Chapter 10
Co-present or Not?

Embodiment, Situatedness and the Mona Lisa Gaze Effect

Jens Edlund, Samer Al Moubayed, and Jonas Beskow

Abstract The interest in *embodying* and *situating* computer programmes took off in the autonomous agents community in the 90s. Today, researchers and designers of programmes that interact with people on human terms endow their systems with humanoid physiognomies for a variety of reasons. In most cases, attempts at achieving this embodiment and situatedness has taken one of two directions: virtual characters and actual physical robots. In addition, a technique that is far from new is gaining ground rapidly: projection of animated faces on head-shaped 3D surfaces. In this chapter, we provide a history of this technique; an overview of its pros and cons; and an in-depth description of the cause and mechanics of the main drawback of 2D displays of 3D faces (and objects): the Mona Liza gaze effect. We conclude with a description of an experimental paradigm that measures perceived directionality in general and the Mona Lisa gaze effect in particular.

10.1 Introduction

The interest in *embodying* and *situating* computer programmes took off in the autonomous agents community in the 90s (e.g. Steels and Brooks 1995). Today, researchers and designers of programmes that interact with people on human terms—most notably using speech in human-machine dialogue and computer-mediated human-human dialogue—endow their systems with humanoid physiognomies for a variety of reasons, ranging from a hope to exploit the purported benefits of humanlike dialogue as a human-machine interface—people know how to speak and many of us are most comfortable communicating face—to a desire to use speech technology and working models of human dialogue to gain deeper understanding of how people communicate (Traum 2008; Edlund 2011).

In most cases, attempts at achieving this embodiment and situatedness has taken one of two directions. The first is to implement virtual characters, often referred to as virtual humans (e.g. Traum and Rickel 2002) or embodied conversational agents

J. Edlund (✉) · S. Al Moubayed · J. Beskow
KTH Speech, Music and Hearing, Lindstedtsvägen 24, 100 44 Stockholm, Sweden
e-mail: edlund@speech.kth.se

Y.I. Nakano et al. (eds.), *Eye Gaze in Intelligent User Interfaces*,
DOI 10.1007/978-1-4471-4784-8_10,

(ECAs; e.g. Cassel et al. 2000). We will adhere to the latter terminology and use ECA here. In principal, an ECA is a 2D or 3D model of a character in virtual space, that is displayed as a 2D rendition on a monitor in physical space. The relationship between the ECA and its virtual space, the monitor, and the humans watching the ECA can be portrayed in several ways by the ECA designer. Common images include the ECA living its life in another world, which is displayed to the onlookers as if it were a movie, as exemplified by Cloddy Hans and Karen in the NICE project (Boye and Gustafson 2005); the ECA again living in its virtual reality but peering out through a window (the monitor) through which the onlookers peer back in, as exemplified by Ville in the DEAL system (Hjalmarsson et al. 2007); and the ECA not living in a virtual world at all, but rather sharing the same physical world as the onlookers, as exemplified by MACK (Cassell et al. 2002) and the characters of the Gunslinger project (Hartholt et al. 2009).

The other main direction is to implement actual physical robots which imitate people, such as MIT's Cog and Kismet (Breazeal and Scassellati 2001). Honda Research's robot Asimo is also merits mention in this context. Although not chiefly developed for communication studies, recent work on retargeting of motion captured from humans allows Asimo to reproduce human gestures quite closely in real-time, which opens up new possibilities for investigations into human communicative gesture (Dariush et al. 2008).

In studies of face-to-face communication, the head and face are often given centre stage. This is particularly true for head pose and gaze, as these are associated with a number of important communicative functions such as turn-taking and grounding. In addition, a number of other features of the head and face—for example facial expressions, eye brow movements, and lip synchronization—frequently receive special attention. In this chapter, we focus on a particular type of face and head embodiment which is becoming increasingly popular—a combination of a virtual talking head and a physical robotic head: projection of animated faces on head-shaped 3D surfaces.

The following section provides a history of the technique. After that, the next section holds a brief overview of the pros and cons of the technique compared to other methods of embodiment, followed by a section providing an in-depth description of the main drawback of 2D displays of 3D faces (and objects): the Mona Liza gaze effect. This section includes a proposed explanation of the cause and mechanics of the effect; an examination of its consequences for face-to-face communication; a description of an experimental paradigm that measures perceived directionality in general and the Mona Lisa gaze effect in particular. Finally, the chapter is summed up with an account of how projection on head-shaped 3D surfaces completely cancels the effect.

10.2 Face Projection on 3D Surfaces

The method of embodiment that is our focus here is sometimes called *relief projection*, and the result is sometimes called a *projection augmented model*. In short,

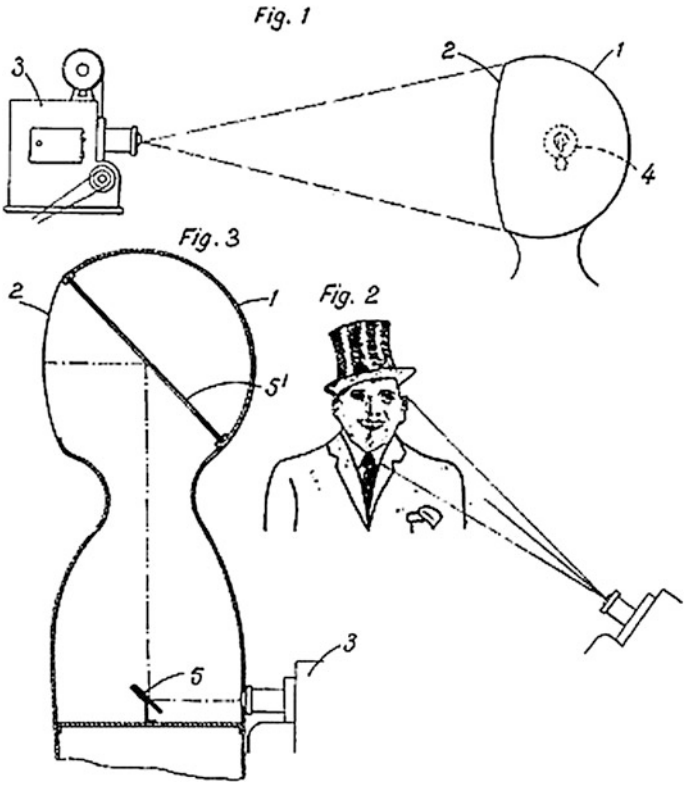

Fig. 10.1 Drawings of front and back projected faces taken from US Patent 1653180 of 1925. The drawing is in the public domain and copyright free, like all patents issued in the United States

a photographic image of an object is projected on a physical, three-dimensional model of the same shape as the object with the aim of creating a realistic-looking 3D object with properties that can be changed by manipulating the image. In the cases we are interested in, the image is a moving image, either a film of a person's face or a generated face such as those used for ECAs.

