Visual Attention towards Gestures in Face-to-Face Interaction vs. on Screen

Marianne Gullberg[1], Kenneth Holmqvist[2]

[1] Max Planck Institute for Psycholinguistics, PO Box 310, NL-6500 AH Nijmegen, The Netherlands
marianne.gullberg@mpi.nl
[2] Lund University Cognitive Science, Kungshuset Lundagård, S-22222 Lund, Sweden
kenneth@lucs.lu.se

Corresponding address:
Marianne Gullberg
Max Planck Institute for Psycholinguistics
PO Box 310
NL-6500 AH Nijmegen
The Netherlands

Abstract. Previous eye-tracking studies of whether recipients look at speakers' gestures have yielded conflicting results but have also differed in method. This study aims to isolate the effect of the medium of presentation on recipients' fixation behaviour towards speakers' gestures by comparing fixations of the same gestures either performed live in a face-to-face condition or presented on video ceteris paribus. The results show that fixation behaviour towards gestures is largely similar across conditions, although fewer gestures are fixated on video (non-significant). In discussing the effect of the absence of a live interlocutor vs. the projection size as a source for this reduction, we touch on some underlying mechanisms governing gesture fixations. The results are pertinent to man-machine interface issues as well as to the ecological validity of video-based paradigms needed to study the relationship between visual and cognitive attention to gestures and Sign Language.

Keywords: Gesture, interaction, perception, attention, vision, eye-tracking, methodology

Visual Attention towards Gestures in Face-to-Face Interaction vs. on Screen*

Marianne Gullberg[1], Kenneth Holmqvist[2]

[1] Max Planck Institute for Psycholinguistics, PO Box 310, NL-6500 AH Nijmegen, The Netherlands
marianne.gullberg@mpi.nl
[2] Lund University Cognitive Science, Kungshuset Lundagård, S-22222 Lund, Sweden
kenneth@lucs.lu.se

Abstract. Previous eye-tracking studies of whether recipients look at speakers' gestures have yielded conflicting results but have also differed in method. This study aims to isolate the effect of the medium of presentation on recipients' fixation behaviour towards speakers' gestures by comparing fixations of the same gestures either performed live in a face-to-face condition or presented on video ceteris paribus. The results show that fixation behaviour towards gestures is largely similar across conditions, although fewer gestures are fixated on video (non-significant). In discussing the effect of the absence of a live interlocutor vs. the projection size as a source for this reduction, we touch on some underlying mechanisms governing gesture fixations. The results are pertinent to man-machine interface issues as well s to the ecological validity of video-based paradigms needed to study the relationship between visual and cognitive attention to gestures and Sign Language.

1 Introduction

Previous studies of visual attention to gestures based on eye-tracking have yielded very different results concerning if and when recipients look at gestures. In a study of gesture fixations in live face-to-face interaction, Gullberg & Holmqvist [1] found that recipients fixated only a minority of speakers' gestures (8.8%), chiefly perceiving gestures through peripheral vision and instead maintaining eye or face contact. Recipients only fixated gestures if they were performed in peripheral gesture space or were fixated by speakers themselves (autofixation). The dominance of the face

* We gratefully acknowledge the financial support of Birgit and Gad Rausing's Foundation for Research in the Humanities, and of the Max Planck Society.

corroborated findings from earlier studies of gaze in interaction conducted without eye-trackers [2, 3, 4]. We hypothesised that gestural performance features compete with social norms for maintained eye contact in determining whether gestures are fixated in live interaction. In the absence of any social pressure for eye contact, as in a video setting, fixation behaviour towards gestures might therefore change. And indeed, Nobe et al. [5, 6] found that recipients who were shown gestures on video fixated a majority of gestures (70-75%).

However, the differences found in these studies are difficult to assess given that the studies differ on a range of methodological points, including the medium of presentation (live face-to-face vs. on screen), the naturalness of gestures (natural vs. synthesised or otherwise manipulated gestures), the embeddedness of gestures in discourse (sustained discourse vs. isolated gestures listed), the type of gesture (spontaneous co-speech gestures vs. emblematic gestures), and agent (human vs. anthropomorphic agents). This study therefore aims to isolate the effect of the medium of presentation by comparing fixations of naturally occurring, spontaneous co-speech gestures [7] in story retellings under two conditions keeping all other features constant. We expect the comparison between a live condition and a video condition to yield two possible types of effects. The presence/absence of the interlocutor will reveal the interactional, social effects on fixation behaviour towards gestures. In contrast, the reduced projection size will reveal more mechanical effects related to the capacities of peripheral vision. We asked the following specific questions:

- do recipients fixate the speaker's face less often on video than in a live condition?
- do recipients fixate more gestures overall on video than live?
- do recipients fixate different gestures on video than live? Or put differently, is the impact of the gestural performance features that determine fixations different on video than in a live condition?

