Chapter 32

Eye Movements and Gestures in Human Face-to-face Interaction

Marianne Gullberg

Gestures are visuospatial events, meaning carriers, and social interactional phenomena. As such they constitute a particularly favourable area for investigating visual attention in a complex everyday situation under conditions of competitive processing. This chapter discusses visual attention to spontaneous gestures in human face-to-face interaction as explored with eye-tracking. Some basic fixation patterns are described, live and video-based settings are compared, and preliminary results on the relationship between fixations and information processing are outlined.

Introduction

This paper is concerned with visual attention to gestures in face-to-face interaction. It addresses a few seemingly simple questions: do addressees look at speakers' gestures in interaction? If they do, at which ones do they look, and why? Despite the widespread interest in the visual perception of hands in neurology (e.g. Decety & Grèzes, 1999; Goldenberg, 2001; Grèzes *et al.*, 1999; Hermsdörfer *et al.*, 2001; Neville *et al.*, 1997; Peigneux *et al.*, 2000; Perani *et al.*, 2001; Rizzolatti, *et al.*, 2001), in studies of Sign Language perception (e.g. Bavelier *et al.*, 2001; Corina *et al.*, 1996; Neville *et al.*, 1997; Rettenbach *et al.*, 1999; Swisher, 1993), and in gesture recognition in man-machine interfaces (e.g. Braffort, *et al.*, 1999; Wachsmuth & Sowa, 2002), we know surprisingly little about the attention afforded to gestures in human interaction. In gesture research the interest in these questions is motivated by an ongoing debate regarding whether or not gestural information is communicatively relevant to participants in interaction. It will be suggested in this chapter that these questions are also of relevance for eye movement research. In particular, the answers to these question may inform research concerning task-specific behaviour and visual attention in complex, natural situations.

The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research

ISBN: 0–444–51020–6

The gestures we are concerned with here are the (mainly manual) movements speakers perform unwittingly while they speak as part of the expressive effort (Kendon, 1993; McNeill, 1992). This definition excludes functional actions, manual object manipulations and movements such as scratching or playing with hair, since these movements are not part of the speaker's intended message. The narrow definition still leaves a broad spectrum of movements to consider. It covers conventionalised gestures like the OK-sign. It also includes a wide range of spontaneous, non-conventional movements that imitate real actions, iconically represent, or indicate entities talked about (Kendon, 1986; McNeill, 1998). Gestures thus defined are symbolic movements, and they are closely temporally and semantically related to language and speech. Simply put, gestures and speech tend to express the same meaning at the same time. However, even if gestural information often is redundant with regard to speech, gestures can also express additional information not present in concomitant speech.

It is partly this latter property that motivates an ongoing debate in gesture research regarding their communicative relevance. Because gestures can encode additional information to that expressed in speech, the need to attend to this information may not be essential to addressees in spoken face-to-face interaction. Whilst a number of studies have been concerned with gaze towards gestures (Goodwin, 1986; Kendon, 1990; Streeck, 1993, 1994; Streeck & Knapp, 1992; Tuite, 1993), there is a conspicuous lack of precise perceptual data concerning attention to gestures. Furthermore, there has been no systematic investigation of the relationship between visual attention to gestures and the processing and integration of the gestural information. This empirical gap has largely motivated the studies outlined in this chapter.

From the point of view of vision research, gestures in interaction also offer a challenging opportunity to study what catches attention in complex natural settings, and the influence of specific tasks and activities on fixation patterns. Gestures constitute a potential locus for visual attention by virtue of being visuospatial phenomena that represent movement in the visual field. Gestures could also be the locus of visual attention because they are symbolic movements that encode meaning closely related to but not necessarily identical to that expressed in language and speech. Gestures could thus be visually attended to for low-level perceptual reasons or for reasons related to higher cognitive processes such as information extraction. Finally, gestures are (mostly) interactional, social phenomena. As such, their occurrence in situations that are governed by socially and culturally determined norms for behaviour is likely to modulate visual behaviour towards them. There is thus potential tension between different mechanisms governing visual attention: the tendency to attend to movement, the need to look at what you are seeking information about, and the social conventions that govern gaze.

This paper will outline the results from a number of recent and forthcoming studies that exploit eye tracking to investigate the attention addressees allocate to gestures in interaction. Since some of these studies are presently unpublished, and the methodology is novel, I will describe the general procedure in some detail and present some key results. This chapter will cover three main foci: (1) the basic fixation patterns in face-to-face interaction; (2) the effect of the medium of presentation, comparing live vs. video settings; and (3) the relationship between fixation and information uptake in a complex setting. In a final section, I will discuss the findings and relate them to some issues raised in eye movement research, particularly with respect to the generalisability of findings from lab settings to naturalistic contexts. Specifically, I will discuss the effect of the task and context on gaze patterns.

Eye Movements and Gestures in Live Interaction

Social, dyadic face-to-face interaction is one of the most primary and common types of human activity. Gaze is of fundamental importance to such interaction and has previously received considerable attention in various disciplines. Numerous studies in social psychology, e.g., have investigated the influence of personality, gender, culture, interactional style and setting, mental health, etc., on gaze patterns in interaction (for comprehensive overviews, see Argyle & Cook, 1976; Fehr & Exline, 1987; Kendon, 1990). However, none of these studies have been based on precise measurements of eye movements. We therefore know very little about the exact patterns that arise from this particular activity, and virtually nothing about how behaviour in this setting relates to findings for other natural activities such as driving, or tea- or sandwich-making (Hayhoe, 2000; Land & Hayhoe, 2001; Land *et al.*, 1999; Shinoda *et al.*, 2001).

Despite the lack of precise measurements, a number of claims have been made in the literature regarding participants' gaze patterns in interaction. For instance, a frequently reported observation is that the speaker's face dominates as a target for addressees (Argyle & Cook, 1976; Fehr & Exline, 1987; Kendon, 1990). It is generally assumed that this is a reflection of a (culture-specific) politeness norm for maintained mutual gaze signalling sustained interest and attention.

