

Meaningful interactions can enhance visual discrimination of human agents

Peter Neri, Jennifer Y Luu & Dennis M Levi

The ability to interpret and predict other people's actions is highly evolved in humans and is believed to play a central role in their cognitive behavior. However, there is no direct evidence that this ability confers a tangible benefit to sensory processing. Our quantitative behavioral experiments show that visual discrimination of a human agent is influenced by the presence of a second agent. This effect depended on whether the two agents interacted (by fighting or dancing) in a meaningful synchronized fashion that allowed the actions of one agent to serve as predictors for the expected actions of the other agent, even though synchronization was irrelevant to the visual discrimination task. Our results demonstrate that action understanding has a pervasive impact on the human ability to extract visual information from the actions of other humans, providing quantitative evidence of its significance for sensory performance.

The ability to detect, analyze, interpret and predict the actions of others is perhaps the most remarkable and sophisticated form of visual processing that the human brain is capable of. Within a fraction of a second, we are able to identify a potentially threatening action and react to it appropriately by taking into account its predicted trajectory. This process must involve a large cortical network spanning early visual cortex in the occipital lobe, regions responsive to body and gaze movement in the parietal lobe¹, action-selective areas in the temporal lobe² and the mirror neuron system in the frontal³ and parietal lobes⁴.

Despite its paramount role in our higher-level cognition, we have no quantitative evidence that this ability actually provides a measurable benefit to human processing of sensory information. Visual discrimination of simple human actions such as walking involves stages that are subsequent to local motion extraction⁵, and these stages can bias low-level vision by top-down control⁶. However, the role of action interpretation and action prediction has not been addressed. Hard data on this are lacking because of the complexity of the subject, which makes it difficult to tackle using quantitative psychophysical measurements and carefully controlled experiments.

Three main features of our experiments made it possible to achieve this goal. First, we created a large dataset of motion-captured human actions. In contrast to previous studies that used simple, stereotyped and short action sequences^{6–8}, our dataset contained enough variety to allow us to present human observers with visual stimuli that approximate those encountered in real-life situations. Second, for the first time we were able to present and experimentally manipulate two interacting agents, rather than just one as in all previous studies. This feature of our dataset allowed us to study action recognition by altering the interaction between the two agents, which resulted in a new methodology that bypassed the difficulties associated with asking similar questions in one-agent sequences. Third, we obtained true

and unbiased measures of sensory performance. This was achieved by designing our experiments according to rigorous psychophysical techniques and by carrying out several control experiments and cross-checks to ensure consistency in our data.

Our results show that, in the presence of impoverished visual information, the human visual system relies on the semantic interaction between the two agents to retrieve information relating to each agent individually. This finding demonstrates an unsuspected degree of sophistication in the ability of the cortex to learn, store and use information about the natural statistics of human action and human interaction⁹. Moreover, it exposes a degree of complexity in this system that is not present in any current model of human visual processing^{5,10}.

RESULTS

Synchronized is better than desynchronized fighting

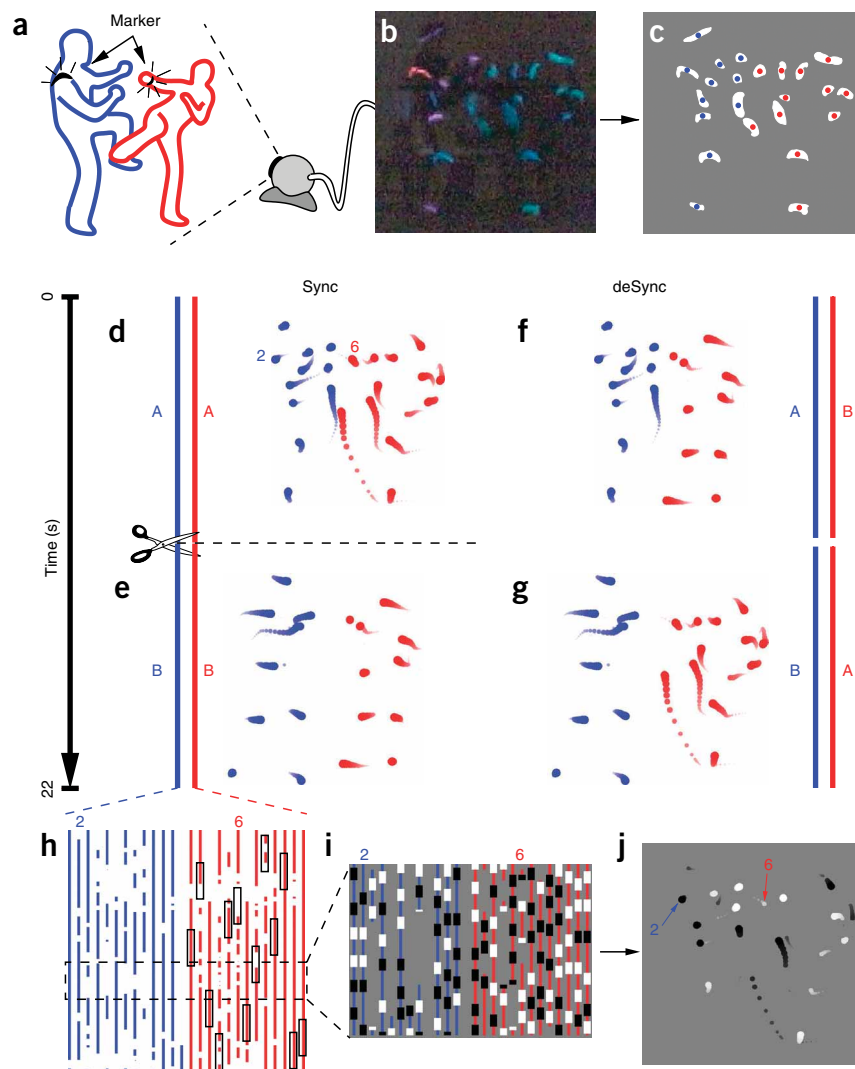
We obtained a natural sample of human fighting by filming two athletes who performed a mixture of kicking and boxing, and were in close-body contact (**Fig. 1a**). We then extracted motion information by tracking the fighters' joints throughout the entire sample (**Fig. 1b**) and converted it into a 22-s-long point-light movie (**Fig. 1c** and **Supplementary Video 1** online). This gave us full control over the trajectories of individual joints for both agents and eliminated visual information unrelated to motion (for example, garment color or body shape). We assembled two sets of sequences from the point-light sample. In the 'Sync' set, we simply split the original 22-s sample into two 11-s-long synchronized sequences (**Fig. 1d,e**). In the 'deSync' set, we cross-paired the actions performed by agent 1 (blue) during the first 11-s sequence (A) with the actions performed by agent 2 (red) during the second 11-s sequence (B) to obtain one desynchronized sequence AB, and vice versa to obtain the other