The earliest attestation of this technique is a patent application submitted by Georges Jalbert in December 1924 (France) and May 1925 (US; Jalbert 1925). The application describes both front projected faces and faces projected from the inside of a translucent bust, as seen in Fig. 10.1. Another patent (Liljegren and Foster 1989) specifically adds fibre optics as the means of transferring the images to within the bust.

The first modern well-documented implementation of face projection on 3D surfaces is the ghosts performing Grim Grinning Ghosts at the Disneyland Haunted Mansion ride. The ride was built in the 60s and opened in 1969, and the technology was described in a 1970 behind-the-scenes TV feature called Disneyland

Showtime,[1] which incidentally also features facial animatronics that seemingly measures up to MIT's Kismet. The ghosts are created by projecting films of strongly lit, highly contrasted faces against a black background onto relatively featureless white busts with shapes matching the faces in the films. Disney's ghosts, as well as another Haunted Mansion character produced with a similar technique, Madame Leota, are popular projections in private Halloween installations featuring face projection on busts. Films showing such installations dating from some time into the 2000s and onwards are easily found on the Internet, and a number of amateur special effects makers claim having produced them as early as the early 80s. Although there are several claims of proof in the form of footage and films, we have not been able to find any of these materials.

Another early and well-attested creation is the talking head projection of MIT's Architecture Machine Group in the early 80s. Inspired by Disney's Haunted Mansion creations. One of the creators of the MIT talking head projection, Michael Naimark, observed visitors at one of the talking heads in the Haunted Mansion at length. He concludes: "It was clear that as the woman spoke, the image of her moving lips would mis-register from the mask-shaped screen, but to most everyone viewing it briefly from their dark-ride car, this anomaly went unnoticed. Most people seemed convinced that they had just seen a full color, moving hologram (which, of course, is nonsense)" (Naimark 2005). The experience led Naimark and colleagues to develop the MIT talking head projection, an elaborate contraption which recorded not only image and sound, but also motion. The film was back projected to a head shaped mask moving in sync with the recorded person (Naimark 2005).[2] MIT Media Lab presented similar display in a tribute to the original experiments at their Defy Gravity exhibition in 2010.

In more recent years, a number of groups have put together 3D projection surfaces that are intended to be used with computer animated faces. From the more obscure prototype system HyperMask, which aims to project a face onto a mask worn by an actor who is free to move around in a room (Morishima et al. 2002) to more mundane ideas of embodying computer persons or improving telepresence. Hashimoto and Morooka (2006) use a spherical translucent projection surface and a back projected image of a humanoid face in combination with a robot neck. Light-Head of University of Plymouth (Delaunay et al. 2009) is more elaborately shaped, but maintains a stylized quality, while Technische Universität München's Mask-bot (Kuratate et al. 2011) and KTH Royal Technical Institute's Furhat (Al Moubayed et al. 2011) aim for higher degrees of human-likeness and project realistic faces into masks more closely resembling the human anatomy. It is worth mentioning that there are other ways to go as well, such as simplistic robot heads with small monitors for eyes and lips or the life-like mechatronic design of Hanson Robotics. And for the future, flexible and curved displays such as the spherical OLED displayed at

[1] Walt Disney's Wonderful World of Color, Season 16, Episode 20, Walt Disney Productions.

[2] Michael Naimark has made a film showing the talking head projection in action available at http://www.naimark.net/projects/head.html.

the Museum of Science and Education in Tokyo, and display techniques that utilize reflected light only, such as the Kindle, hold promise for development.

10.3 Pros and Cons of Face Projection

This section provides an overview of salient differences between face projection on 3D surfaces and the two main alternatives, physical robotic faces and 2D displays of 3D models.

10.3.1 Compared to Physical Robotic Faces

Compared to physical robot heads with moving parts for lips, eye brows, eyes, and other facial features, a projected face has a several advantages. To begin with, it is considerably cheaper to develop, and even more so to modify, as long as the modification can be done using the projection, rather than modifying the actual mask and other hardware. Development and in particular modification and adaptation is not only cheaper, but much faster, making face projection a much more feasible alternative for rapid development and experimentation, regardless of budget.

Another advantage is that projected movements—eye gaze shifts, brow raises, lip movements, and so on—are soundless in the projected face, whereas the hydraulics usually used in robotic components make noises that risks countering the humanlike impression of the robot. The projection is also able to make these movements more rapidly than robot actuators, at a speed that can easily match that with which a human produces them. An example of this is the SynFace lip synchronization (Beskow et al. 2009) that can be used with Furhat, allowing it to function as a remote representation of a human using the original voice of the human, but its own lip synchronization based on analysis of the acoustic signal, eliminating the need for a video stream in order to acquire lip movements.

As for disadvantages, there are two major drawbacks compared to robotic faces. The first one is that the light conditions needed for the back projection to be efficient are restrictive—even though the type of pico projector used in Furhat and other back projected talking heads are rapidly getting stronger at ever-better prices, with the current technology it is unlikely that the face will ever work well in direct sunlight. The second has to do with the inflexibility of the projection surface. Although the small misalignments caused by for example speaking are barely noticeable, as noted by Naimark (2005), larger jaw movements such as wide yawning will cause the face to look clearly out of order, with a projected jaw line clearly missing the jaw line of the projection surface.

