Do the eyes really have it? Dynamic allocation of attention when viewing moving faces

Melissa L.-H. Võ Harvard Medical School, Cambridge, MA, USA

Tim J. Smith Birkbeck, University of London, UK

Parag K. Mital Goldsmiths, University of London, UK

John M. Henderson University of South Carolina, USA

What controls gaze allocation during dynamic face perception? We monitored participants' eye movements while they watched videos featuring close-ups of pedestrians engaged in interviews. Contrary to previous findings using static displays, we observed no general preference to fixate eyes. Instead, gaze was dynamically directed to the eyes, nose, or mouth in response to the currently depicted event. Fixations to the eyes increased when a depicted face made eye contact with the camera, while fixations to the mouth increased when the face was speaking. When a face moved quickly, fixations concentrated on the nose, suggesting that it served as a spatial anchor. To better understand the influence of auditory speech during dynamic face perception, we presented participants with a second version of the same video, in which the audio speech track had been removed, leaving just the background music. Removing the speech signal modulated gaze allocation by decreasing fixations to faces generally and the mouth specifically. Since the task was to simply rate the likeability of the videos, the decrease of attention allocation to the mouth region implies a reduction of the functional benefits of mouth fixations given that speech comprehension was not required. Together, these results argue against a general prioritization of the eyes and support a more functional, information-seeking use of gaze allocation during dynamic face viewing.

Keywords: dynamic faces, eye movements, optimal viewing position, attention allocation, information seeking

Citation: Võ, M. L.-H., Smith, T. J., Mital, P. K., & Henderson, J. M. (2012). Do the eyes really have it? Dynamic allocation of attention when viewing moving faces. *Journal of Vision*, *12*(13):3, 1–14, http://www.journalofvision.org/content/12/13/3, doi:10.1167/12.13.3.

Introduction

Looking at a face provides a variety of types of information, including identity, emotional state, and potentially intentions and goals. Face processing is therefore crucial to social interaction. The eyes as the "window to the soul" have received special interest, since the perception and interpretation of gaze in social cognition promotes our understanding of what another person is currently attending to, thinking about, and feeling.

In early studies, Buswell (1935) and Yarbus (1967) demonstrated that people tend to look at faces in scenes, and more specifically at the face's eyes, a finding that has been replicated many times (e.g., Birmingham, Bischof, & Kingstone, 2008a, 2008b, 2009a, 2009b; Emery, 2000; Friesen & Kingstone, 1998; Henderson, Williams, & Falk, 2005; Itier, Villate, & Ryan, 2007; Walker-Smith, Gale, & Findlay, 1977; for a review see Birmingham & Kingstone, 2009; Langton, Watt, & Bruce, 2000). Williams and Henderson (2007), for example, tracked eye movements of participants viewing upright and inverted faces in both the study and recognition phase of a memory task. They found that although inverted faces were more difficult to recognize than upright faces, up to 80% of all fixations were biased towards the eyes both during learning and recognition, regardless of whether the faces were presented upright or inverted (see also Henderson et al., 2005).

Is the reported eye bias based on visual saliency, social interest, or an information acquisition strategy? Birmingham et al. (2009b) presented evidence against a purely saliency based account. They showed static images of faces embedded in complex social contexts and found that observers' first fixations were biased toward faces in general and the eyes in particular, despite the fact that the eyes and heads were generally nonsalient (according to the Itti & Koch, 2000, saliency model). Birmingham et al.

doi: 10.1167/12.13.3 Received July 24, 2012; published December 3, 2012 ISSN 1534-7362

(2008b) provided further evidence against a saliency explanation. They found that fixations to eyes increased as the social content of a scene (defined as the number of people) increased. The authors concluded that eyes are fixated not because of their visual salience, but because they are a rich source of social information.

The social relevance of the eyes is also suggested by studies investigating the importance of direct eye contact. Humans orient to eye contact preferentially (e.g., Senju & Hasegawa, 2005; Senju, Hasegawa, & Tojo, 2005; von Grünau & Anston, 1995), and seem to do so early in life (e.g., Farroni, Csibra, Simion, & Johnson, 2002). Direct gaze holds attention (e.g., Senju & Hasegawa, 2005) and is processed automatically, even outside of conscious awareness, implying enhanced unconscious representation of faces with direct gaze (Stein, Senju, Peelen, & Sterzer, 2011; for reviews on the eye-contact effect see Kleinke, 1986; Senju & Johnson, 2009). Although eyes certainly are of great social importance, there is evidence that the eyes are not always prioritized during face viewing. Instead, it may be that the deployment of attention to certain parts of a face depends on the current task (e.g., Buchan, Paré, & Munhall, 2007; Eisenbarth & Alpers, 2011; Gosselin & Schyns, 2001; Lansing & McConkie, 1999). That is, where people look in a face seems to be dependent on which parts of a face provide the information necessary to pursue the current goal, such as trying to identify the person, understand what is being said, or determine the person's emotional state.

Using "Bubbles," a reverse-correlation technique that can assign the credit of human categorization performance to specific visual information, Gosselin and Schyns (2001) found that information near the eyes supports gender discrimination, whereas information near the mouth helps to determine whether a face is expressive or not. Subsequent "Bubbles" studies revealed that different emotions are identified using different parts of the face (e.g., Smith, Cottrell, Gosselin, & Schyns, 2005). For example, information around the mouth is mostly used to identify happiness, whereas the eyes are used to identify a fearful expression.

To test whether a given task directly modulates attention and gaze allocation, Buchan et al. (2007) tracked participants' eye movements while they performed either a speech recognition or an emotion judgment task. When participants performed emotion judgments, they directed their gaze toward the eyes, whereas they looked more toward the mouth in the speech recognition task. Interestingly, when noise was added to the acoustic speech signal, gaze in both tasks was directed more to the center of the face. The authors argued that attention allocation is sensitive to the distribution of information in the face, but can also be influenced by strategies aimed at maximizing the amount of visual information processed (see also Buchan, Paré, & Munhall, 2008). Degrading the speech signal by adding noise has previously been shown to decrease the number of transitions between areas of the face, suggesting that the availability of the verbal information may affect how information is gathered (Vatikiotis-Bateson, Eigsti, Yano, & Munhall, 1998). Manipulating the auditory speech signal while having people watch dynamic faces therefore provides a method for testing an information-seeking account while keeping visual information unchanged.

