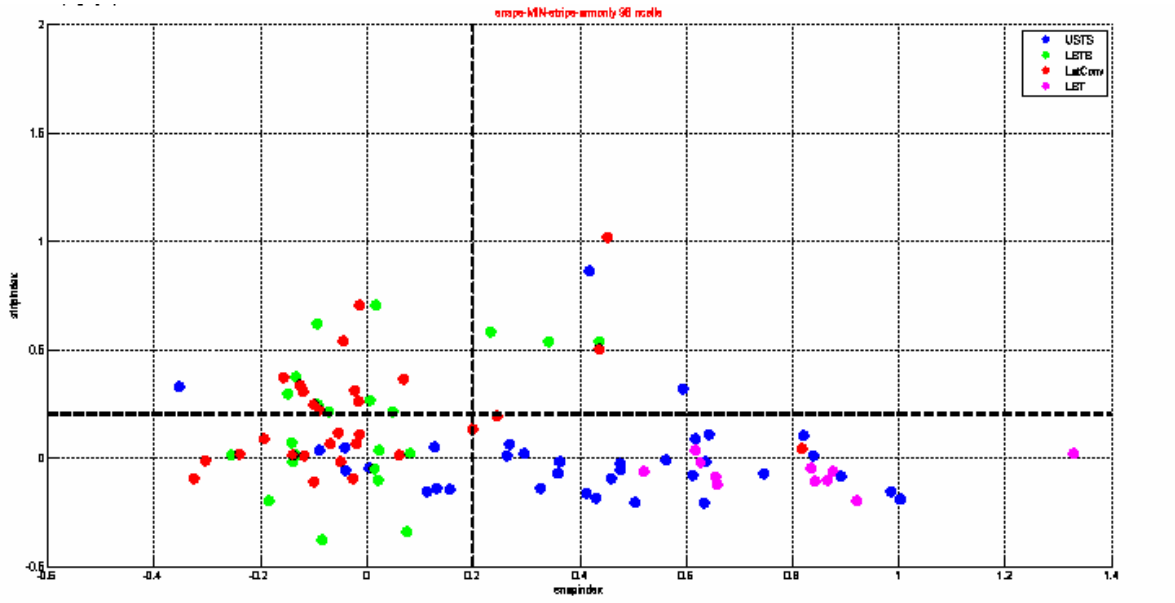


Figure 10. Relating Snapshot Indices (x axis) and Reduction Indices (y axis) of the neurons tested in both test. The different colors indicate the recording region (blue: dorsal STS; green: ventral STS; red: lateral IT convexity; purple: putative LST.

3.7 Snapshot and Motion Neurons Compared: Representation of Action Space

Given that the difference in response to the static snapshot and action movies separated well the different regions and given the theoretical importance of this distinction (i.e. responding to form versus motion information), we distinguished two populations of neurons using a *Snapshot index* of 0.20 as criterion (stippled vertical line in Figure 10). We do not want to imply that there is no continuum in the degree of responsiveness to the static snapshot presentations, but making this distinction proved to be instructive. Thus, in all further analyses we will use the terms "snapshot" and "motion" neurons to refer to neurons that have a *Snapshot index* lower or higher than 0.20, respectively. Note that this distinction corresponds to a regional difference since 90% of the ventral STS/lateral IT neurons (N=67) were Snapshot neurons and 80% of the dorsal STS neurons (N=54) were motion neurons, a difference that was highly significant (Chi Square, $p < 0.0001$).

To examine whether there is a difference in the representation of the parametric action space between the snapshot and motion neurons, we performed the non-linear multi-dimensional

scaling (ISOMAP) of the responses to the 21 shapes for each of the two population of neurons. As shown in Figure 5, the motion neurons represented the action space more faithfully than the snapshot neurons. First, the Scree plot showed that in both populations the two dimensional configuration provides a much better fit than a one dimensional configuration (explained variance < . 85%), but that the two dimensional configuration provided a better fit to the observed data for the motion neurons (explained variance 95%) than for the snapshot neurons (explained variance 89%). Second, inspection of Figures 5C and D indicates that the ordinal relationships among the actions are preserved to a greater extent for the motion than for the snapshot neurons. Third, Procrustes rotation towards the parametric space resulted in a better fit for the motion neurons (Procrustes Distance = 0.15) compared to the snapshot neurons (Procrustes Distance = 0.34). It should be noted that the less faithful representation for the snapshot compared to the motion neurons does not result from differences in the number of neurons between the two samples of neurons: in fact, the number of sampled snapshot neurons (N= 71) was larger than the number of motion neurons (N=50).

As was the case in the full population of neurons, the majority of the snapshot as well as of the motion neurons preferred the real actions over the blends (Figure 6). However, this preference bias was more pronounced for the motion compared to the snapshot neurons: 37% (expected given a uniform distribution: 14% (3/21)) and 66% of the snapshot and motion neurons, respectively, preferred the real actions. The stronger average responses and preferences for the real actions compared to the blends was highly significant for both populations of neurons (response strength: ANOVAs: Ps < 0.0005; Preferences: Chi Square: Ps< 0.001).

To quantify the degree of selectivity within the action space, we employed the omega-square selectivity index which captures both differences in mean responses and trial to trial variability (see Methods). The mean omega-square values were 0.21 and 0.09 for the motion and snapshot neurons, respectively. The on average larger selectivity for the motion compared to the snapshot neurons was highly significant (Mann Whitney U test; $p < 0.000001$). This greater selectivity for the motion compared to the snapshot neurons might explain why snapshot neurons do not represent the action space as faithful as the motion neurons.

3.8 Snapshot and Motion Neurons Compared: Responses During the Course of an Action

A striking feature of the responses to the action movies was that the responses changed during the course of the action sequence. Perhaps this is not surprising given that the action develops during the course of the movie. Thus far, the above reported analyses were performed averaging the responses or taking the peak firing rate for the entire duration of the action. Given the in some cases profound changes in the response during the action, we performed a series of analyses that took into account response changes during the course of the action.