10.3.2 Compared to 2D Displays of 3D Models

The most salient difference is the manner in which eye gaze is perceived, which we go through in detail in the next section. Besides that, the most obvious difference is how the projection is interpreted compared to a monitor. Whereas an ECA displayed on a monitor requires interpretation—it is not obvious whether it should best be interpreted as something in another world seen through a window or peeking out of a window or if it is supposed to be viewed as sharing the same space as its onlooker, it is immediately clear that the latter is the case for the chase minor advantages in being even more inexpensive and adaptable than the other method, and its gaze characteristics can be utilized as an advantage for specific purposes.

10.4 The Mona Liza Gaze Effect

The perception of 2D renditions of 3D scenes is notoriously riddled with artefacts and illusions—for an overview, see Gregory (1997). The most important of these for embodiment is *the Mona Lisa gaze effect*, commonly described as an effect that makes it appear as if the Mona Lisa's gaze rests steadily on the viewer as the viewer moves through the room (Fig. 10.2). Although the reference to the Mona Lisa is a modern invention, documentation of the effect dates back at least as far as Ptolemy in around 100AD "[...] the image of a face painted on panels follows the gaze of moving viewers to some extent even though there is no motion in the image itself" (Smith 1996).

10.4.1 Mechanics of the Mona Lisa Gaze Effect

The Mona Lisa gaze effect has earned frequent enough mention, and a number of more or less detailed explanations have been presented from Ptolemy and onwards (e.g. Smith 1996; Cuijpers et al. 2010), but these do not provide an explanation that satisfies the requirements of a designer of embodied computer programmes. In Al Moubayed et al. (2012a, 2012b), we propose a model that explains Mona Lisa stare effects as well as other observations with a minimum of complexity, and verified its predictions experimentally. The model is based a number of observations, which are described in the following, before the model in itself is presented.

Our first, seemingly trivial observation is that *in order to judge gaze direction, it is not sufficient to know the angle of the eyes relative to the head*—which can be estimated for example by means of relative pupil position within the sclera (e.g. Cuijpers et al. 2010). An estimation of *the position and angle of the head is also required*. The background of Fig. 10.3 shows the Wollaston effect (Wollaston 1824), in which two pair of identical eyes appear to gaze at different points when drawn in to heads that have different angles. This "effect" seems to result from an insistence to view our interpretation of depicted eyes as somehow isolated from the head

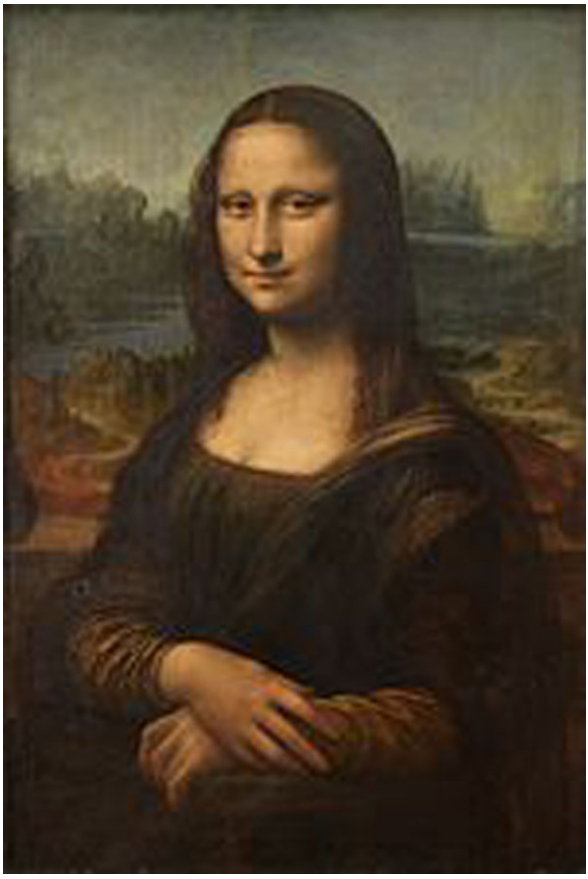

Fig. 10.2 Leonardo da Vinci's *Mona Lisa*. Mona Lisa appears to be looking straight at the viewer, regardless of viewing angle. The painting is the public domain and copyright free

in which they are lodged. If we, like Todorović (2006), instead assume that head and eyes are interpreted in relation to each other and to the space they are depicted in, the Wollaston effect is not only accounted for, but rather ceases being an effect, as illustrated in the foreground of the figure. Todorović's account relates eyes and head pose in virtual space directly to perceived gaze direction in physical space. We generalize this account by means of simplification, and speak exclusively of gaze direction within the same (virtual or physical) space: *the perceived gaze direction within a space, virtual or physical, of a creature within that same space, is a function of the perceived angle of the gazing creature's head within that space, and the perceived angle of her eyes, relative her head.*

The second observation, illustrated in Fig. 10.4, is that *the Mona Lisa gaze effect is not restricted to eye gaze,* but generalizes to anything pointing out from a picture, such as an outstretched index finger. Most viewers feel that complete Uncle Sam in the left pane of the figure and faceless Uncle Sam in its right pane both point straight at them, regardless of viewing angle. This means that although eye and pupil position clearly affects how we perceive gaze direction, they cannot hold the

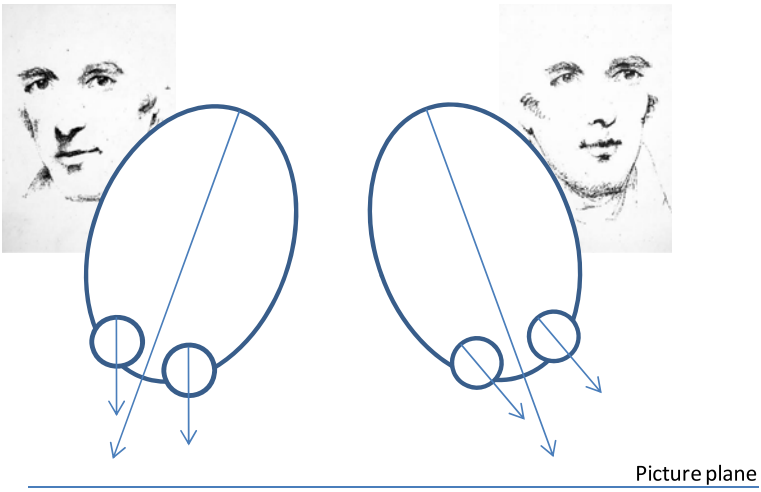

Fig. 10.3 The Wollaston effect is seen in the two drawings: gaze direction is perceived differently although the eyes are identical, and only the head shape differs. The *ovals* with two *circles* represent a possible interpretation of the drawings as seen from above in virtual space. The two drawings are from Wollaston (1824), and are in the public domain and are copyright free