School of Optometry, University of California at Berkeley, Berkeley, California 94720-2020, USA. Correspondence should be addressed to P.N. (pn@white.stanford.edu).

Received 12 May; accepted 7 August; published online 27 August 2006; doi:10.1038/nn1759

Figure 1 Motion capture and stimulus design.

(a–c) Two fighters wearing light-emitting markers (a) were filmed in a dim room (b), and marker location was tracked frame by frame (c). (d–g) Blue and red lines represent the motion of the two fighters across time (static approximations to corresponding movie fragments are depicted by dot streaks, where the size and contrast of each dot decrease as they refer to more remote frames). We cut the sequence (scissors) into two halves, A and B, obtaining a Sync set of two synchronized sequences, AA (d) and BB (e). For the deSync set we cross-paired A and B for the two fighters, obtaining the two desynchronized sequences AB (f) and BA (g). AA is magnified in h to show all 26 marker trajectories (13 per fighter), with line interruptions indicating disappearance of individual markers owing to occlusion. We randomly selected a segment of 1.5 s duration (dashed box in h) for presentation in target intervals, and sampled the trajectories using black and white dots that had a lifetime of 120 ms (indicated by black and white rectangles in i). The outcome of this sampling procedure is schematically depicted in j (dot size and contrast are varied here for the purpose of providing a static depiction of a moving sequence; no such manipulations were present in the real stimulus). In the nontarget interval, the displayed 1.5-s segment was selected independently for each joint of one fighter (solid rectangles in h), effectively scrambling its trajectories⁸. The observers' task was to identify the target interval. The two markers depicted in a are tracked throughout the figure as '2' (right shoulder) on the blue agent and '6' (right wrist) on the red agent.



sequence BA (Fig. 1f,g). Sync and deSync sets therefore contained identical trajectories on the whole, but consisted of different pairings between the two agents so that their actions were meaningfully related only in the synchronized set (Supplementary Video 2 online).

This distinction between 'synchronized' and 'desynchronized' only makes sense to a device that understands the semantics of fighting. By semantics, we mean not only the ability to interpret the actions of individual agents (for example, agent 1 is kicking rather than punching), but, most importantly, the ability to predict how the actions of one agent should relate to the actions of the other agent (for example, if agent 1 is kicking, agent 2 should defend the kick). Without the latter ability, no difference exists between the Sync and deSync sets of sequences. Indeed, this distinction is immaterial to a machine with no knowledge of how agents interact during natural fighting, because both sets are in fact synchronized in the sense that the actions of one agent are always correlated with the actions of the other agent (albeit shifted by 11 s in the deSync set). Provided that there are appropriate controls for low-level differences between the two sets of sequences, any difference between them must be ascribed to human semantic processing of the interaction between the two agents.

Each trial was of either the Sync or the deSync type, depending on which set of sequences was used for stimulus generation. Observers saw two intervals on each trial. In the 'target' interval, we randomly selected 1.5 s of fighting from either set (dashed box in Fig. 1h). In the 'nontarget' interval, we selected another 1.5 s from the same set, but

we scrambled one of the two agents by allowing each joint on that agent to come from any point in the sequence (ref. 8 and Fig. 1h; this ensured that, averaged across trials, the same motion trajectories were sampled in both intervals; see Methods). Following this manipulation, the nontarget interval effectively contained only one agent (the other one being scrambled). Observers were asked to select the interval that contained two agents as opposed to one (Supplementary Video 3 online). From the viewpoint of information content, it was irrelevant for this task whether the sequences used to generate the stimuli belonged to the Sync or deSync sets. However, we analyzed the two types of trials separately to determine whether this manipulation had an effect on observers' efficiency for discriminating between target and nontarget.