Information-seeking gaze control strategies are also evident when viewing other complex stimuli such as static scenes. Najemnik and Geisler (2005, 2008), for example, investigated optimal eye movement strategies of observers searching 1/f noise background regions for predefined targets. They found that observers use sophisticated search mechanisms that maximize the information collected across fixations. Similarly, the cognitive relevance framework proposes that the weight given to a particular object or visual feature in a scene is determined by the current cognitive information-gathering needs rather than by visual salience (Henderson, Malcolm, & Schandl, 2009). Thus, human gaze control is intelligent in that it draws not only on currently available visual input but also on cognitive knowledge structures (for a review, see Henderson, 2006). Moreover, when viewing arrays of objects or static scenes, gaze is preferentially distributed to the center of objects (e.g., Foulsham & Underwood, 2009; Henderson, 1993; Nuthmann & Henderson, 2010). Such preferred viewing locations (Rayner, 1979) are thought to optimize the encoding of information given visual acuity limitations away from the center of gaze (e.g., McConkie, Kerr, Reddix, & Zola, 1988). Similar optimization behaviors may also be observed during face perception. In static face recognition, initial fixations are preferentially directed to the center of the stimulus, in this case, the nose (Hsiao & Cottrell, 2008). Compared to the eyes, the nose may coincide with the "center of the information," where the information is balanced in all directions and thus might be the optimal viewing position for rapid face recognition.

The majority of previous research on face viewing has used static images, but the faces we interact with in the world are dynamic. The motion of dynamic faces is important in that it contains information about social status, identity, and emotion not present in static faces (e.g., Ambadar, Schooler, & Cohn, 2005; Foulsham, Cheng, Tracy, Henrich, & Kingston, 2010; Hill & Johnson, 2001; Knappmeyer, Thornton, & Bülthoff, 2003; Lander & Bruce, 2000; O'Toole, Roark, & Abdi, 2002). There is evidence that the temporal characteristics of facial motion may be represented by a sparse distribution of dynamic points that can enhance phonetic perception (Rosenblum, Johnson, & Saldaña,

1996). Thus, facial dynamics form the core of social interactions, such as looking at someone's eyes to better understand their momentary emotional state, following a person's gaze as an indicator of their current focus of interest, or supporting speech perception by sampling mouth movements (e.g., Buchan et al., 2007). Each of these subcomponents of social interaction might require different, dynamically adjusted viewing strategies that maximize information uptake given a specific task.

Similar to Buchan et al. (2007), Lansing and McConkie (2003) asked observers to watch talking faces while they tried to understand what was being said. These videos were either played in silence or with low-intensity sound. In addition, Lansing and McConkie added a 1-s still image of the first and the last frame of the video to the beginning and end of the trial. This provided a still-image control condition. They found that during still-image periods, an observer's gaze was biased toward the talker's eyes compared to other facial regions. This replicated the many studies that have found an "eye primacy effect" when looking at photographs of faces (for reviews see Birmingham & Kingstone, 2009; Langton et al., 2000). However, once the faces started moving, the gaze shifted toward the mouth. This "information source attraction effect" was increased when the sound was reduced from low-intensity to silence, arguing for even greater effort to gather visual information to support sentence comprehension through lip reading. Subsequent analysis of periods in which participants were fixating the eyes during speech did not, however, reveal a detriment in their ability to understand the speech. Lansing and McConkie suggested that the viewers' belief in the utility of fixating on the mouth for lip reading may be misguided or that parafoveal and peripheral vision is sufficient for identifying mouth movements associated with speech sounds.

In this study, we used highly engaging videos of people in real world settings and monitored eye movements in response to dynamic events, such as making eye contact, talking, or performing head movements. Previous findings give rise to two competing hypotheses: First, if viewers generally prioritize the eyes, as suggested by much of the static face-viewing literature, then there should be no significant modulation of gaze behavior as a function of depicted dynamic events, such as a person talking or making eye contact. Alternatively, if ongoing dynamic events require sampling different visual features depending on the viewer's goals, we should observe a strong modulation of gaze dependent on the depicted dynamic event. If the latter hypothesis is supported, then we can make several supplementary predictions: (a) Eye contact of an actor with the camera (the viewer) should evoke increased attention toward the face (e.g., Stein et al., 2011) and possibly the eyes; (b) A talking face should draw gaze toward the mouth partly due to increased movement of visual features that are known to attract attention (see Itti, 2005; Mital, Smith, Hill & Henderson, 2011), but especially due to sampling of mouth movements to support speech comprehension (Buchan et al., 2007; Lansing & McConkie, 2003); and (c) Finally, we further hypothesize that the informational content of face regions, not the mere saliency of, for example, eyes or moving mouths, predicts gaze behavior (e.g., Birmingham et al., 2009b; Lansing & McConkie, 2003). If true, then removing the speech signal from the videos should change preferred fixation locations within a face, despite unchanged visual information. Lansing and McConkie (2003) found that fixations toward the mouth increased when participants watched movies in silence compared to watching movies with low-intensity speech. This was likely due to the specific task instructions in this study, where participants' main aim was to reproduce the words they had seen or heard being uttered. Participants in our study, however, were merely asked to rate each video on a likeability scale. We therefore expected to find a decrease of mouth fixations when the acoustic speech signal was completely removed, since accurate speech perception was not necessary to fulfill the task.

To test these hypotheses, we recorded eye movements while participants watched video clips of pedestrians being interviewed on the streets of Brooklyn and London. The videos used a fixed camera vantage point, maintaining a close-medium camera shot focused on the faces of interviewees in the video throughout the entire clip (see Figure 1).

These clips were chosen for their vividness and their close-up shots of socially interacting faces. The videos also contained various dynamic events that were of interest for further analysis: (a) people looking at the camera and thereby creating a sense of eye contact with the viewer, (b) people talking in reply to the interviewer's question "Where do you want to wake up tomorrow?," and (c) instances of rapid head movements that might require yet another strategic positioning of gaze to keep track of the face as a whole. In order to be able to differentiate between visual saliency-based control of attention and more cognitively driven guidance in the processing of faces, we played the videos either with or without speech, which allowed us to investigate whether information-seeking modulates attentional allocation in dynamic face viewing.

Method

Participants

Eighty-eight students (vocal condition: 44 total, 25 female; mute condition: 44 total, 27 female) from the

Figure 1. Example frames taken from the two videos that were used in this study. The two upper rows depict scenes from interviews in London, while the two lower rows show scenes from interviews in Brooklyn.

University of Edinburgh ranging in age between 18 and 30 years (vocal condition: $M = 20.77$, $SD = 1.91$; mute condition: $M = 21.36$, $SD = 2.68$) participated in this study for payment. We randomly assigned each of the two experimental conditions (vocal vs. mute) to a group of 44 participants. All participants reported normal or corrected-to-normal vision.

Stimulus material

The stimulus material consisted of two movie clips taken from the Dynamic Images and Eye Movements (DIEM) database (http://thediemproject.wordpress.com/), in which pedestrians in London and Brooklyn were asked where they would wish to wake up tomorrow (see "50 People One Question" at http://fiftypeopleonequestion.com/). The Brooklyn video clip was 122 s in length and consisted of 3,706 frames; the London video was 128 s in length and consisted of 3,878 frames. Both were encoded using the Xvid MPEG-4 codec (SR Research, Ontario, Canada) in an AVI container at 30 frames/s, and depicted close-ups of either one or two faces of pedestrians engaged in an interview. Analysis in the present study focused on video segments containing only one face (as seen in Figure 1).