The answers to these questions are pertinent to the study of the relationship between visual and cognitive attention to gestures and Sign Language. Although there is growing evidence that recipients attend to and retain information expressed in gestures (e.g. [8, 9, 10]), the relationship between visual attention to gestures and cognitive uptake of gestural information is poorly understood. This issue needs to be studied in video-based experimental designs that allow the repeated display of the same gestures to different recipients. However, video-based designs risk compromising ecological validity as long as we do not know what effect the medium of presentation has on recipients' fixation patterns. Moreover, the answers to these questions should be of interest to the field of man-machine interaction. The construction of interfaces with agents gesticulating on screen should benefit from insights into how humans attend to other humans gesticulating on screen as a first step.

2 Procedure

In the live condition, 20 native speakers of Swedish, unacquainted prior to the experiment, were randomly assigned the role of speaker or recipient. The speakers memorised a printed cartoon and were then told to convey the story as well as they could to the recipients who would have to answer questions about it later. Similarly, the recipients were asked to make sure they understood the story, since they would have to answer questions about it later. The recipients were seated 180 cm away from the speakers (measured back to back) facing them, and were wearing a head-mounted SMI iView© eye-tracker (see Figure 1). The output data from the eye-tracker consist of a video recording of the recipient's field of vision (i.e. the speaker), and a superimposed video recording of the recipient's fixations (see Figure 2).

Fig.1. The SMI iView© headmounted eye-tracker

Fig. 2. Example of output from the eye-tracker: the recipient's field of vision (=the speaker) and the recipient's fixation marker (=the white circle) superimposed

While retelling the stories to the recipients, the speakers were simultaneously video recorded with a separate video camera placed behind the recipients. These video recordings of the speakers served as stimuli in the video condition. 20 new Swedish recipients were shown these video recordings on a video screen, 2 new recipients per original speaker. Note that this design allowed us to collect fixation data for exactly 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 recipient; the gestures were thus not performed 'for the camera'. The new recipients were seated 110 cm away from a 28" video screen. The projection size and the distance between recipient and screen decreased all angels by 52.3% of the original

size. The SMI iView© remote set was placed between the recipient and the video screen. The instructions to the recipients were the same as in the live condition.

A post-test-questionnaire was distributed to all subjects to ensure that gesture was not identified as the target of study. The eye-tracker appears not to have disrupted interaction (cf. [1]). All subjects, speakers and recipients alike, declared that the equipment did not disturb them. The speakers' speech and gestural behaviour did not differ quantitatively or qualitatively from data collected in an identical situation without eye-trackers [11]. Recipients' fixation data include fixations of body parts that the subjects might have avoided to fixate had they been concerned about the equipment. This in turn suggests that the recipients tended to forget about the apparatus and behaved naturally.

3 Data Treatment

Speech was transcribed and checked for demonstrative expressions referring directly or indirectly to the gestures, e.g. 'he held it like this'. Such demonstrative expressions function as an interactional deictic device by which speakers can direct recipients' attentions towards the gestures (e.g.[5, 12]). No such deictic expressions were present in the data.

Both a temporal and a spatial criterion were used to identify fixations. The fixation marker had to remain in an area the size of the marker itself for at least 120 ms (=3 video frames) in order to count as a fixation. The fixation data were also coded for the target object fixated, e.g. the right hand gesturing, the shoulder, an object in the room, etc. (see Figure 3).

The gesture data were coded for the three performance features that have been found to attract fixations in previous studies:

a) place of articulation in gesture space using McNeill's [7] schema of space (e.g. centre-centre, peripheral right). For the purpose of calculations, all cases of centre-centre and centre were collapsed, as were all peripherals, leaving two broad categories Central and Peripheral.

b) autofixation, or whether or not speakers fixate their own gestures (see Figure 4). All cases of autofixation were enacted gestures (character viewpoint in McNeill's terms, [7]), where the speakers acted as a character in the story.

c) presence vs. absence of hold, i.e. a momentary cessation of movement in a gesture [13,14]. Post-stroke holds, i.e. cessations of movement after the hand has reached the endpoint of a trajectory, were found to attract fixations in the video-based studies by Nobe et al. [5]. We wanted to test their relevance in a live setting. Note

that non-hold means all other phases of the gesture phrase, i.e. preparation, stroke, or retraction.

