Mediated Faces

Judith Donath
MIT Media Lab

Abstract. Incorporating faces into mediated discussions is a complex design problem. The face conveys social and personal identity; it reports fleeting changes of emotion and the cumulative effects of often repeated expressions. The face both expresses and betrays: it shows what the person wishes to convey – and much more. We are highly attuned to recognizing and interpreting faces (though these interpretations are very subjective). Incorporating faces into mediated environments can be quite desirable: it helps the participants gain a stronger sense of their community and can potentially provide finely nuanced expression. Yet there are significant disadvantages and difficulties. The immediate identifying markers revealed by the face, e.g. race, gender, age, are not necessarily the initial information one wants to have of others in an ideal society. And much can be lost in the path from user's thought to input device to output rendering. This essay discusses key social, cognitive and technical issues involved in incorporating faces in mediated communication.

1 Introduction

The face is essential in real world social interactions: we read character and expression in the face, we recognize people by their face, the face indicates where one's attention lies. Yet the face is mostly absent from online interactions – and this is in part why many people find cyberspace to be only a pale substitute for real world contact.

Today's fast graphics cards and high bandwidth connections have eliminated many of technical barriers to making the virtual world as fully visaged as the real world. Yet the problem goes beyond perfecting geometric models of facial structure, for there are complex social and cognitive aspects to how the face is used in communication that cannot be directly transplanted to a mediated environment. Furthermore, the desirability of faces cannot be assumed for all interfaces -- some online communities have thrived because of the absence of faces and their immediate revelation of race, gender, age and identity.

Bringing the face to the interface requires radically reinventing the notion of personal appearance, while remaining grounded in the cognitive and cultural meanings of the familiar face. It requires analyzing applications to understand what aspect of the face they need to convey - personal identity? level of attentiveness? emotional expression? - and finding intuitive ways both to input and express this information. In some cases, the best interface is as realistic as possible, in others it has no face at all, while others may be best served by a synthetically rendered image that selectively conveys social information.

Faces are used in many ways in computer interfaces, representing both people and machines. This paper focuses on the role of the face in computer-mediated human interactions in which the face represents a particular individual, communicating with other people in a real-time, online discussion.

Unlike much of the research in computer-mediated communication, we do not assume that the ultimate goal is to recreate reality as faithfully as possible. The computer makes it possible to go “beyond being there”[21] – to create environments that have features and abilities beyond what is possible in the ordinary everyday world. We can create environments in which the face shows expression, but does not reveal the user’s identity; we can create worlds in which traces of the user’s history of actions are sketched into the lines of the face. Yet introducing faces into a mediated communication system must be done carefully, for the face is replete with social cues and subtle signals; a poorly designed facial interface sends unintended, inaccurate messages, doing more harm than good.

2 Why use faces in mediated human to human communication?

There are many reasons to uses faces in mediated communication. The face is very cognitively rich and holds great fascination for us. Even newborn babies, a few hours old, will gaze longer at a face-like image than at a random array [24]. An environment filled with faces can be endlessly interesting to observe. People-watching is a perennially favorite pastime, in which we scan the surrounding scene for familiar faces and imagine the identity of the individual behind a stranger’s visage [48]. An online environment populated with “people” with faces may seem more sociable, friendly, intriguing than a textual or purely abstract space.

Faces convey important social information about who you are and what you are thinking. We are cognitively wired to recognize and remember faces and your individual identity is uniquely represented by your face. The face also conveys social identity, with features that indicate basic categories such as age and gender as well as those that are associated with particular personality types. The face conveys emotional state and intent, displaying a wide range of expressions, from puzzlement to terror, fury to delight.

The faces helps to moderate and choreograph conversations. We use gaze to indicate attentiveness, to direct our remarks at an individual, to hold and yield the floor. Facial expressions soften our words, expressing humor, regret, etc. The face is very important in conveying responses, to show understanding, agreement, etc.

People behave more “socially”, that is, more politely and with greater restraint, when interacting with a face. Sproull et al. [39] found that people responded quite differently to questions posed by computers when they were presented as text or as facial displays. For instance, when asked questions about themselves via text they answered with little embellishment but when queried by a facial display they attempted to present themselves in the best possible light.

Some of these reasons for using faces in mediated communication are advantages only in certain circumstances. The “social” responses that Sproull et al. detected can make a friendly discussion forum more sociable, but may be detrimental at other times. Isaacs and Tang [22] noted that non-facial interfaces could be more efficient, since the participants attended to the problems at hand, rather than to the time-consuming rituals of greetings and small-talk that ordinary politeness requires; Sproull and Kiesler [38] found that hierarchical distinctions were flattened in text-only discussions – it is plausible (though untested) that such distinctions would regain prominence in a mediated environment populated with visible faces (the desirability of maintaining or flattening such distinctions is context dependent).

The face allows us to characterize people at a glance. In the real world, the first things one learns about another are such social categories as age, gender and race, for the cues for these categories are embodied in the face. In an ideal world, would that necessarily be one's first impression? The online world has been touted as a place where one is identified first by one's words and ideas, free from the stereotypes imposed by such categorization; online spaces in which one's face is visible afford no such freedom. There is no simple metric for measuring the desirability of conveying this information, with numerous factors such as the purpose of the forum and the background of the participants affecting the evaluation. What we can do is understand fully what social cues the face does convey and use that knowledge to help determine where a facial display is appropriate.

Including faces in the interface is very difficult to do well. This is to a large extent due to the fact that the face is so expressive, so subtle, so filled with meaning. We ascribe character to and read emotion in any face, especially a realistically rendered one. There is no truly "neutral" face. A face in the interface is replete with social messages, but a poorly designed one will send many unintended ones.

