What speakers do and what addressees look at: Visual attention to gestures in human interaction live and on video

Gullberg, Marianne; Holmqvist, Kenneth

Published in:
Pragmatics & Cognition

DOI:
10.1075/pc.14.1.05gul

2006

What speakers do and what addressees look at. Visual attention to gestures in human interaction live and on video

Marianne Gullberg[1] and Kenneth Holmqvist[2]

[1] Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands

[2] Lund University Cognitive Science, Lund, Sweden

(Corresponding author)
Marianne Gullberg
Max Planck Institute for Psycholinguistics
PO Box 310
NL-6500 AH Nijmegen
The Netherlands

Email: marianne.gullberg@mpi.nl
http://www.mpi.nl/Members/MarianneGullberg

Kenneth Holmqvist
Lund University Cognitive Science
Kungshuset, Lundagård
S-222 22 Lund
Sweden

Email: kenneth@lucs.lu.se
http://www.lucs.lu.se/People/Kenneth.Holmqvist/

Acknowledgements

We gratefully acknowledge the support of Birgit and Gad Rausing's Foundation for Research in the Humanities through a grant to the first author, as well as financial and technical support from the Max Planck Institute for Psycholinguistics.

Biographical note

Marianne Gullberg is a Scientific Staff Member at the Max Planck Institute for Psycholinguistics, Nijmegen (NL). She heads a project on the processing of second and third languages, one part of which focuses on the production and comprehension of gestures in a cross-linguistic perspective. Together with Asli Özyürek she also co-ordinates the Nijmegen Gesture Centre.

Kenneth Holmqvist is Associate professor of Cognitive Science at Lund University (SE). His research interests involve eye-tracking and attention research in a wide range of applications.

Abstract

This study investigates whether addressees visually attend to speakers' gestures in interaction and whether attention is modulated by changes in social setting and display size. We compare a live face-to-face setting to two video conditions. In all conditions, the face dominates as a fixation target and only a minority of gestures draw fixations. The social and size parameters affect gaze mainly when combined and in the opposite direction from the predicted with fewer gestures fixated on video than live. Gestural holds and speakers' gaze at their own gestures reliably attract addressees' fixations in all conditions. The attraction force of Holds is unaffected by changes in social and size parameters, suggesting a bottom-up response, whereas speaker-fixated gestures draw significantly less attention in both video conditions, suggesting a social effect for overt gaze-following and visual joint attention. The study provides and validates a video-based paradigm enabling further experimental but ecologically valid explorations of cross-modal information processing.

Keywords: gesture, interaction, eye gaze, cross-modal information processing

Introduction

This paper addresses two seemingly simple questions: do addressees attend to speakers' gestures in interaction, and is attention to gestures modulated by variation in social setting and/or by variation in the physical properties of the visual display? Why study these questions?

Gestures — defined as the (mainly manual) movements speakers perform unwittingly while they speak as part of the expressive effort (Kendon 2004; McNeill 1992) — are an integral part of human communication, and of human face-to-face interaction. Gestures thus defined are symbolic movements that encode meaning in their direction, orientation, and shape. They form an integrated system with speech to which they are semantically and temporally linked (e.g. Kendon 1980; 2004; McNeill 1992; 1998). The meaning they encode is closely related to, but not necessarily identical with, that expressed in language and speech (Melinger and Levelt 2004; Slama-Cazacu 1976). The complementary distribution of information across speech and gesture has prompted a debate on the communicative value of gestures for addressees (e.g. Kendon 1994; Krauss, Chen and Chawla 1996). Disregarding deictic or pointing gestures (for overviews, see Kita 2003), there is growing evidence that addressees process information in representational gestures. These are gestures that iconically represent some aspect of what is being talked about such as shape or direction. For instance, information about events and objects expressed only in such gestures re-surfaces in retellings, either as speech, as gesture, or both (Cassell, McNeill and McCullough 1999; McNeill, Cassell and McCullough 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 and Shovelton 1999a, b), and addressees interpret indirect requests more accurately in the presence of gestures than in their absence (Kelly, Barr, Breckinridge Church and Lynch 1999). Stroop-test designs also show cross-modal interference effects in the processing of gestural and spoken information (Langton and Bruce 2000; Langton, O'Malley and Bruce 1996).

While the evidence that addressees process gesture information is therefore accumulating, it is less clear *how* gesture information comes to be integrated into representations

of meaning. Gesture information is by definition visual in nature and the integration of gestural information therefore calls for cross-modal information processing, i.e. for processing of information from different sources that are potentially competing for attentional resources. Gestures could compete for attention with the face in face-to-face interaction, both because they constitute a source of competing information, and because social rules for overt gaze allocation may favour the face.

As a starting point for a better understanding of how gestural information is integrated, this study examines the allocation of visual attention to gestures and specifically the competition between bottom-up, stimulus-related and top-down, pragmatic factors driving such allocation. We investigate viewers' or addressees' overt visual attention to speakers' gestures in interaction. In particular, we examine the effect of social setting and display size on addressees' eye movements. The variation in social setting is operationalised as the presence (=live) or absence (=video) of a real interlocutor, and the variation in display size is operationalised as life-sized video projection or video on a small screen. The establishment of what factors determine visual attention allocation in human interaction is a necessary first step in order to allow for rigorous experimental investigations of the relationship between fixation, information processing, and cognitive representations.

Theoretical background

Many disciplines have taken an interest in the visual perception of hands, signs of Sign Language, and gestures (e.g. Bavelier, Brozinsky, Tomann, Mitchell, Neville and Liu 2001; Decety and Grèzes 1999; Hermsdörfer, Goldenberg, Wachsmuth, Conrad, Ceballos-Baumann, Bartenstein, Schwaiger and Boecker 2001; Peigneux, Salmon, van der Linden, Garraux, Aerts, Delfiore, Degueldre, Luxen, Orban and Franck 2000; Perani, Fazio, Borghese, Tettamanti, Ferrari, Decety and Gilardi 2001; Rettenbach, Diller and Sireteanu 1999; Rizzolatti, Fogassi and Gallese 2001; Swisher 1993), as well as in gaze behaviour in interaction (for comprehensive overviews, see Argyle and Cook 1976; Fehr and Exline 1987; Kendon 1990; Kleinke 1986). Despite this widespread interest, we know surprisingly little about the attention afforded to

gestures in human interaction (but see Goodwin 1986; Kendon 1990; Streeck 1993; Streeck 1994; Streeck and Knapp 1992).

Recently, cross-modal information processing has drawn a lot of interest, i.e. the processing of multi-sensory information that results from producing and perceiving auditory, visual, tactile, and other types of information. A central notion in these studies is that of competition between information sources (e.g. Callan, Jones, Munhall, Callan, Kroos and Vatikiotis-Bateson 2003; Thompson, Malmberg, Goodell and Boring 2004; Vatikiotis-Bateson, Eigsti, Yano and Munhall 1998). As gestural information is visual, the properties of the visual system are crucial to understanding the relationship between visual attention to gestures and the processing and integration of the gestural information. The relationship can usefully be construed in terms of cross-modal competition, in particular the difference between foveal and peripheral vision. Optimal image quality with detailed texture and colour information is achieved if a target such as a gesture is directly fixated, i.e. if the eye is directed such that the image falls directly on the small central fovea. Outside of the fovea, parafoveal or peripheral vision gives much less detailed and fine-structured information (Bruce and Green 1985; Latham and Whitaker 1996). Fixating an information source therefore ensures the best information quality.