In a first analysis, we divided the entire action sequence into 8 successive 250 ms long segments. The responses were averaged within each of the segments. In order to take into account response latency, the first segment started 50 ms post stimulus onset and the last segment lasted until 50 msec post stimulus offset. We performed ISOMAP analyses for all neurons, the snapshot and the motion neurons for each of the 8 segments. When inspecting the solutions of all neurons (N=240), it was clear that the most faithful representations were for the 2^{nd} (300-550 ms) and 7^{th} (1550 – 1800 ms) segments, both having Procrustes Distances less than 0.15. The motion neurons had the most faithful representation (Procrustes Distance = 0.16) for the 7^{th} segment. Overall, the snapshot neurons produced erratic configurations, especially for the second part of the action (all Procrustes Distances > .30 except for segment 3 (Procrustes Distance = 0.27).

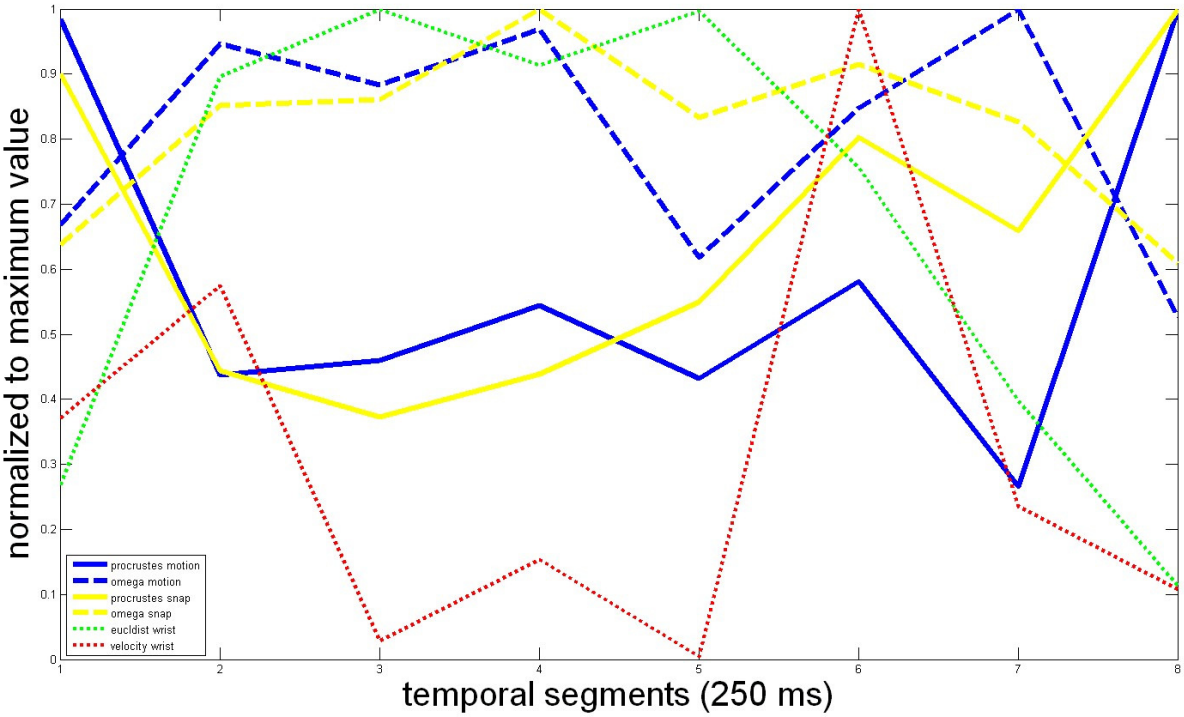


Figure 11. Relating stimulus measures (mean velocity differences among the 21 actions (red line), Euclidean distance between spatial position of the end-effector across the 21 actions (green line)), action selectivity (omega-square : motion cells (stippled blue line) and snapshot cells (stippled yellow line)) and Procrustes Distance of the two dimensional ISOMAP solution to the triangular parametric space (motion cells: blue line; snapshot neurons: yellow line). Each of these parameters were computed for each of the 8 action segments (250 ms long; see text). All values were normalized to their maximum across the *8 segments.*

Figure 11 plots the Procrustes Distances for the snapshot and motion neurons for the 8 segments. Also plotted are the mean Euclidean distances (averaged across the 210 possible stimulus pairs) in the speed and (x,y) position of the end-effector, as well as the degree of action selectivity (omega-squares) measured for the responses in each of the 8 segments. The speed and spatial 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. All values of a single measure were normalized to their maximum.

Correlation analyses of the Procrustes Distances with the different stimulus and neuronal properties produced the following results. First, the variation in the omega-square values during the course of the action correlated significantly with the Procrustes Distances for both the motion ($r = -0.73$; $p < 0.04$; $n = 8$) and the snapshot neurons ($r = -0.75$; $p < 0.03$; $n = 8$). These significant correlations indicate that the variations of Procrustes distances and of selectivity are real and not just noise. In addition, they show that the greater the selectivity of the neurons the more faithful these neurons represent the action space. Note that this need not necessarily be the case since in order to represent the action space, neurons need not only to be selective but also their response to a stimulus should vary as a function of the distance – in the action space – between their preferred action and the stimulus.

Second, the differences in the spatial position of the end-effector across the 8 segments correlated significantly with the Procrustes Distances for the snapshot ($r = -0.87$; $p < 0.005$; $n = 8$) but not for the motion neurons ($r = -0.67$; n.s.; $n = 8$). A similar result was obtained when we computed the Euclidean, spatial position difference for the wrist, elbow and shoulder

points (snapshot neurons: r = - 0.93; p < 0.001; n = 8; motion neurons; r = -.0.70; n.s.; n = 8). This suggests that the snapshot neurons, but less so the motion neurons, are sensitive to the spatial differences between the stimulus configuration.

Third, no significant correlations were found between the speed differences and the Procrustes Distances (or omega-square selectivity measure) for the snapshot (r = 0.25; n.s., n=8) or motion neurons (r = 0.02; n.s.; n=8). Although this suggests that the motion neurons do not merely respond to the speed differences between the actions, one should be careful since in this analysis speeds were averaged in bins of 250 ms, which might have obscured possible correlations between speed and response.