With respect to gestures, three candidates for direct fixation can be derived from the literature. First, gestures performed in peripheral gesture space are often presumed to attract overt visual attention. Gesture space can be divided into central and peripheral gesture space (cf. McNeill, 1992). Central space refers to a shallow disc of space in front of the speaker's body, delimited by the elbows, the shoulders, and the lower abdomen. This area is outlined by a rectangle in Figure 32.1. Peripheral gesture space is everything outside this area. The majority of a speaker's gestures are performed in central gesture space. If an addressee is fixating the speaker's face, then all gestures in principle occur in the addressee's peripheral visual space. Nonetheless, a gesture performed in the speaker's peripheral gesture space occurs in the addressee's extreme peripheral visual field. It is therefore assumed that it could attract fixation, as it would otherwise be too challenging for peripheral vision. Second, pointing gestures are suggested as potential targets for fixation. Pointing gestures direct attention to the target they are indicating. However, it is often presumed that it is necessary to look at the gesture first in order to compute the trajectory towards the target of the pointing gesture. Finally, gestures that speakers themselves look at have also been suggested as candidates for fixation. When speakers look at their own gestures, they are assumed to intentionally direct the addressee's attention to the target of their gaze (Goodwin, 1986; Streeck, 1993, 1994; Streeck & Knapp, 1992; Tuite, 1993). Speakers' gaze shifts are thus claimed to have the same deictic function as pointing.

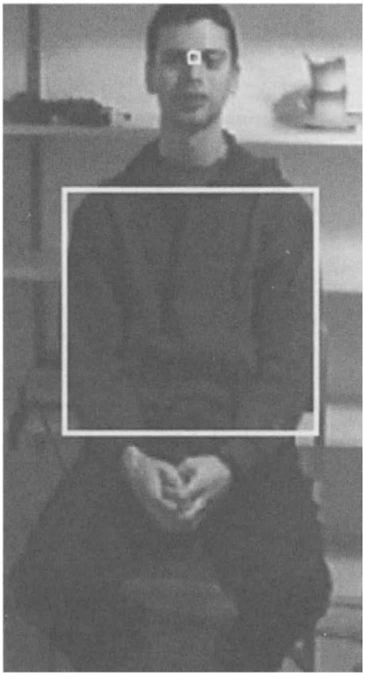

Figure 32.1: The speaker's central gesture space as a rectangle. Everything outside the rectangle represents the speaker's peripheral gesture space. The addressee's fixation as a small white circle at its default location in interaction.

By exploring the possibilities of eye-tracking in face-to-face interaction, we have attempted to investigate these proposals and also to charter the basic fixation patterns in face-to-face interaction (Gullberg & Holmqvist, 1999, 2002). Fifteen pairs of Swedish participants, unacquainted prior to the experiment, were randomly assigned the role as "speaker" or "addressee" (i.e. the wearer of the eye-tracker). In order to allow for spontaneous gesturing while maintaining control over the gestural content, a story-retelling task was used. Speakers memorised a printed cartoon and were instructed to convey the story as well as possible to the addressees who would have to answer questions about it. Addressees were instructed to make sure they understood the story and were encouraged to ask questions and engage in the interaction. Addressees were fitted with a head-mounted (HED) SMI iView© eye-tracker, a monocular 50 Hz pupil and corneal reflex video imaging system. This eye-tracker is well suited to interactional studies as both participants have an unobstructed face view of each other. The device samples data to an average spatial accuracy of 1 degree. Each addressee was calibrated using a nine-point matrix on the wall. The location and size of the matrix was equivalent to the area that the speaker would later occupy. After calibration of the addressees, speakers were introduced into the room and seated 180 cm away from the addres-sees (measured back to back) facing them. The speakers' stories generated natural narratives and a range of spontaneous gestures, all of which were analysed and considered as potential targets for addressees' fixations.

During the story-retelling interaction, the addressees' eye movements were recorded with the corneal reflex camera. The eye-tracker also has a scene-camera on the headband. The iView software creates a circle overlay indicating the gaze position that is then merged with the video of the scene image. Since the scene-camera moves with the head, the eye-in-head signal indicates the gaze point with respect to the world. Head movements therefore appear on the video as full-field image motion. Given that both the target stimulus (the speaker) and the field of vision itself moved, the merged video data of the subject's gaze position on the scene image were analysed frame-by-frame.

Fixations were defined as instances where the gaze marker remained for at least 120 ms directly on a fixated object.

Post-test questionnaires showed that subjects did not identify gestures as the target of the study and were not disturbed by the equipment. In fact, the speech and gestural behaviour of speakers did not differ quantitatively or qualitatively from data collected in an identical situation without eye-trackers (Gullberg, 1998). Addressees' gaze data include fixations of socially unacceptable areas that they might have avoided had they been concerned about the equipment. We interpret this as meaning that the apparatus did not interfere with the addressees' natural behaviour. The ecological validity of the data is therefore not compromised.

Basic Fixation Patterns in Interaction

In both studies the default location of the addressee's fixations is the speaker's face, viz. the nose bridge or eye area as seen in Figure 32.2 (Gullberg & Holmqvist, 1999, 2002).[1] In a context where the background does not represent any particular interest and there are no objects present relevant to the ongoing talk (cf. Argyle & Graham, 1976), addressees on average devote as much as 96% of their total viewing time to the face. This finding is consistent with earlier reports on addressees' gaze in interaction. The remainder of the time is spent fixating objects in the room behind the speaker or the speaker's immobile body parts. Only 0.5% of the total viewing time is devoted to gestures. In terms of number of fixations, only a minority of gestures are overtly fixated by addressees, or on average 7% of all gestures.

There is very little continuous scanning of the scene as a whole in these data and almost no cases of smooth pursuit of gestures. Saccades to fixation locations outside the face, including gestures, are direct and accurate despite the distances involved (cf. Land *et al.*, 1999). Typically, the eye moves from the face directly to a gesture in progress, stays on this target for an average fixation duration of 458 ms (SD = 436 ms), then returns directly back to the default location, i.e. the face. Note that fixations on gestures are spatially unambiguous. In all cases of gesture fixation the entire fixation marker is clearly located directly on the hand, not tangential to it, and not in the vicinity of a gesture in progress. This is also true for landing sites such as objects in the room and immobile body parts.

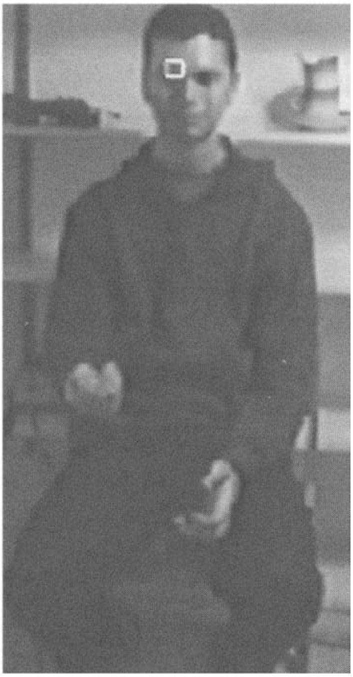

Figure 32.2: Example of the data. The addressee is fixating the default location, the speaker's face, despite the gesture in progress.