To investigate the effect of acoustic speech on dynamic gaze control during face viewing, we created two versions of the same visual stimulus material: the "vocal" condition, in which the video clips were accompanied by the original audio including dialogue and background music ("Don't See the Sorrow" by Au Revoir Simone in the Brooklyn clip and "Don't Make Say Think" by Chinatown in the London clip), and the "mute" condition, in which we retained background music but removed the speech stream from the audio.

Apparatus

Eye movements of both the left and the right eye were recorded with a binocular EyeLink2000 desktop mount system (SR Research, Ontario, Canada) at a sampling rate of 1000 Hz for each eye. Videos were displayed in their native resolutions and centered on a 2100 Viewsonic monitor with desktop resolution 1280 × 960 at 120 Hz, viewed at a distance of 90 cm

Figure 2. Example of dynamic regions of interest (face = black, eyes = blue, nose = red, mouth = green) for illustration purposes. Observers never actually saw these.

subtending visual angles of 25.66° (horizontal) and 19.23° (vertical). Standard stereo desktop speakers delivered the audio media component. Experimental sessions were carried out on a computer running Windows XP. Stimulus presentation and response recording was controlled by Experiment Builder (SR Research, Ontario, Canada).

Procedure

The two target videos were inserted randomly into a series of other video clips taken from a broad selection of television clips, news reports, music videos, movie trailers, and naturalistic videos. Participants were informed that they would watch a series of short, unconnected video clips. Following each clip, instructions appeared on the screen asking them to rate (by pressing the relevant button on the joypad) how much they had liked it on a scale from one to four. This procedure ensured some interactivity without interfering with the free-viewing task. The order of the clips was randomized across participants (for more details, see Mital, Smith, Hill, & Henderson, 2011).

A chin and headrest (unrestrained) were used throughout. A 13-point binocular calibration preceded the experiment. Central fixation accuracy was tested prior to each trial, with a full calibration repeated when necessary. The central fixation marker also served as a cue for the participant and offered an optional break in the procedure. After checking for a central fixation, the experimenter manually triggered the start of each trial. The experiment lasted 45–60 min.

Data analyses

To analyze the eye movement data as a function of dynamic regions of interest and dynamic events depicted in the video clips, each frame was marked to characterize different events (see event tagging below).

Dynamic regions of interest

Each face was encompassed by an oval region of interest used to define fixations on the face versus the rest of the scene. Since we were mainly interested in which parts of depicted faces observers looked at, we created the following regions of interest for all of our analyses: (a) elliptical regions for each eye (for analysis purposes these are treated as a single "eyes" region), (b) an elliptical region for the nose, and (c) a rectangular region that included the mouth (see Figure 2).

The size of each region was set to encompass the critical features at all times, i.e., the mouth region was chosen to include the mouth both when opened and closed. Because the faces were constantly moving, the x/y positions of the regions of interest had to be dynamically adapted for each frame of the video. An in-house analysis tool, Gazeatron, built in Shockwave (Adobe Ltd., San Jose, CA), was used to puppet the dynamic regions of interest (dROI) over the videos and ensure dynamic translation. Gazeatron allows both the creation of dROIs on videos and the matching of fixations to region to create region-of-interest reports. Overlaps between eye, nose, and mouth regions were avoided where possible. When fixations fell on overlapping dROIs, these were added to the fixation count of both areas.

Event tagging

In addition to marking dynamic regions of interest, each video frame was tagged according to the following dynamic events related to the people depicted in the videos: (a) TALKING: yes versus no, (b) EYE CONTACT: gaze directed at the camera versus away from the camera, and (c) HEAD MOVEMENT: yes versus no. These events were separately coded for each face. Note that the three types of events may co-occur for each face. For example, a face might be talking, looking at the camera, and moving at the same time. However, to simplify our analyses, we will only discuss one event at a time. Analysis of the interacting effects

Movie 1. The Brooklyn video clip taken from the Dynamic Images and Eye Movements (DIEM) database with dynamic heat map overlays that represent the distribution of fixation points of all 44 observers at any given frame. The upper panel shows fixation distributions in the vocal condition, while the lower panel shows fixation distributions in the mute condition (no speech stream).

Movie 2. The London video clip taken from the Dynamic Images and Eye Movements (DIEM) database with dynamic heat map overlays that represent the distribution of fixation points of all 44 observers at any given frame. As in Movie 1, the upper panel shows fixation distributions in the vocal condition, while the lower panel shows fixation distributions in the mute condition (no speech stream).

of multiple events will be investigated in future research.

Results

The two movies below show the Brooklyn (Movie 1) and the London (Movie 2) video clips with dynamic heat maps generated on the basis of fixation distribution from our 44 observers. The upper panels show gaze behavior in the vocal condition, which includes dialogue and background music, while the lower panel shows fixation distributions in the mute condition, which was visually identical to the vocal condition, but contained no speech stream. The differences in dynamic heat maps between upper and lower panels indicate differences in our observers' attention allocation as a function of vocal information being either present or absent from the movie clips.

In the following analyses we report the percentage of fixations directed to the dROIs (eyes, nose, and mouth) separately for each audio condition (vocal vs. mute) during instances of different dynamic events. Each analysis of variance (ANOVA) therefore included regions and dynamic events as within-subject factors. For each dynamic event, we first report ANOVAs on vocal conditions followed by ANOVAs on mute conditions. For better comparison with previous studies, event analyses were based on instances in which only one face was present in a scene (73% of all frames). Analyses of overall looking behavior were followed by more fine-grained analyses of three dynamic events: (a) talking, (b) eye contact, and (c) head movements.

Overall gaze distribution

The percentage of all fixations falling within the three dROIs (eyes, nose, and mouth) was calculated for all participants in both the vocal and mute conditions (see Figure 3). Performing a repeated-measures ANOVA between the three fixation percentages and across audio conditions did not result in an overall main effect of region, $F(2, 43) < 1$. This lack of a difference across regions can be seen in Figure 3, with no evidence of an overall prioritization of the eyes (shown in blue) when

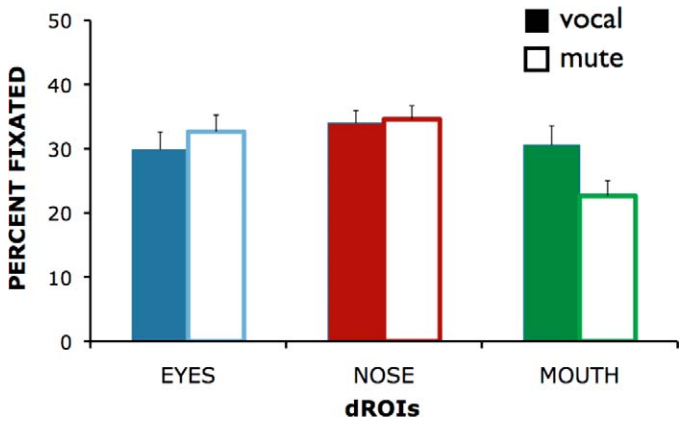

Figure 3. Overall gaze distributions in percent (*SE*) to dynamic regions of interest (eyes = blue, nose = red, mouth = green) as a function of vocal (filled) and mute (outlined) conditions.

viewing dynamic faces. The following analyses further investigate overall gaze distributions separately for vocal and mute conditions.