In real world social situations we are constantly adjusting our face to do the appropriate thing – to hide or show our feelings and to gaze (or not) in the proper direction. We expect the same from mediated faces and when they elide a particular social protocol, we read an unintended message in the absence of a required expression or the accidental invoking of an inappropriate one. Making the "right" expression is extremely complex, for it is not a single motion, but a precisely timed choreography of multiple movements: a smile that flashes briefly conveys a different message than a smile that lingers.

One of the goals of this paper is to better understand what the fundamental limits are using mediated faces. Can the problems with mediated faces sending unintended messages be ameliorated with better input sensors and better renderings? Are the aspects of the face's social role that cannot be transferred to the mediated world? We will address these questions by first looking more closely at what social information the face conveys and then examining the technologies through which we bring these features to the mediated world.

3 What does the face convey?

Almost every aspect of the face provides some sort of social cue and we are very adept at perceiving minute details of its configuration[1]. Knowing how to act toward someone and what to expect from them is fundamental to social interaction, and this knowledge depends upon being able to distinguish men from women, expressions of anger from those of joy, children from adults, friends from strangers – information that we read in the face. Our societal structures and mores have developed with the assumption that this face-conveyed information is available as the context for interaction.

The face conveys information through its structure, its dynamics, and its decorations[49]. The structural qualities include the overall head shape, the size and placement of the eyes and other features, the lines and texture of the skin, the color and quantity of scalp and facial hair. From these, viewers assess personality and make classifications such as race, gender and age. The dynamic qualities include gaze direction, pupil dilation, blushing, smiling, squinting, and frowning. From these, viewers read emotional expression and attention. Decorations include eyeglasses, cosmetics and hairstyle from which viewers read cultural cues, ranging from large scale group membership to subtleties of class distinctions and subcultural membership. There is also considerable interplay in how these qualities convey social cues. Haircuts affect the assessment of age, cultural mores modify the production and interpretation of emotional expressions, gender determination based on structural cues impacts the cultural interpretation of fabricated elements such as very short hair or lipstick. Recognition is primarily structural, though many times one will not recognize an acquaintance who has grown a beard or is shown in a photograph with an uncharacteristic expression.

The face conveys four major types of social information: individual identity, social identity, expression, and gaze. (This is not an all-inclusive list, for there are important functions that fall outside the scope of this paper, such as, as any lip-reader knows, displaying the words one is saying). These types may seem unbalanced: social identity is a broad conglomeration of all sorts information about one’s gender, genetics, and geniality, whereas gaze is really a means by which the faces conveys information (such as conversational turn openings and attention). Yet this division is useful for thinking about mediated interactions, for addressing these communicative functions independently brings a great deal of flexibility and creative possibilities to the design of the interface.

3.1 Individual identity

We are very adept at recognizing people. We recognize them at a distance, from various viewpoints, with different expressions and as they change with age [49]. We can find a familiar face in a crowd with remarkable speed, especially considering how complex this task is: one’s mental construct of the sought face is compared to each of the visible faces, all of which are quite similar in overall structure and are seen from different angles, in a range of lighting conditions, and feature different expression. There is strong evidence for specific neurological bases for recognizing faces. For example, injury to a particular area of the brain (the occipitotemporal section of the

[1] Our ability to distinguish minute differences among faces is so acute that Chernoff proposed taking advantage of this ability to do multivariate statistical visualization with faces as the graphical representation: “Chernoff faces” map data to facial features as nose length, eye tilt, head shape, etc. [6]. The resulting faces may look happy, sad, surprised or pained - but the underlying data is independent of the interpreted social meaning of the face.

central visual system) leaves people with their vision intact, but nearly unable to recognize faces, a condition known as prosopagnosia[7]. Indeed, our notion of personal identity is based on our recognition of people by their face. To be faceless is to be, according to the Oxford English Dictionary, “anonymous, characterless, without identity.”

In today’s online, text-based worlds, facelessness is the norm and the extent to which participants are identified or left anonymous is a design feature of the various environments. Both anonymous and named forums exist and flourish, though each produces a different tone and is suited for a different purpose [10]. Anonymous or pseudonymous spaces provide an arena for exploring alternate personas and a safe haven for discussing highly sensitive subjects; they are also more likely to devolve into an endless exchange of flames or spam. Named forums bring the weight of one’s real world reputation to the online world; in general, people behave in them more as they would in real life.

Online forums in which the participants’ real faces are featured – as in, for example, a videoconference – are essentially named environments. Much of the discussion about the desirability of video as a medium focuses on issues such as bandwidth requirements and the common gaze problem (discussed below). The fact that it makes the forum into a public sphere in which everyone is seen and known needs to also be kept in mind, for it has a deep effect on the mores of the space.

3.2 Social identity and character

We recognize people not only as individuals, but also as types. Based on the cues we see in the face we quickly categorize people according to gender, ethnicity and age and make judgements about their character and personality.

These classifications tell us how to act toward the other, what behaviors to expect from them, how to interpret their words and actions. In many languages, it is difficult to construct a grammatically (or at least culturally) correct sentence without knowing the other’s age, gender or relative social status. Such distinctions are also the basis of prejudice, with significant biases are found even among people who consciously decry race or gender based stereotypes[2]. More subtle but perhaps even more pervasive biases can be found in character judgements made on the basis of facial structure, e.g. a person with a baby-ish facial structure (large eyes, small nose, large forehead, small chin) will be judged to be more child-like in nature - trusting, naive, kind, weak [49]. This, and many other character judgements based on the face derive from “over generalization effects”. According to Zebrowitz [49], we have very strong responses to cues for important attributes such as health, age, anger etc., so strong that they get over generalized to people whose faces merely resemble those with that attribute or emotion.

Cyberspace (the text version) has been touted as an ideal realm where the visual absence of these cues means that people are known and judged by their words, rather than by their gender, race, or attractiveness. Yet it is not simply a matter of text=good, face-based classification=bad. The cues we gather from the face are basic to much of our established social interactions, and many people find that they need to “put a face to a name" to go beyond a certain level of familiarity or comfort. Furthermore, simply eliminating the face does not eliminate the underlying cultural differences.