Gestures thus receive remarkably little overt visual attention. If addressees attend to gestures, they appear to do so in peripheral vision. Gestures that do attract direct fixations tend to display one or both of two features. They are either gestures with so-called post-stroke holds or gestures that speakers themselves have first looked at, speaker-fixated gestures. A hold is the momentary cessation of movement in a gesture while the hand is maintained immobile in gesture space (Kendon, 1972, 1980). Gestures that stop moving thus attract direct fixations to a greater extent than moving gestures. This finding was unexpected (but see Nobe *et al.,* 1998), as the standard assumption is that movement — rather than offset of movement — attracts attention. However, it is possible that the movement of inalienable body parts is not "salient" enough to draw overt fixations in this particular context (cf. Raymond, 2000). Since gestures are pervasive in interaction during speech, they represent almost constant movement in the addressees' visual field. It is conceivable that the cessation or offset of movement in gesture space instead represents a sudden change in the visual field and that this then evokes fixations. It may also be relevant that the cessation of movement takes place in gesture space, i.e. in symbolic space, and not anywhere in the visual field. A gesture that stops moving because the speaker drops her hands into the lap, i.e. effectively stops gesturing, does not attract fixation.

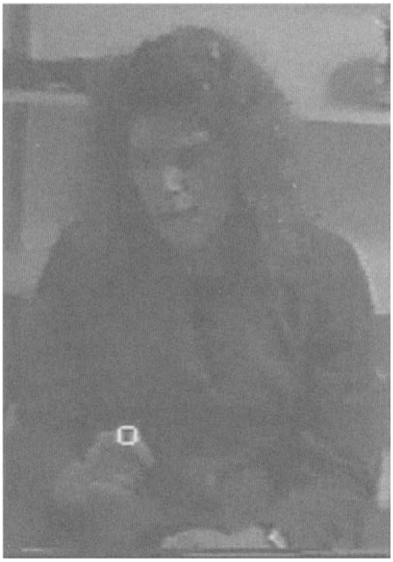

Figure 32.3: Example of a speaker-fixated gesture that is also fixated by the addressee (= white circle).

Speaker-fixated gestures also tend to be fixated by addressees in interaction as exemplified in Figure 32.3. This result is consistent with the claims in the literature. It is also in accordance with what is known about the powerful effect of speakers' gaze and head orientation on joint attention (Deák *et al.*, 2000; Doherty & Anders, 1999; Driver *et al.*, 1999; Gibson & Pick, 1963; Langton, 2000; Langton & Bruce, 1999; Langton *et al.*, 1996; Langton *et al.*, 2000; Moore & Dunham, 1995). The novelty is that gestures themselves can be the target of such joint attention, not just serve as indicators. Also, it has been claimed that speakers' gaze shifts induce automatic or reflexive shift of attention in addressees (cf. Langton *et al.*, 2000). An important observation in these data, however, is that speakers' gaze at their own gestures does not lead to automatic *overt* attention shift in addressees, i.e. to rapid saccades to and fixations of the target. The data do not, of course, tell us anything about the *covert* attention shifts (Hoffman, 1998; Johnson, 1995; Posner, 1980). Nevertheless, it is noteworthy that not all but only 23% of all speaker gaze shifts to gestures lead to overt fixation by addressees.

The results for the two other predicted gesture types, pointing gestures and gestures performed in peripheral gesture space, vary between the studies, and are inconclusive. This variation is presumably due to the fact that the relevant features tend to cluster in spontaneous gesture data. The individual attraction force of these features is therefore difficult to assess. Gullberg and Kita (forthcoming) consequently attempted to establish the individual effect of location of gesture performance (in central/peripheral gesture space), hold (presence/absence), and speaker-fixation (presence/absence) (see also pp. 695–697). Addressees were shown video clips of speakers retelling short stories. Each video clip contained only one gesture with the relevant feature. This gesture was designated as the target gesture. It was embedded in sequences of other gestures so as not to draw attention as a singleton. The target gesture had not been manipulated, but natural examples had been carefully selected from a bigger database such that each target gesture displayed the desired feature. The videos were projected life-sized on a wall and the eye movements of the addressees were recorded using the same set-up and task as described above.

The results showed that the location of the gesture in central or peripheral gesture space had no impact on addressees' fixations. In contrast, holds and speaker-fixation both individually attracted fixations significantly more often than gestures without these features. 12.5% of holds and 12.5% of speaker-fixated gestures were fixated as opposed to 0% of the gestures without these features. We also found a significant difference in saccade onset latencies to gestures with different features. Holds were fixated on average 108 ms after the onset of the gestural hold. In contrast, speaker-fixated gestures were fixated on average 808 ms after the onset of the speaker-fixation, i.e. after the speakers had directed their gaze towards their own gesture. Note that addressees were fixating the speaker's eye region at the moment of the onset both of holds and of speaker's fixations of their own gestures. This means that, in the case of holds, addressees responded very quickly to the cessation of movement perceived peripherally. In contrast, even though they were fixating the speaker's eyes and therefore should have detected the gaze shift immediately on the onset, saccades to the target gestures were only initiated after a considerable time delay. These findings suggest that gestures may be fixated for different reasons. We propose that fixations of holds are stimulus-driven or bottom-up guided whilst fixations of speaker-fixated gestures appear to be goal-driven (cf. Yantis, 1998). At this point, we cannot exclude the possibility that the gestural movement immediately preceding the holds does not initiate some form of saccadic planning. However, it seems unlikely given that gestural movements not followed by holds do not lead to saccades to gestures.[2]