Vocal

Overall, 87% of fixations during the video clips targeted the face region. Further, there was no bias to preferentially look at one face region more than another. That is, eyes, nose, and mouth regions were looked at equally often, $F(2, 43) < 1$.

Mute

In contrast to the vocal condition, gaze distributions significantly differed across face regions, $F(2, 43) = 5.93$, $p < 0.01$, $p\eta^2 = 0.12$. Fixations to the mouth were significantly lower compared to the eyes, $t(43) = 2.15$, $p < 0.05$, as well as to the nose, $t(43) = 3.71$, $p < 0.01$, whereas fixations to the eyes and the nose did not differ, $t(43) < 1$.

Vocal versus mute

When acoustic speech was removed from the audio stream, overall face fixations dropped from 87% to 82% in the mute condition, $t(86) = 2.89$, $p < 0.01$, implying increased fixations to the scene background compared to the vocal condition. Contrary to Lansing and McConkie (2003), the absence of vocal information resulted in decreased attention to the mouth region, $t(86) = 2.08$, $p < 0.05$. As we hypothesized, our instructions to rate the likeability of each video might have decreased the importance of the mouth region especially when no auditory information was present. Fixations to the eyes and nose, on the other hand, did not significantly diminish when vocal information was missing, both ts < 1.

These overall gaze distributions clearly show an equal allocation of attention across all face regions rather than a prioritization of the eyes. In the absence of acoustic speech information, faces are looked at slightly less, mainly due to a decrease in mouth fixations, which implies that in our study the mouth lost its importance as a source of information when no speech was audible.

Eye contact

In the following analyses we looked at the effect of observed eye contact. We compared fixation distributions as a function of whether the depicted face was making eye contact with the observer by means of directly looking at the camera, or whether the face was looking elsewhere. An example of face gaze distributions as a function of eye contact can be seen in Figure 4.

Vocal

There was no significant difference in gaze allocation between regions, $F(1, 43) < 1$, but there was a significant increase in fixations on face regions in general with eye contact compared to without eye contact, $F(1, 43) = 4.94$, $p < 0.05$, $p\eta^2 = 0.10$, as well as a significant interaction, $F(2, 43) = 20.09$, $p < 0.01$, $p\eta^2 = 0.32$ (see Figure 5a). The interaction was characterized by an increase in fixations to the eyes when the depicted face made eye contact, $t(43) = 4.93$, $p < 0.01$, whereas fixations to the mouth region decreased, $t(43) = 4.32$, $p < 0.01$, and fixations on the nose remained unaffected, $t(43) = 1.69$, $p = 0.95$.

Mute

In contrast to the vocal condition, we observed significant differences in gaze allocations across regions in the mute condition, $F(1, 43) = 5.22$, $p < 0.01$, $p\eta^2 = 0.11$. Mouth regions were fixated less than both eye and nose regions, $t(43) = 2.02$, $p < 0.05$ and $t(43) = 3.43$, $p < 0.01$. There was also a significant decrease of fixations with eye contact, $F(1, 43) = 18.36$, $p < 0.01$, $p\eta^2 = 0.30$, as well as a significant interaction, $F(2, 43) = 8.19$, $p < 0.01$, $p\eta^2 = 0.16$ (see Figure 5b). The interaction was characterized by a decrease of fixations to the mouth, $t(43) = 5.87$, $p < 0.01$, whereas fixations to the nose and eye regions remained unaffected, both ts < 1, when the depicted face made eye contact.

Together these findings show that eye contact had a strong effect on the allocation of gaze for both the eyes and mouth regions: Fixations to the mouth decreased with increasing fixations to the eyes when the observed face gazed at the camera making eye contact with the viewer. While a decrease of mouth fixations was also

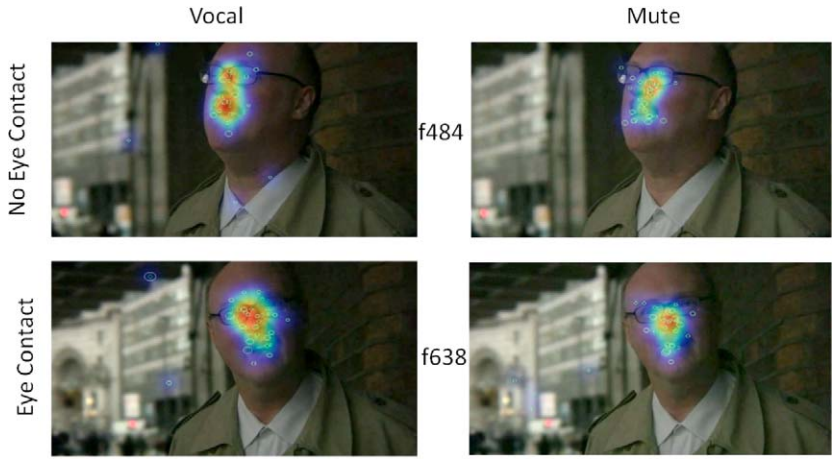

Figure 4. Example fixation heat maps of 44 observers viewing a face making eye contact versus no eye contact in either the vocal or mute condition. Heat map represents the distribution of fixation points across a particular frame: top = frame 484 of the London video; bottom = frame 638. Warmer colors indicate greater clustering.

observed in the mute condition, eye fixations remained unaffected by eye contact when speech information was absent. Eye contact, therefore, seems to exhibit stronger modulations of gaze distributions when observers are able to understand what the depicted face is saying.

Talking

The following analyses contrast instances when a face was talking as part of the interview versus when the depicted face was not talking. For example, a talking face might draw observers' gaze to the mouth region to support speech comprehension. In addition, we examined whether the visual talking modulation is differentially affected by the presence or absence of the acoustic speech signal. Lansing and McConkie (2003) found an increase of mouth fixations when sound was taken away, probably due to the high motivation of the participants to try to understand what was being uttered by increased lip reading. Our study did not require accurate speech perception. We therefore hypothesized that gaze should increase to the mouth region when acoustic speech is present, while we should see less modulation when speech is absent. An example of face distributions as a function of eye contact can be seen in Figure 6.