Vision research has posited two opposing principles that drive allocation of visual attention, i.e. the selection of what target to fixate next. Entities can draw fixations for stimulus-based, low-level perceptual reasons or for task-related, social pragmatic reasons connected to higher cognitive processes (Posner 1980; Yantis 1998). Examples of low-level factors include motion, abrupt onset, and contrast, whereas high-level factors include goals and intentions such as deliberate information retrieval. Agents' goals and properties of the scene are potentially competing factors and interact in determining 'gaze control' (Henderson 2003), as evidenced from findings of eye movement studies in real world settings other than interaction (Hayhoe 2000; Hayhoe and Ballard 2005; Land, Mennie and Rusted 1999; Land and Hayhoe 2001; O'Regan and Noë 2001; Shinoda, Hayhoe and Shrivastava 2001; Turano, Geruschat and Baker 2003).

In this perspective, attention to gestures in their natural habitat, face-to-face interaction, thus becomes an issue of processing under competition. There is competition between the

auditory (speech) and the visual (gesture) modalities. There is also competition within the visual modality. First, the visuo-spatial properties of gestures as movement in the peripheral visual field could make gestures a prime target for more mechanical, bottom-up selection of overt visual attention in the form of fixation (for overviews, see Hoffman 1998; Wolfe 1998; Yantis 1998). At the same time, however, gestures could potentially be in competition with the face, as a source of speech-related information. Gestures encode information related to ongoing speech, whereas the mouth provides detailed linguistic-phonetic information (e.g. Thompson, Malmberg, Goodell and Boring 2004; Vatikiotis-Bateson, Eigsti, Yano and Munhall 1998). Because gestures encode linguistic information, they could thus be targets of task-related gaze direction and attract fixations as part of addressees' strategic, goal-directed attempts to retrieve information. Finally, gestures could be in competition with the face for social, pragmatic reasons. The human face has been shown to draw a lot of attention both on static pictures (Yarbus 1967) and in face-to-face interaction (Fehr and Exline 1987; Kleinke 1986; Rutter 1984). In particular, *addressees* look more intently and continuously at the speaker's face than vice versa (e.g. Argyle and Cook 1976; Argyle and Graham 1976; Bavelas, Coates and Johnson 2002; Kendon 1967; 1973; 1990; Rutter 1984). It has been suggested that the face dominance reflects a socially and culturally determined norm for maintaining eye contact or mutual gaze in face-to-face interaction to signal attention, interest and engagement (Argyle and Cook 1976; Fehr and Exline 1987; Goodwin 1981; Kendon 1990; Kleinke 1986; Watson 1970). Under such a view, maintaining gaze on the face qualifies as a scripted schema for behaviour in interaction (Schank and Abelson 1977) or a task-related strategy for gaze control. Such a strategy could be related to 'scene-schema knowledge', i.e. generic semantic and spatial knowledge about objects and regularities in a specific scene such as human interaction (Henderson 2003). Such a strategy constitutes a top-down factor directing attention away from gestures and towards the face.

In sum, there is potential competition between different mechanisms governing visual attention to gestures: the tendency to attend to movement, the need to look at what you are seeking information about, and the social conventions that govern gaze away from gesture and towards the speaker's face.

Which gestures when and why

In a previous study of visual attention to gestures in face-to-face interaction addressees afforded very little direct attention to gestures (Gullberg and Holmqvist 1999). [1] Addressees fixated only a minority (9%) of gestures and instead mainly fixated the speaker's face, specifically the eye/nose bridge area. The gestures that did draw fixations were of two kinds. First, addressees tended to fixate gestures performed in the speaker's peripheral gesture space. A similar bias for fixating gestures in peripheral gesture space was found in studies of viewers looking at gestures performed by an anthropomorphic agent on a computer screen (Nobe, Hayamizu, Hasegawa and Takahashi 1998; 2000). A speaker's gesture space can be divided into central and peripheral gesture space (cf. McNeill 1992). Central space refers to a shallow disk 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 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 occur in the addressee's peripheral vision. However, gestures performed in peripheral gesture space will be projected even further away from the fovea if the addressee is fixating the speaker's face. Both Nobe et al. and Gullberg & Holmqvist therefore hypothesised that overt fixation on peripherally performed gestures was prompted by their occurrence in the addressee's extreme peripheral visual field.

[INSERT FIGURE 1]

Second, these studies also found that addressees were more likely to fixate gestures at which speakers themselves had looked, as illustrated in Figure 2. This finding is consistent with interactional claims about how speakers use gaze deictically to direct their addressee's attention to their gestures as a target of attention (Goodwin 1986; Streeck 1993; 1994; Streeck and Knapp 1992). It is also in accordance with more general claims about speakers' gaze as a cue to more or less automatic gaze-following and to joint attention (Deák, Flom and Pick 2000; Doherty and Anders 1999; Driver, Davis, Ricciardelli, Kidd, Maxwell and Baron-Cohen 1999; Gibson and Pick 1963; Langton 2000; Langton and Bruce 1999; Langton, O'Malley and Bruce 1996; Langton, Watt and Bruce 2000; Moore and Dunham 1995).

[INSERT FIGURE 2]

In addition, Nobe et al. (1998) found that gestures with so called holds drew fixations. A gestural hold is a momentary cessation of gestural movement in gesture space before the gesture proceeds (Kendon 1972). Nobe et al. found that gestures that cease to move and are held attract fixations significantly more often than gestures that move. They surmised that holds attract fixations because they occur when the gesture 'waits' for the relevant unit in speech to be completed. Why this should lead to fixation is unclear at this point.

The fixation patterns reported in these studies thus suggest that top-down and bottom-up processes may be in competition to allocate attention to gestures and to the face in interaction. For instance, the face dominance and the reduced number of gesture fixations found by Gullberg & Holmqvist could be motivated by a social norm for maintained mutual gaze in face-to-face interaction, a top-down socio-pragmatic strategy over-riding more low-level processes to look at movement. Conversely, the tendency to fixate peripheral gestures seems to be a stimulus-driven bottom-up response by the visual system to a challenge to peripheral vision. Finally, the tendency to fixate gestures that speakers themselves have looked at could be an automatic response to the gaze-direction of con-specifics, and as such, a bottom-up driven process.

Although the accounts for the patterns just outlined seem plausible, the findings do not elucidate how low-level stimulus-related and social-pragmatic top-down constraints interact and compete to modulate viewers' attention to gestures. When do top-down processes over-ride bottom-up processes, and vice versa? Moreover, some accounts seem contradictory. For instance, both Gullberg & Holmqvist and Nobe et al. found that gestures in peripheral space were fixated more than central gestures. Both also hypothesised that the reason was that the gestures challenged the viewer's peripheral vision. However, while peripheral vision may have been challenged in the live setting, given the distance in face-to-face interaction between the foveal fixation on the speaker's face and the gestures in peripheral gesture space, it seems less likely in the case of gestures showed on a computer-screen. The distance between the agent's face and its gestures should not have been great enough for the viewer's peripheral vision to be challenged in this setting.

In order to test which gestures addressees look at when and why, the constraints governing top-down and bottom-up processes need to be systematically varied. Two dimensions pertinent to visual behaviour can be manipulated: the social setting (top-down related effects) and the size of the visual display (bottom-up related effects). Differences in social setting can be operationalised as having a real speaker present or not. By showing speakers on video, the social rules of behaviour should be neutralised. If the general face dominance is socially driven, it should be sensitive to such manipulations. With top-down social constraints removed, fixation patterns may change and reveal more bottom-up driven eye movements. One direct consequence of such a change would be that more gestures are fixated when top-down constraints no longer over-ride bottom-up responses to movement in the visual field.

With regard to bottom-up processes, the visual system is sensitive to changes in angles and distances in the visual field. Changes in display size could therefore affect bottom-up driven fixations. Differences in display size are easily implemented on video. Overall, if a video setting leads to more gestures being fixated than a live setting, a *small* screen video display should lead to fewer gestures being fixated, since gesture detection by peripheral vision is facilitated. Moreover, if the tendency to fixate gestures in peripheral gesture space is indeed driven by challenges to peripheral vision, then this tendency should be modulated by differences in display size. Fixations of gestures in peripheral space should be less likely when projected on a small screen.