The distinction between structural, dynamic and decorative facial features is especially useful when thinking about mediated faces, for not only do these features serve different social purposes, they may also be electively and separately implemented. For instance, the decorative features – glasses, hairstyle, makeup, etc. – reflect one's choices and circumstances. This can be re-created in the decoration of online self-representations and indeed graphical MUDs and games such as the popular *Asheron's Call* feature avatars whose appearance derives from both the player's taste (today I wish to appear as a purple alligator) and role (but because I have not registered I may only choose between being a yellow or green smiley-face). While such simplistic decorations are far from the subtle social messages we communicate via our personal decorations in the real world, the potential certainly exists for these online decorations to become increasingly sophisticated as the mediated world evolves [41][40].

The dynamic features are also separable: there are motion capture facial animation programs that track the dynamic facial movements of a live actor and use them to animate a synthetic face [14][42]. The synthesized face can be that of the original actor (a technique used to achieve low bit-rate transmission of facial expressions [16]) or of any appropriately modelled face. While such techniques are used primarily to convey expression independently of other features, it is important to note that more information about social identity may be imparted this way than one might think: people can detect cues for age and gender in the dynamics of the face alone, as has been demonstrated with point-light experiments in which key points of the face are marked with dots and the rest is made invisible so that observers see only the moving dots [49].

The structural features are the most problematic in terms of stereotyping. It is the use of genetically determined features such as bone structure and skin color to assess someone's personality, morality, intelligence, etc. that raises the biggest concerns about unfair bias based on facial features. Cyberspace (the text version) has been touted as an ideal world in which such prejudice is eliminated because the initial cues by which such stereotypes are made are invisible. From this viewpoint, an interface that brings one's real face into cyberspace destroys this utopia, reintroducing the mundane world's bias-inducing cues. In practice the situation is more complex. For instance, gender differences permeate our use of language, and men and women are socialized to use apologetics, imperatives, etc. quite differently. Hiding one's gender online requires more than simply declaring oneself to be of the other gender: one must adapt one's entire tone and wording to the often subtle mores of the other. Thus, gender that is hidden online can be uncovered by writing style, albeit more slowly than such identification is made in the face to face world [10]. Furthermore, a lack of cues as to social identity does not lead to people thinking of each other as ciphers; rather, categorization still occurs, but with a high likelihood of error - an error which can have further consequences. For instance, if I mistakenly assume someone is a man who is actually a woman, and "he" uses locutions that would seem ordinary if spoken by a woman but coming from a man seem very passive and accommodating, I not only see him as a man, but as a particular type of man, timid and sensitive. Thus we see that while removing the face from the interface does remove some immediate social categorization cues, it does not eliminate such categorization entirely, and the ambiguity that ensues introduces new social problems.

3.3 Expression

One of the most important - and most controversial - communicative aspects of the face is its ability to convey emotion. We see someone smiling and know they are happy, we see someone frowning and know they are angry – or are they? Perhaps the smile was forced, a deliberate attempt to appear happy while feeling quite the opposite, and perhaps the frown indicates deep concentration, not anger at all. Although we are surrounded by expressive faces, there is still considerable controversy about how they communicate and what they really reveal.

Debate surrounds questions about whether the face reveals emotions subconsciously or whether it is primarily a source of intentional communication. Debate surrounds questions of whether our interpretation of emotions as revealed by face is innate, and thus cross-cultural, or learned, and thus subject to cultural variation [13][35]. Debate even exists about what emotions are [18] and whether they even exist or if they are a non-scientific construct, cobbling together disparate features ranging from physiological state to intent [17].

The most prevalent conceptualization of the relationship between the face and emotions is what Russell and Fernández-Dols call the Facial Expression Program [35], which has roots in Darwin’s writings about the face [8] and is elucidated in the work of Izard [23], Ekman and others. The key ideas in this model are that there are a number of basic, universal emotions (7 is an often cited number: anger, contempt, disgust, fear, happiness, sadness and surprise), that the face reveals one’s internal emotional state, though one may attempt to hide or distort this expressive view and that observers of the face generally are able to correctly read the underlying emotion from the facial expression [35]. Ekman’s work has been quite influential in the computer graphics field, and this conceptualization of the relationship between emotions and facial expression underlies much research in facial animation (e.g. [47]).

In the context of designing face-based interfaces for mediated communication systems the debate about emotional expression vs. the communication of intent is especially relevant. Ekman’s work emphasizes the expressive, often subconscious, revelatory side of facial expressions - indeed, one major branch of his research is the study of deception and the involuntary cues in facial expression and gesture that reveal that one is lying [12]. From this perspective, the advantage to the face is for the receiver, who may gain a truer sense of the other’s intent by the involuntary cues revealed by the face (as well as gesture, tone of voice, etc.) than from the more deliberately controlled words. This model is rejected by Fridlund, who claims that the face’s communicative functions must be to the advantage of the face’s owner for if expression revealed information to the advantage of the receiver and the disadvantage of the owner, it would be evolutionarily untenable [17].

As a design problem, the issue becomes one of control – is the facial display controlled deliberately by the user or is it driven by other measurements of the user’s affective state? If the display is the user’s actual face (e.g. video) then the question is moot, it is the face, which may be displaying affective state or intentionality or both, but the system does not change this. If, however, the expressions on the facial display are driven by something else, the decision about what that something is important. To take two extremes, a very deliberate facial expression model is implemented when the face is controlled by pressing a button ("The Mood command opens a cascading menu from which you can select the facial expression of your avatar. Alternatively you can change the mood of your avatar by pressing one of the function keys listed in the cascading menu or use the mood-buttons in the toolbar."), as opposed to one in which the face's expression was driven by affective data gathered from sensors measuring blood pressure, heart rate, breathing rate, and galvanic skin response – bodily reactions that provide cues about one's affective state [32].