Vocal

While there was no significant difference in gaze allocation across regions, $F(2, 43) < 1$, and no main effect of talking, $F(1, 43) < 1$, these factors significantly

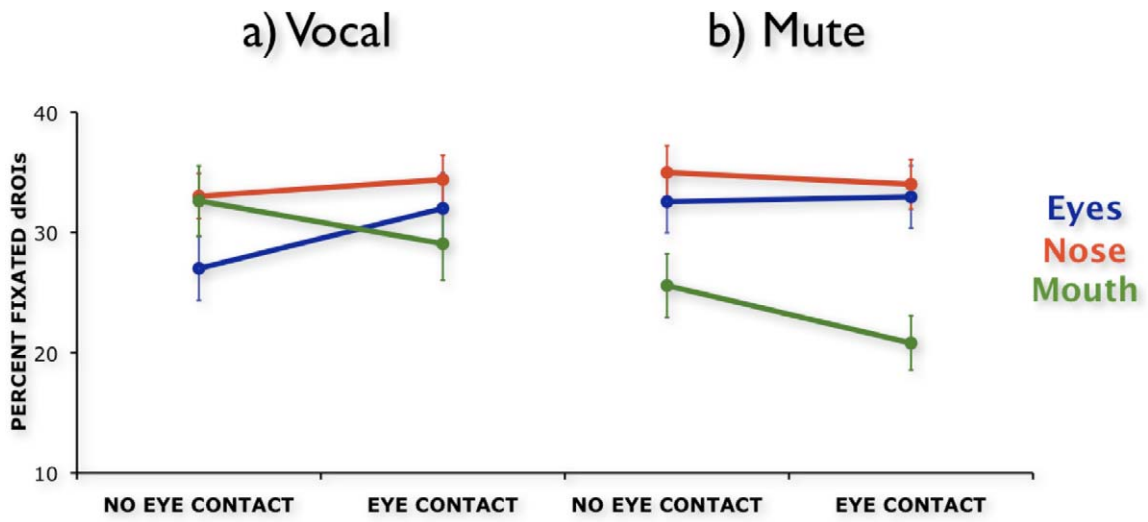

Figure 5. Mean percentage of fixations to dynamic regions of interest (eyes = blue, nose = red, mouth = green) as a function of eye contact or no eye contact for the (a) vocal and (b) mute condition.

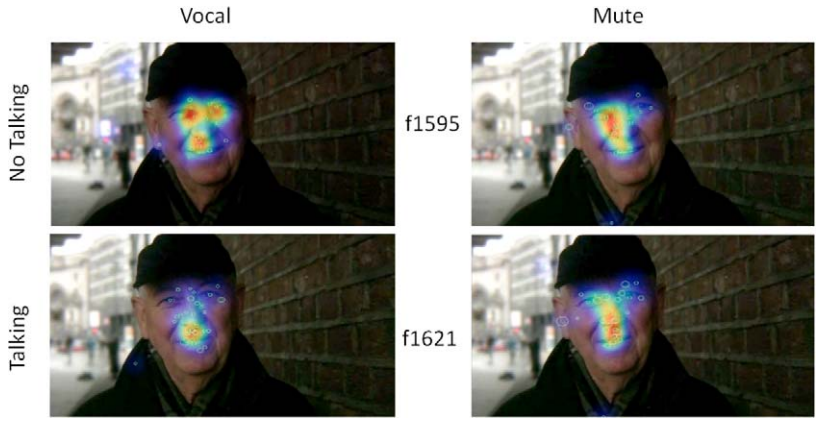

Figure 6. Example of fixation heat maps of 44 observers viewing a face either talking or not talking in the vocal versus the mute condition. Heat map represents the distribution of fixation points across a particular frame: top = frame 1595 of the London video; bottom = frame 1621. Warmer colors indicate greater clustering.

interacted, $F(2, 43) = 13.31$, $p < 0.01$. As can be seen in Figure 7a, the interaction was characterized by an increase in fixations to the mouth, $t(43) = 4.01$, $p < 0.01$, along with a decrease in fixations to the eyes, $t(43) = 4.31$, $p < 0.01$, when the depicted face was talking, whereas fixations to the nose remained constant, $t(43) < 1$.

Mute

In contrast to the vocal condition, we observed significant differences in gaze allocations across regions, $F(1, 43) = 6.44$, $p < 0.01$, $p\eta^2 = 0.13$. Mouth regions were fixated less than both eye and nose regions, $t(43) = 2.32$, $p < 0.05$ and $t(43) = 3.79$, $p < 0.01$. There was also a small but significant decrease in fixations when the depicted person was talking (no talking: 31% vs. talking 29%), $F(1, 43) = 15.36$, $p < 0.01$, $p\eta^2 = 0.26$, as well as a significant interaction, $F(2, 43) = 4.44$, $p < 0.01$, $p\eta^2 = 0.09$ (see Figure 7b). The interaction was characterized by a decrease in fixations to the eyes, $t(43) = 4.22$, $p < 0.01$, and the nose, $t(43) = 2.02$, $p < 0.05$, whereas fixations to the mouth region remained unaffected, $t < 1$.

In sum, comparing instances of when a face was talking during the interview to instances when a face was not talking showed great differences in the distribution of gaze across face regions: Fixations to the eyes decreased with increasing fixations to the mouth when the observed face was talking. We saw no such modulations when the interviews were watched without audible speech, which argues against a purely visual saliency-driven control of eye movements, but instead underlines the functional use of looking at a face's mouth to support speech comprehension when speech information is present.

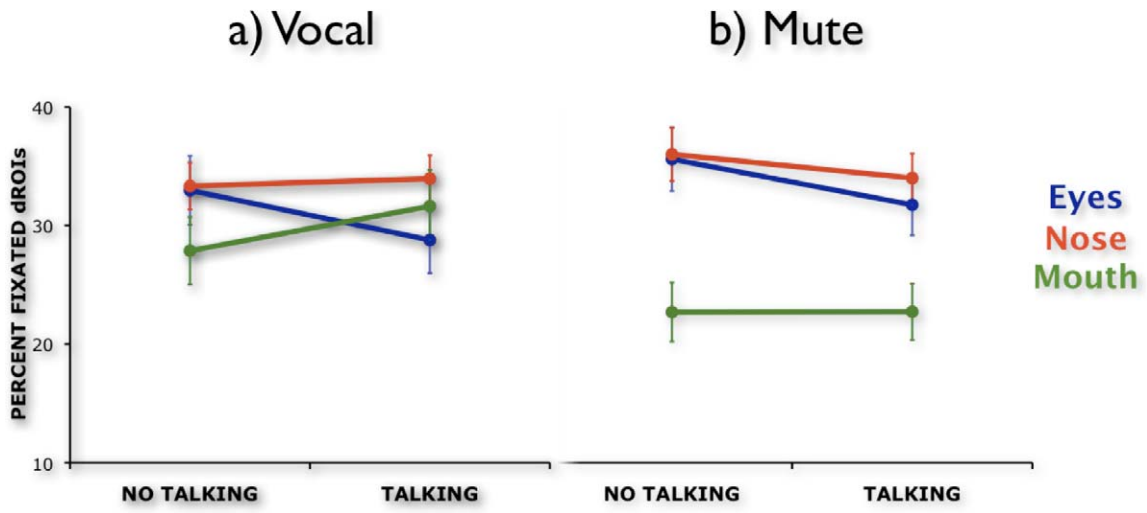

Figure 7. Mean percentage of fixations to dynamic regions of interest (eyes = blue, nose = red, mouth = green) as a function of talking or no talking for the (a) vocal and (b) mute condition.