More on Context: Live vs. Video, Social and Size Related Effects

The basic gaze pattern outlined above thus displays a number of properties that appear to be specific to the context of human social interaction. These findings have been challenged, however, in two studies of addressees' attention to the manual gestures of an anthropomorphic agent presented on a computer screen (Nobe *et al.*, 1998, 2000). In these studies, addressees fixated the vast majority of gestures (70–75%). This is in stark contrast to the mere 7% fixated in live interaction. The reduced number of gesture fixations and possibly the dominance of the face in the live setting could therefore be motivated by a social norm for maintained mutual gaze in face-to-face interaction. In the absence of any social pressure for eye contact, as in a video setting for instance, fixation behaviour towards speakers and their gestures may therefore look different. However, the studies by Nobe *et al.* differed from ours on a range of parameters other than the presence/absence of a live interlocutor. Most importantly, they used a different type of agent (anthropomorphic vs. human), a different size of the visual scene (computer screen vs. life size), and different types of gestures (conventionalised vs. spontaneous). To enable an evaluation of these differences, Gullberg & Holmqvist (2002) therefore undertook to specifically study the social effect of the presence/absence of a live interlocutor and the effect of the size of the display on gaze behaviour *ceteris paribus*. We compared fixation behaviour towards the same speakers and their gestures in three conditions: face-to-face live, on a 28-inch video screen, and on life-sized video projected on a wall. The task and the equipment were the same as outlined above, and the procedure for the live condition was identical. In addition to being filmed by the scene-camera on the eye-tracker, speakers in the live condition were filmed with a separate video camera. The resulting video served as the stimulus in the video conditions, where it was projected to two new sets of addressees. The design thus yielded fixation data for exactly the same gestures in three conditions.

Based on the assumption that the pattern for live interaction is motivated by social norms for sustained eye contact, we hypothesised that the absence of a live interlocutor might lead to less time spent on the face, more fixations on gestures and other targets, as well as more scene scanning behaviour. However, we also hypothesised that the smaller display size might result in fewer gestures being fixated. The logic here is that with the reduced angles on a small screen, gesture detection would be possible even if the fixation marker remained on the speaker's face.

The socially motivated hypothesis was only minimally borne out. The results showed that the face dominated as a target overwhelmingly in all three conditions. The average viewing time on the face did drop on video, and more so in the small screen condition than the life-sized condition (from 96% live to 94% on life-sized and 91% on small screen video), but not significantly so. The number of gestures fixated decreased in both video conditions, although only significantly so on small screen video (from 7% of all gestures fixated live to 4.5% on life-sized and 3% on small screen video). Notice that this finding is in accordance (only) with the size-related hypothesis. With reduced viewing times on the face and fewer gesture fixations, the video conditions instead showed more fixations on body parts and empty space. This was partially predicted by the social hypothesis. The number of fixations of immobile body parts increased significantly in the small screen video condition (from 35% live and 32% on life-sized video to 57% on small screen video). Fixations on empty space increased significantly in the life-sized video condition (from 5% live and on small screen video to 21% on life-sized video). While there was a somewhat increased tendency for scene scanning in the small screen video condition, in principle, the typical saccade pattern outlined above for the live condition held in both video conditions as well. In sum, gaze behaviour in the live and life-sized video conditions was very similar overall. The small screen video condition showed the greatest number of differences from the live condition, even if most did not reach significance in this study. We interpret this as meaning that the reduced display size had a greater impact on general gaze behaviour than the absence of a live interlocutor.

Despite the overall reduction in fixation rate of gestures on video, by and large, the same gestures were fixated across the conditions. Specifically, the (proportional) attraction effect of holds and speaker-fixated gestures was maintained on video. In other words, holds and speaker-fixated gestures were fixated significantly more often than gestures without these features in all conditions. Importantly, speaker-fixated gestures were clearly affected by the absence of a live interlocutor. The number of fixations on speaker-fixated gestures was significantly reduced in both video conditions (from 23% live to 8% in both video conditions). This is consistent with the view that the speaker's gaze is essentially a social cue to joint attention, and that it is more powerful in a fully social setting than on video. There was also an overall if non-significant decrease in fixation of holds across the conditions (from 33% live to 20% on life-sized and 15% on small screen video). This decrease appears to be motivated by both social and size related factors, but these findings are more difficult to interpret. If holds attract fixations largely as a response to a sudden change in the visual field, then this reaction should be maintained even on video. These questions must be studied in more detail.

To summarise, in all conditions the face dominates as a fixation target and there is very little scanning behaviour of the rest of the scene. Gestures are only given a minimal amount of overt visual attention, and the same gestural features attract fixations across conditions. The social effects appear to be less powerful overall than size effects with the exception of the impact on speaker-fixation. The pattern outlined for the interactional situation is thus not exclusively socially conditioned. Gaze behaviour towards a human interlocutor on video is clearly more similar to behaviour towards an interlocutor live than to other types of video stimuli. This seems to reflect the particular status of the human face as an inherent focus of attention. This well-documented bias seems to have a biological basis in neural circuitry dedicated to face processing (for an overview, see Farah, 2000), and is manifest very early in infants' preference for faces as targets of attention (Valenza *et al.*, 1996).

Gestures, Fixations, and Information Uptake

Up until this point, gestures have mainly been considered as visuospatial phenomena. However, as shown in the introduction, gestures are also symbolic movements that carry meaning. The issue of whether addressees attend to and process gestural information is controversial in the field of gesture studies. Simplifying matters somewhat, it is sometimes argued that gestures have little real communicative value since addressees cannot reliably assign meaning to gestures in the absence of speech (e.g. Krauss *et al.*, 1991; Krauss *et al.*, 1996). However, there is a growing body of evidence showing that addressees do process gestural information (cf. Kendon, 1994). For instance, information expressed only in gestures re-surfaces in retellings, either as speech, as gesture, or both (Cassell *et al.*, 1999; McNeill *et al.*, 1994). Questions about the size and relative position of objects are better answered when gestures are part of the description than when gestures are absent (Beattie & Shovelton, 1999a; 1999b). Stroop-test designs also show cross-modal interference effects in the processing of gestural and spoken information (Langton & Bruce, 2000; Langton *et al.*, 1996). When subjects are shown static pictures of a person pointing either up, down, left, or right, and hear or see an incongruent corresponding word, their responses to words are affected by gestures, and responses to gestures by words. These data suggest that gesture processing is automatic and occurs in parallel to processing of speech.

Given that addressees direct very little overt visual attention to gestures in interaction, gestures must generally be attended to covertly. However, in contrast to Sign Language, we cannot automatically assume that non-fixated gestures are nonetheless attended to peripherally in the sense that the information encoded is processed and integrated into a representation of meaning. In order to investigate the division of labour between foveal and peripheral processing of gestural information, we need to study the uptake of information and the allocation of visual attention simultaneously. The question then arises whether there is any evidence that fixations or overt visual attention lead to better uptake of gestural information. Put differently, are any of the gesture fixations that do occur driven by the need to extract information, as often assumed in the non-technical literature?