This study

This study investigates how differences in the social setting and display size may influence addressees' fixation behaviour towards human speakers and their gestures. We manipulate the social setting and display size by comparing visual behaviour towards speakers and their gestures in three exposure conditions, viz. live face-to-face, on life-sized video, and on video presented on a TV screen. We explore three things: 1) the amount of overt visual attention to the face across conditions; 2) the amount of overt visual attention to gestures, globally; and 3) the amount of overt visual attention to specific gestures. For the live condition, we expect to replicate

the general patterns found in the previous study of face-to-face interaction (Gullberg and Holmqvist 1999).

On the basis of the previous findings, we hypothesise the following:

If there is a social norm for maintaining eye contact in social interaction that governs overt visual attention toward the face and away from gestures, this norm will be neutralised on video. Addressees are therefore likely to spend less time looking at speakers' faces on video than live. Instead, they will spend more time looking at speakers' gestures and directly fixate more gestures on video than live.

We also expect addressees to look at fewer gestures on small-screen video than on life-sized video. With the reduced angles on a small screen, gesture detection should be possible even if the fixation marker remains on the speaker's face given that the gesture will be projected closer to the fovea.

With regard to which specific gestures draw fixations, we expect addressees to look at fewer gestures performed in peripheral gesture space on a small-scale video screen than on life-size video, relying on the same logic as above. We also expect addressees to look at gestures first looked at by speakers to an equal degree regardless of differences in the social setting or of display size. We are assuming that a shift in speaker gaze induces a more or less automatic shift of attention in addressees as suggested in some studies (e.g. Driver, Davis, Ricciardelli, Kidd, Maxwell and Baron-Cohen 1999; Langton and Bruce 1999). We have no predictions for gestural holds since they have only been found to attract fixations in the studies by Nobe et al. and there is no comparable data live. The predictions are summarised in Table 1.

[INSERT TABLE 1]

Method

Apparatus

We used the head-mounted as well as the remote set of the SMI iView© eye-tracker, which is 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

output data from the eye-tracker consist of a merged video recording showing the addressee's field of vision (i.e. the speaker), and an overlaid video recording of the addressee's fixations as a circle overlay (see Figure 3). 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. The fixation marker represents the foveal fixation and covers a visual angle of 2°. The output video data allow us to analyse both gesture and eye movements with a temporal accuracy of 40 ms.[2]

[INSERT FIGURE 3]

Participants

Participants were 60 native speakers of Swedish who were students at Lund University. They were recruited using advertisements on campus and were paid for their participation. They were all unacquainted before the experiment.

Stimuli and Procedure

The social setting could be Live, meaning that the observer was face-to-face with a live speaker; or Video, in which case the observer viewed a videotape of the speaker. The display size was either Life-sized, as in the Live condition or a life-sized video projection; or VideoScreen, in which case the video was displayed on a 28" TV screen. Three exposure conditions were thus created: the Live condition, the VideoLife-size condition, and the VideoScreen condition.[3]

Live condition. In the *Live* condition, 20 native speakers of Swedish were randomly assigned the role of speaker or addressee forming 10 pairs. In order to allow for spontaneous gesture production while maintaining control over the gestural content, a story telling task was used. Speakers memorised a printed cartoon and were then told to convey the story as well as they could to the addressees who would have to answer questions about it later. Addressees were instructed to make sure they understood the story, and were encouraged to ask questions and engage in the interaction. The instructions thus elicited conversational narratives and focused the addressees on the content of the story. Addressees were fitted with the head-mounted SMI iView© eye-tracker. They were calibrated using a nine-point matrix on the wall. After calibration,

speakers were introduced into the room and were seated 180 cm away from the speakers (measured back to back) facing them. They were given final instructions and then retold the story. The pairs were tested individually. The task generated natural narratives and a range of spontaneous gestures.

While retelling the stories to the addressees, the 10 speakers were simultaneously video recorded with a separate video camera placed behind the addressees. These 10 video recordings of the speakers served as stimuli in the two video conditions. This design allowed us to collect fixation data for the same gestures presented live and on video. In addition, the design ensured that the gestures shown on video were 'natural' since they were performed by speakers facing a live addressee; the gestures were thus not performed 'for the camera'.

VideoLife-size condition. In the *VideoLife-size* condition, 20 new Swedish addressees were shown video recordings of the original live speakers. Each new addressee saw one video recording, such that each video of an original speaker was viewed by two new recipients. The recordings were projected life-sized against a wall. The addressees were seated 180 cm from the wall. The SMI iView© remote set was placed between the addressee and the wall. The videotape contained a calibration screen that was first displayed before the actual video of the speaker began.

VideoScreen condition. In the *VideoScreen* condition, 20 new addressees were shown one video recording each on a TV screen. The addressees were seated 110 cm away from a 28" TV screen on which the video was projected. The projection size and the distance between addressees and screen decreased all angels by 52.3% of the original size. The SMI iView© remote set was placed between the addressee and the video screen. As in the VideoLife-size condition, the videotape contained a calibration screen.

The instructions to the addressees in the two video conditions were identical to the instructions in the live condition with the exception of encouraging questions, naturally.

Data Treatment

Eye movements

The eye movement data were retrieved from the digitised video output from the eye-tracker. Fixations were defined as instances where the fixation marker remained for at least 120 ms (=three video frames) directly on a fixated object (cf. Melcher and Kowler 2001). 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.

The eye movement data were coded for duration and the location of a fixation, i.e. the object fixated. Fixation targets include the face, left or right hand or arm gesturing, resting or immobile body parts, objects in the room. A number of fixations also rest on empty space. A final category, Other, only occurs in the video conditions. Fixations on Other represent the location on the video from which the voice of the original addressee comes. Fixations on this target are an artefact of the data collection set-up. They have been included descriptively in the interest of completeness, but have not been considered in the quantitative comparisons. Note that fixations on all objects including gestures are spatially unambiguous. In all cases of gesture fixation the entire fixation marker is clearly located directly on the hand or arm. The marker is not tangential to the gesturing body part and is not found near a gesture in progress.

Two measures are relevant to the study. First, the mean accumulated fixation time on various objects in the scene during the interaction gives a general overview of interaction as a scene type (cf. Buswell 1935; Chun 2000; Yarbus 1967). Second, the proportion of fixated gestures overall as well as the proportion of fixated gestures displaying a particular feature addresses more specific questions regarding the status of gestures as visuo-spatial objects and fixation targets in the scene type. In this latter case, we consider the distribution of the fixations on gestures to be binomial. Each gesture is either fixated or not.

Speech

Speech from the Live condition was transcribed verbatim and checked for the presence of demonstrative expressions referring directly or indirectly to the gestures, e.g. 'he held it like this'.

Such demonstrative expressions were considered undesirable as they function as a deictic device by which speakers can direct addressees' attention towards the gestures (cf. Nobe, Hayamizu, Hasegawa and Takahashi 1998; Streeck and Knapp 1992). No such deictic expressions were present in the data.

Gestures

The video output from the eye-tracker was digitised and coded in software for video annotation (Mediatagger 3.1, Brugman and Kita 1995). All gestures in the data were identified and coded for the three features that have been found to attract fixations in previous studies:

1. place of articulation in gesture space. We used McNeill's (1992) schema of gesture space that include areas like centre-centre, peripheral right, etc. All cases of centre-centre and centre were collapsed into one category. Similarly, all cases of peripheral were collapsed, leaving two broad categories Central and Peripheral, as shown in Figure 1.
2. speaker-fixation, i.e. whether or not speakers look at their own gestures.
3. presence vs. absence of hold, i.e. a momentary cessation of movement in a gesture (Kendon 1972; 1980). The data have been specifically coded for post-stroke holds, i.e. cessation of movement after the hand has reached the endpoint of a trajectory. Note that non-hold includes all other phases of the gesture phrase, i.e. preparation, stroke, or retraction (Kendon, 1980).