The data is considered to have a binomial distribution. Proportions of fixated gestures are calculated relative to produced gestures in the various categories, whereby the quantitative variations in gesture production are neutralised. We employ a test of significance of differences between proportions that is mathematically equivalent to the chi-square test under one degree of freedom.

Fig. 3. Example of a fixation coded for target Body part (low abdomen)

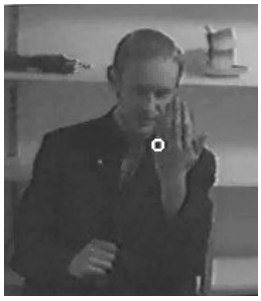

Fig. 4. Example of autofixated gesture. The gesture is also fixated by the recipient (=white circle)

4 Results

- Do Recipients Fixate the Speaker's Face less Often on Video than Live?

Table 1 shows that recipients spend significantly less time fixating the face in the video condition than in the live condition (from on average 95.6% of the time in the live condition to 88.7% in the video condition). Despite this decrease, the face nevertheless clearly dominates as a fixation target in both conditions.

Table 1. Average time in percent spent fixating face vs. outside face across conditions.

	Live	Video	Comparison
Average time on face %	95.6	88.7	$p \leq .01$

- Do Recipients Fixate More Gestures Overall on Video than Live?

Table 2 indicates that only a minority of gestures are fixated in either condition, on average 7.4% vs. 3% of all gestures. The amount of fixated gestures in the live condition corresponds well to the findings in Gullberg & Holmqvist ([1]; 8.8%).

Contrary to expectations, however, recipients fixate *fewer* gestures in the video condition than in the live condition, although the difference between the conditions is not significant.

Table 2. Proportion of fixated gestures across conditions.

	Live	Video	Comparison
# gestures	364	734*	
# fixated gestures	27	22	
Fixated gestures %	7.4	3	n.s.

The combination of more time spent outside the face but less time on gestures in the video condition means that other targets receive more fixations. Table 3 lists the significant changes in extra-facial fixation targets between the conditions. In the video condition, the amount of extra-facial fixations landing on gestures decreases significantly (from 13.4% to 3.4%), as well as fixations of objects in the room (from 45.9% to 21.2%). In contrast, fixations of body parts other than gesticulating hands or arms, i.e. immobile body parts such as the abdomen, the chest area, etc., increase significantly (from 35.2% to 58%).

Table 3. Proportions of extra-facial fixation on three targets across conditions.

Extra-facial fixation targets %	Live	Video	Comparison
Gestures	13.4	3.4	$p \leq .025$
Objects in the room	45.9	21.2	$p \leq .01$
Immobile body parts	35.2	58	$p \leq .05$

- Do Recipients Fixate Different Gestures on Video than Live?

Table 4 shows the extent to which the three gestural performance features attract fixations across the conditions. Despite the (non-significant) overall decrease in gesture fixations in the video condition, the proportional attraction strength of the features is maintained such that the gestural performance features can be said to operate across conditions. As can be seen in Table 4, there is no effect for Place of articulation in either condition, meaning that Peripheral gestures are *not* fixated more often than

* Note that the total number of gestures in the video condition exceeds the number of gestures in the live condition times two ((364*2)=728 vs. 734). There was a technical problem with the recording of one of the live speakers, such that it could not be projected in the video condition. Instead, we projected the recording of another live speaker, who produced roughly the same amount of gestures in the live condition as the original speaker.

central gestures. In contrast, both Holds and Autofixations attract significantly more fixations than other gestures in both conditions.

Table 4. Fixated gestures in % with respect to performance features across conditions.

Performance features	Live		Video		Live vs. video
	fix %		fix %		
Peripheral/Central	8/7	n.s.	4/2	n.s.	n.s.
Hold/no Hold	33/4	p≤.01	15/2	p≤.025	n.s.
Autofix/no Autofix	23/5	p≤.05	8/2	p≤.05	n.s.

5 General Discussion and Conclusion

The results show that fixation behaviour is both similar and different across the conditions. The similarities include the face dominance and the tendency to fixate only a minority of gestures. Similarly, the two articulatory features that reliably attract fixations, Holds and Autofixations, operate in both conditions. The differences consist of a significant reduction of face fixations and a significant increase in fixations of body parts in the video condition. In addition, there is an unexpected overall (but non-significant) reduction in gesture fixation rate. This decrease is the only relevant difference in fixation behaviour towards gestures across the conditions. The question then is what causes it.