How universal vs. subjective is the interpretation of facial expression is also controversial. Even the smile, which seems to be the most universally recognized and agreed upon expressions, is used quite differently in different cultures. When it is appropriate to smile, for how long, etc. is culturally dependent. Much of the meaning we read in an expression has to do with minute timings and motions – what makes a smile seem like a smirk?

Context is also essential for understanding facial expression. Fernández-Dols and Carroll [15]caution that most studies of facial expression have been carried out without taking context into consideration, referring not just to broad cultural contexts, but the ubiquitous immediate context of any interaction. They point out that facial expressions carry multiple meanings and that the observer uses contextual information to interpret them. This is an important feature to keep in mind in understanding mediated faces, for mediated discussions occur in complex, bifurcated settings, where each participant is simultaneously present in an immediate and a mediated context. The smile I perceive may be one you directed at me – or it may have been triggered by an event in your space which I am not privy to. Such mixing of contexts occurs in real life too, for one's thoughts, as well as one's surroundings, constitute a context: "What are you smiling about?" "Oh nothing, I was just remembering something..." But in a mediated situation, with its multiple contexts, the observation of expressions triggered by and intended for other contexts may be a common occurrence.

3.4 Gaze

Gaze – where one is looking – is an important channel of social information[1][4][22][44]. We are quite adept at perceiving gaze direction (aided by the strong contrast between the eye's white cornea and colored iris) and use it, along with other contextual information, to infer other people's state of mind. Gaze is used in conversation, to determine whether someone is turning the floor over to another or is thinking about what to say next. Gaze is used to disambiguate language: I'm talking to "you", you're welcome to "that". Gaze is both input and output: we look at something or someone because we are interested in them and our interest is revealed by the visible direction of our gaze.

The rules that govern how gaze is used in communication are complex and culturally dependent. Studies of gaze in conversation (see, for instance [1] or [22]) show an intricate ballet of words, gestures, and eye-movements that taken together are used to negotiate turn-taking, establish social control, reflect levels of intimacy, and indicate understanding and attention [4]. Research on gaze often focuses on its role as an indicator of attention. Yet in social communication, gaze has many functions – and averted eyes may not be an indication of averted attention. In a typical conversation, the speaker looks at the listeners to monitor their level of agreement and understanding, to direct an utterance at particular individuals, to command attention or persuade. The speaker may look away from the listeners in order to concentrate on a complex cognitive task, such as thinking about what to say next, or from embarrassment or discomfort (typically, speakers look at the listeners about 30-40% of the time [1]). Listeners look at the speaker more (about 60-70% of the time) and gaze directed at the speaker may signal agreement or it be an attempt to gain a turn. The listener's averted gaze may indicate very close concentration – or complete lack of attention. Furthermore, the length of time it is socially comfortable for two people to look at each other depends on their relationship: strangers look at each other more briefly and less frequently than acquaintances do, and prolonged mutual gaze is a sign of romance and intimacy [1].

There have been numerous attempts to bring gaze to computer mediated conversations. The problem – to show where each person is looking – is deceptively simple, but remains imperfectly solved. Some interfaces, such as many avatar-based graphical chats and current multi-party videoconferencing systems, simply ignore the problem, leaving the avatars to gaze off in random directions and the videoconference participants to appear in separate windows, each appearing to look intently at a spot just beyond the viewer's shoulder. Some interfaces take a very simplistic approach to gaze, using it to broadly indicate attention (e.g. [9]) but ignoring the myriad other social cues gaze provides. Some interfaces do attempt to recreate meaningful gaze in a mediated environment, but these quickly become immense and baroque systems: Hydra[37], a relatively simple system, requires *n*(n-1)* cameras and monitors (where *n* is the number of participants) and Lanier describes an immersive approach [26] that uses numerous cameras, fast processors and more bandwidth than is available even at high-speed research hubs to facilitate a casual conversation in not-quite-real time.

Bringing gaze to the mediated world is difficult because gaze bridges the space between people – and the people in a mediated conversation are not in the same space. Addressing this problems requires not only developing a means for the participants to signal meaningful gaze patterns but creating a common, virtual space for them to gaze across.

Addressing this problem means finding some way to create a common, virtual space, as well as finding a way for the participants to control their gaze, whether algorithmically (as in [46]) or by detecting where they are actually looking (as in [26]).

With videoconferencing, the basic problem is that no common space is shared by the participants. With a two person system, the camera can (more or less) function as a stand-in for the one's conversational partner: when one looks at the camera, it will appear as if one were looking at the other person. The camera must be appropriately located; ideally, it is coincident with the video image of the other's eyes – and challenges are generated by both the opacity of video screens and the mobility of people's heads. Once there are more than two participants, the problem becomes far more difficult, for a single camera cannot stand-in for more than one person.

With avatar systems, the problem is that the user must somehow convey where he would like his avatar to be depicted gazing. Here, the act of indicating gaze is separated from the process of looking; the challenge is to motivate the user to provide this attention indicating information.

The face is highly expressive and informative, but it is not a quantitative graph. Almost everything it conveys is somewhat ambiguous and subjective, open to a range of interpretations and strongly colored by the observer's context. I may find a particular person's face to seem very warm and friendly, with a touch of mischievous humor – and much of that interpretation may be because of a strong resemblance of that person's structural features to those of a friend of mine, whose personality I then ascribe to the new acquaintance. Even something as seemingly objective as gaze is subjectively interpreted. If you are looking at me from a video window and you appear to glance over my shoulder, I may instinctively interpret this as meaning your attention is drawn to the activity occurring behind me, rather than to the activity in your own space beyond the camera.