We measured this efficiency quantitatively by masking each stimulus with noise dots scattered all over the monitor, and established the threshold number of noise dots that could be tolerated before the discrimination became impossible (Supplementary Video 4 online). Stimulus trajectories were sampled by similar dots that moved for 120 ms at a time (Fig. 1i). This procedure is well established in the motion literature^{7,11}, and generated reliable thresholds. We plotted noise tolerance for Sync versus deSync trials (Fig. 2a) and found that all five observers displayed significantly higher noise tolerance for Sync

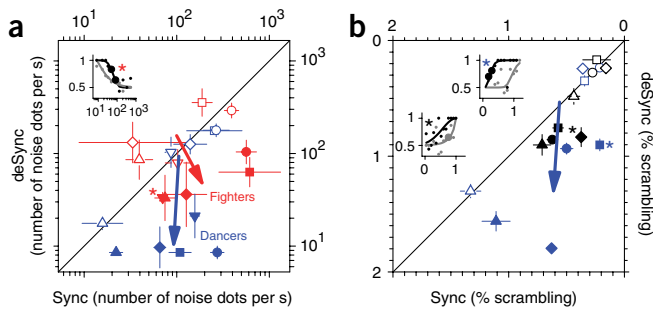


Figure 2 Sensitivity is greater for Sync as opposed to deSync stimuli. (a) Tolerance to masking noise dots for deSync versus Sync trials (thresholds are plotted in units of number of noise dots per s). Solid symbols show data for fighting and dancing sequences. All points fall below the unity line, indicating larger tolerance for Sync versus deSync conditions. Open symbols, control experiment in which one agent was flipped upside down. Arrows point from average of open symbols to average of solid symbols for fighters and dancers separately. Inset, two psychometric functions (black for Sync and gray for deSync) corresponding to the data point indicated by the asterisk (naïve subject). Lines are Weibull fits; dot size roughly scales with number of trials (each point was weighted by the square root of this number for fitting). (b) Plots of scrambling sensitivity. Similar conventions as in a. Smaller percent scrambling thresholds correspond to better sensitivity (axes have been accordingly reversed to make this panel directly comparable to a). Black symbols refer to the occlusion control experiment (solid) and to the unlimited-lifetime condition (open). Symbol shape refers to different subjects. Two authors (J.Y.L., squares; and P.N., circles) participated in the experiment; all other subjects were naïve. Error bars indicate s.e.m. (smaller than symbol when not visible).

trials than for deSync trials (paired *t*-test for Sync versus deSync across observers, $P < 0.01$). Because we used a forced-choice procedure (observers were forced to select one of two intervals and received feedback on every trial), our measurements were uncontaminated by changes in the observers' willingness to report on the configuration of one type of trial as opposed to the other type (criterion changes). Rather, they reflect genuine differences in sensory processing.

Control experiment to rule out low-level cues

We worried that Sync and deSync sequences may have differed in some trivial manner. For example, it is possible that dots tended to be farther apart or closer to each other in one set of sequences than in the other, or that they were moving more coherently in the Sync set as opposed to the deSync one (or vice versa). We performed several simulations with full statistical knowledge of low-level aspects of the position and motion content in the two sets of sequences, but in none of our simulations could the two sets be distinguished even in the absence of any visual noise and sampling (a few examples of the metrics we used are shown in **Supplementary Video 2**). These simulations all indicated that it was very unlikely that potential differences in low-level spatio-temporal properties of our stimuli may have caused the effect described above. However, we wished to address this issue more directly with data. We therefore ran the following control experiment. In both target and nontarget intervals, we inverted the nonscrambled agent (in the target interval both agents were not scrambled, so we picked one at random). It is well known that upside-down point-light agents are poorly perceived as meaningful human actors^{12,13}, so this manipulation effectively eliminated the semantic contribution of one of the two agents from the fighting sequence. However, it preserved almost all low-level structure, such as the average distance between dots and motion coherence. If low-level structure resulted in the differential effect between Sync and deSync (**Fig. 2a**), then the effect should not be

affected by inversion of one agent. If, however, the effect was caused by the semantics of fighting between the two agents, inverting one agent should eliminate any differential effect between Sync and deSync conditions. For this experiment, observers were asked to indicate the interval containing one agent as opposed to no agent. The results (**Fig. 2a**) showed that, clearly, low-level structure was not the source of the differential effect ($P = 0.23$ for Sync versus deSync).

Generalizability to other forms of human interaction

We wondered about the generality of the Sync versus deSync effect (**Fig. 2a**). For example, is it specific to fighting or does it apply to other types of human interaction? To address this question, we repeated our measurements using an action (dancing) that lies at the opposite extreme from fighting in the conflictual/cooperative spectrum (**Supplementary Video 1**). In dancing, the two agents cooperate with each other, rather than contesting each other as they do in fighting. Moreover, visual discrimination of dancing has an important role in human mate selection¹⁴.

We filmed two competition dancers performing a rumba-like routine for 24 s and motion-captured the entire sample as was done for fighting. We repeated our measurements on the dancing sequences (**Fig. 2a**; stimulus duration was increased from 1.5 s to 3 s for dancers) and found that the differential effect between Sync and deSync trials ($P < 0.01$) was even more pronounced than for fighting. It is tempting to speculate that the strength of the effect might be related to the degree to which the synchronized actions were cooperative or antagonistic. As for fighting, the inverted control experiment showed that this effect did not result from potential low-level differences between the two sequences ($P = 0.38$). We concluded that our results generalize to more than one type of action.

Generalizability to other modes and performance metrics

Although the Sync versus deSync effect seems to generalize across different types of human interaction, it is possible that it may not generalize to other forms of disruption or modes of presentation besides embedding in visual noise. In other words, it is possible that the effect is only observed when using noise dots to measure it and/or that the presence of disrupting visual noise *per se* is crucial for observing this effect. For example, the spatial location of the agents is unambiguous in the absence of noise, but can be uncertain when noise is added to the stimulus. To address this issue of generality, we repeated our experiments in conditions where our agents were displayed on a clear background with no noise dots. We also chose a different measure of sensory efficiency than noise tolerance. In the nontarget interval, we manipulated the amount of scrambling applied to the scrambled agent and determined the threshold amount of scrambling that observers needed in order to be able to discriminate between target and nontarget (**Methods and Supplementary Video 4**).