We originally hypothesised that the absence of a live interlocutor on video would lead to relaxed social pressure for maintained eye contact with a reduction of face fixations and an increase in gesture fixations as a direct result. The first prediction was borne out, since the results indicate that the norm for face fixations is indeed relaxed. However, recipients at liberty to fixate whatever they want do not fixate more gestures. Instead they fixate more immobile body parts, i.e. they allow themselves to fixate areas that are probably not ‘socially acceptable’ fixation targets in a live condition. The general change in face and body fixations thus seems clearly socially motivated. A tentative social explanation for the decrease in gesture fixations, related to the tendency to fixate only ‘socially acceptable’ targets in a live condition, could be that fixations of gestures are the only ‘legitimate’ reason to fixate something other than the speaker’s face in a live condition. In a video situation without any social constraints on fixation behaviour, recipients can ignore gestures to a greater extent and only fixate their ‘real’ targets of interest.

However, the decrease in fixation rate may also be related to the reduced projection size of the video condition. The arguments related to this 'mechanical' factor are associated with the capacities of peripheral vision by which fixations occur when the target is too far removed from the fixation point. Under this reading, the video condition would lead to fewer gesture fixations simply because the reduced visual field means that recipients can rely on their peripheral vision to a greater extent than live. In the video condition at hand, the projection size decreased all angels by 52.3% of the original size. If the reduction in gesture fixations were determined by size projection alone, the reduction in fixation rate between conditions should match this number [15]. However, the general reduction in gesture fixation between the conditions is 67.9% (from 7.4% of all gestures fixated live to 3% of all gestures fixated on video), which leaves 15.3% of the reduction unexplained. When the individual gestural performance features are considered, we find that the reduction in fixation rate for Peripherals (50.2%) and Holds (55.7%) is close to the value of the size reduction. In contrast, for Autofixation the reduction in fixation rate is somewhat greater (64%), leaving 11.7% of the reduction unexplained. These figures suggest that projection size accounts for the brunt of the reduction, but that different gestures are affected differently by the move from a live to a video condition such that some gesture types may be affected by social factors as well. Ultimately, this differential effect may reflect the fact that different gestures are fixated for different reasons.

As seen above, the reduction of fixations of Holds is well accounted for by the size reduction. This suggests that the tendency to fixate Holds is associated with the capacities of peripheral vision. In a live condition, Holds in fact represent a challenge to peripheral vision, which is good at motion detection but bad at fine-grained detail. As long as gestures are moving, peripheral vision is sufficient for detecting (and processing) the broader gestural information (location, direction, or size) even when gestures are performed in the periphery. This is, incidentally, the likeliest explanation for why peripherally performed gestures are in fact not fixated more often than centrally performed ones in this study. However, when a gesture ceases to move as in a hold, peripheral vision is challenged. In a live condition where recipients mainly fixate the speaker's face, their peripheral vision may tell them that the speaker's hand is not in a resting position. Peripheral vision is insufficient to retrieve information from a gesture in hold, as it can pick up neither motion nor configurational detail such as hand shape from it. Speakers therefore have to fixate holds in order to retrieve any gestural information at all. In the video condition, on the other hand, the distance between the gesture and the fixation of the face on the screen is presumably short enough for peripheral vision to operate efficiently despite the lack of movement. As holds are

fixated for reasons of limitations to peripheral vision, so they are affected by size changes in the video condition.

In contrast, the size reduction did not account for the entire decrease in fixation rate for Autofixations. Autofixation is not a purely articulatory feature, and is probably fixated for different reasons than Holds. It is not the gesture itself that attracts the fixation, but the fact that the speaker's gaze is on it. Speakers' gaze serves as a powerful social cue to joint attention [16, 17, 18]. As a consequence, Autofixation may be perceived as a means for the speaker to direct recipients' attention towards gestures (e.g. [19, 20]). In this sense, Autofixation is a social phenomenon. Not to co-fixate an autofixated gesture in a live condition would be socially inept. It is in fact common for speakers who autofixate their gestures to look back up on the recipient to ensure that joint attention has indeed been established. In a video condition, there is no such social pressure to follow an Autofixation. The decrease in fixations of Autofixations on video may thus be partly due to the increased capacities of peripheral vision following from the reduced projection size, and partly to the absence of a live interlocutor.