4 Ways of bringing the face to the interface

Once one decides to create a mediated social environment that includes faces, there are many ways of bringing the face to the interface. The face may be a photographic likeness of the person it represents, or it may be a cartoon visage, conveying expressions but not identity. The face may be still or in motion, and its actions may be controlled by the user's deliberate input or by autonomous algorithms. Each of these design decisions has an impact on the technological requirements and complexity of the system and significantly changes the social dynamics of the interface.

Bringing the face to the interface is a difficult problem and all of today's systems are steps towards achieving an ultimate goal, with many more steps yet to go.

For many researchers, the ultimate goal is to achieve verisimilitude, to make the mediated encounter as much like the experience of actually being in the same place as possible. Most work in video-based conferencing shares this goal, especially research in computationally sophisticated approaches such tele-immersion [26], in which multiple distant participants interact in a common virtual space. Some of the problems in this domain, such as today's poor image quality and lag, can be solved through increased bandwidth and computational power. Yet there are still immense challenges here; in particular, the need to create a common virtual space for the interaction while simultaneously depicting the subtle expressive shifts of the participants.

Yet verisimilitude is not the only goal. Hollan and Stornetta [21] termed reproducing reality as "being there" and urged designers to go "beyond being there", to develop new forms of mediated interaction that enable people to communicate in unprecedented ways that aim at being "better than reality". For example, we may wish to have an interface that uses an expressive face with gaze to provide the sense of immediacy, presence, and the floor control that we get in real life, but which does not reveal the user's identity. We may wish to have faces that change expression in response to the user's deliberate commands or, conversely, in direct response to the user's affective state as analyzed by various sensors. We may wish to have faces that function as a visualization of one's interaction history, an online (and hopefully benign) version of Wilde's *Picture of Dorian Gray.* Or faces that start as blank ciphers and slowly reveal identity cues as acquaintances grow closer. Some of these possible interfaces are relatively simple to implement, others are even more difficult than attempting verisimilitude. And they present a further design challenge, which is to know which, out of the universe of possible designs, are the useful, intriguing, intuitive designs.

4.1 Video and the quest for verisimilitude

Video technology makes it possible to transmit one's image across a network, to be displayed at a distant location. Video has the advantage of letting one's natural face be the mediated face. A slight smile, a fleeting frown, raised brows – expressive nuances are transmitted directly. Video reveals personal and social identity: you appear as your recognizable self.

Video can make people self-conscious. In real life, we speak, act, gesture without seeing ourselves; videoconferences often feature a window showing you how you appear to others. Also, online discussions may be recorded. The combination of appearing as oneself and seeing oneself in a possibly archived discussion can greatly constrain one's behavior. The desirability of this restraint depends on the purpose of the forum; it is neither inherently good or bad.

Contemporary videoconferencing technology has one camera per participant and each participant's image and audio is transmitted to all the others. The quality of the transmission is often poor, due to limited bandwidth. As we discuss the advantages and drawbacks of video as a conversational interface, we will attempt to separate problems that are solvable with increased computational power and faster networks from those that are inherent in the medium.

Video reveals identity, but it is not the same as being there. Studies indicate that although the face's identity cues are transmitted via video, something is lost in the process. Rocco [34] observed that people often need an initial face to face meeting to establish the trust needed to communicate well online, whether using text or video. This may be primarily due to the poor quality of today's video channel, which loses and distorts social cues by introducing delays and rendering gaze off axis. For instance, given limited bandwidth, it is known that given limited bandwidth, reducing audio lag is most important and that eliminating motion lag is more important than reproducing spatial detail[31], yet many social cues, such as subtle expressions, may be lost without this detail. The timing delays that do exist are jarring and can give a distorted sense of the other's responsiveness, interest, etc. While the delays may be measurably slight, they are perceptually significant, potentially creating a quite misleading (and generally not terribly flattering) impression of the other, an impression that might be interpreted as awkward, unfriendly, shifty, etc. - but is purely an artefact of the technology.

Video does improve social interactions, as compared with audio-only conferencing. Isaacs and Tang's research comparing collaboration via videoconferencing with audio conferencing and with face to face meetings has many interesting observations about the social role of the mediated face[22]. They found the greatest advantage of video to be making the interactions more subtle, natural and easier. They point out that while it may not make a group of people do a task more quickly (the sort of metric that has often been used to measure the usefulness of the video channel), it provides an important channel for social messages. For instance, it helps to convey one's level of understanding and agreement: people nod their heads to indicate they are following an argument, and may lift their eyebrows to show doubt, tilt their heads to indicate skepticism or frown to indicate confusion. Video is useful in managing pauses: one can see whether the other person is struggling to find the right phrase or has been interrupted by another activity. Video, they said, "adds or improves the ability to show understanding, forecast responses, give non-verbal information, enhance verbal descriptions, manage pauses and express attitudes... Simply put, the video interactions were markedly richer, subtler and easier than the telephone interactions."

Yet video also has some inherent drawbacks. Isaacs and Tang [22] enumerated a number of videoconferencing weaknesses, noting that it was "difficult or impossible for participants to: manage turn-taking, control the floor through body position and eye gaze, notice motion through peripheral vision, have side conversations, point at things in each other's space or manipulate real-world objects." These drawbacks arise because the participants do not share a common space.

Isaac's and Tang found these problems even in two person videoconferences. A key problem is gaze awareness: if I look at your image, I am not looking at the camera and the image you see appears to be gazing elsewhere. While this can be addressed with clever use of half-silvered mirrors and integrated camera, the gaze does not match our real world expectations. Indeed, being close may be worse, for once the awareness of the camera is lost, we attribute any oddity of gaze behavior to intent, rather than to the technology.

These problems are exacerbated once there are more than two participants. With two people, it is theoretically possible for the camera's to transmit from at least an approximately correct point of view; with more, it is not, at least not without more cameras. There have been a number of experimental designs made to address this problem. These fall into two categories: one can use multiple cameras and displays to extend the one-to-one videoconference model (e.g. Hydra [37]) or one can use a combination of three-D modelling and head-tracking gear to create a video driven synthetic space (e.g. tele-immersion [26]).