The relationship between attention in a broad sense, information processing, and fixations is not a trivial one in complex naturalistic settings. Recent studies on change blindness (see e.g. the section on this topic in Hyönä, *et al.*, 2002) show that subjects occasionally fail to detect startling changes in a visual scene. Provided that their conscious attention is directed at a specific task, subjects fail to notice the switch of the main protagonist in a story or men in gorilla suits walking across the scene (e.g. Mack & Rock, 1998; Simons, 2000; Simons *et al.*, 2002; Simons & Levin, 1998). Complex scene perception is particularly challenging as the fixation marker is no straightforward indicator of what aspect of a scene is being attended to. As remarked by O'Regan *et al.*: "what an observer 'sees' at any moment in a scene is not the *location* he or she is directly fixating with the eyes, but the *aspect* of the scene he or she is currently attending to, that is, presumably, what he or she is processing with a view to encoding for storage into memory." (O'Regan *et al.*, 2000: 209). What aspects are relevant and what information is extracted (and retained) is still an empirical question (e.g. Aginsky & Tarr, 2000; Tatler, 2002).

Gullberg and Kita (forthcoming) investigated the relationship between addressees' information uptake from gestures and their fixations of gestures by exposing subjects to video clips of speakers retelling short stories and performing spontaneous gestures in face-to-face interaction in Dutch (Kita, 1996). Each video clip contained one relevant gesture (the target gesture) that was embedded among other spontaneous gestures. The target gestures represented motion of a protagonist left or right. This information (left or right direction) was only present in the target gesture and not in concurrent speech. The information could not be inferred from surrounding gestures. Each target gesture also displayed one of the features assumed to attract fixations: articulation in peripheral gesture space, hold, and speaker-fixation (see p. 692). Subjects watched four video clips of four different speakers retelling stories and gesturing in sequence. The videos were projected life-sized against a wall. Subjects' eye movements were recorded with the SMI iView head-mounted eye-tracker as outlined above. After watching the four videos, subjects answered questions about the target events by drawing pictures of the protagonists of the story. The data were coded for fixation on target gesture and for matched reply. A target gesture was coded as fixated if the fixation marker was immobile on the gesture for a minimum of 120 ms. As explained above, fixations on gestures were spatially unambiguous. Either a gesture was clearly fixated, or the fixation marker stayed on the speaker's face. A drawing was coded as a matched reply if the direction in the drawing matched the direction of the gesture as seen on video from the addressee perspective (see Figures 32.4 and 32.5).[3]

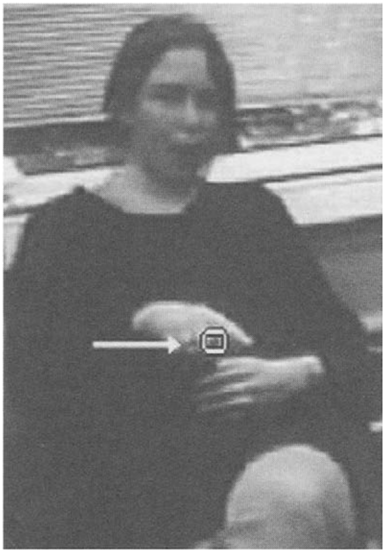

Figure 32.4: Example of a speaker performing a directional gesture that is also fixated by the addressee (= white circle).

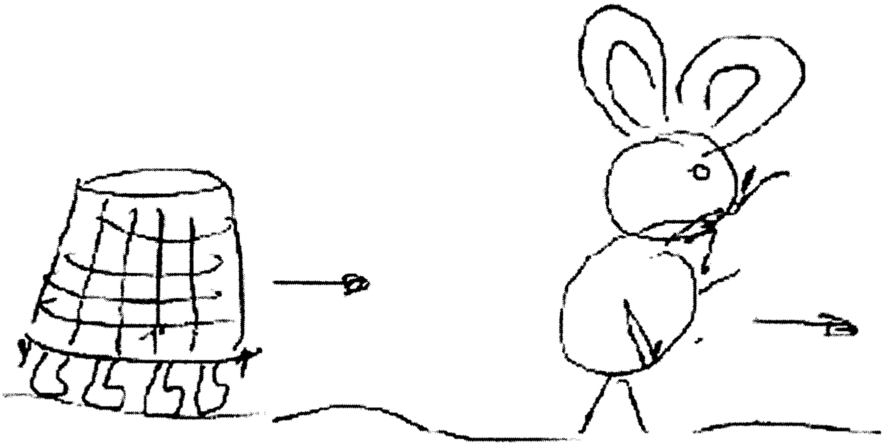

Figure 32.5: Example of the addressee's drawing matching the direction of the target gesture in Figure 32.4 as seen from the addressee-perspective.

As seen in the second section, gestures with holds and speaker-fixated gestures attracted fixations in 12.5% of the cases, respectively. However, addressees picked up the gestural information reliably above chance only from speaker-fixated gestures (82.5%), i.e. when speakers themselves had first looked at the gestures. In other words, addressees fixated some gestures (with hold) whose information they did not process. Conversely, addressees processed information from many gestures that they did not fixate (speaker-fixated), provided that speakers themselves had first look at the gestures. There was no significant difference in uptake for fixated vs. non-fixated gestures, and no effect of fixation duration.

The lack of information uptake above chance from the hold fixations could simply be due to the fact that the directional information is not present once the gesture has stopped moving. In contrast, the reliable uptake effect from speaker-fixated gestures even when these were *not* overtly fixated by the addressees is striking. This finding supports the claims that speakers' gaze shifts direct addressees' covert attention semi-automatically to the target of the gaze (cf. Langton & Bruce, 1999; Langton *et al.*, 2000). It also confirms that peripheral vision is sufficient to process gestural information. It does not, however, tell us whether gestures are fixated for the purpose of information extraction. In fact, the observation that uptake is not determined by direct fixations rather seems to suggest that overt fixations on speaker-fixated gestures are essentially socially motivated. The overt following of a speaker's gaze shift to the target of that gaze may be determined by social norms for joint attention. This finding is not per se in contradiction with statements to the effect that foveating gives an advantage to information extraction (e.g. Tatler, 2002). However, it does show that fine-grained information extraction is possible in complex situations even without direct fixation provided that covert attention has been directed to a specific target by speaker's gaze.