Figure 8. Example of fixation heat maps of 44 observers viewing a face that is either moving or not moving in the vocal versus the mute condition. Heat map represents the distribution of fixation points across a particular frame: top = frame 2815 of the London video; bottom = frame 2830. Warmer colors indicate greater clustering.

Head movement

In order to test whether fixation distributions are modulated by head movements, we analyzed gaze distributions across face regions as a function of instances in which the depicted face was moving or not. Examples of fixation distributions on faces that were either moving their heads or not are shown in Figure 8.

Vocal

There was no main effect of gaze allocation across regions, $F(2, 43) < 1$, whereas fixations on face regions overall increased during head movement, $F(1, 43) = 20.22$, $p < 0.01$, $p\eta^2 = 0.32$. As can be seen in Figure 9a, there was a tendency for an interaction, $F(2, 43) = 2.86$, $p = 0.06$, $p\eta^2 = 0.05$, due to increased nose fixations during head movements, $t(43) = 3.39$, $p < 0.01$, evident to a lesser degree to the mouth, $t(43) = 2.00$, $p = 0.05$, whereas fixations to the eyes remained unaffected, $t < 1$.

Mute

While we found significantly different degrees of gaze allocation across regions, $F(2, 43) = 6.41$, $p < 0.01$, $p\eta^2 = 0.13$, there was no main effect of head movement and no interaction, both Fs < 1 (see Figure 9b).

While nose fixations remained unaffected during instances of eye contact or talking, head movements led to an increase of fixations to the nose. Adopting a center bias during rapid movements of the face might support the tracking of the face as a whole to maintain an optimal viewing position from which to target either the eyes or the mouth depending on the subsequent

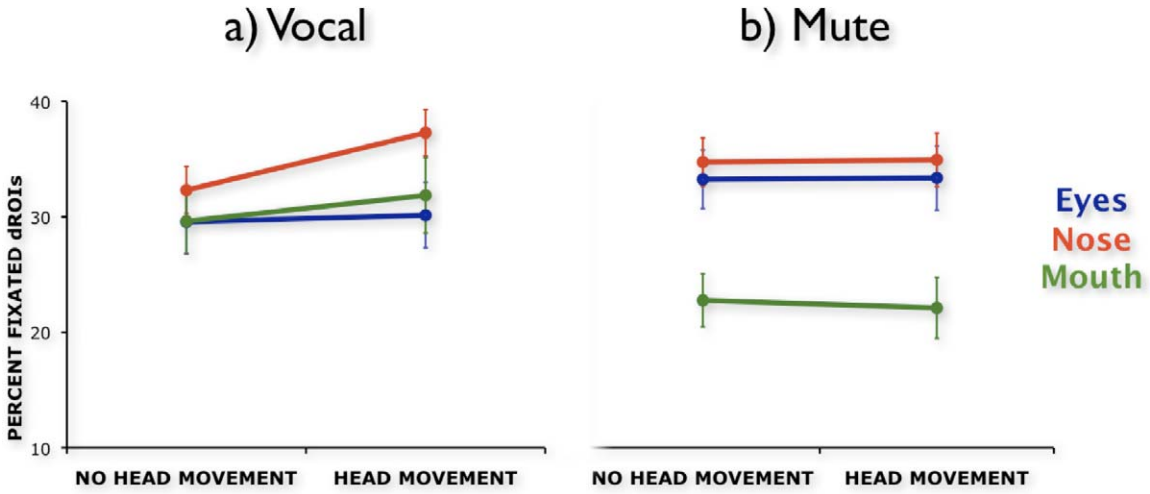

Figure 9. Mean percentage of fixations to dynamic regions of interest (eyes = blue, nose = red, mouth = green) as a function of whether the head was moving or not moving for the a) vocal and b) mute condition.

event. The lack of such modulations in the mute condition implies that keeping track of a moving face might be especially functional when following a conversation.

Discussion

Is there a general bias to look at someone's eyes? According to our results the short answer would have to be "no." Rather than a general bias to look at someone's eyes (see Birmingham et al., 2008a, 2008b, 2009a, 2009b; Emery, 2000; Friesen & Kingstone, 1998; Henderson et al., 2005; Itier et al., 2007; Walker-Smith et al., 1977; for reviews, see Birmingham & Kingstone, 2009; Langton et al., 2000), gaze is differentially allocated toward different parts of a face depending on current behavior such as making eye contact, moving, and speaking. Especially in dynamically changing situations, attention appears to be directed to locations that provide useful information on a moment-by-moment basis.

Optimal viewing positions during dynamic face viewing

Our visual apparatus has evolved to become a system highly optimized for efficient information uptake that can be adapted to many different tasks. When performing real-world tasks, such as making tea or a sandwich, gaze precedes actions to allow for relevant information uptake "just in time" (see Ballard, Hayhoe, & Pelz, 1995; Hayhoe & Ballard, 2005).

From word recognition, reading, and scene viewing studies, we know that words (Conrad, Võ, Schneider, & Jacobs, 2011; McConkie et al., 1988; O'Regan, Lévy-Schoen, Pynte, & Brugaillère, 1984; Rayner, 1979) as well as objects (Foulsham & Underwood, 2009; Henderson, 1993; Nuthmann & Henderson, 2010) have preferential viewing locations close to their center. Similarly, faces seem to be initially fixated at their center, i.e., the nose, before gaze is moved to other locations of interest (Hsiao & Cottrell, 2008). However, optimal sampling of facial features obviously depends on the information that is sought.

Our study provides evidence for efficient and dynamically adjusted sampling of facial features when viewing socially interacting faces. Rather than identifying *one* optimal viewing position, there are several quickly changing positions that are preferentially chosen for fixation during face viewing, the position of which greatly depends on the currently depicted dynamic event. For instance, we found that as soon as eye contact is made, the viewer tends to look more at the person's eyes, whereas gaze is drawn away from the eyes and directed toward the mouth when someone starts talking. This dynamic adjustment to either prioritize the eyes or the mouth has previously been reported by Buchan et al. (2007), who found that observers looked more at the eyes when their task was to make emotional judgments, but looked to the mouth when asked to report what the depicted face had uttered. Similarly, Lansing and McConkie (2003) argued that attention is drawn to the face region that is the richest information source. They referred to this strategy as the "information source attraction effect." In the attempt to support speech perception, a perceiver often (1) disengages attention from the talker's eyes, (2) moves gaze to a new, information rich location (in this case, the mouth), and (3) engages attention at this new location. Which parts of a face contain the most beneficial information depends on the raw sensory features of the face at a particular moment and the current agenda or task of the observer (e.g., Gosselin & Schyns, 2001).