To examine the possible relationship between speed and response in more detail, we correlated the smoothed, instantaneous firing rates of each neuron – in bins of 2 ms – with the log speed of the motion of the end effector. The Pearson correlations were performed separately for the first and second phase of the action (see Methods) and were computed for a set of time delays between the motion and instantaneous firing rate. For each neuron we choose that time delay for which the explained variance of log speed and the response was the greatest. The median explained variances of the instantaneous response by the log speed were 0.16 and 0.27 for the snapshot and motion neurons, respectively, a difference that was highly significant (Mann Whitney U test; p < 0.0001). These are medians of the maximum explained variance of the two phases (directions) of the actions. Figure 12 shows scatterplots of the explained variance of the response by speed for the two phases of the action and this for motion and snapshot neurons separately. It is clear that the relationship between speed and instantaneous firing rate was relatively weak for the snapshot neurons. Also, the speed dependency of the motion neurons was significantly stronger for the second part of the action than for the first part (Wilcoxon matched pairs test; P < 0.01). These results show that the responses of the motion neurons were modulated by the speed of the end-effector.

In the above analysis, the correlations between log speed and response were computed across the 21 actions. When considering the changes in responses within an action, there were also striking differences between the snapshot and motion neurons, as is illustrated in Figure 13 for 7 different actions (the same as those of Figure 1A). The average response of the motion neurons follows the speed profile of the end-effector (and of the shoulder, elbow and wrist) point more closely than the average response of the snapshot neurons. However, it should be stressed that speed is not the only determinant of the response of the motion neurons since otherwise ISOMAP would have produced a 1 and not 2 dimensional solution (speed is a one dimensional parameter). Anyway, it is clear from the average responses of the motion neurons plotted in Figure 13 that the peak responses of the neurons during the action are not uniformly distributed but favor action segments in which there is motion, i.e. the actor is acting.

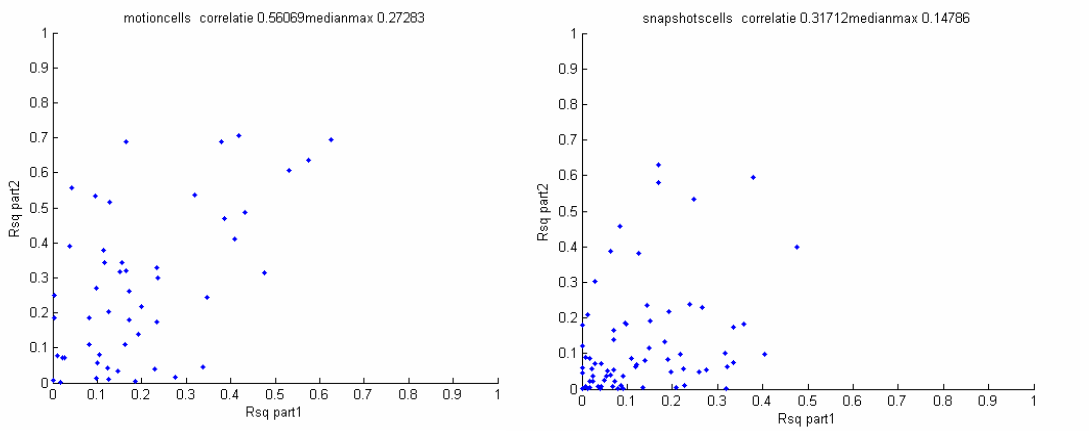


Figure 12. Scatterplots of r2 values (explained variance) for the correlations between log speed of the

end-effector and instantaneous firing rate for the first (horizontal axis) and second part of the action sequence (vertical axis). Each dot corresponds to data of single neuron. Left panel: motion neurons; Right panel: Snapshot neurons.

The average responses of the snapshot neurons shown in Figure 13 display only a small overall modulation during the course of the action sequences (when ignoring the initial transient related to stimulus onset). However, it should be noted that these are responses averaged across neurons. A neuron by neuron analysis showed that many snapshot neurons do show strong modulations of the response during the action sequences – i,e. do not fire continually during the action. An example of such a neuron is shown in Figure 4D. We quantified for each neuron the depth of the modulation by subtracting, after smoothing, the minimum peak firing rate from the maximum peak firing rate, divided by their sum, using an analysis window of 200 to 2050 ms post-stimulus onset. This was done for the response to the effective action sequence. The median within-action modulation index was 0.73 indicating considerable modulation during the course of an action, even for the snapshot neurons. The analyses reported above showed that these modulations correlate little with speed for the snapshot neurons, thus the question of the underlying cause of these within-action response modulations for these neurons. One obvious possibility that we examined is whether these within-action response modulations relate to the selectivity for different snapshots. Indeed one might expect that neurons that show different responses to different snapshots, when the latter are presented statically, will also show a relative high degree of within-action modulation. This was indeed the case: there was a significant correlation ($r = 0.54$; $p < 0.0005$; Figure 14) between the within-action modulation and the degree of selectivity for static presentations of the snapshots (the *Snapshot selectivity index;* see Methods*)*. Interestingly, Figure 14 shows an asymmetric relationship between the *Snapshot selectivity index* and the degree of within-action modulation: neurons with a low within-action modulation do rarely show a high snapshot selectivity, while neurons with a low static snapshot selectivity can also show strong within-action modulations. This suggests that even neurons that respond similarly to static presentations of different snapshots can show a selectivity for the segments of an action, i.e. show a sensitivity to the sequence in which the snapshots are presented. Thus the snapshot selectivity is one determinant of the within-action response modulation but not the only one.