Validity and Reliability

A post test questionnaire was distributed to all subjects to ensure that gesture was not identified as the target of study. The questionnaire also contained questions regarding the experience with the eye-tracker, since the ecological validity of the data is of some concern, especially in the live condition. All subjects, speakers and addressees alike, declared that the equipment did not disturb them (cf. Gullberg and Holmqvist 1999). Speakers' speech and gestural behaviour did not differ quantitatively or qualitatively from data collected in an identical situation without eye-trackers (Gullberg 1998). Moreover, addressees' eye movement data in the live condition include

fixations of body parts that the subjects might have avoided to fixate had they been concerned about the equipment. We interpret this as meaning that the apparatus did not interfere with the addressees' natural behaviour.

All data sets were coded by two scorers. The eye movement data were coded by an expert (scorer 1) and by two student scorers with minimal scoring experience (scorers 2 and 3, collapsed into scorer A) to identify objects fixated. The inter-rater reliability for identification of fixated objects was 95.5%. Similarly, the gesture data were coded by an expert (scorer 1), and by two student scorers with minimal scoring experience (scorers 2 and 4, collapsed into scorer B). The inter-rater reliability for gesture identification was 93%, and for the three gesture features Location, Speaker-fixation, and Hold, 91%, 96%, and 91.5%, respectively.

Results

Accumulated fixation times and proportion of fixated gestures

The mean accumulated fixation time on objects in the scene gives an overall view of the general fixation patterns towards speakers and gestures in the setting. Figure 4 shows a typical scanpath plot from one of the VideoLife-size recordings. The plot shows not only locations of fixations but also an overview of accumulated viewing time. Each circle represents a fixation, and its diameter is proportional to the fixation time. The vast majority of circles, which are also generally very large, are centred over the speaker's face. An inspection of the video data indicates that the nose bridge and eye area attracted the most and longest fixations in the face. This location constitutes a sort of default location for visual attention in this setting (cf. Vatikiotis-Bateson, Eigsti, Yano and Munhall 1998 for similar findings). Excursions from this default location are seen as saccades with brief fixations on elements in the periphery of these circles. The video data reveal that these fixation sites correspond to gestures or to objects in the room around the speaker. Typically, there is no continuous scanning of the scene as a whole and virtually no cases of smooth pursuit of gestures. Saccades to landing sites outside the face are direct and accurate, despite the distances involved (cf. Land, Mennie and Rusted 1999). Typically, the eye moves from the face directly to a gesture in progress, then returns directly back to the default location. 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, Mennie and Rusted 1999). Typically, the eye moves from the face directly to a gesture in progress, stays briefly on this target, then returns directly back to the default location, i.e. the face.

[INSERT FIGURE 4]

Table 2 shows the average time the addressees spent looking at the face and at all other objects in the scene including gestures as a percentage of the overall recording time.

[INSERT TABLE 2]

In all three conditions, more than 90% of the time is spent on the face. Although somewhat less time is spent on the face in the video conditions, the difference between the conditions is not statistically significant ($F(2)=1.761$, $p=0.183$). Neither social setting nor size seems to influence the tendency to fixate the face.

Also, in all three conditions, less than 0.5% of the time is spent fixating gestures. Again, although less time is spent fixating gestures in the two video conditions than in the Live condition, the difference between the conditions is not statistically significant ($F(2)=0.692$, $p=0.506$).

The only significant difference in fixation time for different objects between the conditions is the increase in fixation time on the speakers' body parts, i.e. on parts other than the face and gestures, in the VideoScreen condition ($F(2)=3.504$, $p=0.038$). A post hoc LSD test shows that only the Life-size condition is significantly different from the VideoScreen condition ($p=0.02$). There appears to be an effect for size alone, but no individual effect for difference in the social setting.

[INSERT TABLE 3]

Next, we consider the proportion of fixated gestures across conditions (summarised in Table 3). Only a minority of all gestures are fixated overall, and fewer gestures are fixated in the two video conditions than live. However, only the difference between the Live and the VideoScreen condition is statistically significant ($\chi^2(1)=9.11$, $p=0.0025$), suggesting that only the combination of

a change in social setting and display size yields a significant decrease in gesture fixation. There is no independent effect of social setting or size.

The findings for the Live condition replicate the findings from the previous study of live interaction (Gullberg and Holmqvist 1999). The face dominates as a default locus of visual attention. Very little time is devoted to gestures, and only a small proportion of all gestures are fixated. Interestingly, the predictions regarding the effect of social setting and display size were not borne out. The face overwhelmingly dominates as a target in all three conditions with no measurable effect of variation in the social setting or the display size. Moreover, contrary to predictions, there was also no effect on the time spent on gestures. Only viewing times of immobile body parts were influenced, and only by the difference in display size. Addressees spent significantly more time looking at body parts in the VideoScreen condition than in the VideoLife-size condition. This size-related effect is somewhat surprising. An increase in fixations of body parts was expected to be mainly socially determined and therefore to be visible in the VideoLife-size condition where the social norm for overt gaze behaviour is neutralised. It seems unlikely that the increase in proportion of fixations be determined by projection size per se.

The proportion of fixated gestures in the Live condition closely matches the findings in Gullberg and Holmqvist (1999) (7% vs. 9%). As in the case of accumulated fixation time, there is no evidence that the variation in social setting or in display size individually affect addressees' tendency to fixate gestures. Only the combination of smaller screen size and lack of live interlocutor has an impact on fixation behaviour to gestures. Moreover, the effect goes in the opposite direction from the prediction with a *decrease* in proportion of gesture fixations, not an increase.

Distribution of fixations across gesture features

Next we analyse the effect of the three gestural features on fixation behaviour: location of articulation in gesture space (Central or Peripheral), presence/absence of post-stroke Hold, and presence/absence of Speaker-fixation on the gestures. Table 4 summarises the results. First, a comparison is made between fixations of gestures with vs. without the specific features. There is

no effect for Place of articulation (Central vs. Peripheral) in any condition, meaning that Peripheral gestures are not fixated more often than Central gestures (Live: $\chi^2(1)=0.02$, p=0.8875; VideoLife-size $\chi^2(1)=2.29$, p=0.1302; VideoScreen $\chi^2(1)=1.28$, p=0.2579). In contrast, in all conditions gestures with Holds (+Hold) are fixated significantly more often than gestures without Hold (-Hold) (Live: $\chi^2(1)=30.09$, p<0.0001; VideoLife-size $\chi^2(1)=44.79$, p<0.0001; VideoScreen $\chi^2(1)=23.37$, p<0.0001). Similarly, gestures that have been looked at by speakers (+Speaker-fixation) are fixated significantly more often than gestures that have not (-Speaker-fixation) in all conditions (Live: $\chi^2(1)=14.11$, p=0.0002; VideoLife-size $\chi^2(1)=4.34$, p=0.0372; VideoScreen $\chi^2(1)=12.47$, p=0.0004).

[INSERT TABLE 4]

Second, a comparison is made between proportions of fixations of gestures with the features +Hold and +Speaker-fixation, i.e. the features that draw fixation reliably, across the conditions. Although fewer Holds are fixated in the video conditions than Live, there is no effect for variation in social setting or display size ($\chi^2(2)=3.22$, p=0.1999). In contrast, Speaker-fixated gestures are affected by the variation in social setting such that fewer gestures are fixated on video (Live vs. VideoLife-size ($\chi^2(1)=5.2$, p=0.0226)). There is no effect for display size alone (VideoLife-size vs. VideoScreen ($\chi^2(1)=0.04$, p=0.8415)), but the combined absence of a live interlocutor and small display size also significantly reduces the number of fixated gestures with +Speaker-fixation (Live vs. VideoScreen ($\chi^2(1)=5.03$, p=0.0249)).