With the former approach, multiple cameras and displays are placed throughout one's space. Each participant is seen in his or her individual monitor and the setup is replicated at each site. For instance, a camera/monitor setup can be placed in each seat at a conference table, with each camera facing the one live person in the room. The video from the camera associated with your image at every node needs to be sent to you, as it then shows that person from the correct angle, as if you were looking at them from your seat. If implemented correctly, this method allows multiple participants to indicate attention by looking at each other and to share a common space, at least to the extent that the physical environment is replicated at each site. This approach requires multiple installations and (N)*(N-1) cameras and monitors. It provides little flexibility (e.g. one cannot leave one's seat to chat quietly with another person[2]). In the reduced case of N=2 participants, it is indistinguishable from one on one video conferencing, and thus shares the aforementioned advantages and disadvantages.

The latter approach attempts to create an environment that seamlessly blends the local and the remote in a common virtual space. Multiple video cameras capture the actions of each participant and using location information from various sensors and a considerable amount of computational power, each participant is mapped into a common virtual world. Such a system is far from implementation today and Lanier's estimates of the computational and network requirements for even minimally acceptable levels of detail put it at least 10 years in the future [26]. Furthermore, the quantities of gear required – cameras, head-tracker, eye-trackers, etc. – make the experience far from the seamless de-spatialization of daily experience that is the goal.

Ten years – or even twenty or fifty years – is a long time off, but it is not forever. We can assume that something like a seamless tele-immersive environment will one day exist, realistic enough to be just like being there. We will then have mediated environments in which the face, with all its expressive and revelatory powers, exists much as it does in daily life. We turn now to considering approaches to the mediated face that go beyond being there.

4.2 Avatars and the quest for expression

There are numerous and varied ways of bringing faces to the interface that do not attempt to fully imitate real life. There are simple graphical avatars and intelligently animated agents. There are video windows in virtual space and sensor-driven cartoons. A simple photograph replicates the user's appearance, but does not convey dynamically changing expression and gaze. A cartoon avatar may have a fictional visage while deriving its expression from an analysis of the user's speech.

There are a number of reasons why one would want to use a synthetic face. First, it supports interaction among large numbers of people in a common virtual space. The difficulty with video-based systems is integrating a number of separate spaces into a common environment; once one is no longer trying to bring in disparate real world elements, the common space problem disappears. Second, it allows for communication without necessarily conveying identity. Text-based online discussions support the full spectrum of identity presentation, from authenticated veracity to absolute anonymity: synthetic images can provide the same range within a graphical context (a synthetic image may be entirely fictional or it can be derived from photographic and range data of the real person).

The goal with many systems is to bring the expressive qualities of the face to a virtual world; the challenge is sensing and producing expression in a socially meaningful

[2] An interesting solution to this problem is Paulos and Canny's work on personal tele-embodiment using remote controlled mobile robotic devices that incorporate two-way video communication [30].

way. Such systems are still at the very early stages of development. Commonly used avatar programs have only the most primitive style of expressive input (and output): expression buttons and keyboard shortcuts that let the user change the avatar's face to sport a smile, frown, etc. [19].

While these systems are simple, I will argue here that simplicity alone is not a problem, nor is complexity always desirable. Rather, the key is a balance between the information provided and the message that is sent. If minimal information is provided, a minimal message should be sent. The problem with many face-based interfaces is that they are sending too complex a message upon the receipt of too little data. The face is so highly expressive, and we are so adept at reading (and reading into) it, that any level of detail in its rendering is likely to provoke the interpretation of various social messages; if these messages are unintentional, the face is arguably hindering communication more than it is helping.

One solution is to stick with very simple faces. The ubiquitous "emoticons" – typed symbols that resemble sideways faces, e.g. the smile :-) the frown :-< and the wink ;-) – are extremely simple, yet function quite well at helping to communicate expressive information that clarifies the sender's intention. E-mail is notorious for generating anger due to miscommunication of irony, sympathy etc. Emoticons can make it clear that a statement is meant in jest, or that a writer is deploring, rather than celebrating, the incident they are reporting. Essentially new forms of punctuations, emoticons spread quickly because they were intuitive as well as needed. Their reference to familiar iconic facial expression makes them immediately accessible to readers[3].

Creating an avatar that is even somewhat reminiscent of a human being brings into play numerous requirements about its behavior. For instance, if I use a plain circle as the user's representation (see [45] for an example), I can move this circles across the screen by sliding it, and the movement seems perfectly reasonable. If I decide to use a more human-like representation and create an avatar with legs, then sliding it across the screen seems awkward – the avatar appears passive and inert. The legs make me want to have it walk, and to do so, one may either have the user painstakingly render each step, or have an automatic walking algorithm. The hand rendered one, far from being more expressively communicative, puts an onerous burden on the user, who must expend so much attention getting the avatar to put one foot in front of the other, that he or she has little time left over for actually communicating with others. So, one equips the avatar with automated walking algorithms. A simple interface might ask the user for a destination and would take care of getting the avatar there. Now, a behavior such as walking has some social information in it: we read moods, such as whether one is buoyant or dejected, from gait, as well as characteristics ranging from athleticism to sexual attractiveness By providing the avatar with legs we then require it to walk, and walking is inherently expressive. All that the user has indicated is an endpoint, but via the avatar, has communicated much more.

[3] Although cultural differences occur even here. Japanese emoticons differ from Western ones. For instance, in Japan, women are not supposed to show their teeth when smiling, as is depicted in the female emoticon smile (.) And the second most popular icon is the cold sweat (;), with no clear Western equivalent [33]

The same is true of the face. Once there is a representational avatar, it requires behaviors and behaviors are expressive, introducing the big question of whether it is expressing what the person behind it wishes to express.