General Discussion and Conclusions

The studies reviewed in this chapter show that the human face overwhelmingly dominates as a target for overt visual attention in live interaction as well as on video. Gestures, in contrast, attract very few direct fixations both live and on video. They are fixated mainly if they momentarily cease to move as in gestural holds, or if speakers themselves have looked at them first. Holds are fixated quickly after their onset, suggesting a bottom-up response, whereas speaker-fixated gestures are fixated after a considerable delay, suggesting a top-down mechanism. Gestures thus appear to be fixated in their capacities as visuospatial entities, due to change in the visual field (holds), and possibly also for social reasons related to joint attention (speaker-fixated gestures). We know less about whether gestures are directly fixated for the purpose of extracting information. Most information processing appears to be covert or to be done in peripheral vision. At the very least, there is no simple relationship between fixations on a gesture and uptake of the gestural information in this complex setting.

Some of these results seem challenging in view of what is known about eye movements from studies in lab settings. However, they are not incompatible with recent findings from studies of eye movements in "conditions of competitive, parallel processing" (Tatler, 2002) as present in the real world. A number of studies have investigated subjects' eye movements as they perform natural tasks such as driving, making tea or sandwiches (Hayhoe, 2000; Land & Hayhoe, 2001; Land *et al.*, 1999; Shinoda *et al.*, 2001, Tatler, 2001), copying blocks (Pelz *et al.*, 2001), or drawing portraits (Tchalenko, this volume). In these activities the eye is typically directed to different areas according to the requirements of the current task, and many of the studies are concerned with eye-hand co-ordination. The results generally point in the direction of task and context sensitivity of fixation patterns. The studies reviewed above confirm the strong influence of the task, the activity type, and the context on gaze behaviour. They also highlight the contextual constraints on overt responses to attentional processes. Subjects do not automatically overtly respond to anything that catches their attention in interaction. Their responses are constrained by factors such as social norms for interaction, the status of the human face, and even kinaesthetic knowledge about body movements. Specifically, these factors interact in complex ways to influence behaviour. Finally, the results also suggest that the temporal resolution of overt behaviour is affected by the context. For instance, addressees' slow overt responses to speakers' gaze shifts are in stark contrast to the claims in the literature regarding the automatic effect on attention of such gaze shifts. Taken together, these issues form an important cluster to consider in discussions of the generalisability of findings from lab contexts to more complex, naturalistic settings.

Many issues need further investigation. Most pressing is perhaps the need to take the dynamic aspects of gesture performance into account when considering what triggers addressees' fixations on gestures. Despite the interactional perspective, we have treated gestures as individual events, isolated from each other in time and space, that can be fixated or not. However, since gestures are de facto linked to each other in gesture units that unfold over time (Kendon, 1972), a given gesture fixation is equally likely to be influenced by the number and the nature of preceding gestures as by the properties of any individual gesture. The influence of properties of speech should also be considered. For instance, noise in the speech signal or lowered comprehensibility of speech may also lead to increased attention to gestures. The assumption is that when the speech channel is compromised, addressees will rely more on gestures for decoding the message (Rimé *et al.*, 1988; Rogers, 1978). Other potentially influential factors include deictic expressions directing attention explicitly to gestures (e.g. "it was this big"), as well as interruptions and dysfluencies (Seyfeddinipur & Kita, 2001).

Many questions and unsolved puzzles thus remain. Hopefully the findings outlined here nonetheless show the importance of studying complex, dynamic and interactive contexts where inherent foci of interest (the human face), knowledge of the world (how bodies move), and social factors (principles of joint attention) conspire and influence behaviours at spatial and temporal levels alike. By considering such contexts, we hope to contribute to a multi-faceted picture of how visual attention works in the real world.

Acknowledgements

I thank Jukka Hyönä and an anonymous reviewer for helpful and constructive comments on an earlier version of this chapter. I also gratefully acknowledge the financial support of Birgit and Gad Rausing's Foundation for Research in the Humanities.

Notes

[1] A similar pattern has been suggested in American Sign Language (ASL) interaction. Signers' default visual focus in face-to-face signing is reported to be the face. Interestingly, however, the default locus of attention appears to be a region about the *lower* half or the chin region of the signer's face rather than the eye region (Corina *et al.*, 1996). This shift to the lower region of the signer's face could be a modification to enable foveal perception of oral grammatical components while simultaneously allowing for good sign perception in peripheral vision. For linguistic reasons, signers need to ensure that both motion detection and hand configuration details are adequately processed. Little is known, however, about actual fixation patterns during sign interaction.

[2] We are currently investigating the effect of the preceding movement vs. the hold itself by comparing fixation behaviour to the same gestures without and with artificially introduced holds.

[3] There is no evidence that addressees reversed the directions in the drawings in order to represent the direction as expressed from the speaker's viewpoint. Had addressees been reversing the viewpoints, we would have expected within-subject consistency of such reversals. There is no such consistency in the data, however.

References

Aginsky, V., & Tarr, M. J. (2000). How are different properties of a scene encoded in visual memory? *Visual Cognition*, *7*, 147–162.

Argyle, M., & Cook, M. (1976). *Gaze and Mutual Gaze*. Cambridge: Cambridge University Press.

Argyle, M., & Graham, J. A. (1976). The Central Europe experiment: Looking at persons and looking at things. *Journal of Environmental Psychology and Nonverbal Behavior*, *1*, 6–16.

Bavelier, D., Brozinsky, C., Tomann, A., Mitchell, T., Neville, H., & Liu, G. (2001). Impact of early deafness and early exposure to Sign Language on the cerebral organization for motion processing. *Journal of Neuroscience*, *21*, 8931–8942.

Beattie, G., & Shovelton, H. (1999a). Do iconic hand gestures really contribute anything to the semantic information conveyed by speech? *Semiotica*, *123*, 1–30.

Beattie, G., & Shovelton, H. (1999b). Mapping the range of information contained in the iconic hand gestures that accompany spontaneous speech. *Journal of Language and Social Psychology*, *18*, 438–462.

Braffort, A., Gherbi, R., Gibet, S., Richardson, J., & Teil, D. (eds) (1999). *Gesture-based Communication in Human–Computer Interaction*. (Vol. 1739). Berlin: Springer Verlag.

Cassell, J., McNeill, D., & McCullough, K.-E. (1999). Speech-gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information. *Pragmatics & Cognition*, *7*, 1–33.

Corina, D., Kritchevsky, M., & Bellugi, U. (1996). Visual language processing and unilateral neglect: Evidence from American Sign Language. *Cognitive Neuropsychology, 13*, 321–356.