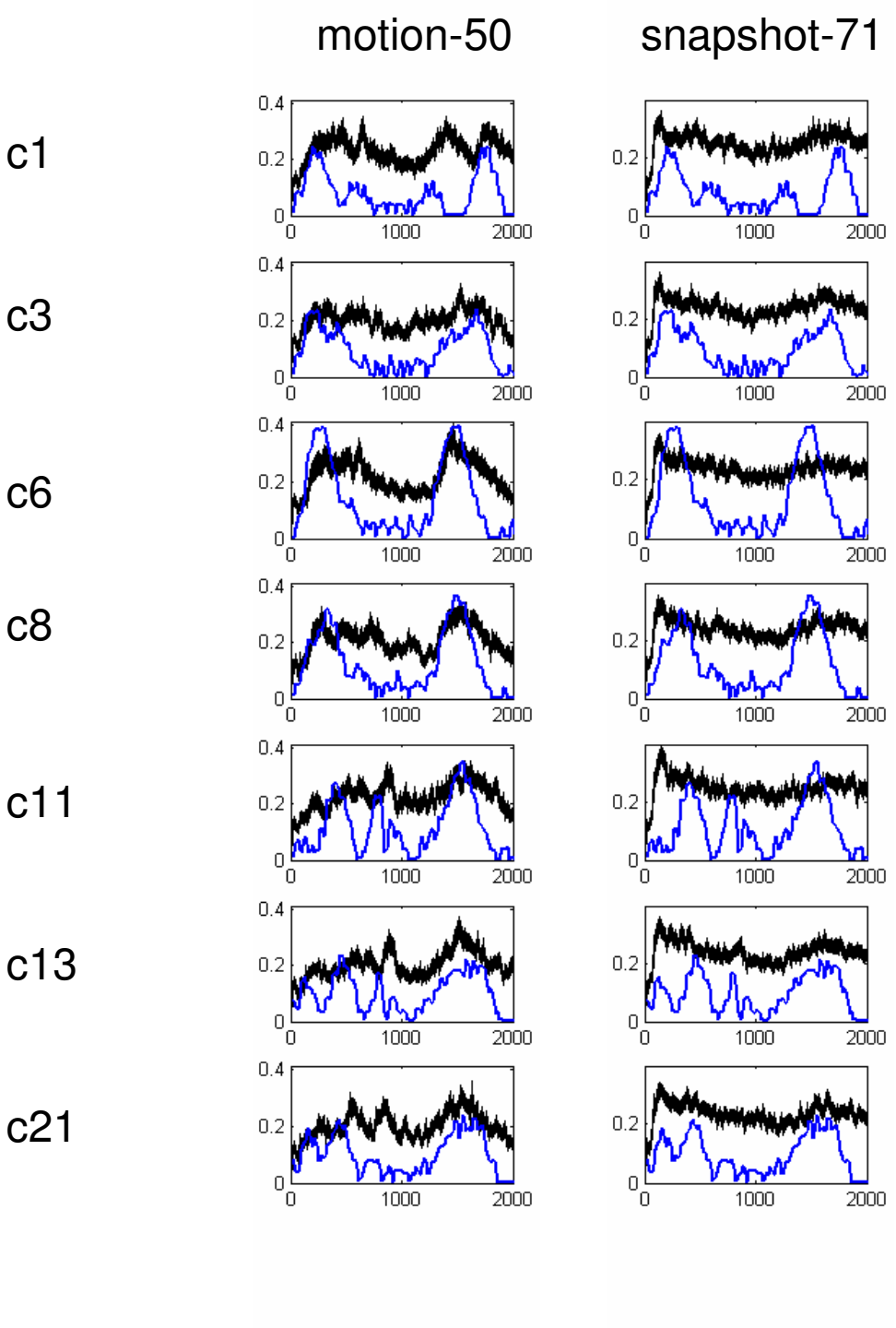


Figure 13. Population Peri Stimulus Time Histograms of motion neurons (middle column) and snapshot neurons (right column) compared to speed of the end-effector during the course of the action. Black bands: mean normalized responses plus and minus one standard error of the mean (binwidth 16 ms – one frame) . Blue lines: instantaneous speed of end effector (see Figure 2). Each row corresponds to a different action as indicated by its condition number (see Figures 1B and 5E).

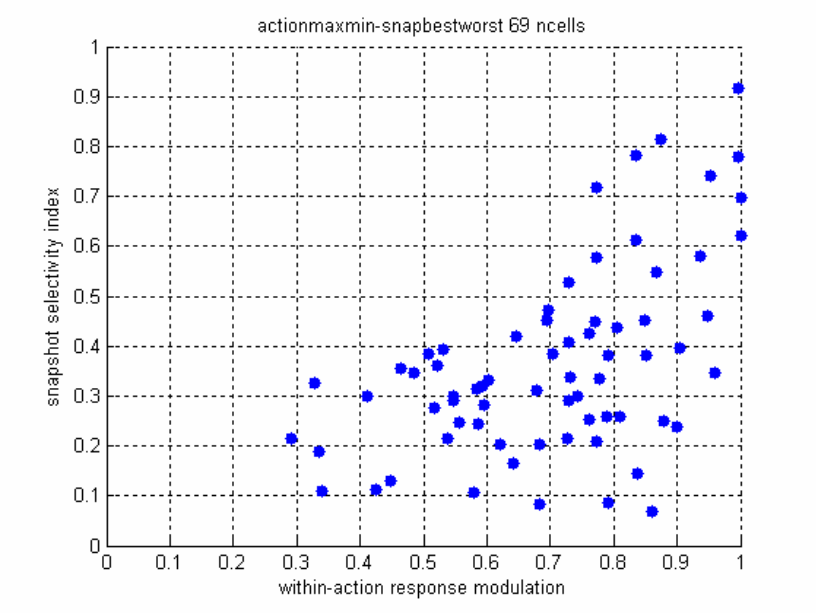


Figure 14. Scatterplot of within-action response modulation and the Snapshot selectivity index for snapshot neurons.

3.9 Spatial Position Dependency of Responses

We tested the responses of the neurons for different spatial positions of the action stimulus. During the course of the experiments we employed different variations of such position tests, with the most eccentric positions ranging from 4.2 to 7.8 deg. A full analysis of the responses in the position tests are beyond the scope of the present report and here we will focus on two issues: position invariance of action selectivity and position invariance of the response peak during the course of the action.

Similarly to previous analyses of position invariance of static shapes (e.g. Vogels, 1999) we ranked for each selective neuron ((best-worst)/ (best + worst) > 0.20 at foveal position) the two actions according to their peak firing rate at the foveal position. Then the same ranking of the two actions was applied for the other position at which the neuron still responded with at least 10 spikes/sec net peak firing rate, followed by an averaging of the responses for rank 1 and rank 2 across positions and neurons. The results of this analysis indicated that the average action ranking was preserved at the peripheral stimulus positions (Wilcoxon matched pairs test on responses to first and second rank at peripheral position; $p < 0.05$) indicating a position invariance of the action selectivity.