While eyes and mouth provide ample emotional and verbal information, the nose does not contain information per se. However, coinciding with the center of the face and equidistant from eyes and mouth, the nose seems to be a strategically optimal vantage point from which gaze can rapidly be moved to either the eyes or the mouth. Also when fixating the nose, the visual parafovea encompasses both eyes and mouth, which is not the case when fixating the eyes. Our data show that fixations to eye and mouth regions were greatly affected by the current event, for example eye contact or talking, but nose fixations remained mostly unaffected. This result supports the notion that the nose is used as a default fixation point for moving to the eyes and mouth. However, we also found that rapid head movements selectively attracted overt attention to the nose, whereas eyes and mouth did not receive more fixations during these movements. This result suggests that the nose may serve as a fixation anchor when faces move rapidly (see Buchan et al., 2007). Noses are in the center, hard to conceal, relatively invariant to facial expression (Moorhouse, Evans, Atkinson, Sun, & Smith, 2009), and might therefore be especially suitable for tracking moving faces.

In sum, it appears that gaze allocation during face viewing is dynamically adjusted for the purpose of seeking information on an event-to-event basis.

Modulations of face viewing by auditory information

Since dynamic visual stimuli exhibit strong directional and motion cues, it has not been clear to what degree attention allocation in dynamic face viewing is driven by visually salient features like the simple movements of the mouth when talking (see Itti, 2005).

Changing the auditory information present during face viewing enabled us to manipulate the presence of speech while holding the visual input constant. If the presence of speech plays a crucial role in face viewing, manipulations of the acoustic input should greatly affect observers' gaze. Buchan et al. (2007) had people watch videos of expressive talking faces and found that when the intelligibility of the speech was decreased by the addition of acoustic noise, observers adopted a vantage point centered on the face, i.e., they reduced the frequency of gaze fixations on the eyes and lengthened their fixation durations on the nose and the mouth. Research by Vatikiotis-Bateson et al. (1998) has also shown that the number of transitions between areas of the face decreases in the presence of noise during a speech task, suggesting that the availability of the verbal information may affect how the information is gathered.

In our study, we completely removed acoustic speech. We found that when the speech sound was not present, observers looked less at the depicted faces and more to the scene background. Given our very lightly constrained free-viewing task, the lack of acoustic speech apparently rendered the faces less important, as reflected in an overall drop of fixations from 87% to 82% when the acoustic speech signal was absent.

Within a face, fixations to the mouth region decreased from 31% when viewing videos with the speech signal present to only 23% when acoustic speech was absent. Observers moved their gaze away from the mouth and towards the eyes and nose. It seems that the mouth region no longer provided sufficiently important information to attract gaze. This runs counter to findings by Lansing and McConkie (2003), who reported that a video watched in silence led observers to further increase their gaze toward the mouth as compared to periods of low-level speech. This was probably due to differences in task demands. Participants in the Lansing and McConkie study were highly motivated to retrieve any information from the images that could support their main task of speech reproduction. In addition, Lansing and McConkie's participants had at least some natural proficiency in visual speech perception. Therefore, looking at the mouth provided an additional source of information to support speech perception. This slightly differs from Buchan et al.'s (2007) result, which found that adding noise to the speech stream led to greater central fixation biases rather than increased fixations to the mouth as would have been predicted by the increased necessity for lip reading. Buchan and colleagues argued that a reason for this difference in gaze allocation was due to the fact that their sentences were emotionally engaging and might therefore have resulted in a greater bias to monitor the speaker's eyes while trying to understand what was being said.

Together with previous studies, our results illustrate that the movement of the mouth does not alone attract attention, as would be predicted by an explanation based on capture by visual salience (Itti & Koch, 2000). Instead, fixations on the mouth are modulated by the degree to which they provide task-relevant information.

Conclusions

Living in a complex, dynamically changing world requires a visual system that is able to effectively gather information from a constantly changing environment. With a highly engaging and realistic set of videos showing people in the real world and the inclusion of dynamic events, such as making eye contact and performing head movements, we were able to monitor visual attention in response to moment-to-moment changes in facial dynamics. Our results show that during dynamic face viewing, rather than being predominantly directed towards the eyes, "gaze follows function" by rapidly directing attention to different face regions on the basis of information-seeking control processes in interaction with dynamic events.

Acknowledgments

This research was supported by grant Ref F/00-158/BZ from the Leverhulme Trust and grant BCS-1151358 from the National Science Foundation to JMH, and DFG:VO 1683/1-1 to MLV. We'd like to thank Robin Hill and Laura Speed for data collection and both Keith Rayner and an anonymous reviewer for their helpful comments on this paper.

Commercial relationships: none.
Corresponding author: Melissa Le-Hoa Võ.
Email: mlvo@search.bwh.harvard.edu.
Address: Visual Attention Lab, Harvard Medical School, Cambridge, MA, USA.

References

Ambadar, Z., Schooler, J. W., & Cohn, J. F. (2005). Deciphering the enigmatic face: The importance of facial dynamics in interpreting subtle facial expressions. *Psychological Science*, *16*, 403–410.

Ballard, D. H., Hayhoe, M. M., & Pelz, J. B. (1995).

Memory representations in natural tasks. *Journal of Cognitive Neuroscience*, *7*(1), 66–80.

Birmingham, E., Bischof, W. F., & Kingstone, A. (2008a). Gaze selection in complex social scenes. *Visual Cognition*, *16*(2/3), 341–355.

Birmingham, E., Bischof, W. F., & Kingstone, A. (2008b). Social attention and real world scenes: The roles of action, competition, and social content. *Quarterly Journal of Experimental Psychology*, *61*(7), 986–998.

Birmingham, E., Bischof, W., & Kingstone, A. (2009a). Get real! Resolving the debate about equivalent social stimuli. *Visual Cognition*, *17*, 904–924.

Birmingham, E., Bischof, W. F., & Kingstone, A. (2009b). Saliency does not account for fixations to eyes within social scenes. *Vision Research*, *49*, 2992–3000.

Birmingham, E., & Kingstone, A. (2009). Human social attention. *Annals of the New York Academy of Sciences*, *1156*(1), 118–140.

Buchan, J. N., Paré, M., & Munhall, K. G. (2007). Spatial statistics of gaze fixations during dynamic face processing. *Social Neuroscience*, *2*, 1–13.

Buchan, J. N., Paré, M., & Munhall, K. G. (2008). The effect of varying talker identity and listening conditions on gaze behavior during audiovisual speech perception. *Brain Research*, *1242*, 162–171.

Buswell, G. T. (1935). *How people look at pictures: A study of the psychology and perception in art*. Chicago: University of Chicago Press.

Conrad, M., Võ, M. L.-H., Schneider, D., & Jacobs, A. M. (2011). Syllable structure is modulating the optimal viewing position in visual word recognition. *Fonética*, *31*(1), 2–13.

Eisenbarth, H., & Alpers, G. W. (2011). Happy mouth and sad eyes: Scanning emotional facial expressions. *Emotion*, *11*(4), 860–865.