If the Sync versus deSync effect we measured using noise dots reflects a general difference in sensory efficiency between Sync and deSync trials, we expect to see a similar difference in sensitivity to scrambling (here we expect deSync > Sync because smaller scrambling thresholds correspond to better sensitivity). Indeed, we observed a marked difference (**Fig. 2b**) for the dancing sequences ($P < 0.03$; axes have been reversed). We also repeated the inverted control experiment, and, again, this experiment showed that potential low-level properties of the stimulus do not explain the Sync versus deSync effect ($P = 0.41$), even when measured using a completely different metric. This control experiment provides very strong evidence in this respect, because any such low-level properties would be substantially different between noisy and clear displays.

Control to exclude a role for interagent occlusion

A potentially relevant difference between Sync and deSync sequences was that the disappearance of one agent due to occlusion by the other agent was meaningfully represented in the Sync sequence, but not in the deSync one (for example, one agent may disappear owing to occlusion by the other agent in the original sample, but in the desynchronized version, the other agent may be somewhere else doing something else). This difference could potentially be the source of the differential effect reported above (Fig. 2). We believe this is highly unlikely for a number of important reasons (listed in the caption for **Supplementary Video 3**). For example, the effect was stronger for dancing than for fighting, but interagent occlusion was more pronounced in the fighting sequence. Moreover, occlusion has little meaning when dots appear and disappear with a limited lifetime. However, to convince ourselves further, we ran the following control experiment.

We identified all occurrences of interagent occlusion throughout the entire 24-s-long dancing sample and were able to pick out two 3-s-long segments (one from the first half and one from the second half of the sample) that contained virtually no interagent occlusion (Methods). We then assembled these segments in the same manner used for the two 12-s halves (described previously) and repeated our measurements on occlusion-devoid sequences. If interagent occlusion was the source of the Sync versus deSync differential effect, this effect should be absent for the new sequences. The results (Fig. 2b) clearly showed that interagent occlusion was not the source of the effect, as it persisted when occlusion was eliminated ($P < 0.03$). We observed a slight reduction in the effect, but this was most probably due to the shorter stimulus duration that we were forced to adopt for this experiment (1.5 s as opposed to 3 s; Methods).

deSync/Sync ratio correlates with synchronicity detection

If the Sync versus deSync effect was indeed a consequence of interagent synchronization, we would expect the size of the effect to correlate with the perceptual impact of synchronization. To test this prediction, we made independent measurements of how efficiently each subject could detect the presence of synchronization in each main condition we tested. Specifically, we presented a synchronized stimulus in one interval and a desynchronized stimulus in the other interval, and asked our observers to identify the interval in which the two agents interacted in a meaningful way (the Sync interval). Before running this last experiment, we asked the three naïve observers whether they had ever been aware that, during the preceding experiments, some trials in fact contained sequences in which the two agents were desynchronized. None of the three reported being aware of this manipulation at any point during testing (nevertheless, they all showed the Sync versus deSync effect we report here). Similarly, in the presence of threshold noise levels, the two non-naïve subjects were unaware of whether a given trial was of the Sync or deSync type.

We plotted the ratio between deSync and Sync thresholds (which reflects the magnitude of the effect) versus the ability to detect synchronization (Fig. 3), for all observers and in all the main conditions tested (for scrambling thresholds, we plotted Sync versus deSync). We expected a negative correlation between these two quantities—that is, the more detectable the synchronization (large values on the abscissa), the larger the reduction in deSync thresholds as opposed to Sync thresholds (small values on the ordinate). Indeed, we observed a correlation coefficient of -0.84 for the noise thresholds and -0.78 for the scrambling thresholds (emphasizing the reliability of our threshold estimates). This independent assessment of the impact of synchronization strengthened our conclusion that interagent synchronization is indeed the relevant variable in our experiments. The similarity in

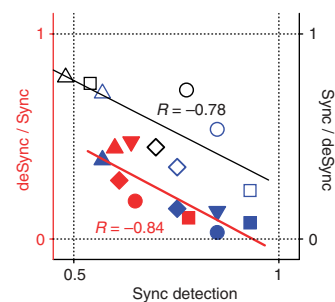


Figure 3 The Sync versus deSync effect correlates with the detectability of synchronization. The magnitude of the Sync versus deSync effect reported in **Figure 2** is plotted on the ordinate as the ratio between deSync and Sync thresholds for noise tolerance (solid) and the ratio between Sync and deSync thresholds for scrambling sensitivity (open). A value of 1 indicates no effect; a value of 0 indicates maximum effect. Red, fighting sequences; blue, dancing; black, occlusion control experiment. The percentage of correct identifications of Sync versus deSync intervals for the synchronization detection experiment is plotted on the abscissa (0.5 is chance, 1 is 100% correct). Each symbol refers to a different subject (same convention as in **Fig. 2**). Linear fits to solid and open symbols are shown by red and black lines respectively (correlation coefficients are indicated next to lines). Slopes for the fits are -1.06 and -1.02 .

correlation coefficient for the two independent sets of measurements (noise versus scrambling thresholds) was also accompanied by identical slopes (-1) for the linear fits, supporting the notion that these two performance metrics were targeting the same mechanism.

In summary, we observed enhanced visual discrimination of a human agent when the agent was embedded in a fighting or dancing sequence with another agent who was acting synchronously as opposed to asynchronously, even though synchronicity was irrelevant to the discrimination task. As demonstrated by our control experiments, this effect was not explained by low-level motion information or by interagent occlusion. Our synchronized and desynchronized stimuli were matched in all respects apart from interagent relationships. The average raw motion content was identical across trials. The structure-from-motion content was also matched, in that individual agents were preserved in both conditions and structure-from-motion only extends to an individual agent, not across agents (there is no structure between agents). Finally, the actions of each agent taken individually were identical. We therefore conclude that the synchronicity effect is a consequence of human implicit knowledge about the implications of an agent's action for the other agent. We now turn to a more detailed interpretation of this unexpected result.