Despite the overall reduction in fixation rates of gestures on video, by and large, the same gestures were fixated across the exposure conditions. The tendency for addressees not to fixate gestures articulated in peripheral gesture space was unexpected given the results from the previous studies. However, the earlier findings could reflect the fact that gesture features tend to cluster in natural gesture production. For instance, gestures can simultaneously be articulated in peripheral gesture space and be held. Fixations interpreted as being caused by peripheral articulation could have been attracted by a gestural hold in the same gesture that went uncoded. In general, then, there is no evidence for the assumption that addressees fixate gestures projected in the extreme peripheral visual field because peripheral vision is challenged. Moreover,

the related prediction that addressees fixate fewer gestures articulated in peripheral gesture space when projected on a small screen was also not borne out.

The significantly greater tendency for speaker-fixated gestures to attract fixations in general does replicate earlier results. Contrary to predictions, however, fixations on these gestures are clearly affected by differences in the social setting alone, such that the absence of a live interlocutor leads to a decrease in gesture fixations.

Finally, the addressees' inclination to fixate gestural Holds, i.e. offset of gestural motion, confirms Nobe et al.'s results (Nobe, Hayamizu, Hasegawa and Takahashi 1998). Moreover, this result is unaffected by the social setting as well as by display size and even by the combination of these.

General Discussion and Conclusions

This study compared addressees' fixation behaviour towards speakers and their gestures in three conditions, Live, VideoLife-size, and VideoScreen, manipulating the social setting or the presence/absence of a live interlocutor, and the display size from life-sized to small screen. The results showed that the face dominates as a fixation target in all conditions and that only a minority of gestures attract fixations in all conditions, both measured in terms of total viewing time and in terms of proportion of fixated gestures. The socially motivated hypotheses were only minimally borne out. No independent effect could be found for social setting or size on the number of fixated gestures. Only when combined did these factors have an impact such that there was a significant *decrease* in the number of fixated gestures in the VideoScreen condition. Two gestural features reliably attracted fixations in all conditions, viz. gestural Holds and speakers' fixations of their own gestures (Speaker-fixation). The attraction force of Holds was unaffected by changes in the social setting and the display size. In contrast, the number of fixations on speaker-fixated gestures was significantly lower in both video conditions, suggesting a social effect for this feature. The overall decrease in amount of gesture fixation in the video conditions was mainly carried by this social effect.

The first main finding is that the face dominates as a target for addressees' visual attention regardless of variation in social setting and display size. The overall dominance suggests that the favoured status of the face in a live setting is not exclusively caused by a socio-cultural norm for maintained mutual gaze. Instead, the finding is more in line with accounts suggesting that the human face is a particular type of stimulus and a potentially biologically inherent focus of attention. Neonates' preference for faces and the existence of neural circuitry dedicated to face processing support this interpretation (for overviews, see Farah 2000; Valenza, Simion, Macchi Cassia and Umilta 1996). However, an alternative possibility is that the face dominance is a task-based effect related to the addressees' goals and intentions. The average viewing time on the face is somewhat greater than in studies made without eye-trackers. Argyle & Graham (1976), for instance, found that addressees who were engaged in conversation spent roughly 77% of the viewing time looking at their interlocutor in the absence of an interesting background or relevant surrounding objects. However, the task for our subjects was not conversation but to memorise a story well enough to retell it. This task and the general situation therefore called for greater attentiveness to the speaker, which is likely to have caused the very strong focus on the face in all conditions. This, in turn, may reflect an 'attentional control setting' (Folk and Remington 1998; Yantis 2000) to attend to the face as the main source of information, especially of linguistic information relevant to speech perception (Vatikiotis-Bateson, Eigsti, Yano and Munhall 1998). At this point, neither of the two accounts of the face dominance can be ruled out, and they need not be mutually exclusive.

The second main finding is that gestures draw very few fixations in general and that, contrary to the predictions, even fewer gestures are overtly attended to on (small) video than live. This observation raises obvious challenges to standard assumptions about attention allocation as a bottom-up driven response. The visual search paradigm typically stresses the importance of motion, abrupt onset, and size as factors that operate pre-attentively in a low-level, bottom-up fashion (e.g. Theeuwes, Atchley and Kramer 2000; Wolfe 1998). This is in contrast to the scene perception paradigms where goals and intentions are considered to be important allocation mechanisms (e.g. Henderson and Hollingworth 1999; Klein and Shore 2000; Theeuwes 1994;

Yantis 1998; 2000). The view that unspecified motion functions as a pre-attentive attraction factor is challenged in two ways. First, the onset of gestural movement clearly does not attract fixation per se in this setting since the vast majority of gestures are not fixated at all. It is possible that the movement of an inalienable body part is not salient enough to draw overt attention in this context, if by salient we mean different from the surrounding context on some relevant parameter. Addressees' knowledge of the human brachial and manual motor patterns renders gestures fairly predictable. Moreover, since gestural movement is pervasive in interaction, gestures may be considered to be part of the visual "background elements" of the scene (cf. Henderson and Hollingworth 1999). Gestures are a kind of 'visual noise' of constant motion, and as such, they do not draw much overt visual attention. Gestural motion would thus appear to be qualitatively different from motion in general. The observation that Holds, i.e. cessation of gestural motion, attract fixations in all conditions fits with such a view. The cessation of gestural motion could be a low-level, bottom-up, reason for gesture fixation. If gestural movement is visual noise, then Holds represent 'sudden change' in the visual field in terms of sudden offset of motion and could evoke fixations for this reason.[4,5] We return to this issue below.

There is no evidence in this study for the assumption that the tendency not to fixate gestural movement is socially motivated. However, the effect could still be task-specific and top-down driven. Vision studies have shown that if motion is irrelevant to a given task, it does not necessarily attract fixation. Attention capture is generally dependent on attentional control settings, such that stimuli in natural environments become relevant "by virtue of their role in ongoing behavior" (Pelz, Hayhoe and Loeber 2001: 266) and not necessarily or exclusively as a function of their physiological parameters (Folk and Remington 1998; Folk, Remington and Johnston 1992; Ludwig and Gilchrist 2002; Raymond 2000).

The findings that the face dominates visual attention and that few gestures are fixated regardless of setting or display size are both in direct opposition to results from Nobe et al.'s studies. In these studies viewers who watched gestures performed by an anthropomorphic agent on a computer-screen fixated the vast majority of the agent's gestures (70-75%) (Nobe, Hayamizu, Hasegawa and Takahashi 1998; 2000). The discrepancy between their results and the

findings in Gullberg & Holmqvist (1999) could have been explained by the difference in setting and size, as their studies took place in a non-social, small-screen video setting. However, the findings from this study leave their results unexplained in terms of social setting or size. Explanations for the discrepancy must instead be sought in one of the other differences between their and our studies. A plausible candidate is the difference between human agents, as in this study, and non-human anthropomorphic agents, as in theirs. Recent studies have suggested that visual processing of human and virtual reality hands activate somewhat different brain areas (Perani, Fazio, Borghese, Tettamanti, Ferrari, Decety and Gilardi 2001). This difference could carry over to actual eye movement behaviour.

A third major finding is that different gestures are fixated for different reasons and are therefore also differentially affected by the variation in social setting and display size. We have already suggested above that Holds may be fixated in a bottom-up fashion because they represent sudden change in the visual field. Nobe et al. suggested that Holds are fixated because they occur in cases of asynchrony between speech and gesture (Kendon 1980; Kita 1990; McNeill, Levy and Pedelty 1990). Their assumption was that addressees consider cases where gesture 'waits' for speech as 'relevant' in some sense. The precise relevance of this waiting is unclear, however. A vision related bottom-up account seems more plausible. In addition to representing sudden change, Holds constitute a challenge to peripheral vision. If addressees mainly fixate the speaker's face, they process gestures in peripheral vision, which is good at motion detection (cf. Swisher 1990). However, Holds, which by definition do not move, would not be processable by peripheral vision and so would lead to fixation if addressees wanted to extract information. Under this view, fewer Holds should be fixated on small screen than life-sized. At this point, it is not clear whether such an account is compatible with the fact that fixations to holds are not sensitive to variation in display size. The attraction force of Holds clearly needs further research.