Many neurons, especially the motion neurons, showed strong modulation of their response during the action. One possible explanation of this within-action modulation (in addition to end-effector speed; see above) is the presence of inhomogeneities of the neuron's RF through which the arm moves. If the latter explanation holds, then the timing of the peak response within the action would vary with the spatial position of the stimulus, since the trajectory of the arm will differ relative to the supposed RF heterogeneity.

Figure 15 shows the responses of one motion neuron in a position test (step size 1.5 deg in the horizontal and vertical dimension) for the effective action. Two effects are noteworthy: first, the response of the neuron varies with the spatial position of the action stimulus, and second, the timings of the peak responses are invariant with spatial position.

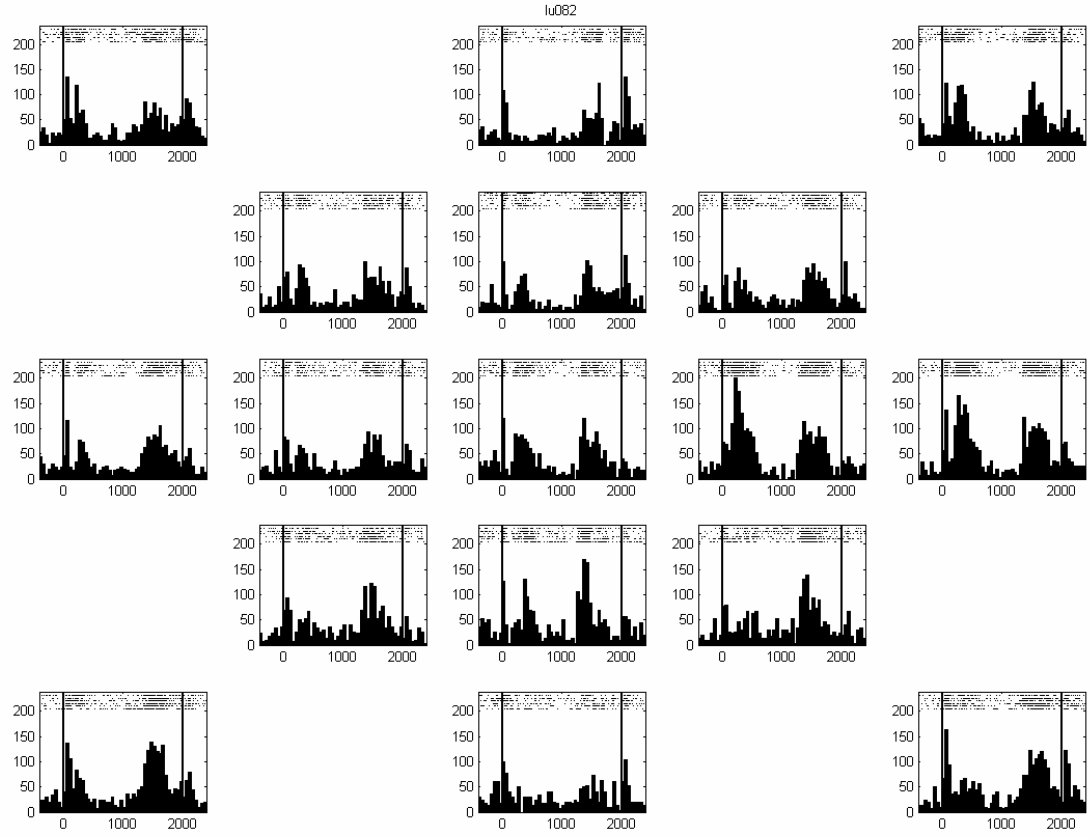


Figure 15**. Responses in the position test of a single motion neuron. The effective action was presented at 17 different positions. The minimum distance between two positions along the vertical and horizontal axis was 1.5 deg. The centre position corresponds to the foveal position. The contralateral visual field is at the right.**

We examined the position invariance of the timing of the peak response across the population of neurons that were tested at different spatial position by determining for each neuron the time, relative to action onset, of the smoothed peak firing rate for the effective action at the foveal position. Then we aligned the responses at the other spatial positions with the time of the peak firing rate determined at the foveal position, and these responses were averaged across neurons. The results for the 23 motion neurons of which we had a position test are shown in Figure 16. Note that not all neurons were tested at all the positions shown in Figure 16: only 7 neurons were tested at the outer 5.3 and 7.5 deg eccentric positions. The population responses of the tested neurons of Figure 16 showed the same two trends as the example neuron of Figure 15: a variation of the peak response with the spatial position of the stimulus but an invariance of the timing of the peak response with spatial position. Note that the arm movement has a maximum amplitude in the x and y direction of about 1.5 and 2.5 deg. If the within-action modulations of responses would have resulted from inhomogeneities or hot spots within the RF, then, given the range of the arm trajectories, the peaks should have shifted considerably or even have disappeared for the eccentricities tested in Figure 16. Thus, we conclude that the within-action modulation does not merely result from RF inhomogeneities. On the other hand, the neurons are sensitive to the spatial position of the overall stimulus, which can contribute to the coding of the arm trajectory.