DISCUSSION

Time-locking versus meaningful interaction

How can synchronicity between agents affect visual discrimination of individual agents? In our task, target and nontarget intervals only differed with respect to one agent (the scrambled one). From the point of view of an ideal detector with full knowledge of the statistics of the signal on every interval, the presence of the second agent does not provide additional benefit for discriminating between target and nontarget. However, a nonideal detector with internal noise could benefit from the second agent, because the actions of the two agents are time-locked. If the relevant agent is encoded in space-time with some uncertainty due to internal noise, time-locking provides an implicit temporal cue and the additional agent can be used to predict the expected trajectory of the relevant agent with better precision. In other words, if action A for agent 1 is always coupled with action B for agent 2,

the two actions will statistically reinforce each other and allow for more robust coding of the occurrence of either one. However, our stimuli were time-locked in both Sync and deSync conditions, so both conditions should be equally affected by the additional agent and no difference between them should be observed. Clearly, human visual processing requires not only that the two agents are time-locked, but that they are time-locked in a meaningful way according to patterns that are observed in natural behavior.

Enhancement by Sync versus disruption by deSync

The Sync > deSync effect that we report may have arisen in one of two ways: enhancement by synchronization or disruption by desynchronization (or both). We can distinguish between these two possibilities by studying how data points from the inverted control experiment shifted toward those obtained in our main experiments, where the control served as a proxy for baseline (arrows show the average shift across subjects). A rightward shift would correspond to enhancement by synchronization. Similarly, a downward shift would indicate disruption by desynchronization. There was indicative evidence of enhancement by synchronization only in the fighting experiment, where the rightward shift was very close to significant ($P = 0.055$). However, disruption by desynchronization was evident and highly significant in all conditions ($P < 0.001$ and $P < 0.005$ for fighters and dancers in Fig. 2a, and $P < 0.04$ for dancers in Fig. 2b), indicating that the effect of synchronicity may be mainly due to disruption from nonsynchronous action rather than enhancement from synchronous action.

Incidentally, this analysis validates the notion that the inverted control experiment did preserve all potential low-level cues that may have aided performance in the Sync condition. If such low-level cues were at all present, and if they had been lost by the inversion of one agent, performance in the inverted control experiment should have been lower than that in the main experiment for Sync trials. This would correspond to a rightward shift in our plots. As detailed above, such a shift was barely present for the fighting sequence and totally absent for the dancing sequence, implying that the inverted control experiment did preserve all putative low-level cues that may have been present in the Sync sequence.

Role of undersampling and fragmentary information

In light of the above discussion and experimental evidence, we propose that performance in the desynchronized condition is impaired because the human visual system attempts to interpret the visual stimulus using an implicit model (the synchronized one) that does not capture the desynchronized signal. In this framework, the actions of one agent are used to guide the processing of the actions performed by the other agent, according to patterns expected from natural vision. When the two agents fight or dance according to such patterns, this predictive strategy successfully matches the signal. When their interaction is no longer natural due to desynchronization, this strategy leads to errors because the prediction does not match the signal.

If this interpretation is correct, we expect that the effect of desynchronization would be most pronounced in the presence of a fragmentary signal like the undersampled stimuli we used in our experiments. Undersampling prompts the visual system to interpolate missing information, and interpolation is dependent upon the predictive process just described. We expect that when the signal is not fragmentary, the effect of desynchronization should be largely reduced because predictive interpolation should have a minimal role in processing an optimally sampled signal.

We tested this prediction by measuring scrambling thresholds for unlimited lifetime of the sampling dots. In this condition, all joints

were constantly sampled throughout the entire presentation (Supplementary Video 3), leaving little room for the predictive processing of an agent's action by projection of the other agent's actions. The differential Sync versus deSync effect was entirely eliminated (Fig. 2b)—that is, the observers were equally efficient on Sync and deSync trials ($P = 0.7$)—confirming our expectations. This experiment also added evidence to the notion that the reduced performance on deSync trials cannot be attributed to either interagent occlusion or to observers being 'put off' by desynchronization and refusing to perform the task altogether (details in caption to Supplementary Video 3).

The fragmentary stimuli we created for our experiments resemble the impoverished signals that may be encountered in natural vision. For example, a sampling effect akin to the one we obtained using the limited-lifetime technique may be caused by patchy occlusion from foliage or mist. The changing polarity (bright/dark) of the sampled segments may result from fragmented shadow patterns projected by surrounding vegetation. The ability to extract signals from the motion of potential predators is at its highest premium in these impoverished conditions: once the predator is in full view and visual information about its configuration is unambiguous, it is typically too late to act successfully. Sampling has a much more prominent role in affecting the discrimination of structured biological motion as opposed to simple translational motion⁷. In combination with this, these results suggest that heavily sampled and fragmentary information may be the most fruitful signal-to-noise regime for gaining insights into the processing of complex action patterns.

Relations to physiology and high-level cognition

Individual neurons within the superior temporal polysensory area, located in the superior temporal sulcus (STS), respond selectively to body movements such as walking¹⁵. Recent functional magnetic resonance imaging (fMRI) measurements in monkey cortex have exposed action-selective responses along the entire STS extending to inferior temporal cortex². Human STS has been similarly implicated in action processing (refs. 16,17; review in ref. 1) and responds selectively to the percept of visual 'animacy' induced by the motion of simple geometrical shapes^{18–20}. Clearly, action detection in our tasks must involve these cortical regions. However, a satisfactory explanation of the Sync versus deSync effect requires processing beyond action detection—namely, action interpretation and its implications for the interaction between two third-party agents.