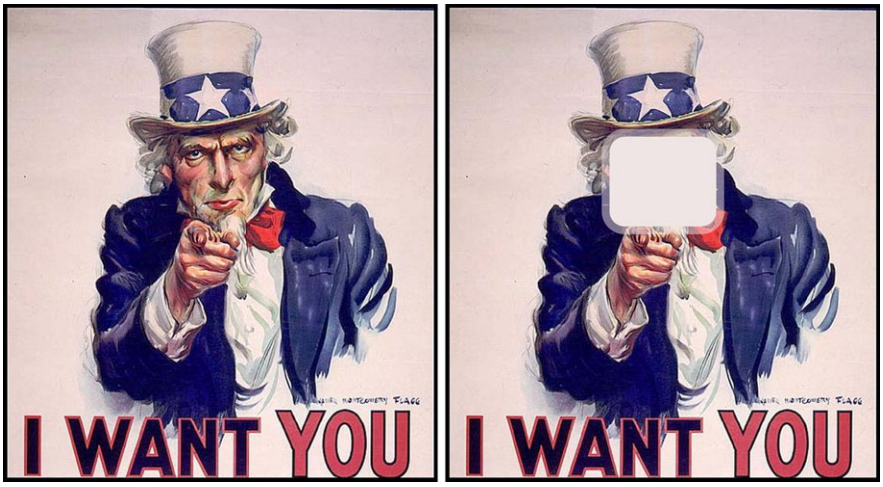

Fig. 10.4 *I want you for the U.S. Army nearest recruiting station*, commissioned by the US federal government and painted by James Montgomery Flagg (cropped in the *left pane*, and cropped and edited in the *right pane*). The painting is in the public domain and copyright free

key to the effect, as the effect is present also when eyes and pupils are not. This observation also allows to generalize our statement from the last paragraph to not concern not only eyes and heads, but any object with a perceived direction contained within another.

Fig. 10.5 *Interiors of the Winter Palace. The Throne Room of Empress Maria Fiodorovna*. Painting by Yevgraf Fyodorovich Krendovsky. The picture is in the public domain and copyright free

The third observation is that *2D images representing 3D objects or scenes are interpreted as having their own virtual 3D space, distinct from physical space.* The axes of this virtual space are oriented along the horizontal and vertical edges of the image (perceived as width and height, respectively), with the third axis perpendicular to its surface (perceived as depth). This is particularly clear when we watch photos or paintings of large rooms with walls, ceiling and floors at right angles, as in Fig. 10.5. The painting in the figure gives a clear impression of a large three-dimensional space with a throne located at the far back. The location of the throne in relation to physical space is ambiguous: if our viewing angle and distance to the painting is varied, the throne's position in the portrayed virtual space is maintained, and its position in physical space remains unclear.

Our forth observation has to do with the high degree of interpretation that goes in to the shapes we perceive in images. The phenomenon, known as *shape constancy*, is well-documented and was described early on. Descartes states in his Dioptrics of 1637: "[…] shape is judged by the knowledge, or opinion, that we have of the position of various parts of the objects, and not by the resemblance of the pictures in the eye; for these pictures usually contain only ovals and diamond shapes, yet they cause us to see circles and squares" (Descartes 1637, p. 107). Phrased differently, *viewers of 2D images perceives the shapes in the images as invariant, even when the viewing angle changes*, as exemplified by the two top right groups of circles in Fig. 10.6. Although the top left group contains exactly the same shapes, it is not necessarily perceived as three circles but rather as a circle and two ovals. This indicates that this perception is indeed dependent on interpretation. Note that when the circular shapes are viewed at a steep angle, most viewers still perceive a circle,

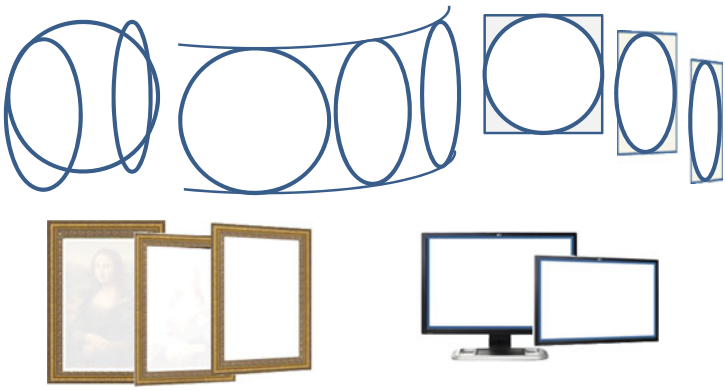

Fig. 10.6 Three groups of three rounded shapes, a group of picture frames, and two computer monitors

although the shape is in fact distorted to one of Descartes' ovals. The figure also illustrates that shape constancy holds true for other shapes, such as the rectangular shape of the frame of the Mona Lisa or the edge of a standard monitor, *both of which are perceived as perfectly rectangular regardless of viewing angle*. For a more detailed account of shape constancy, see Gregory (1997).

We now have all the pieces we need for our explanation of the Mona Lisa gaze effect except one: a description of how viewers of 3D virtual space align that virtual space with their own, physical space. To solve this, we assume that a mechanism similar to shape constancy is at play: *viewers of 2D images depicting 3D objects interpret their position in relation to the virtual 3D space as head-on, perpendicular to the surface plane of the image*.

In addition to the support provided by what is known about shape constancy, this is intuitively pleasing as well. 2D images, at least those that use perspective to depict a 3D space, are created as seen from some vantage point in front of the objects seen in the picture. In the case of photographs or paintings created using camera obscura, this vantage point can be calculated exactly from the geometry of the image and the characteristics of the lens. Paintings allow for artistic license and may leave more ambiguity, but are still generally interpreted as seen head-on. This is again an observation that may seem trivial, but it has bearing as to how we may connect the virtual 3D space depicted in an image to the physical space of our surroundings.