An interesting example is provided by Vilhjálmsson and Cassell's *BodyChat* [46]. Here, humanoid avatars in a chat environment are provided with automated social actions. The user indicates to the system the social actions he or she would like to perform and the avatar then performs a series of visible actions that communicate this intention. For instance, to indicate a desire to break away from a conversation, the user puts a "/" at the beginning of a sentence; the avatar then accompanies those words with a diverted gaze. If the other person responds with a similarly prefixed sentence, the conversation ends with a mutual farewell; if not, the conversation continues, until both parties produce leave-taking sentences. While the developers of *BodyChat* have addressed the whole body problem of avatar physical behavior, their approach – and the issues it raises – can be considered primarily in the realm of the face.

A key issue this highlights is communicative competence. The social signals that I send when I greet someone or take leave are not simply informative actions, but also displays of communicative competence. Let's compare the input and the output in this situation. In the real world, I decide I've had enough of the conversation - perhaps I am bored, perhaps I am late for another appointment, perhaps I sense that the other person needs to go and I don't want to detain them, perhaps a combination of all three. In each of these cases, the gestures I make to indicate leave-taking may be quite different – I may look around for a distraction, I may glance at my watch, or I may look directly at the other person as I take my leave. Each of these conveys a different message and each also expresses a different level of politeness and competence. If I am leaving because I sense the impatience of the other, the impression I convey will be quite different if I look down at my shoes, mumble goodbye and flee, or if I graciously and warmly shake hands, say some pleasant farewells, and go. My actions upon taking leave are modified by both my immediate motivations and my underlying social knowledge and style. As a participant in a conversation, I gather a lot of information from the leave-taking behaviors, only one bit of which is that the other intends to leave. I also get a sense of the leave-taker's reasons for leaving, level of concern for my feelings, social sophistication, etc. In the *BodyChat* system, the user conveys only that one bit - the forward slash that says "I intend to leave". The systems expands it into a more complex performance, designed to draw upon our social knowledge – a performance that the receiver interprets as the sender's intent. The problem is, much of that performance has nothing to do with anything that the sender intends. Is it better to have unintentional cues than none at all? The answer depends on the context - it is again a design decision. Vilhjálmsson and Cassell state that their research goals include pushing the limits of autonomous avatar behavior "to see how far we can take the autonomous behavior before the user no longer feels in control". Understanding these limits is an important contribution to understanding how to integrate the face into mediated communications.

There are numerous other approaches to creating mediated faces. Some use as their input the user's writing [28][29] or speech [11] to derive expression and drive the animation. Like *Body Chat* these systems all introduce some unintentional expressivity, for they are all translation systems, transforming their input into a model of the user's inner state or intentionality and then representing that state via an animation. Perhaps, as Neal Stephenson suggests in his novel *Snowcrash*[40], future expressivity will come in our choice of autonomous behavior avatar modules, much as we express ourselves via clothing today.

Systems that use video images or other measurements of the face to animate facial models ([5][14]) are interesting, for they do no such translation. Here, although the rendered face may be completely fictional (or photorealistic - such systems can thus run the gamut from anonymous to identified), its expressions, whether deliberate or subconscious, are derived directly from the user's face; it is the facial expressions themselves that are re-presented, not an implicit state.

5 Conclusion

The key problem in bringing the face to a mediated environment is to balance input and output. In our real world face, there are millions of "inputs" controlling the highly nuanced features, from the genes that determine the basic facial structure to the nerves and muscles that control the lips, eyes, and eyebrows. In the virtual world, the control structure is much coarser. We must understand what is the communicative ability of the system we create, and match the face to it. The face is an extraordinarily rich communication channel and a detailed face conveys a vast amount of subtle information, whether we wish for it to do so or not.

References

1. Argyle, M. and Cook, M.: Gaze and Mutual Gaze. Cambridge University Press, Cambridge (1976)
2. Aronson, E. The Social Animal. Freeman, NY (1988)
3. Ayatsuka, Y., Matsushita, N., Rekimoto, J.: ChatScape: a Visual Informal Communication Tool in Communities. In: CHI 2001 Extended Abstracts (2001) 327-328
4. Bruce, V. & Young, A.: In the eye of the beholder: The science of face perception. Oxford University Press, Oxford UK. (1998)
5. Burford D. and Blake, E.: Real-time facial animation for avatars in collaborative virtual environments. In: South African Telecommunications Networks and Applications Conference '99, (1999) 178-183
6. Chernoff H.: The use of faces to represent points in k-dimensional space graphically. In: Journal of American Statistic Association, Vol. 68 (1973) 331-368
7. Choissier, B.: Face Blind! http://www.choisser.com/faceblind/
8. Darwin, C. and Ekman, P. (ed.): The Expression of the Emotions in Man and Animals. Oxford University Press, Oxford UK (1998)
9. Donath, J.: The illustrated conversation. In: Multimedia Tools and Applications, Vol 1 (1974) 79-88.
10. Donath, J.: Identity and deception in the virtual community. In: Kollock, P. and Smith, M. (eds.): Communities in Cyberspace. Routledge, UK (1998)