Action interpretation is currently thought to result from an implicit simulation process²¹, whereby a viewed action is understood by linking it with its potential execution by the viewer³. The neural substrate for this link has been established by both single-unit^{22,23} and fMRI (ref. 24) recordings in area F5 within monkey premotor cortex, where 'mirror' neurons respond when the monkey performs or views a specific action²⁵. Subsequent studies have located regions of human cortex with analogous properties^{26–30}, defining a large-scale network for the human mirror neuron system (MNS). In an imaging study of immediate relevance to our stimuli, activity within this system for expert dancers was shown to depend upon a match between their motor repertoire and the observed dancing routine³¹. More broadly, the MNS may support 'mentalizing'³² (the process by which we make inferences about mental states). This cognitive operation has been implicated in emotional processing^{33,34} and in autism^{8,35–37}. Although our results are not directly pertinent to emotional processing, point-light figures can convey information about emotional state^{38,39} and have been shown to activate the human amygdala¹⁶, a limbic structure which is also responsive to social 'vignettes' containing geometrical shapes engaged in intentional motion¹⁹.

We speculate that the higher-level processing stages exposed by our experiments extend to the MNS complex and could not be entirely supported by the STS because the meaningfulness of interagent interaction in our stimuli can only be represented once mentalizing about the actions of the individual agents is completed. Although action processing in the STS can display sophisticated properties such as audiovisual integration⁴⁰, there is no electrophysiological evidence that this multimodality extends to motor programs⁴¹, thus making this region an unlikely candidate for action interpretation through visuo-motor simulation (the increase in fMRI activity within human STS in response to both action observation and execution is interpreted as feedback from the MNS; ref. 42). To summarize, a comprehensive neural account of our results is likely to involve a large network of cortical areas spanning both STS and MNS.

We wish to emphasize that the above considerations are of a very speculative nature, not only because our behavioral measurements cannot address any of these issues, but, most importantly, because at present there is no conclusive evidence as to which of these brain structures are directly responsible for which operations. Future investigations should therefore remain open to the possibility that different brain pathways may be involved.

On a more practical note, our study clearly raises the difficult question of how to design performance metrics that permit a quantitative assessment of the impact of ‘social processing’ on behavioral performance. In our study, we combined well-established psychophysical techniques with higher-level stimuli that were reduced in non-semantic complexity to the point of being amenable to measurements using these techniques. An important question for future investigations is whether this strategy can be successfully applied to other and more complex problems in social cognition.

Visual cortex as a model builder

Our experiments demonstrate that the human visual system uses knowledge about the natural statistics of human interactions to guide processing of the motion patterns generated by the actions of individual agents. No existing model of action processing^{5,10} incorporates such a sophisticated degree of top-down control. Indeed, it is surprising that cortex uses such high-level information to guide visual processing of human kinematics. These results provide a notable demonstration of the pervasive nature of predictive coding in visual information retrieval^{9,43}. We provide the first demonstration of how predictive coding at the highest level tested so far in the motion hierarchy (interagent interaction) directly affects sensory processing of behaviorally relevant visual information. Future studies will be necessary to characterize this phenomenon in more detail: to understand its exact computational characteristics, pinpoint its neural substrates, trace its evolutionary origin and determine its dependence on postnatal experience.

METHODS

Motion capture. Actors were recruited from the University of California, Berkeley, Martial Arts and Ballroom Dance teams for fighting and dancing, respectively. They were asked to perform standard routines while wearing sports clothes which we had fitted with battery-driven body lights (ClubThings; Fig. 1b). There were 13 such light markers on each actor: one on the head, and two each on the shoulders, elbows, wrists, hips, knees and feet. We filmed the actors in a dimly lit room using a camera device (Logitech QuickCam) that generated digital audio video interleave (AVI) movies at a resolution of 10 Hz and 640 pixels \times 480 pixels. The movies were processed by customized Matlab software, which we wrote for the specific purpose of computer-assisted motion capture. The program performed basic cluster analysis to identify extensive regions of high luminance corresponding to the body lights and attempted to

place markers that would track the motion of individual clusters throughout all frames in the movie. A graphical user interface allowed us to view the outcome of this automated tracking frame by frame and to correct the numerous errors made by the program. This human/computer mixed procedure allowed us to track each joint in x - y - t (we interpolated the sequence to obtain 30-Hz sampling) and to tag all disappearances caused by occlusion. We tracked 22.7 s of fighting and 24 s of dancing (Supplementary Video 1).