Emery, N. (2000). The eyes have it: The neuroethology, function and evolution of social gaze. *Neuroscience and Biobehavioral Reviews*, *24*, 581–604.

Farroni, T., Csibra, G., Simion, F., & Johnson, M. H. (2002). Eye contact detection in humans from birth. *Proceedings of the National Academy of Sciences USA*, *99*, 9602–9605.

Foulsham, T., Cheng, J. T., Tracy, J. L., Henrich, J., & Kingstone, A. (2010). Gaze allocation in a dynamic situation: Speaking. *Cognition*, *117*, 319–331.

Foulsham, T., & Underwood, G. (2009). Does conspicuity enhance distraction? Saliency and eye landing position when searching for objects. *Quarterly Journal of Experimental Psychology*, *62*(6), 1088–1098.

Friesen, C., & Kingstone, A. (1998). The eyes have it! Reflexive orienting is triggered by nonpredictive gaze. *Psychonomic Bulletin and Review*, *5*(3), 490–495.

Gosselin, F., & Schyns, P. G. (2001). Bubbles: A technique to reveal the use of information in recognition tasks. *Vision Research*, *41*(17), 2261–2271.

Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences*, *9*(4), 188–194.

Henderson, J. M. (1993). Eye movement control during visual object processing: Effects of initial fixation position and semantic constraint. *Canadian Journal of Experimental Psychology*, *47*, 79–98.

Henderson, J. M. (2006). Regarding scenes. *Current Directions in Psychological Science*, *16*(4), 219–222.

Henderson, J. M., Malcolm, G. L., & Schandl, C. (2009). Searching in the dark: Cognitive relevance drives attention in real-world scenes. *Psychonomic Bulletin & Review*, *16*(5), 850–856.

Henderson, J. M., Williams, C. C., & Falk, R. (2005). Eye movements are functional during face learning. *Memory & Cognition*, *33*(1), 98–106.

Hill, H., & Johnson, A. (2001). Categorizing sex and identity from the biological motion of faces. *Current Biology*, *11*, 880–885.

Hsiao, J. H., & Cottrell, G. (2008). Two fixations suffice in face recognition. *Psychological Science*, *19*(10), 998.

Itier, R. J., Villate, C., & Ryan, J. D. (2007). Eyes always attract attention but gaze orienting is task-dependent: Evidence from eye movement monitoring. *Neuropsychologia*, *45*, 1019–1028.

Itti, L. (2005). Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Visual Cognition*, *12*(6), 1093–1123.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, *40*, 1489–1506.

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

Knappmeyer, B., Thornton, I. M., & Bülthoff, H. H. (2003). The use of facial motion and facial form during the processing of identity. *Vision Research*, *43*, 1921–1936.

Lander, K., & Bruce, V. (2000). Recognizing famous faces: Exploring the benefits of facial motion. *Ecological Psychology*, *14*, 385–388.

Langton, S., Watt, R., & Bruce, I. (2000). Do the eyes have it? Cues to the direction of social attention. *Trends in Cognitive Sciences*, *4*(2), 50–59.

Lansing, C. R., & McConkie, G. W. (1999). Attention to facial regions in segmental and prosodic visual speech perception tasks. *Journal of Speech, Language, & Hearing Research*, *42*, 526–539.

Lansing, C. R., & McConkie, G. W. (2003). Word identification and eye fixation locations in visual and visual-plus-auditory presentations of spoken sentences. *Perception & Psychophysics*, *65*, 536–552.

McConkie, G. W., Kerr, P. W., Reddix, M. D., & Zola, D. (1988). Eye movement control during reading: I. The location of the initial eye fixations on words. *Vision Research*, *28*, 1107–1118.

Mital, P. K., Smith, T. J., Hill, R. L., & Henderson, J. M. (2011). Clustering of gaze during dynamic scene viewing is predicted by motion. *Cognitive Computation*, *3*(1), 5–24.

Moorhouse, A., Evans, A. N., Atkinson, G. A., Sun, J., & Smith, M. L., (2009). The nose on your face may not be so plain: Using the nose as a biometric. *Audio, Transactions of the IRE Professional Group on,* 1–6.

Najemnik, J., & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature*, *434*(7031), 387–391.

Najemnik, J., & Geisler, W. S. (2008). Eye movement statistics in humans are consistent with an optimal search strategy. *Journal of Vision*, *8*(3):4, 1–14, http://www.journalofvision.org/content/8/3/4, doi: 10.1167/8.3.4. [PubMed] [Article]

Nuthmann, A., & Henderson, J. M. (2010). Object-based attentional selection in scene viewing. *Journal of Vision*, *10*(8):20, 1–19, http://www.journalofvision.org/content/10/8/20, doi:10.1167/10.8.20. [PubMed] [Article]

O'Regan, J. K., Lévy-Schoen, A., Pynte, J., & Brugaillère, B. (1984). Convenient fixation location within isolated words of different length and structure. *Journal of Experimental Psychology: Human Perception & Performance*, *10*, 250–257.

O'Toole, A. J., Roark, D. A., & Abdi, H. (2002). Recognizing moving faces: A psychological and neural synthesis. *Trends in Cognitive Sciences*, *6*, 261–266.

Rayner, K. (1979). Eye guidance in reading: Fixation locations in words. *Perception*, *8*, 21–30.

Rosenblum, L. D., Johnson, J. A., & Saldaña, H. M. (1996). Visual kinematic information for embellishing speech in noise. *Journal of Speech & Hearing Research*, *39*, 1159–1170.

Senju, A., & Hasegawa, T. (2005). Direct gaze captures visuospatial attention. *Vision Cognition*, *12*, 127–144.

Senju, A., Hasegawa, T., & Tojo, Y. (2005). Does perceived direct gaze boost detection in adults and children with and without autism? The stare-in-the-crowd effect revisited. *Visual Cognition*, *12*, 1474–1496.

Senju, A., & Johnson, M. H. (2009). The eye contact effect: Mechanisms and development. *Trends in Cognitive Sciences*, *13*(3), 127–134.

Smith, M. L., Gosselin, F., Cottrell, G. W., & Schyns, P. G. (2005). Transmitting and decoding facial expressions. *Psychological Science*, *16*, 184–189.

Stein, T., Senju, A., Peelen, M. V., & Sterzer, P. (2011). Eye contact facilitates awareness of faces during interocular suppression. *Cognition*, *119*(2), 307–311.

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

von Grünau, M., & Anston, C. (1995). The detection of direct gaze: A stare-in-the-crowd effect. *Perception*, *24*, 1297–1313.

Walker-Smith, G. J., Gale, A. G., & Findlay, J. M. (1977). Eye movement strategies involved in face perception. *Perception*, *6*, 313–326.

Williams, C. C., & Henderson, J. M. (2007). The face inversion effect is not a consequence of aberrant eye movements. *Memory & Cognition*, *35*(8), 1977–1985.

Yarbus, A. L. (1967). Eye movements and vision (B. Haigh, Trans.). New York: Plenum Press. (Original work published 1965).