It is worth pointing out that provided that we are standing in front of a picture, interpreting the general orientation and left-right position of the objects depicted in it is straightforward, whereas deciding the distance to the objects from the imagined vantage point of the creator can pose more of a problem, as illustrated by Fig. 10.7. It is trivial in both panes to see that all of the animals in the sculpture face left and that the rooster is on top and the horse at the bottom. Size variation, however, is ambiguous in 2D depictions and can be interpreted as deriving from at least three sources: the size of the depicted object, the distance and projection from the object to the position from which it is captured, and the size of the actual 2D image. The

Fig. 10.7 *Bremen Town Band*, Bremen, Germany. The picture was taken in 1990 by Adrian Pingstone and released into the public domain

image in the figure's left pane has been edited to remove the sculpture's surroundings. Without references, it is difficult to judge the size of the sculpture. The right pane contains more clues for the viewer to get a fair sense of distance and size, but the distance from the vantage point of the camera to the sculpture and from the sculpture to the people in the back are difficult to guess, so the size of the sculpture is elusive in that pane too.

We now have all the observations and assumptions we need for our proposed interpretation of the Mona Lisa gaze effect. Combining the assumptions we propose that the directionality of objects in 2D images are interpreted in relation to a virtual 3D space with axes oriented along the horizontal and vertical edges of the image and the third axis perpendicular to it, and that this space is aligned to the physical space of the viewer as if the image were viewed head-on. Shape constancy further allows us to make this interpretation regardless of the actual viewing angle, so that when observed, anything pointing straight out of the picture is perceived as pointing directly at the viewer, regardless of viewing angle. Figure 10.8 illustrates the model. The leftmost pane relates the head and eye of the gazing creature to the 3D space of the virtual space created by the picture, and to the picture plane. The illustration represents a head in virtual 3D space at a 20° angle relative to the depth axis, and eyes at the same but opposite angle relative to the head. The resulting eye direction is parallel with the virtual 3D depth axis and perpendicular to the picture plane. Virtual space is then aligned to physical space along the along their respective depth axes, as illustrated in the centre pane. Finally, shape constancy allows the viewer to view the picture *as if* facing it head-on, regardless of the viewer's position in the room, as

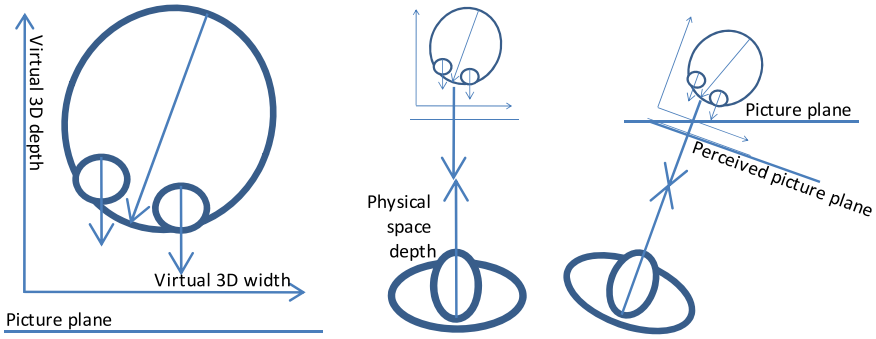

Fig. 10.8 Our observations and assumptions combined into a model of how gaze direction (and the directionality of other objects) in 2D pictures are perceived

in the rightmost pane, causing the Mona Lisa gaze effect to occur. In other words: if something points straight out of a two-dimensional picture, it will be perceived by each on-looker, regardless of position in the room, to point straight at said on-looker. The model predicts that people viewing 2D images of gaze should have no problem judging gaze relative the virtual space coordinates, and should judge gaze direction in physical space *as if they were standing directly in front of the picture*. In other words, any gaze directed straight out of the picture would be perceived as looking straight at an on-looker, as is the case with Mona Lisa. Furthermore, gaze to the left or to the right the depth of the virtual space should always to the left or right, respectively, of an on-looker, and by a constant angle. As it turns out, all of these predictions bear out (Al Moubayed et al. 2012a, 2012b).

The model is very similar to that suggested by Todorović (2006), with the addition of shape constancy to account for the fact that most viewers do not perceive drawings viewed at an angle as distorted. The processing model, in which differences caused by viewing angle are removed initially, has the further advantage that the actual recognition becomes simpler, as there is less variability left to account for.

10.4.2 Impact on Human-Human and Human-Machine Communication

The importance of gaze in social interaction is well-established. From a human communication perspective, Kendon's work on gaze direction in conversation (Kendon 1967) is particularly important in inspiring a wealth of studies that singled out gaze as one of the strongest non-vocal cues in human face-to-face interaction (see e.g. Argyle and Cook 1976; Bavelas et al. 2002). Gaze has been associated with a variety of functions within social interaction—Kleinke's review article from 1986, for example, contains the following list: "(a) provide information, (b) regulate interaction,

(c) express intimacy, (d) exercise social control, and (e) facilitate service and task goals" (Kleinke 1986).

These efforts and findings, in turn, were and are shadowed by an increasing effort in the human-computer interaction community, which recognized the importance of modelling gaze and its social functions such as expressing and communicating attitudes and emotions in embodied conversational agents (ECAs). Examples include Takeuchi and Nagao (1993), Poggi and Pelachaud (2000), Bilvi and Pelachaud (2003), and Lance and Marsella (2008). As multimodal and facial communication with communication devices become more advanced and more popular, the demand for ECAs in control of their gaze behaviour increases. Multimodal interfaces are now able to provide testing and manipulation frameworks for behavioural models of gaze and other non-vocal signals. Such systems have recently been effectively used to investigate and quantify the effects of gaze using controlled experiments (Edlund and Nordstrand 2002; Lance and Marsella 2008; Gu and Badler 2006; Edlund and Beskow 2009; Nordenberg et al. 2005).