Turning to Speaker-fixated gestures, the social effect on fixations of such gestures was unexpected. The finding challenges claims regarding automatic shifts of attention to the target of speakers' gaze (Driver, Davis, Ricciardelli, Kidd, Maxwell and Baron-Cohen 1999; Langton and

Bruce 1999; Langton and Bruce 2000; Langton, Watt and Bruce 2000). The current findings suggest that, while speakers' gaze may trigger automatic shifts in covert attention, *overt* attention shift or gaze following is not automatic. Note that even if speaker-fixated gestures tend to attract fixations to a greater extent than other gestures, it is nevertheless only a minority of these gestures that are fixated. Even in the live condition, only 23% of all speaker-fixations lead to overt addressee-fixations on gestures. In combination with the clear social effect, this finding indicates that overt following of the speaker's gaze in human interaction is a social, and very likely a strategic top-down phenomenon. Not to co-fixate a speaker-fixated gesture in a live condition would be socially inept. It is in fact common for speakers who fixate their own gestures to look back up on the addressee to ensure that joint attention has indeed been established. In a video condition, there is no such social pressure to follow a Speaker-fixation. This dissociation between overt and covert attention to the gaze direction of others, and the finding that overt gaze following is a social phenomenon, is an important modification of claims regarding gaze and joint attention (e.g. Langton, Watt and Bruce 2000; Tomasello 1995).

What are the methodological and theoretical implications of these findings? Although the predictions regarding social and size related effects were not borne out, both factors did affect behaviour but in other ways than predicted. The comparison of behaviours across exposure conditions showed a tendency for the Live and VideoLife-size conditions to be more similar than either the Live and VideoScreen or the VideoLife-size vs. VideoScreen conditions. Put differently, the VideoScreen condition differed from the other two such that display size appeared to interact with the social parameter. We interpret this as meaning that when the stimulus is a human agent 'addressing' the addressee directly, addressees tend to treat speakers on life-sized screens as live interlocutors and behave towards them accordingly, even to the point of avoiding socially unacceptable fixations sites such as body parts. Only the combination of small display size and absent interlocutor seems to engender a sufficient feeling of social distance from the interlocutor to fully neutralise social norms for behaviour. Support for such an interpretation comes from studies of responses to television projections. Lombard (1995) found that large-screen viewing yielded more positive emotional responses to and impressions of speakers. This is presumably

because viewers respond as if to real interlocutors. VideoLife-size projection thus seems to emulate live face-to-face interaction. This is of crucial methodological importance as it suggests that life-sized projections can successfully be used to study attention allocation to human agents, thereby allowing rigorous experimental set-ups. A related observation is the difference between our small VideoScreen condition and the data in the Nobe et al. studies. Our VideoScreen condition generated the fewest number of gesture fixations, in stark contrast to their massive fixation rate of gestures. If the critical difference is indeed that between a human interlocutor vs. an anthropomorphic agent, this is another important methodological point. If the task is to examine human visual behaviour towards humans, then humanoids as stimuli will not do since addressees behave differently towards them.

The theoretical ramifications of this methodological step are crucial. In order to examine issues of cross-modal information processing in general, and integration of gesture information in particular, it is necessary to clarify the relationship between fixation and cognitive representations. A range of issues needs investigation. First, since gesture information partially overlaps with speech, gestures may be attended to differently from other sources of information. Second, peripheral vision may be sufficient to detect most gesture information, meaning that it may not be necessary for addressees to directly fixate gestures to pick up their meaning. The limitations on peripheral information extraction therefore need to be examined in conjunction with systematic variation in information overlap between speech and gesture. Third, visual fixation and visual attention need not coincide. Fixations are physiological events, but attention is a cognitive phenomenon. It is possible for perceivers to dissociate the locus of visual fixation from the locus of their visual attention, as evidenced by the notions of overt and covert attention. You can fixate a visual target without attending to it ("looking without seeing"), and conversely, attend to something without directly fixating it ("seeing without looking"). The dissociation between overt and covert attention may be more important in interaction than in other settings because of socio-pragmatic constraints and interact with the two preceding points. Finally, in this study we have treated gestures as individual events, isolated from each other in time and space, that can be fixated or not. However, gestures are de facto linked to each other in gesture units that unfold

over time, and a given gesture fixation is as likely to be influenced by the nature of the preceding gesture as by the properties of any individual on-going gesture. All of these issues call for an experimental paradigm where identical but natural human gesture stimuli are presented on video. This study has provided validation of such a video-based approach that will ensure that the transition between a live situation and a video presentation of gestures does not distort the phenomena studied.

In conclusion, this study has shown that visual attention to human actors in interaction is constrained by a complex inter-dependence between social norms, the status of the human face, and knowledge about human motor patterns. In answer to the original questions, addressees *do* visually attend to speakers' gestures in interaction, but only to a small number of them, and this attention *is* modulated by social and size related factors, but in unexpected ways. Visual attention to gestures is not guided by movement, nor by the location of gestures in gesture space. However, different gestures are fixated for different reasons: gestures that speakers themselves have fixated (Speaker-fixated gestures) are looked at for social, top-down related reasons, and gestures that stop moving (Holds) are fixated for reasons related to the inner workings of the visual system. Fewer gestures are fixated on video than live, but the transition mainly affects gestures that draw fixations for social reasons. Whether or not human interlocutors are displayed on video or perceived 'live' potentially alters visual behaviour towards them radically and thereby the possible generalisations regarding behaviour. This study has provided a paradigm that allows human interaction to be emulated in a controlled and rigorous setting, thereby opening up the field for further explorations of cross-modal information processing under ecologically valid conditions.

Notes

[1] The terms *addressee* and *speaker* will be used throughout this paper as convenient shorthand to refer to viewers (addressees) and their interlocutors (speakers). In the experiments *addressees* are thus the subjects wearing the eye-tracker and the *speakers* are the stimuli.

[2] Note that although the sampling rate of the eye-tracker is 50 Hz, the video output reduces the temporal granularity of the data to 40 ms, given a video frame rate of 25 frames/second.

[3] The fourth logical possibility, a live condition with small display size, could obviously not be accommodated. Note that a live condition with a speaker situated at a greater distance would have introduced another set of variables. Speakers accommodate to distance in both gesture and speech, and we would therefore not have been studying the same phenomenon had we undertaken this manipulation.

[4] This issue could be elucidated by latency data for fixation onsets in relation to onset of Hold. However, the latency data from the naturally occurring gestures in this study are less than ideal to investigate the role of individual gesture features on fixation since features tend to cluster. In other words, each individual gesture can and often does display more than one feature, e.g. a combination of Hold and performance in peripheral gesture space. When fixated gestures displaying only one feature are considered in this data set, the number of data points is very small. Nevertheless, in this set, the median fixation onset time after Hold onset across all conditions is 440ms. The median fixation onset time after Speaker-fixation onset across all conditions is 1120 ms. Although no statistical analysis can be performed on this set, the numbers do suggest that fixations of Holds are qualitatively different from fixations of Speaker-fixated gestures. We are currently investigating this issue in more detail in a data set where gesture features are controlled for.

[5] At this point, we cannot exclude the possibility that it is the preceding gestural movement that attracts attention and initiates saccadic planning. We would however make a plausibility argument against this assumption. Had it been the preceding movement, more gestures without holds

would surely have attracted fixations. However, it is an empirical question. We are currently investigating the effect of the hold itself vs. the preceding movement by introducing artificial holds in video clips of gestures.

[i] As this landing site is an artefact of the stimulus design, the numbers for this category are not analysed further.

References

Argyle, M. and Cook, M. 1976. *Gaze and mutual gaze.* Cambridge: Cambridge University Press.