11. P.Eisert, S. Chaudhuri and B. Girod.: Speech Driven Synthesis of Talking Head Sequences. In: 3D Image Analysis and Synthesis, Erlangen (1997) pp. 51-56
12. Ekman, P.: Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage. New York: W. W. Norton. (1992)
13. Ekman, P.: Should we call it expression or communication?. In Innovations in Social Science Research, Vol. 10, No. 4 (1997) pp 333-344.
14. Essa, I, Basu, S. Darrell, T. Pentland, A.: Modeling, Tracking and Interactive Animation of Faces and Heads using Input from Video. In: Proceedings of Computer Animation '96 Conference, Geneva, Switzerland, IEEE Computer Society Press (1996)
15. Fernández-Dols, J. M and Carroll, J.M.: Context and Meaning. In: Russell, J.A, Fernández-Dols, J. M. (eds.): The Psychology of Facial Expression. University of Cambridge Press, Cambridge, UK (1997)
16. Forchheimer R. and Fahlander, O.: Low Bit-rate Coding Through Animation. In: Proceedings of Picture Coding Symposium. (March 1983) 113-114
17. Fridlund, A.J.: The new ethology of human facial expression. In: Russell, J.A, Fernández-Dols, J. M. (eds.): The Psychology of Facial Expression. University of Cambridge Press, Cambridge, UK (1997)
18. Frijda, N.H., Tcherkassof, A.: Facial expressions as modes of action readiness. In: Russell, J.A, Fernández-Dols, J. M. (eds.): The Psychology of Facial Expression. University of Cambridge Press, Cambridge, UK (1997)
19. Fujitsu Systems. New World Radio Manual. http://www.vzmembers.com/help/vz/communicate.html (1999)
20. Herring, S.: Gender differences in computer-mediated communication. Miami: American Library Association (1994)
21. Hollan, Jim, and Stornetta, Scott. Beyond Being There. In Proceedings of CHI '92
22. Isaacs, E. & Tang, J.: What Video Can and Can't Do for Collaboration: A Case Study. In: Multimedia Systems, 2, (1994) 63-73.
23. Izard, C.E.: Emotions and facial expressions: A perspective from Differential Emotions Theory. In: Russell, J.A, Fernández-Dols, J. M. (eds.): The Psychology of Facial Expression. University of Cambridge Press, Cambridge, UK (1997)
24. Johnson, M., Dziurawiec, S., Ellis, H. and Morton, J.: Newborns' preferential tracking of face-like stimuli and its subsequent decline. Cognition. 40. (1991)1-19.
25. Kunda, Z.: Social Cognition: Making Sense of People. Cambridge, MA: MIT Press. (1999)
26. Lanier, J.: Virtually there. In: Scientific American, April 2001. pp 66-76 (2001)
27. Nakanishi, H., Yoshida, C., Nishimura, T. and Ishida, T.: FreeWalk: Supporting Casual Meetings in a Network. In: Proceedings of ACM Conference on Computer Supported Cooperative Work CSCW'96, (1996) 308-314
28. Nass, C., Steuer, J. and Tauber, E.: Computers are Social Actors. In: Proceedings for Chi '94., 72-78 (1994)
29. Ostermann, J., Beutnagel, M. Fischer, A., Wang, Y.: Integration of Talking Heads and Text-to-Speech Synthesizers for Visual TTS. In: Proceedings of the International Conference on Speech and Language Processing, Sydney, Australia (1998)
30. Paulos, E. and Canny, J.: Designing Personal Tele-embodiment. In: IEEE International Conference on Robotics and Automation. (1998)
31. Pearson, D. E., and Robinson, J. A.: Visual Communication at Very Low Data Rates. In: Proceedings of the IEEE, vol. 4, (April 1985) 795-812
32. Picard, R. Affective Computing. MIT Press, Cambridge, MA (1997)
33. Pollack, A.: Happy in the East (--) or Smiling :-) in the West. In: The New York Times, (Aug. 12, 1996) Section D page 5

34. Rocco, E.: Trust breaks down in electronic contexts but can be repaired by some initial face-to-face contact. In: Proceedings of CHI ‘98. (1998) 496-502.
35. Russell, J.A, Fernández-Dols, J. M.: What does a facial expression mean? In: Russell, J.A, Fernández-Dols, J. M. (eds.): The Psychology of Facial Expression. University of Cambridge Press, Cambridge, UK (1997)
36. Scheirer, J., Fernandez, J. and Picard, R.: Expression Glasses: A Wearable Device for Facial Expression Recognition. In: Proceedings of CHI '99, Pittsburgh, PA (1999)
37. Sellen, A., Buxton, W. & Arnott, J.: Using spatial cues to improve videoconferencing. In: Proceedings of CHI '92 (1992) 651-652
38. Sproull, L. and Kiesler, S. Connections. Cambridge: MIT Press (1990)
39. Sproull, L, Subramani, R., Walker, J. Kiesler, S. and Waters, K.: When the interface is a face.In: Human Computer Interaction, Vol. 11: (1996) 97-124
40. Stephenson, N. Snow Crash. Bantam, New York (1991)
41. Suler, J.. The psychology of avatars and graphical space. In: The Psychology of Cyberspace, www.rider.edu/users/suler/psycyber/psycyber.html (1999)
42. Terzopoulos D.and Waters, K.: Analysis and synthesis of facial image sequences using physical and anatomical models. In: PAMI, 15(6) (1993) 569--579
43. Valente, S. and Dugelay, J.-L.: Face tracking and Realistic Animations for Telecommunicant Clones. In: IEEE Multimedia Magazine, February (2000)
44. Vertegaal, R.:. The GAZE Groupware System: Mediating Joint Attention in Multiparty Communication and Collaboration. In: Proceedings of CHI ‘99. Pittsburgh, PA. (1999)
45. Viégas, F. and Donath, J.: Chat circles. In: Proceeding of the CHI ‘99 conference on Human factors in computing systems, (1999) 9 - 16
46. Vilhjálmsson, H.H.and Cassell, J.: BodyChat: autonomous communicative behaviors in avatars. In: Proceedings of the second international conference on Autonomous agents. Minneapolis, MN USA. (1998) 269-276
47. Waters, K.: A Muscle Model for Animating Three-Dimensional Facial Expression. In: ACM Computer Graphics, Volume 21, Number 4, (July 1987)
48. Whyte, W.: City. Doubleday, New York. (1988)
49. Zebrowitz, L.: Reading Faces. Westview Press, Boulder, CO. (1997)