Given the importance of gaze, the effects of presenting an ECA which displays a perceivable gaze direction without being able to control this direction are potentially devastating for the communication and for how the ECA is perceived. Oddly enough, there is one clear example when the Mona Lisa gaze effect does not cause this to happen, but rather presents us with the remedy: when our ECA communicates with one single person whose head and face we have no ability to track. Incidentally, this is historically the most common setup for interactional experiments with spoken dialogue systems represented by an ECA.

The way this works is as follows. A key problem with using ECA gaze for communicative purposes is that unless we have access to sensors and head tracking equipment, which were expensive and hard-to-get until rather recently, the system does not know where its human interlocutor's head and eyes are, which makes directing the ECAs gaze at them a feat of magic. In many cases, experimenters have simply hoped that the human interlocutor will stay relatively immobile in front of the mobile, and used a gaze straight out from the monitor as an approximation of "looking at the interlocutor". Whether a case of insight or sheer luck, this method is quite reliable—more so than one would think. As movements by the human interlocutor are negated by the Mona Lisa gaze effect, the system is always perceived as gazing at the interlocutor when it attempt to do so, and never when it attempts to look away. Under these quite restricted but rather common circumstances, harnessing the Mona Lisa gaze effect is the only way to achieve gaze reliably towards the interlocutor, as access to head pose information does not improve the situation—on the contrary, attempting to gaze at the real position of the interlocutors head would have the opposite effect in all cases except when the interlocutors sits straight in front of the monitor.

As soon as there is more than one person in the room, the Mona Lisa gaze effect becomes a very real problem. The system designer has a choice of having the ECA look straight out from the monitor, thus being perceived as looking straight at each person in the room simultaneously (and meeting their gaze of they look back), or look away from every person in the room. Note that having access to head pose data still will not help.

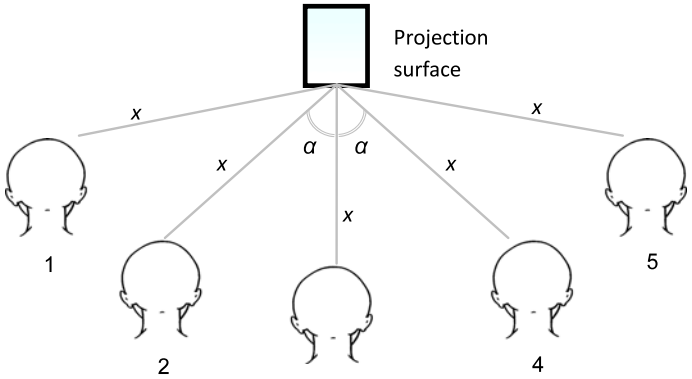

Fig. 10.9 Schematic layout of a subject/target experiment with 5 targets at x meters distance from the stimuli—a projection surface—and with equal distances between adjacent subject/targets

10.4.3 Measuring Perceived Direction

There are many cases where we would want to test how directionality in face-to-face situations is perceived, for example to verify a model such as the one just proposed; to investigate the accuracy of perception, perhaps under adverse conditions; or to train targets of a robot so that they match human perception. There are also ways to present ECAs that make it less than obvious whether the Mona Lisa gaze effect is in place or not, or that are perhaps only partially susceptible to the effect, such as the illusionistic 3D ECA presented by Kipp and Gebhard (2008).

Beskow and Al Moubayed (2010) pioneered an experimental paradigm that was developed to allow experimenters to quickly investigate and gather large amounts of data on human perception of gaze targets/direction. The paradigm is described here in generalized form, allowing it to function as a means of comparing not only gaze targets but arbitrary directional stimuli such as directional audio or verbal descriptions. In recognition of the fact that the key to the effectiveness of the paradigm is to utilize the same people as subjects and targets for the directional stimuli, we will call the method the subject/target paradigm here. The practice of using subjects as targets also adds to the ecological validity of the paradigm, as the distinction between being or not being the person gazed at or spoken to is a salient distinction in face-to-face interaction.

In the subject/target paradigm, a group of N subjects are placed in a circle or semi-circle, so that there is one point at their centre which is equidistant to each subject, from which all stimuli are presented (the *centre*). Subjects positions are numbered P_1 to P_N, and the angle between each subject's position, that of the centre, and that of the subject's closes neighbouring subjects ($A(P_1 P_2) \ldots A(P_N P_1)$) is calculated. Subjects may or may not be equidistant from their closest neighbours. Figure 10.9 shows a subject/target setup with five subjects and stimuli presented on a projection surface.

All subjects double as targets for the directional stimuli. During an experiment, directional stimuli are aimed at each of the subjects. The order is varied systematically, and the number of stimuli is such that each subject is targeted as many times as the others in one set of stimuli. A set of stimuli, then, contains a multiple R of N for a total of R^*N stimuli. Once one set is completed, the subjects rotate—they shift their positions by one step and the process of presenting a set of N^*R stimuli is repeated. The rotation is repeated N times, until each subject has been in each position once, making the total number of stimuli presented in an experiment N^*R^*N.

Each time a stimulus has been presented, each subject is asked to point out the intended target in such a manner that the other subjects cannot see it. This result in N judgements for each stimulus, for a total of $N^*R^*N^*N$ data points in one experiment. If more than one experiment condition is to be tested, the entire process is repeated from the beginning. The manner in which the subject/targets point out the intended target is not prescribed by the paradigm. Methods that have been used to date include jotting the result down on a predesigned form, which requires the full use of hands and eyes (Beskow and Al Moubayed 2010); simply asking the addressee to respond in an interactive test, which yields considerably fewer data points, but may increase ecological validity (Al Moubayed and Skantze 2011); and marking the target with through manual signing, for example by showing different numbers of fingers, which obviates the need for eye sight and so can be used for tests of acoustic directionality—the use of blindfolds would render regular form filling impractical, as the subjects cannot see to write (Edlund et al. 2012).

The experiment results can be analyzed in a number of ways. Subject performance measures such as inter-subject agreement, target accuracy, and average error in degrees are obvious examples, which can be analyzed more finely to show whether the average error is, for example, larger when the target is far away from the subject. Another use of the paradigm that is useful when the exact relation between system internal controls for pointing and perceived reality of the pointing is unknown. By pointing (gazing) at systematically varied spots along the circle of subject/targets and analyzing the resulting judgements, we can find a function connecting the system controls to perceived target angles. The experiments showed clearly that subjects are very good at reliably estimating gaze targets from 3D projected talking heads, but considerably less so from 2D displays.