Stimulus design and psychophysical tasks. The two ‘Sync’ sequences simply corresponded to the first and second halves of the original tracked movie (Fig. 1d,e). We refer to the two collectively as the Sync set. The deSync set consisted of the two sequences obtained by cross-pairing the trajectories of one agent in the Sync sequence with the trajectories of the other agent in the other sequence (Fig. 1f,g and Supplementary Video 2). For all experiments with the exception of synchronicity detection, both intervals on each trial were generated using only one set, whether Sync or deSync. For each interval, we randomly selected a short segment from either one (randomly chosen) of the two sequences in the chosen set. The duration of the segment was 1.5 s for fighters and 3 s for dancers, with the exception of the occlusion control experiment for which we had to use 1.5 s for dancers because the original usable sample was shorter for this condition (it consisted of two 3-s segments in which interagent occlusion was reduced from 6% in the overall sample to 1%). The selected segment was displayed using a limited-lifetime sampling technique (with the exception of the unlimited-lifetime experiment—open black symbols in Fig. 2b—for which no sampling was applied). The 26 trajectories were randomly sampled by 12 dots (4.6 arcmin diameter) whose lifetime was 120 ms (matching the temporal integration window of local motion detectors⁴⁴), after which they sampled a different trajectory. Dot appearance and disappearance was asynchronous across dots in order to avoid motion transients from simultaneous transitions of all sampling dots (ref. 7 and Supplementary Video 3). Dots could be randomly bright (74 cd m⁻²) or dark (0 cd m⁻²) on a gray (37 cd m⁻²) background (which ensured that no change in mean luminance ever took place in our experiments) and did not change color during their lifetime (whether limited or unlimited). The trajectories were sized so that their overall center of mass (across the entire sample) was centered on an Iiyama monitor driven by a VSG graphics card (Cambridge Research Systems), and they did not extend outside a 6.4° \times 6.4° region. Subjects fixated on a central marker located 114 cm from the monitor (fixation was only loosely enforced). In the nontarget interval, we scrambled the 13 trajectories of one (randomly chosen) agent by selecting a different segment from the original sample for each trajectory, similarly to ref. 8. This procedure is schematically illustrated by the solid rectangles in Figure 1h and ensured that the raw motion content averaged across trials was identical for a scrambled and a nonscrambled agent because each individual joint was sampled uniformly in both. We were able to vary the amount of scrambling by varying the temporal width W of the window within which the solid rectangles could sample the original sequence. For example, in Figure 1h the width of the window matches the duration of the entire sequence. We were able to reduce scrambling by forcing the solid rectangles to fall within a temporal window centered on the dashed outline (when the dashed outline fell close to the edges of the sequence, the window was shifted slightly). In Figure 2b, we express scrambling as $(W - S)/S$ (same definition used in ref. 6 for depth scrambling), where S is stimulus duration (the height of the dashed outline in Fig. 1h). For measuring scrambling thresholds (Fig. 2b), we measured the percentage of correct target identifications as a function of scrambling strength (on a two-up one-down staircase; two separate staircases were run in parallel for Sync and deSync trials), and fitted a Weibull function to this psychometric curve to obtain the threshold estimate (α parameter; see insets in Fig. 2 and Supplementary Video 4). For measuring noise thresholds (Fig. 2a), we kept scrambling strength fixed at the largest available value (W equal to the duration of the entire sequence) and disrupted performance by masking the stimuli with a varying number of noise dots (similarly to scrambling, we derived a psychometric curve as a function of number of noise dots and used the Weibull fit to estimate a threshold number). Each noise dot trajectory was generated as if it came from one of the dots sampling the tracked stimulus, except that it was then rotated by 0°, 90°, 180° or 270° randomly for each lifetime and its starting positions could be anywhere within a 7.7° \times 7.7° region. This ensured that the local motion of the noise dots

was identical to that of the dots sampling the stimulus (except for rotation). For measuring detection of synchronization, we did not apply any scrambling or noise dots. In one interval, we presented the stimulus that was originally presented in the target interval of Sync trials, and in the other interval we presented the stimulus that was originally presented in the target interval of deSync trials. Observers had to identify the meaningful (Sync) interval (two-alternative forced choice with feedback), and we express detection as percentage of correct identifications (plotted on the abscissa in Fig. 3). This experiment was run on each subject only after all the other conditions had been tested, including the unlimited-lifetime condition.

Note: Supplementary information is available on the Nature Neuroscience website.

ACKNOWLEDGMENTS

Supported by US National Institutes of Health grant RO1EY01728.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at <http://www.nature.com/natureneuroscience>

Reprints and permissions information is available online at <http://npg.nature.com/reprintsandpermissions/>