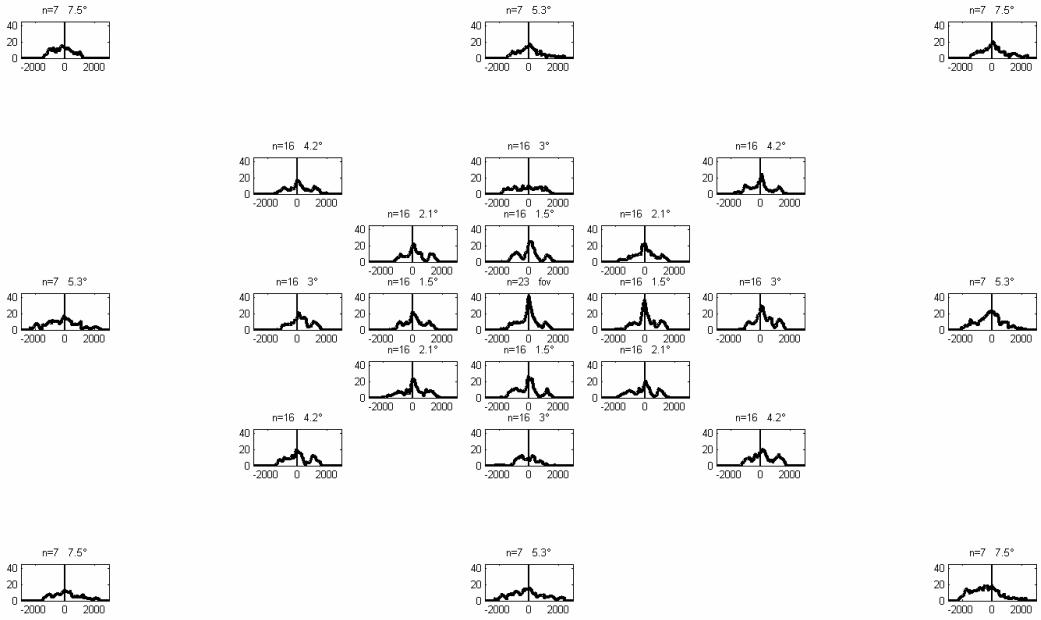


Figure 16: **Response of motion neurons time locked to their peak response at the foveal position. For each spatial location, time locked responses were averaged across the neurons that were tested at that position. Stimulus eccentricity as well as number of averaged neurons is indicated above each histogram. Fov = foveal position. X axis is in ms, and Y axis in spikes/sec.**

4. Discussion and Conclusions

The present study examined the coding of visual actions of stick figure displays 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, which enabled us to examine whether temporal cortical neurons can represent the 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 demonstrates that the output of these neurons can be used for action categorization.

Natural actions differ not only in their kinetics but also in form parameters ("posture"). We examined to what degree rostral temporal cortical neurons respond to the posture versus motion information and found neurons that respond to the action movies but do not respond to static presentations of the snapshots of the same movies ("motion neurons") and neurons that responded equally well to the action movies and the static snapshot presentations ("snapshot neurons"). Of course, as with many other neuronal properties (e.g. the distinction between simple and complex RFs in early visual cortex), there was a continuum between the relative degree of the responses to the snapshots versus actions. However, the motion neurons were found mainly in the fundus and upper bank of the STS while the snapshot neurons were mainly in the lower bank of the STS and then lateral convexity of IT, suggesting an anatomical if not functional dissociation of this response property. The above anatomical dissociation fits previous observations of motion responses in STP (Bruce et al., 1981; Baylis et al., 1987), but unlike in these earlier studies we could show the dissociation using the same stimuli in the different temporal areas.

Our analyses suggest that the response of the motion neurons depend on stimulus speed of the end-effector. On average, stronger speeds produced stronger responses. The correlation between speed and response for the motion neurons might not be that surprising given that high speed action segments are those that contain most of the action and thus are the most informative to code. Indeed, our results show, that at least for these simple action stimuli, STS neurons do not code for the action as such but only for temporal segments of the action. Overall, these temporal segments coincided with those that contain most of the motion. However, it should be stressed that speed is certainly not the only determinant of the response of the motion neurons since we found that the neurons represented the action space in a 2D and not 1D configuration (speed is a one dimensional parameter). Thus, other factors beside speed determined the tuning of the motion neurons to the actions. One possible candidate of such a factor is direction of motion, but most rostral STS neurons were only weakly tuned for motion direction when tested with translating motion of the snapshots, suggesting direction of motion does not contribute much to the action tuning. Another candidate is the position of the moving arm within the RF of the neuron, i.e. a spatial code of arm position. The results of the position test showed that indeed the responses depended on the spatial position of the stimulus within the RF. However, the neurons still responded to the same segments of the action at the different spatial positions, suggesting that mere spatial RF heterogeneities contribute little to the action coding. A third candidate is that the neurons code for the relative motion of the arm segments. The results of the reduction test suggest that such relative motion coding does occur: the selectivity for the actions decreased when reducing the stimulus to a single dot suggesting that relative motion of the arm points or arm segments contribute to action coding.

The snapshot neurons respond to the form of the actor or parts of the actor and thus such neurons, as a population, can contribute to the action coding by signaling the posture of the actor. However, two points are relevant here when discussing the contribution of the snapshot neurons to action coding. First, most snapshot neurons were rather broadly tuned to the different snapshots of a movie, limiting their ability to code for the actions. Although it cannot be excluded that their selectivity is greater when using full body displays instead of stick figures, it does put a limitation on the implementation of pure snapshot based models, such as the template matching model of Lange et al. (2006). Second, if the neurons only respond to the snapshot without taking into consideration the temporal sequence in which the snapshot occurs, their contribution to the action coding will be limited to mere form analysis. However, our results do provide some evidence for the role of temporal information in the coding of actions by snapshot neurons since for some snapshot neurons the depth of within-action modulation was greater than expected from their (broad) snapshot coding. However, it should be noted that action coding by a snapshot-based sequence mechanism, as postulated by Giese and Poggio (2003) for their form-analyzing pathway and by Lange et al. (2006), will only work for highly familiar motion patterns of which the temporal snapshot sequence is well known to the subject. Also such a mechanism has the pitfalls of all template matching algorithms, being their susceptibility to changes in the stimulus due to transformations that are related to the viewpoint of the observer with respect to the actor (position, size and viewpoint invariance).

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 neuronal sample. Thus it is tempting to conclude that the motion neurons contribute more to the action coding than the snapshot neurons. The possible contribution – and its limitations – of the snapshot neurons has been discussed above. We propose that the motion neurons code the effector motion on a moment-by-moment basis, allowing a full reconstruction 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 reduction of the actor to the effector (arm) only, and (3) the (partial) correlation of their response with motion parameters such as speed. Such a motion trajectory coding scheme has the advantage that familiar as well as non-familiar, novel action patterns can be

represented – which is unlike the snapshot-based sequence mechanism which works only for highly familiar patterns. However, it is limited since actions as such are not represented explicitly, since this requires the integration of the information of such motion neurons. A second issue is that it is unclear how these neurons will respond to more complex actions that are not limited to one limb: how will these neurons respond when several limbs move simultaneously inside their RF (as when viewing a walking person for instance)?