Deák, G. O., Flom, R. A., & Pick, A. D. (2000). Effects of gesture and target on 12- and 18-month-olds' joint visual attention to objects in front of or behind them. *Developmental Psychology, 36*, 511–523.

Decety, J., & Grèzes, J. (1999). Neural mechanisms subserving the perception of human actions. *Trends in Cognitive Sciences, 3*, 172–178.

Doherty, M. J., & Anders, J. R. (1999). A new look at gaze: Pre-school children's understanding of eye-direction. *Cognitive Development, 14*, 549–571.

Driver, J., Davis, G., Ricciardelli, P., Kidd, P., Maxwell, E., & Baron-Cohen, S. (1999). Gaze perception triggers reflexive visuospatial orienting. *Visual Cognition, 6*, 509–540.

Farah, M. J. (2000). *The Cognitive Neuroscience of Vision*. Oxford: Blackwells.

Fehr, B. J., & Exline, R. V. (1987). Social visual interaction: A conceptual and literature review. In: A. W. Siegman and S. Feldstein (eds), *Nonverbal Behavior and Communication* (pp. 225–326). Hillsdale, NJ: Erlbaum.

Gibson, J. J., & Pick, A. D. (1963). Perception of another person's looking behavior. *American Journal of Psychology, 76*, 386–394.

Goldenberg, G. (2001). Imitation and matching of hand and finger postures. *NeuroImage, 14*, S132-S136.

Goodwin, C. (1986). Gestures as a resource for the organization of mutual orientation. *Semiotica, 62*, 29–49.

Grèzes, J., Costes, N., & Decety, J. (1999). The effects of learning and intention on the neural network involved in the perception of meaningless actions. *Brain, 122*, 1875–1887.

Gullberg, M. (1998). *Gesture as a Communication Strategy in Second Language Discourse. A Study of Learners of French and Swedish*. Lund: Lund University Press.

Gullberg, M., & Holmqvist, K. (1999). Keeping an eye on gestures: Visual perception of gestures in face-to-face communication. *Pragmatics & Cognition, 7*, 35–63.

Gullberg, M., & Holmqvist, K. (forthcoming). What speakers do and what listeners look at. Visual attention to gestures in face-to-face interaction and on video.

Gullberg, M., & Holmqvist, K. (2002). Visual attention towards gestures in face-to-face interaction vs. on screen. In: I. Wachsmuth and T. Sowa (eds), *Gesture and Sign Language based Human–Computer Interaction* (pp. 206–214). Berlin: Springer Verlag.

Gullberg, M., & Kita, S. (forthcoming). Attention to gestures. Information processing and fixations.

Hayhoe, M. (2000). Vision using routines: A functional account of vision. *Visual Cognition, 7*, 43–64.

Hermsdörfer, J., Goldenberg, G., Wachsmuth, C., Conrad, B., Ceballos-Baumann, O., Bartenstein, P., Schwaiger, M., & Boecker, H. (2001). Cortical correlates of gesture processing: Clues to the cerebral mechanisms underlying apraxia during the imitation of meaningless gestures. *NeuroImage, 14*, 149–161.

Hoffman, J. E. (1998). Visual attention and eye movements. In: H. Pashler (ed.), *Attention* (pp. 119–153). Hove: Psychology Press Ltd.

Hyönä, J., Muñoz, D., Heide, W., & Radach, R. (eds) (2002). *The Brain's Eye: Neurobiological and Clinical Aspects of Oculomotor Research*. Amsterdam: Elsevier.

Johnson, M. H. (1995). The development of visual attention: A cognitive neuroscience perspective. In: M. S. Gazzaniga (ed.), *The Cognitive Neurosciences* (pp. 735–747). Cambridge, MA: MIT Press.

Kendon, A. (1972). Some relationships between body motion and speech: An analysis of an example. In: A. W. Siegman and B. Pope (eds), *Studies in Dyadic Communication* (pp. 177–210). New York: Pergamon.

Kendon, A. (1980). Gesticulation and speech: Two aspects of the process of utterance. In: M. R. Key (ed.), *The Relationship of Verbal and Nonverbal Communication* (pp. 207–227). The Hague: Mouton.

Kendon, A. (1986). Some reasons for studying gesture. *Semiotica*, *62*, 3–28.

Kendon, A. (1990). *Conducting Interaction*. Cambridge: Cambridge University Press.

Kendon, A. (1993). Human gesture. In: K. R. Gibson and T. Ingold (eds), *Tools, Language and Cognition in Human Evolution* (pp. 43–62). Cambridge: Cambridge University Press.

Kendon, A. (1994). Do gestures communicate?: A review. *Research on Language and Social Interaction*, *27*, 175–200.

Kita, S. (1996). Listeners' up-take of gestural information. *MPI Annual Report, 1996*, 78.

Krauss, R. M., Chen, Y., & Chawla, P. (1996). Nonverbal behavior and nonverbal communication: What do conversational hand gestures tell us? *Advances in Experimental Social Psychology*, *28*, 389–450.

Krauss, R. M., Morrel-Samuels, P., & Colasante, C. (1991). Do conversational hand gestures communicate? *Journal of Personality and Social Psychology*, *61*, 743–754.

Land, M. F., & Hayhoe, M. (2001). In what ways do eye movements contribute to everyday activities. *Vision Research*, *41*, 3559–3565.

Land, M. F., Mennie, N., & Rusted, J. (1999). Eye movements and the roles of vision in activities of daily living: Making a cup of tea. *Perception*, *28*, 1311–1328.

Langton, S. R. H. (2000). The mutual influence of gaze and head orientation in the analysis of social attention direction. *Quarterly Journal of Experimental Psychology*, *53*, 825–845.

Langton, S. R. H., & Bruce, V. (1999). Reflexive visual orienting in response to the social attention of others. *Visual Cognition*, *6*, 541–567.

Langton, S. R. H., & Bruce, V. (2000). You must see the point: Automatic processing of cues to the direction of social attention. *Journal of Experimental Psychology: Human Perception and Performance*, *26*, 747–757.

Langton, S. R. H., O'Malley, C., & Bruce, V. (1996). Actions speak no louder than words: Symmetrical cross-modal interference effects in the processing of verbal and gestural information. *Journal of Experimental Psychology: Human Perception and Performance*, *22*, 1357–1375.

Langton, S. R. H., Watt, R. J., & Bruce, V. (2000). Do the eyes have it? Cues to the direction of social attention. *Trends in Cognitive Sciences*, *4*, 50–59.