- Puce, A. & Perrett, D. Electrophysiology and brain imaging of biological motion. *Phil. Trans. R. Soc. Lond. B* **358**, 435–445 (2003).
- Nelissen, K., Vanduffel, W. & Orban, G.A. Charting the lower superior temporal region, a new motion-sensitive region in monkey superior temporal sulcus. *J. Neurosci.* **26**, 5929–5947 (2006).
- Rizzolatti, G. & Craighero, L. The mirror-neuron system. *Annu. Rev. Neurosci.* **27**, 169–192 (2004).
- Gallese, V., Fadiga, L., Fogassi, L. & Rizzolatti, G. Action representation and the inferior parietal lobule. in *Common Mechanisms in Perception and Action: Attention and Performance XIX* (eds. Prinz, W. & Hommel, B.) 247–266 (Oxford Univ. Press, Oxford, 2002).
- Giese, M.A. & Poggio, T. Neural mechanisms for the recognition of biological motion. *Nat. Rev. Neurosci.* **4**, 179–192 (2003).
- Bülthoff, I., Bülthoff, H. & Sinha, P. Top-down influences on stereoscopic depth-perception. *Nat. Neurosci.* **1**, 254–257 (1998).
- Neri, P., Morrone, M.C. & Burr, D.C. Seeing biological motion. *Nature* **395**, 894–896 (1998).
- Blake, R., Turner, L.M., Smoski, M.J., Pozdol, S.L. & Stone, W.L. Visual recognition of biological motion is impaired in children with autism. *Psychol. Sci.* **14**, 151–157 (2003).
- Barlow, H.B. Cerebral cortex as model builder. in *Models of the Visual Cortex* (eds. Rose, D. & Dobson, V.G.) 37–46 (John Wiley & Sons, New York, 1985).
- Lange, J. & Lappe, M. A model of biological motion perception from configural form cues. *J. Neurosci.* **26**, 2894–2906 (2006).
- Morrone, M.C., Burr, D.C. & Vaina, L.M. Two stages of visual processing for radial and circular motion. *Nature* **376**, 507–509 (1995).
- Sumi, S. Upside down presentation of the Johansson moving light spot pattern. *Perception* **13**, 283–286 (1984).
- Pavlova, M. & Sokolov, A. Orientation specificity in biological motion perception. *Percept. Psychophys.* **62**, 889–899 (2000).
- Brown, W.M. *et al.* Dance reveals symmetry especially in young men. *Nature* **438**, 1148–1150 (2005).
- Oram, M. & Perrett, D.I. Responses of anterior superior temporal polysensory (STPa) neurons to 'biological motion' stimuli. *J. Cogn. Neurosci.* **6**, 99–116 (1994).
- Bonda, E., Petrides, M., Ostry, D. & Evans, A. Specific involvement of human parietal systems and the amygdala in the perception of biological motion. *J. Neurosci.* **16**, 3737–3744 (1996).
- Grossman, E.D. & Blake, R. Brain areas active during visual perception of biological motion. *Neuron* **35**, 1167–1175 (2002).
- Castelli, F., Happé, F., Frith, U. & Frith, C. Movement and mind: a functional study of perception and interpretation of complex intentional movement patterns. *Neuroimage* **12**, 314–325 (2000).
- Martin, A. & Weisberg, J. Neural foundations for understanding social and mechanical concepts. *Cogn. Neuropsychol.* **20**, 575–587 (2003).
- Schultz, J., Friston, K.J., O'Doherty, J., Wolpert, D.M. & Frith, C.D. Activation in posterior superior temporal sulcus parallels parameter inducing the percept of animacy. *Neuron* **45**, 625–635 (2005).
- Gallese, V. & Goldman, A. Mirror neurons and the simulation theory of mind-reading. *Trends Cogn. Sci.* **2**, 493–501 (1998).
- Kohler, E. *et al.* Hearing sounds, understanding actions: action representation in mirror neurons. *Science* **297**, 846–848 (2002).
- Umiltà, M.A. *et al.* I know what you are doing: a neurophysiological study. *Neuron* **31**, 155–165 (2001).
- Nelissen, K., Luppino, G., Vanduffel, W., Rizzolatti, G. & Orban, G.A. Observing others: multiple action representation in the frontal lobe. *Science* **310**, 332–336 (2005).
- Rizzolatti, G., Fogassi, L. & Gallese, V. Neurophysiological mechanisms underlying the understanding and imitation of action. *Nat. Rev. Neurosci.* **2**, 661–670 (2001).
- Fadiga, L., Fogassi, L., Pavesi, G. & Rizzolatti, G. Motor facilitation during action observation: a magnetic stimulation study. *J. Neurophysiol.* **73**, 2608–2611 (1995).
- Iacoboni, M. *et al.* Cortical mechanisms of human imitation. *Science* **286**, 2526–2528 (1999).
- Peelen, M.V., Wiggett, A.J. & Downing, P.E. Patterns of fMRI activity dissociate overlapping functional brain areas that respond to biological motion. *Neuron* **49**, 815–822 (2006).
- Nishitani, N. & Hari, R. Temporal dynamics of cortical representation for action. *Proc. Natl. Acad. Sci. USA* **97**, 913–918 (2000).
- Iacoboni, M. *et al.* Grasping the intentions of others with one's own mirror neuron system. *PLoS Biol.* **3**, E79 (2005).
- Calvo-Merino, B., Glaser, D.E., Grzes, J., Passingham, R.E. & Haggard, P. Action observation and acquired motor skills: an fMRI study with expert dancers. *Cereb. Cortex* **15**, 1243–1249 (2005).
- Frith, C.D. & Frith, U. The neural basis of mentalizing. *Neuron* **50**, 531–534 (2006).
- Adolphs, R. Cognitive neuroscience of human social behaviour. *Nat. Rev. Neurosci.* **4**, 165–178 (2003).
- Gallese, V., Keysers, C. & Rizzolatti, G. A unifying view of the basis of social cognition. *Trends Cogn. Sci.* **8**, 396–403 (2004).
- Williams, J.H., Whiten, A., Suddendorf, T. & Perrett, D.I. Imitation, mirror neurons and autism. *Neurosci. Biobehav. Rev.* **25**, 287–295 (2001).
- Oberman, L.M. *et al.* EEG evidence for mirror neuron dysfunction in autism spectrum disorders. *Brain Res. Cogn. Brain Res.* **24**, 190–198 (2005).
- Dapretto, M. *et al.* Understanding emotions in others: mirror neuron dysfunction in children with autism spectrum disorders. *Nat. Neurosci.* **9**, 28–30 (2006).
- Dittrich, W.H., Troscianko, T., Lea, S. & Morgan, D. Perception of emotion from dynamic point-light displays represented in dance. *Perception* **25**, 727–738 (1996).
- Bassili, J.N. Facial motion in the perception of faces and of emotional expression. *J. Exp. Psychol. Hum. Percept. Perform.* **4**, 373–379 (1978).
- Barracough, N.E., Xiao, D., Baker, C.I., Oram, M.W. & Perrett, D.I. Integration of visual and auditory information by superior temporal sulcus neurons responsive to the sight of actions. *J. Cogn. Neurosci.* **17**, 377–391 (2005).
- Keysers, C. & Perrett, D.I. Demystifying social cognition: a Hebbian perspective. *Trends Cogn. Sci.* **8**, 501–507 (2004).
- Iacoboni, M. *et al.* Reafferent copies of imitated actions in the right superior temporal cortex. *Proc. Natl. Acad. Sci. USA* **98**, 13995–13999 (2001).
- Kersten, D., Mamassian, P. & Yuille, A. Object perception as Bayesian inference. *Annu. Rev. Psychol.* **55**, 271–304 (2004).
- Burr, D. Motion smear. *Nature* **284**, 164–165 (1980).