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

Bavelas, J. B., Coates, L. and Johnson, T. 2002. "Listener responses as a collaborative process: The role of gaze". *Journal of Communication* September 2002, 566-580.

Bavelier, D., Brozinsky, C., Tomann, A., Mitchell, T., Neville, H. and Liu, G. 2001. "Impact of early deafness and early exposure to Sign Language on the cerebral organization for motion processing". *Journal of Neuroscience* 21(22): 8931-8942.

Beattie, G. and Shovelton, H. 1999a. "Do iconic hand gestures really contribute anything to the semantic information conveyed by speech?" *Semiotica* 123(1/2): 1-30.

Beattie, G. and 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(4): 438-462.

Bruce, V. and Green, P. 1985. *Visual perception. Physiology, psychology and ecology.* Hillsdale, NJ: Erlbaum.

Brugman, H. and Kita, S. 1995. "Impact of digital video technology on transcription: a case of spontaneous gesture transcription". *KODIKAS/CODE: Ars Semeiotica An international journal of semiotics* 18: 95-112.

Buswell, G. T. 1935. *How people look at pictures.* Chicago: University of Chicago Press.

Callan, D. E., Jones, J. A., Munhall, K., Callan, A. M., Kroos, C. and Vatikiotis-Bateson, E. 2003. "Neural processes underlying perceptual enhancement by visual speech gestures". *Neuroreport* 14(17): 2213-2218.

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

Chun, M. M. 2000. "Contextual cueing of visual attention". *Trends in Cognitive Sciences* 4(5): 170-178.

Deák, G. O., Flom, R. A. and 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(4): 511-523.

Decety, J. and Grèzes, J. 1999. "Neural mechanisms subserving the perception of human actions". *Trends in Cognitive Sciences* 3(5): 172-178.

Doherty, M. J. and 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. and Baron-Cohen, S. 1999. "Gaze perception triggers reflexive visuospatial orienting". *Visual Cognition* 6(5): 509-540.

Farah, M. J. 2000. *The cognitive neuroscience of vision.* Oxford: Blackwells.

Fehr, B. J. and Exline, R. V. 1987. "Social visual interaction: A conceptual and literature review". In A. W. Siegman and S. Feldstein (eds), *Nonverbal behavior and communication.* Hillsdale, NJ: Erlbaum, 225-326.

Folk, C. L. and Remington, R. 1998. "Selectivity in distraction by irrelevant featural singletons: Evidence for two forms of attentional capture". *Journal of Experimental Psychology: Human Perception and Performance* 24(3): 847-858.

Folk, C. L., Remington, R. W. and Johnston, J. C. 1992. "Involuntary covert orienting is contingent on attentional control settings". *Journal of Experimental Psychology: Human Perception and Performance* 18(4): 1030-1044.

Gibson, J. J. and Pick, A. D. 1963. "Perception of another person's looking behavior". *American Journal of Psychology* 76(3): 386-394.

Goodwin, C. 1981. *Conversational organisation: Interaction between speakers and hearers.* New York: Academic Press.

Goodwin, C. 1986. "Gestures as a resource for the organization of mutual orientation". *Semiotica* 62(1/2): 29-49.

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. and Holmqvist, K. 1999. "Keeping an eye on gestures: Visual perception of gestures in face-to-face communication". *Pragmatics & Cognition* 7(1): 35-63.

Hayhoe, M. 2000. "Vision using routines: A functional account of vision". *Visual Cognition* 7(1/2/3): 43-64.

Hayhoe, M. and Ballard, D. H. 2005. "Eye movements in natural behavior". *Trends in Cognitive Sciences* 9(4): 188-194.

Henderson, J. M. 2003. "Human gaze control during real-world scene perception". *Trends in Cognitive Sciences* 7(11): 498-504.

Henderson, J. M. and Hollingworth, A. 1999. "High-level scene perception". *Annual Review of Psychology* 50: 243-271.

Hermsdörfer, J., Goldenberg, G., Wachsmuth, C., Conrad, B., Ceballos-Baumann, O., Bartenstein, P., Schwaiger, M. and 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.* Hove: Psychology Press Ltd, 119-153.

Kelly, S. D., Barr, D. J., Breckinridge Church, R. and Lynch, K. 1999. "Offering a hand to pragmatic understanding: The role of speech and gesture in comprehension and memory". *Journal of Memory and Language* 40(4): 577-592.

Kendon, A. 1967. "Some functions of gaze direction in two-person conversations". *Acta Psychologica* 26: 22-63.

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.* New York: Pergamon, 177-210.

Kendon, A. 1973. "The role of visible behaviour in the organization of social interaction". In M. von Cranach and I. Vine (eds), *Social communication and movement. Studies of interaction and expression in man and chimpanzee.* New York: Academic Press, 29-74.

Kendon, A. 1980. "Gesticulation and speech: Two aspects of the process of utterance". In M. R. Key (eds), *The relationship of verbal and nonverbal communication.* The Hague: Mouton, 207-227.

Kendon, A. 1990. *Conducting interaction.* Cambridge: Cambridge University Press.

Kendon, A. 1994. "Do gestures communicate?: A review". *Research on Language and Social Interaction* 27(3): 175-200.

Kendon, A. 2004. *Gesture. Visible action as utterance.* Cambridge: Cambridge University Press.

Kita, S. (ed). 2003. *Pointing: Where language, culture, and cognition meet.* Mahwah, NJ: Erlbaum.

Klein, R. M. and Shore, D. I. 2000. "Relations among modes of visual orienting". In S. Monsell and J. Driver (eds), *Attention and performance XVIII.* Cambridge, MA: MIT Press, 195-208.

Kleinke, C. L. 1986. "Gaze and eye contact: A research review". *Psychological Bulletin* 100(1): 78-100.

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

Land, M., Mennie, N. and Rusted, J. 1999. "The roles of vision and eye movements in the control of activities of daily living". *Perception* 28(11): 1311-1328.

Land, M. F. and Hayhoe, M. 2001. "In what ways do eye movements contribute to everyday activities". *Vision Research* 41(25-26): 3559-3565.

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(3): 825-845.

Langton, S. R. H. and Bruce, V. 1999. "Reflexive visual orienting in response to the social attention of others". *Visual Cognition* 6(5): 541-567.

Langton, S. R. H. and 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(2): 747-757.

Langton, S. R. H., O'Malley, C. and 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(6): 1357-1375.

Langton, S. R. H., Watt, R. J. and Bruce, V. 2000. "Do the eyes have it? Cues to the direction of social attention". *Trends in Cognitive Sciences* 4(2): 50-59.

Latham, k. and Whitaker, D. 1996. "A comparison of word recognition and reading performance in foveal and peripheral vision". *Vision Research* 37: 2665-2674.

Lombard, M. 1995. "Direct responses to people on the screen: Television and personal space". *Communication Research* 22(3): 288-324.

Ludwig, C. J. H. and Gilchrist, I. D. 2002. "Stimulus-driven and goal-driven control over visual selection". *Journal of Experimental Psychology: Human Perception and Performance* 28(4): 902-912.

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. M. Iverson and S. Goldin-Meadow (eds), *The nature and functions of gesture in children's communication.* San Francisco: Jossey-Bass, 11-27.

McNeill, D., Cassell, J. and McCullough, K.-E. 1994. "Communicative effects of speech mismatched gestures". *Research on Language and Social Interaction* 27(3): 223-237.

McNeill, D., Levy, E. T. and Pedelty, L. L. 1990. "Speech and gesture". In G. R. Hammond (eds), *Cerebral control of speech and limb movements.* Amsterdam: North Holland, 203-256.

Melcher, D. and Kowler, E. 2001. "Visual scene memory and the guidance of saccadic eye movements". *Vision Research* 41: 3597-3611.

Melinger, A. and Levelt, W. J. M. 2004. "Gesture and the communicative intention of the speaker". *Gesture* 4(2): 119-141.