Given our observation that the motion neurons respond only to segments of the action, one would expect that if one takes into account the temporal evolution of the neuronal 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 50 ms long analysis windows, the obtained configuration was greatly improved (Figure 17; motion neurons: Procrustes Distance: 0.07; snapshot neurons : Procrusted Distance: 0.15).

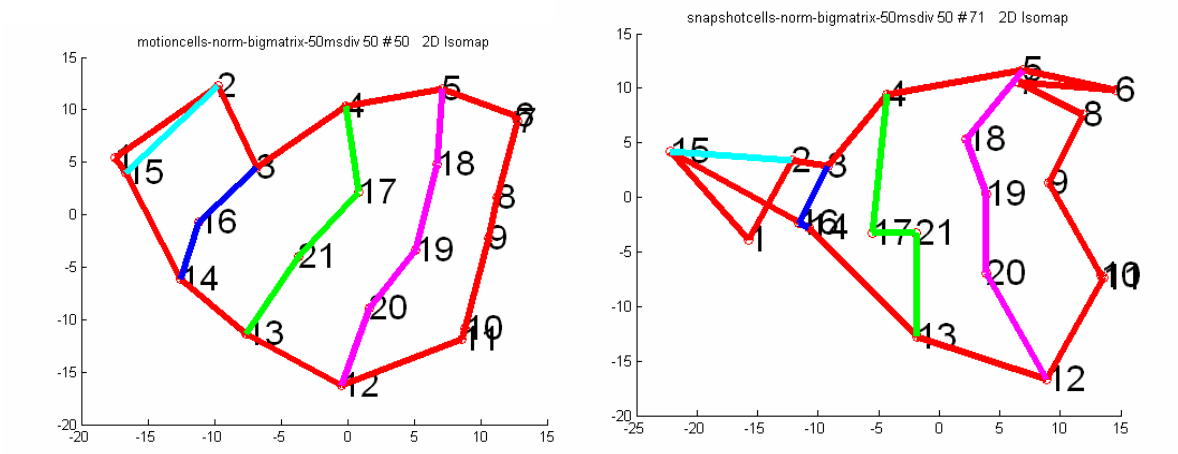


Figure 17**. ISOMAP solution for motion neurons (left) and snapshot neurons (right) when taking into account the temporal evolution of the responses within each action (50 ms binwidth).**

The configuration obtained using the temporal information from the motion neurons was excellent, especially when considering the relatively low number of neurons (N = 50) involved. This supports our view that especially the motion neurons can contribute to the coding of these simple arm actions and do this by computing the motion of the effector on a moment-by-moment basis. Thus these neurons do not code for the action as such – in terms of action semantics such as "lifting" vs "knocking" – but code only for motion trajectories. However, as ISOMAP shows, the information these motion neurons provide, when integrated properly, allows an excellent reconstruction of the similarities between different actions which can form the basis of the categorization of novel actions (according to their similarity with learned actions).

A surprising finding was that the average activity of the neurons, especially the motion neurons, was larger for the "real actions" than for the blended actions. This shows unequivocally that a simple motion parameter does not explain the responses of these neurons – otherwise the responses to the blends should not always be lower than those to the real actions. It is tempting to relate this finding to the tuning of static shapes for 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 are apparent and this suggests that it reflects a learning- or exposure-based tuning for the extremities of a parametric space. One possible mechanism underlying this tuning for extremities is

adaptation: the blends are more similar to the other stimuli than the real actions which lie on the extremities of the space and thus one might expect stronger similarity-based adaptation for the former compared to the latter stimuli. The strength of the tuning for extremities effect for the motion neurons would suggest that such similarity based adaptation also occurs for neurons that only respond to motion and it would need to be based on a adaptation to motion trajectory since the range of local motions strongly overlap for the different actions. Another possible explanation for the stronger responses to the real actions is that the neurons respond stronger to natural than to unnatural, i.e. the blended, action patterns. However, since the blends appear rather natural – at least to human observers – we find this explanation implausible.

We found a patch of neurons in the posterior part of our recording range that we tentatively labeled as belonging to the putative LST region. We analyzed the neurons of this region were separately from those of the other regions. Our putative LST region is located in the fundus of the STS at an anterior-posterior level that is similar to the LST region defined by Nelissen et al. (2006). These putative LST neurons responded strongly to motion but much less so to static snapshots. So these putative LST neurons are motion neurons as defined in this report (but we did not include these in our sample of 50 motion neurons). Several putative LST neurons showed strong direction selective responses, and their direction selectivity was on average stronger than those of the more rostrally located motion neurons. A monkey fMRI adaptation study has suggested that LST neurons are direction selective (Nelissen et al., 2006) and our single cell data agree with this proposal. It is likely that the LST region projects to the more rostrally located upper bank/fundus STS regions from which we recorded the other motion neurons that were analyzed in the Results. It is possible that the LST region contributes to action coding by analyzing motion patterns (Nelissen et al., 2006). It should be noted that the LST neurons typically showed strong responses to the second part of the action – the return phase of the arm - and appear to be somewhat more restricted in their action coding than the more rostral motion neurons that we examined.

Our results support an action coding scheme in which the motion of an end-effector is analyzed by motion neurons that do not respond to static presentations of snapshots but require motion. At the population level, the information contained in the responses of these motion neurons are sufficient to compute the similarities among novel, unfamiliar actions but the neurons themselves do not represent actions as such since they respond only to segments of an action sequence. Thus, further integration of these responses is needed to obtain a full action code. These motion neurons are predominantly located in the dorsal bank and fundus of the STS, while neurons in the more ventral and lateral parts of the visual temporal cortex respond to static snapshots as well as to actions. We found that these snapshot neurons represent the similarities among the actions to a lesser degree than the motion neurons. Further research using more complex, multi-limb actions as well involving a comparison of the sensitivity of the neural and behavioral responses is underway to understand the contribution of these different neurons to action coding.