A final example of analysis takes us back to the Mona Lisa gaze effect. We have stated that this effect ought not ruin a person's ability to judge directions altogether, but merely change the way these directions are mapped into the physical world of the person. This allows us to remap the subjects responses from an absolute target to a target relative to the subject's position. If the Mona Lisa gaze effect is in place, the re-mapped responses should be as accurate, or almost as accurate if we allow for some loss in translation, as the absolute targets when the Mona Lisa gaze effect is not in place. Al Moubayed et al. (2012a, 2012b) show in that this is indeed the case: the original mappings (as stated above) yield good results for the 3D projected talking heads and less so from 2D displays, while the remapped responses yield the opposite result.

We can take this reasoning one step further. The Mona Lisa gaze effect is in place when the gazing creature is perceived as being present in a separate space, such as a painting or the virtual reality inhibited by ECAs. And when the Mona Lisa gaze effect is in place, we get high accuracy of subject/target experiments once the results are re-mappen into subject-relative terms. On the other hand, the Mona Lisa gaze effect is not present when the gazing creature is perceived as sharing the same space as the subject—that is when it is *co-present* with the subject. In these cases, we get high accuracy of subject/target experiments with the original results. This suggests that by comparing the score of re-mapped, relative accuracy with original, absolute accuracy, we may be able to get a bearing on to what extent the subjects perceive an embodiment as co-present, as suggested by preliminary results presented in Edlund et al. (2011).

10.5 Summary

Hashimoto and Morooka (2006) state that "a curved surface image has a dependable direction of observation and presence in actual space". To quantify this, the results of several studies of the accuracy and inter-subject agreement of perceived gaze targets of Furhat show unequivocally that the use of a front or back projected talking head onto a surface of similar shape completely cancels the Mona Lisa gaze effect (Beskow and Al Moubayed 2010; Al Moubayed and Skantze 2011; Al Moubayed et al. 2012a, 2012b). Preliminary results from multi-party conversations with Furhat also suggest that its gaze characteristics are suitable for turn-taking and addressee selection (Al Moubayed et al., in press).

What, exactly, is it that causes this? We have established that the Mona Lisa gaze effect is derived from human interpretation and is a result of a person aligning the coordinate system of a perceived virtual space with that of the physical space in which the human resides in such a manner that the human places herself in a position straight ahead of the image or movie that portrays the virtual space. We have even suggested that a measure based on a comparison of absolute (non-Mona Lisa) direction accuracy versus relative (Mona Lisa) gaze accuracy may give us some insight as to the extent to which an embodied computer programme is perceived as co-present with the viewer.

We suggest that in the end, it boils down to a simple matter on whether the viewer interprets the person gazing or the finger pointing as being present in the same room, or as being portrayed through a "window" onto another space. We leave this narration with an image—two pen drawings of two pairs of eyes on two plain sheets of A4 paper. When the drawings are viewed flat—as images on a rectangular piece of paper—they display the Mona Lisa gaze effect to its fullest. When, on the other hand, they are curved into cylinders, as in Fig. 10.10, their gaze is easily perceived as having an absolute target in the room—even though the area behind the eyes is forced flat by a piece of cardboard behind the eyes. Clearly, more study is needed to learn exactly what is needed to turn our perception from through-the-looking-glass to co-present mode. As it stands, it may well be a question of

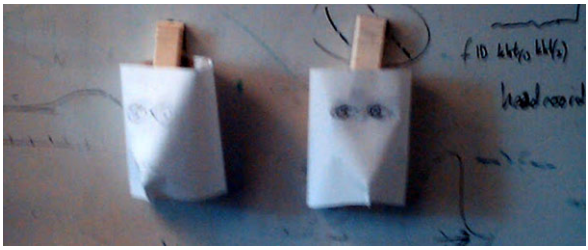

Fig. 10.10 Two pen drawings on rolled-up A4 sheets of paper

whether the features of a face appear to be drawn *inside* their own space on a piece of paper, or *outside* the object boundaries outlined by the same piece of curved paper.

References

Al Moubayed S, Skantze G (2011) Effects of 2D and 3D displays on turn-taking behavior in multiparty human-computer dialog. In: Proceedings of SemDial, Los Angeles, pp 192–193

Al Moubayed S, Alexanderson S, Beskow J, Granström B (2011) A robotic head using projected animated faces. In: Salvi G, Beskow J, Engwall O, Al Moubayed S (eds) Proceedings of AVSP2011, p 69

Al Moubayed S, Beskow J, Granström B, Gustafson J, Mirning N, Skantze G, Tscheligi M (2012a) Furhat goes to Robotville: a large-scale multiparty human-robot interaction data collection in a public space. In: Proceedings of LREC workshop on multimodal corpora, Istanbul, Turkey

Al Moubayed S, Edlund J, Beskow J (2012b) Taming Mona Lisa: communicating gaze faithfully in 2D and 3D facial projections. ACM Trans Interact Intell Syst 1(2):25

Argyle M, Cook M (1976) Gaze and mutual gaze. Science 194(4260):54–55

Bavelas J, Coates L, Johnson T (2002) Listener responses as a collaborative process: the role of gaze. J Commun 52(3):566–580

Beskow J, Al Moubayed S (2010) Perception of gaze direction in 2D and 3D facial projections. In: The ACM/SSPNET 2nd international symposium on facial analysis and animation, Edinburgh, UK

Beskow J, Salvi G, Al Moubayed S (2009) SynFace—verbal and non-verbal face animation from audio. In: Proceedings of the international conference on auditory-visual speech processing, AVSP'09, Norwich, England

Bilvi M, Pelachaud C (2003) Communicative and statistical eye gaze predictions. In: Proceedings of international conference on autonomous agents and multi-agent systems (AAMAS), Melbourne, Australia

Boye J, Gustafson J (2005) How to do dialogue in a fairy-tale world. In: 6th SIGdial workshop on discourse and discourse

Breazeal C, Scassellati B (2001) Challenges in building robots that imitate people. In: Dautenhahn K, Nehaniv CL (eds) Imitation in animals and artifacts. MIT Press, Boston, pp 363–390