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

Nobe, S., Hayamizu, S., Hasegawa, O. and 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.* Berlin: Springer, 49-59.

Nobe, S., Hayamizu, S., Hasegawa, O. and Takahashi, H. 2000. "Hand gestures of an anthropomorphic agent: Listeners' eye fixation and comprehension". *Cognitive Studies. Bulletin of the Japanese Cognitive Science Society* 7(1): 86-92.

O'Regan, J. K. and Noë, A. 2001. "A sensorimotor account of vision and visual consciousness". *Behavioral and Brain Sciences* 24(5): 939-1031.

Peigneux, P., Salmon, E., van der Linden, M., Garraux, G., Aerts, J., Delfiore, G., Degueldre, C., Luxen, A., Orban, G. and 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. and 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. and 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(1): 3-25.

Raymond, J. E. 2000. "Attentional modulation of visual motion perception". *Trends in Cognitive Sciences* 4(2): 42-50.

Rettenbach, R., Diller, G. and 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(5): 560-583.

Rizzolatti, G., Fogassi, L. and Gallese, V. 2001. "Neurophysiological mechanisms underlying the understanding and imitation of action". *Nature Reviews Neuroscience* 2: 661-670.

Rutter, D. R. 1984. *Looking and seeing: The role of visual communication in social interaction.* Chichester: Wileys.

Schank, R. S. and Abelson, R. P. 1977. *Scripts, plans, goals and understanding: An inquiry into human knowledge structures.* Hillsdale, N.J.: Erlbaum.

Shinoda, H., Hayhoe, M. M. and Shrivastava, A. 2001. "What controls attention in natural environments?" *Vision Research* 41(25-26): 3535-3545.

Slama-Cazacu, T. 1976. "Nonverbal components in message sequence: "Mixed syntax"". In W. C. McCormack and S. A. Wurm (eds), *Language and man: Anthropological issues.* The Hague: Mouton, 217-227.

Streeck, J. 1993. "Gesture as communication I: Its coordination with gaze and speech". *Communication Monographs* 60(4): 275-299.

Streeck, J. 1994. "Gesture as communication II: The audience as co-author". *Research on Language and Social Interaction* 27(3): 239-267.

Streeck, J. and 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.* Amsterdam: Benjamins, 3-23.

Swisher, M. V. 1990. "Developmental effects on the reception of signs in peripheral vision". *Sign Language Studies* 66: 45-60.

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.* Hillsdale: Erlbaum, 209-228.

Theeuwes, J. 1994. "Endogenous and exogenous control of visual selection". *Perception* 23(4): 429-440.

Theeuwes, J., Atchley, P. and Kramer, A. F. 2000. "On the time course of top-down and bottom-up control of visual attention". In S. Monsell and J. Driver (eds), *Attention and performance XVIII.* Cambridge, MA: MIT Press, 105-124.

Thompson, L. A., Malmberg, J., Goodell, N. K. and Boring, R. L. 2004. "The distribution of attention across a talker's face". *Discourse Processes* 38(1): 145-168.

Tomasello, M. 1995. "Joint attention as social cognition". In C. Moore and P. J. Dunham (eds), *Joint attention.* Hillsdale: Erlbaum, 103-130.

Turano, K. A., Geruschat, D. R. and Baker, F. H. 2003. "Oculomotor strategies for the direction of gaze tested with a real-world activity". *Vision Research* 43(3): 333-346.

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

Vatikiotis-Bateson, E., Eigsti, I.-M., Yano, S. and Munhall, K. G. 1998. "Eye movement of perceivers during audiovisual speech perception". *Perception and Psychophysics* 60(6): 926-940.

Watson, O. M. 1970. *Proxemic behavior: A cross-cultural study.* The Hague: Mouton.

Wolfe, J. M. 1998. "Visual search". In H. Pashler (ed), *Attention.* Hove: Psychology Press Ltd, 13-73.

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

Yantis, S. 2000. "Goal-directed and stimulus-driven determinants of attentional control". In S. Monsell and J. Driver (eds), *Attention and performance XVIII.* Cambridge, MA: MIT Press, 73-103.

Yarbus, A. 1967. *Eye movements and vision.* New York: Plenum Press.

Figure 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 white circle at its default location in interaction.

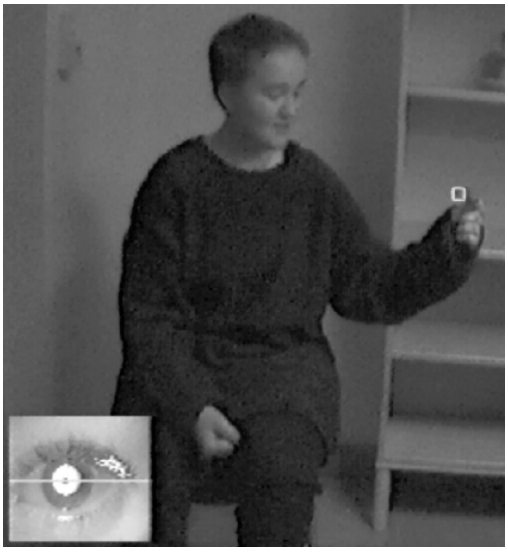

Figure. 2. Example of a speaker-fixated gesture that is also fixated by the addressee (=white circle). The addressee's eye is shown as a picture-in-picture.

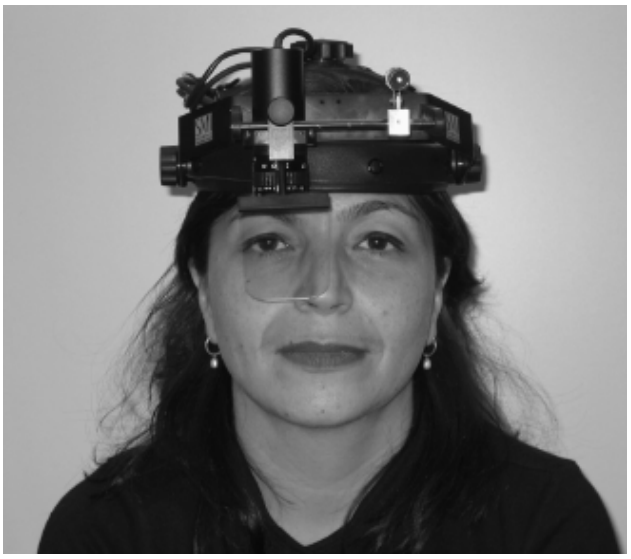

Figure. 3. The head-mounted SMI iView© eye-tracker.

Figure 4. The saccades, fixation locations, and fixation durations of an addressee looking at a speaker. The diameter of the circles indicates the duration of the fixation. The fixation comparison point in the bottom right corner equals one second.

Table 1. Summary of predictions.

Social effects	less face, more gestures on video than live
Size effects	fewer gestures on small-scale video screen than full-sized video, especially gestures in peripheral gesture space

Table 2. Average viewing times in percent on targets across conditions.

Fixated objects	Live	VideoLife-size	VideoScreen
Face	95.6	94.2	90.8
Gestures	0.5	0.4	0.2
Body parts	1.3	1.4	5.6
Object in room	2.1	1.4	2.7
Empty space	0.4	1.1	0.3
Other (the original live addressee)[i]	0	1.4	0.4
	100	100	100

Table 3. Proportion of fixated gestures of the total number of gestures across conditions.

	Live	VideoLife-size	VideoScreen
fixated gestures %	7.4	4.5	3

Table 4. Proportion of fixated gestures with a specific gestural feature across conditions.

Gesture features	Live %	VideoLife-size %	VideoScreen %	
Central	7	6	2	
Peripheral	8	3	4	
	n.s.	n.s.	n.s.	
+Hold	33	20	15	n.s.
-Hold	4	2	2	
	***	***	***	
+Speaker-fixation	23	8	8	*
-Speaker-fixation	5	4	2	
	***	*	***	

1

2

3

4

5