References

K. Anderson and R. Siegel. Optic flow selectivity in the anterior superior temporal polysensory area, STPa, of the behaving monkey. *Journal of Neuroscience, 19, 2681-2692,* 1999.

K. Anderson and R. Siegel. Three-dimensional structure-from-motion selectivity in the anterior superior temporal polysensory area, STPa, of the behaving monkey. *Cerebral Cortex, 15, 1299-1307,* 2005.

F. Ashby and N. Perrin. Toward a unified theory of similarity and recognition. *Psychological Review, 95, 124-150,* 1988.

N. Barraclough, D. Xiao, C. Baker, M. Oram, and D. Perrett. Integration of visual and auditory information by superior temporal sulcus neurons responsive to the sight of actions. *Journal of Cognitive Neuroscience, 17, 377-391,* 2005.

N. Barraclough, D. Xiao, M. Oram, and D. Perrett. The sensitivity of primate STS neurons to walking sequences and to the degree of articulation in static images. *Progressive Brain Research, 154, 135-148,* 2006.

G. Baylis, E. Rolls, and C. Leonard. Functional subdivisions of the temporal lobe neocortex. *Journal of Neuroscience, 7, 330-342,* 1987.

C. Bruce, R. Desimone, and C. Gross**.** Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. *Journal of Neurophysiology, 46, 369-384,* 1981.

W. De Baene, E. Premereur, and R. Vogels. Properties of shape tuning of macaque inferior temporal neurons examined using rapid serial visual presentation. *Journal of Neurophysiology, 97, 2900-2916,* 2007.

S. Edelman. Represenation and recognition in vision. *MIT Press, Cambridge, Massachusetts,* 1999.

D. Freedman, M. Riesenhuber, T. Poggio, and E. Miller. A comparison of primate prefrontal and inferior temporal cortices during visual categorization. *The Journal of Neuroscience, 23,* 2003.

M. Giese and T. Poggio. Neural mechanisms for the recognition of biological movements. *Nature Reviews Neuroscience, 4, 179-192,* 2003.

T. Jellema, G. Maassen, and D. Perrett. Single cell integration of animate form, motion and location in the superior temporal cortex of the macaque monkey. *Cerebral Cortex, 14, 781-790,* 2004.

T. Jellema and D. Perrett. Cells in monkey STS responsive to articulated body motions and consequent static posture: a case of implied motion. *Neuropsychologia, 41,* 2003.

T. Jellema and D. Perrett. Neural representations of perceived bodily actions using a categorical frame of reference. *Neuropsychologua, 44, 1535-1546,* 2006.

G. Johansson. Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics, 14, 201-211,* 1973.

G. Kayaert, I. Biederman, H. Op de Beeck, and R. Vogels. Tuning for shape dimensions in macaque inferior temporal cortex. *European Journal of Neuroscience, 22, 212-224,* 2005.

G. Keysers and D. Perrett. Demystifying social cognition: a Hebbian perspective. *Trends in Cognitive Science, 8, 501-507,* 2004.

R. Kirk. Experimental design: procedures for the behavioral sciences. *Brooks/Cole, Belmont, California,* 1968.

L. Kovar and M. Gleicher. Flexible automatic motion blending with registration curves. *ACM SIGGRAPH Symposium on Computer Animation,* 2003.

J. Lange and M. Lappe. A model of biological motion perception from configural form cues. *Journal of Neuroscience, 26, 2894-2906,* 2006.

K. Nelissen, W. Vanduffel, and G. Orban. Charting the lower superior temporal region, a new motion-sensitive region in monkey superior temporal sulcus. *Journal of Neuroscience, 26, 5929-5947,* 2006.

R. Nosofsky. Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory and Cognition, 10, 104-114,* 1984.

H. Op de Beeck, J. Wagemans, and R. Vogels. Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. *Nature Neuroscience, 4, 1244-1252,* 2001.

M. Oram and D. Perrett. Responses of anterior superior temporal polysensory (STPa) neurons to "biological motion" stimuli. *Journal of Cognitive Neuroscience, 6, 99-116,* 1994.

M. Oram and D. Perrett. Integration of form and motion in the anterior superior temporal polysensory area (STPa) of the macaque monkey. *Journal of Neurophysiology, 76, 109-129,* 1996.

M. Oram, D. Perrett, and J. Hietanen. Directional tuning of motion-sensitive cells in the anterior superior temporal polysensory area of the macaque. *Experimental Brain Research, 97, 274-294,* 1993.

T. Palmeri and I. Gauthier. Visual object understanding. *Nature Reviews Neuroscience, 5, 291-303,* 2004.

A. Pasupathy and C. Connor. Shape representation in area V4: position-specific tuning for boundary conformation. *Journal of Neurophysiology, 86, 2505-2519,* 2001.

D. Perrett, M. Harries, R. Bevan, S. Thomas, P. Benson, A. Mistlin, A. Chitty, J. Hietanen, and J. Ortega. Frameworks of analysis for the neural representation of animate objects and actions. *Journal of Experimental Biology, 146, 87-113,* 1989.

D. Perrett, M. Harries, A. Mistlin, J. Hietanen, P. Benson, P. Bevan, S. Thomas, M. Oram, J. Ortega, and K. Brierley. Social signals analyzed at the single cell level: Someone is looking at me, something touched me, something moved! *International Journal of Computational Psychology, 4, 25-55,* 1990.

D. Perrett, P. Smith, A. Mistlin, A. Chitty, A. Head, D. Potter, R. Broennimann, A. Milner, and M. Jeeves. Visual analysis of body movements by neurones in the temporal cortex

of the macaque monkey: a preliminary report. *Behavioural Brain Research, 16, 153-170,* 1985.

F. Pollick, P. McAleer, M. Gleicher, J. Vangeneugden, and R. Vogels. Human recognition of action blends. *Vision Sciences Society Abstract, 653,* 2007.

K. Tanaka (1996). Inferotemporal cortex and object vision. *Annual Reviews of Neuroscience, 19, 109-139,* 1996.

J. Tenenbaum, V. Da Silva, and J. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science, 290, 2319-2323,* 2000.

R. Vogels. Categorization of complex visual images by rhesus monkeys. Part 2: single-cell study. *European Journal of Neuroscience, 11, 1239-1255,* 1999.