Mack, A., & Rock, I. (1998). *Inattentional Blindness*. Cambridge, MA: MIT Press.

McNeill, D. (1992). *Hand and Mind. What the Hands Reveal about Thought*. Chicago: Chicago University Press.

McNeill, D. (1998). Speech and gesture integration. In: J. Iverson and S. Goldin-Meadows (eds), *The Nature and Functions of Gesture in Children's Communication* (pp. 11–27). San Francisco: Jossey-Bass.

McNeill, D., Cassell, J., & McCullough, K.-E. (1994). Communicative effects of speech mismatched gestures. *Research on Language and Social Interaction*, *27*, 223–237.

Moore, C., & Dunham, P. J. (eds) (1995). *Joint Attention*. Hillsdale, NJ: Erlbaum.

Neville, H. J., Coffe, S. A., Lawson, D. S., Fischer, A., Emmorey, K., & Bellugi, U. (1997). Neural systems mediating American Sign Language: Effects of sensory experience and age of acquisition. *Brain and Language*, *57*, 285–308.

Nobe, S., Hayamizu, S., Hasegawa, O., & Takahashi, H. (1998). Are listeners paying attention to the hand gestures of an anthropomorphic agent? An evaluation using a gaze tracking method. In: I. Wachsmuth and M. Fröhlich (eds), *Gesture and Sign Language in Human–Computer Interaction* (pp. 49–59). Berlin: Springer.

Nobe, S., Hayamizu, S., Hasegawa, O., & Takahashi, H. (2000). Hand gestures of an anthropomorphic agent: Listeners' eye fixation and comprehension. *Cognitive Studies. Bulletin of the Japanese Cognitive Science Society*, *7*, 86–92.

O'Regan, J. K., Deubel, H., Clark, J. J., & Rensink, R. A. (2000). Picture changes during blinks: Looking without seeing and seeing without looking. *Visual Cognition*, *7*, 191–211.

Peigneux, P., Salmon, E., van der Linden, M., Garraux, G., Aerts, J., Delfiore, G., Degueldre, C., Luxen, A., Orban, G., & Franck, G. (2000). The role of lateral occipitotemporal junction and area MT/V5 in the visual analysis of upper-limb postures. *NeuroImage*, *11*, 644–655.

Pelz, J., Hayhoe, M., & Loeber, R. (2001). The coordination of eye, head, and hand movements in a natural task. *Experimental Brain Research*, *139*, 266–277.

Perani, D., Fazio, F., Borghese, N. A., Tettamanti, M., Ferrari, S., Decety, J., & Gilardi, M. C. (2001). Different brain correlates for watching real and virtual hand actions. *NeuroImage*, *14*, 749–758.

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, *32*, 3–25.

Raymond, J. E. (2000). Attentional modulation of visual motion perception. *Trends in Cognitive Sciences*, *4*, 42–50.

Rettenbach, R., Diller, G., & Sireteanu, R. (1999). Do deaf people see better? Texture segmentation and visual search compensate in adult but not in juvenile subjects. *Journal of Cognitive Neuroscience*, *11*, 560–583.

Rimé, B., Boulanger, B., & d'Ydewalle, G. (1988). Visual attention to the communicator's nonverbal behavior as a function of the intelligibility of the message. Paper presented at the Symposium on TV Behavior, 24th International Congress of Psychology, Sydney, Australia, 28 August–2 September 1988.

Rizzolatti, G., Fogassi, L., & Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews Neuroscience*, *2*, 661–670.

Rogers, W. T. (1978). The contribution of kinesic illustrators toward the comprehension of verbal behavior within utterances. *Human Communication Research*, *5*, 54–62.

Seyfeddinipur, M., & Kita, S. (2001). Gesture and dysfluencies in speech. In: C. Cavé, I. Guaïtella and S. Santi (eds), *Oralité et Gestualité: Interactions et Comportements Multimodaux Dans la Communication* (pp. 266–270). Paris: L'Harmattan.

Shinoda, H., Hayhoe, M. M., & Shrivastava, A. (2001). What controls attention in natural environments? *Vision Research*, *41*, 3535–3545.

Simons, D. J. (2000). Attentional capture and inattentional blindness. *Trends in Cognitive Sciences*, *4*, 147–155.

Simons, D. J., Chabris, C. F., Schnur, T., & Levin, D. T. (2002). Evidence for preserved representations in change blindness. *Consciousness and Cognition*, *11*, 78–97.

Simons, D. J., & Levin, D. T. (1998). Failure to detect changes to people during real-world interaction. *Psychonomic Bulletin and Review*, *5*, 644–649.

Streeck, J. (1993). Gesture as communication I: Its coordination with gaze and speech. *Communication Monographs*, *60*, 275–299.

Streeck, J. (1994). Gesture as communication II: The audience as co-author. *Research on Language and Social Interaction*, *27*, 239–267.

Streeck, J., & Knapp, M. L. (1992). The interaction of visual and verbal features in human communication. In: F. Poyatos (ed.), *Advances in Nonverbal Communication: Interdisciplinary Approaches Through the Social and Clinical Sciences, Literature and the Arts* (pp. 3–23). Amsterdam: Benjamins.

Swisher, M. V. (1993). Perceptual and cognitive aspects of recognition of signs in peripheral vision. In: M. Marschark and M. D. Clark (eds), *Psychological Perspectives on Deafness* (pp. 209–228). Hillsdale: Erlbaum.

Tatler, B. W. (2001). Characterising the visual buffer: real-world evidence for overwriting early in each fixation. *Perception*, 30, 993–1006.

Tatler, B. W. (2002). What information survives saccades in the real world? In: J. Hyönä, D. Muñoz, W. Heide and R. R. (eds), *The Brain's Eyes: Neurobiological and Clinical Aspects of Oculomotor Research*. Amsterdam: Elsevier.

Tuite, K. (1993). The production of gesture. *Semiotica*, *93*, 83–105.

Valenza, E., Simion, F., Macchi Cassia, V., & Umilta, C. (1996). Face preference at birth. *Journal of Experimental Psychology: Human Perception and Performance*, *22*, 892–903.

Wachsmuth, I., & Sowa, T. (eds) (2002). *Gesture and Sign Language in Human–Computer Interaction*. (Vol. 2298). Berlin: Springer Verlag.

Yantis, S. (1998). Control of visual attention. In: H. Pashler (ed.), *Attention* (pp. 223–256). Hove: Psychology Press Ltd.