What are the implications of these findings for the ecological validity of a video-based paradigm used to study gesture perception? The social change between conditions, i.e. the absence of a live interlocutor, mainly affects face and body fixations, and not (or only partly) gesture fixations. In contrast, the mechanical change, i.e. the reduction in projection size, does affect gesture fixations, leading to fewer gestures being fixated overall, although this change is not significant. A life-size projection should eliminate most of this reduction in gesture fixation rate. We therefore suggest that, as a first evaluation, a video-based paradigm need *not* compromise ecological validity. However, in order to isolate the true domains of the social and the mechanical size effect for all gesture types, and especially for autofixated gestures, a direct comparison should be made between a live condition and a video condition where stimuli are projected life-size. Such a study should also reveal more about the general underlying mechanisms governing gesture fixations.

References

1. Gullberg, M., Holmqvist, K.: Keeping an Eye on Gestures: Visual Perception of Gestures in Face-to-Face Communication. Pragmatics & Cognition 7 (1999) 35-63
2. Argyle, M.: The Psychology of Interpersonal Behaviour. Penguin, Harmondsworth (1976)
3. Kendon, A.: Conducting Interaction. Cambridge University Press, Cambridge (1990)
4. Nielsen, G.: Studies in Self Confrontation: Viewing a Sound Motion Picture of Self and Another Person in a Stressful Dyadic Interaction. Munksgaard, Copenhagen (1962

5. Nobe, S., Hayamizu, S., Hasegawa, O., Takahashi, H.: Are Listeners Paying Attention to the Hand Gestures of an Anthropomorphic Agent? An Evaluation Using a Gaze Tracking Method. In: Wachsmuth, I., Fröhlich, M. (eds.): Gesture and Sign Language in Human-Computer Interaction. Springer Verlag, Berlin (1998) 49-59
6. Nobe, S., Hayamizu, S., Hasegawa, O., Takahashi, H.: Hand Gestures of an Anthropomorphic Agent: Listeners' Eye Fixation and Comprehension. Cognitive Studies. Bulletin of the Japanese Cognitive Science Society 7 (2000) 86-92
7. McNeill, D.: Hand and Mind. Chicago University Press, Chicago (1992)
8. Cassell, J., McNeill, D., McCullough, K.E.: Speech-Gesture Mismatches: Evidence for one Underlying Representation of Linguistic and Nonlinguistic Information. Pragmatics & Cognition 7 (1999) 1-33
9. Beattie, G., Shovelton, H.: Do Iconic Hand Gestures Really Contribute Anything to the Semantic Information Conveyed by Speech? Semiotica 123 (1999) 1-30
10. Beattie, G., Shovelton, H.: Mapping the Range of Information Contained in the Iconic Hand Gestures that Accompany Spontaneous Speech. J. of Lang. and Social Psy. 18 (1999) 438-462
11. Gullberg, M.: Gesture as a Communication Strategy in Second Language Discourse. Lund University Press, Lund (1998)
12. Streeck, J., Knapp, M.L.: The Interaction of Visual and Verbal Features in Human Communication. In: Poyatos, F. (ed.): Advances in Nonverbal Communication. Benjamins, Amsterdam (1992) 3-23
13. Kendon, A.: Some Relationships between Body Motion and Speech: An Analysis of an Example. In: Siegman, A.W., Pope, B. (eds.): Studies in Dyadic Communication. Pergamon, New York (1972) 177-210
14. Kendon, A.: Gesticulation and Speech: Two Aspects of the Process of Utterance. In: Key, M. (ed.): The Relationship of Verbal and Nonverbal Communication. Mouton, The Hague (1980) 207-227
15. Latham, K., Whitaker, D.: A Comparison of Word Recognition and Reading Performance in Foveal and Peripheral Vision. Vision Research 37 (1996) 2665-2674
16. Baron-Cohen, S.: The Eye Direction Detector (EDD) and the Shared Attention Mechanism (SAM): Two cases for evolutionary Psychology. In: Moore, C., Dunham, P.J. (eds.): Joint attention. Erlbaum, Hillsdale (1995) 41-59
17. Langton, S.R.H., Watt, R.J., Bruce, V.: Do the eyes have it? Cues to the direction of social attention. Trends in Cognitive Sciences 4 (2000) 50-59
18. Tomasello, M.: Joint attention as social cognition. In: Moore, C., Dunham, P.J. (eds.): Joint attention. Erlbaum, Hillsdale (1995) 103-130
19. Streeck, J.: Gesture as Communication I: Its Coordination with Gaze and Speech. Communication Monographs 60 (1993) 275-299
20. Goodwin, C.: Gestures as a resource for the organization of mutual orientation. Semiotica 62 (1986) 29-49