Cassel J, Sullivan J, Prevost S, Churchill EE (2000) Embodied conversational agents. MIT Press, Cambridge

Cassell J, Stocky T, Bickmore T, Gao Y, Nakano Y, Ryokai K (2002) MACK: media lab autonomous conversational kiosk. In: Proceedings of Imagina02, Monte Carlo

Cuijpers RH, van der Pol D, Meesters LMJ (2010) Mediated eye-contact is determined by relative pupil position within the sclera. In: Perception ECVP abstract supplement, p 129

Dariush B, Gienger M, Arumbakkam A, Goerick C, Zhu Y, Fujimura K (2008) Online and markerless motion retargeting with kinematic constraints. In: Proceedings of the IEEE/RSJ international conference on intelligent robots and systems (IROS 2008), pp 191–198

Delaunay F, de Greeff J, Belpaeme T (2009) Towards retro-projected robot faces: an alternative to mechatronic and android faces. In: Proceedings of the international symposium on robot and human interactive communication (RO-MAN), Toyama, Japan

Descartes R (1637) Dioptrics. In: Discourse on method, optics, geometry, and meteorology. Hackett, Indianapolis, pp 65–162

Edlund J (2011) In search of the conversational homunculus—serving to understand spoken human face-to-face interaction. Doctoral dissertation, KTH

Edlund J, Beskow J (2009) MushyPeek—a framework for online investigation of audiovisual dialogue phenomena. Lang Speech 52(2–3):351–367

Edlund J, Nordstrand M (2002) Turn-taking gestures and hour-glasses in a multi-modal dialogue system. In: Proceedings of ISCA workshop on multi-modal dialogue in mobile environments, Kloster Irsee, Germany

Edlund J, Al Moubayed S, Beskow J (2011) The Mona Lisa gaze effect as an objective metric for perceived cospatiality. In: Vilhjálmsson HH, Kopp S, Marsella S, Thórisson KR (eds) Proceedings of the 10th international conference on intelligent virtual agents (IVA 2011), Reykjavík. Springer, Berlin, pp 439–440

Edlund J, Heldner M, Gustafson J (2012) Who am I speaking at?—perceiving the head orientation of speakers from acoustic cues alone. In: Proceedings of LREC workshop on multimodal corpora 2012, Istanbul, Turkey

Gregory R (1997) Eye and brain: the psychology of seeing. Princeton University Press, Princeton

Gu E, Badler N (2006) Visual attention and eye gaze during multiparty conversations with distractions. In: Proceedings of the international conference on intelligent virtual agents

Hartholt A, Gratch J, Weiss L, Leuski A, Morency L-P, Marsella S, Liewer M, Thiebaux M, Doraiswamy P, Tsiartas A (2009) At the virtual frontier: introducing Gunslinger, a multi-character, mixed-reality, story-driven experience. In: Proceedings of the 9th international conference on intelligent virtual agents (IVA'09). Springer, Berlin, pp 500–501

Hashimoto M, Morooka D (2006) Robotic facial expression using a curved surface display. J Robot Mechatron 18(4):504–505

Hjalmarsson A, Wik P, Brusk J (2007) Dealing with DEAL: a dialogue system for conversation training. In: Proceedings of SIGdial, Antwerp, Belgium, pp 132–135

Jalbert G (1925) Lay figure. Technical report, US Patent 1653180

Kendon A (1967) Some functions of gaze direction in social interaction. Acta Psychol 26:22–63

Kipp M, Gebhard P (2008) IGaze: studying reactive gaze behavior in semi-immersive human-avatar interactions. In: Proceedings of the 8th international conference on intelligent virtual agents (IVA'08), Tokyo, Japan

Kleinke CL (1986) Gaze and eye contact: a research review. Psychol Bull 100:78–100

Kuratate T, Matsusaka Y, Pierce B, Cheng G (2011) Mask-bot: a life-size robot head using talking head animation for human-robot communication. In: Proceedings of the 11th IEEE-RAS international conference on humanoid robots (humanoids), pp 99–104

Lance B, Marsella S (2008) A model of gaze for the purpose of emotional expression in virtual embodied agents. In: Proceedings of the 7th international conference on autonomous agents and multiagent systems, pp 199–206

Liljegren GE, Foster EL (1989) Figure with back projected image using fiber optics. Technical report, US Patent 4978216

Morishima S, Yotsukura T, Binsted K, Nielsen F, Pinhanez C (2002) HyperMask: talking head projected onto real objects. Vis Comput 18(2):111–120

Naimark M (2005) Two unusual projection spaces. Presence 14(5):597–605

Nordenberg M, Svanfeldt G, Wik P (2005) Artificial gaze—perception experiment of eye gaze in synthetic faces. In: Proceedings from the second Nordic conference on multimodal communication

Poggi I, Pelachaud C (2000) Emotional meaning and expression in performative faces. In: Paiva A (ed) Affective interactions: towards a new generation of computer interfaces, pp 182–195

Smith AM (1996) Ptolemy's theory of visual perception: an English translation of the "Optics" with introduction and commentary. Am. Philos. Soc., Philadelphia

Steels L, Brooks R (eds) (1995) The artificial life route to artificial intelligence: building embodied, situated agents. Lawrence Erlbaum Associates, Hillsdale

Takeuchi A, Nagao K (1993) Communicative facial displays as a new conversational modality. In: Proceedings of the INTERACT'93 and CHI'93 conference on human factors in computing systems

Todorović D (2006) Geometrical basis of perception of gaze direction. Vis Res 45(21):3549–3562

Traum D (2008) Talking to virtual humans: dialogue models and methodologies for embodied conversational agent. In: Wachsmuth I, Knoblich G (eds) Modeling communication with robots and virtual humans. Springer, Berlin, pp 296–309

Traum D, Rickel J (2002) Embodied agents for multi-party dialogue in immersive virtual worlds. In: Proceedings of the first international joint conference on autonomous agents and multiagent systems (AAMAS 02). ACM, New York

Wollaston WH (1824) On the apparent direction of eyes in a portrait. Philos Trans R Soc Lond B 114:247–260