Mediating human interaction with video: the fallacy of the talking heads

Harro Vonsa,*, Roel Vertegaalb, Gerrit van der Veerc

aBaan Development, Barneveld, The Netherlands
bErgonomics Department, Twente University, Twente, The Netherlands
cVrije Universiteit, Amsterdam, The Netherlands

Abstract

Given that communication network bandwidth remains at a premium, the design of multiparty mediated systems is a problem of conveying the least redundant information first. We investigated the redundancy of gaze direction in relation to typical full-motion visual cues in performing a triadic abstract office communication task. The results of the experiment indicate that the use of still-video does not necessarily have a negative impact on task performance. Gaze directional cues may provide extra regulative information, but with regard to task performance in small, well-organized groups this information seems to be largely redundant.

Keywords: Interaction design; Collaborative work; Videoconferencing

1. Introduction

Face-to-face communication is generally regarded as an optimal form of human communication. It was thought that by adding video images to telephony-based applications, video mediated communication (VCM) would approximate a face-to-face situation more closely.

However, although high quality full-motion video images add a rich subset of visual information to audio-only communication, it is less clear what the benefit of such relatively expensive information is to the efficiency of, for example, office communication tasks. According to Short [17], many of the visual cues conveyed by current-day VMC systems are also redundantly coded in speech. Humans are so flexible in dealing with redundantly coded information that the removal of, for example, information about facial expressions may be compensated relatively easily. Indeed, several studies have shown that in terms of performance and other measures, there appears to be surprisingly little difference between groups working together by means of an audio-only communication link and groups working together by means of VMC [16].

In addition to this, usability studies indicate that current VMC systems do not seem to solve a problem typical to multiparty mediated communication: knowing who is talking to whom [9,23]. Isaacs and Tang [9] suggest this may be due to the lack of information about the gaze direction of others, implying that this cue carries information not coded by means other than explicit verbal references.

So there seems to be a mismatch in functionality of current-day VMC systems with regard to everyday communication needs. Before engaging in a substantial investment in terms of the computer hardware and communication networks, it is important to obtain a closer understanding of the functional requirements of such systems. In order to assess the minimum requirements for VMC systems for joint tasks, we analysed the visual cues conveyed in typical VMC systems in terms of their functionality, and conducted an experiment. In this experiment, we attempted to gauge the relative importance of conveying full-motion visual cues and gaze directional cues in a multiparty mediated communication task.

2. The visual channel in VMC

2.1. Human communication with the visual channel

Usually, VMC systems convey people’s faces and the upper part of their torsos. This means the following visual cues can be used to express a message and regulate interaction:

1. gaze direction (the absolute direction in which someone looks), and related to it;
2. eye contact (two people aiming their gaze direction at each other’s eyes);
3. lip movements;
4. facial expressions and head movements; and
5. physical appearance.

2.1.1. Gaze direction
In expressing things, gaze direction is important for giving deictic references (indicating what is meant by words like ‘this’ and ‘that’) [2] and to indicate the focus of attention [22]. For regulation of interactions, gaze direction serves several purposes, including [2,7];
1. turn-taking (changing the role of speaker and listener): when a speaker just pauses and wants to keep the turn, he may indicate this by looking away from his listeners;
2. back channels: people typically look at their listeners at the end of a long utterance, to know whether they are still attending;
3. communicating emotions: someone telling a happy experience will look at the listeners a lot, while he will cast his eyes mainly downwards when telling a sad experience; and
4. communicating the nature of relationships: people who like each other, or are in love, are seen to have more frequent and longer eye contact than others.

Eye contact occurs when two people aim their gaze direction at each other’s eyes. It is seen to be used for turn-taking, to communicate emotions, and to establish a sense of social presence [2,12].

In conventional VMC systems, gaze direction and eye contact are rarely conveyed. To mediate gaze, people will have to look into the camera so that their participants perceive they are looking at them. Since most systems use one camera per participant, it is not possible to gaze at one individual at a time. Eye contact is also difficult because people will look at the screen instead of into the camera to look at someone’s eyes. This results in asymmetry: person A tries to establish eye contact with person B, who does not perceive this cue and will not respond to it by gazing at A.

2.1.2. Lip movements
Lip movements contribute to the perception of speech. People are seen to use the movements of the speaker’s lips in interpreting the speech sounds. Misunderstandings occur when these movements do not match the speech sound. In such cases, people typically seem to hear a speech sound which is a compromise between the visual and auditory information (the McGurk effect). No problems seem to occur when lips are shown that do not move at all [10]. Since VMC systems convey moving lips with video and speech with audio, delays between the audio and video signals may result in lip movements that differ from speech sounds. According to Steinmetz, people start noticing delays between audio and video signals when the time difference between these signals becomes more than 80 ms. The perception of these delays does not seem to be influenced by the size of the lips on the screen. People seeing only the speaker’s head, his head and shoulders or his whole body all show the same sensibility to delays [18].

2.1.3. Facial expressions and head movements
Since VMC generally shows the participants’ faces with motion video, movements by the head and of the face influence the communication. Facial expressions may be used to communicate emotions and attitudes such as liking or hostility. Head movements such as shaking the head or nodding can be used instead of saying ‘yes’ or ‘no’ [2]. A lot of expressions are possible in the facial region: some 250,000, many of which are very much alike, others meaningless. Altogether there seem to be about 30 distinguishable combinations of head movements and facial expressions which can be used during conversations. These are mostly used for connotating the contents of messages [4].

2.1.4. Physical appearance
The way people appear to others, their physical appearance, is seen to influence the interaction between them. Argyle found a number of features people use to judge attractiveness and personal attitudes of their conversational partners, some of which can be mediated by VMC. These involve [2]:
1. the regular features of the face: people seem to find full lips, a soft skin, smiling expressions and dilated pupils [11] attractive;
2. hair style: people use this to judge social class, political class, political conviction and personality;
3. clothes and accessories carry similar non-verbal signals.
   They signal age, status and importance. Together with hair style, they are strongly gender dependent.

Physical appearance is used in grounding (judging the mutual knowledge and background relevant to the interaction [6]), and is mainly seen to influence an interaction at the beginning of a meeting. Beyond the introductory phase, physical appearance as such only plays a minor role in the interaction.

2.2. Mediating the visual channel in VMC
Although we now know a great deal about the functions of visual cues in VMC systems, we cannot simply predict the effectiveness of a communication system by listing the number of cues conveyed by its channels. In communication, people use both the visual and the auditory channel to convey information in different ways, ways which may be more or less appropriate to express a particular message, or to regulate the exchange of that message. However, a typical strategy for ensuring that a message is received correctly is the use of multiple encodings of similar or related information. Such redundant coding schemes are not only useful when there is noise in the channel (e.g. deaf people can
seems to be a form of redundant coding of the underlying motion video, yet much of the information they provide and lip movement belong to this category. They require full-conventional VMC systems. Cues such as facial expressions set of redundantly coded cues that typically is the effect on multiparty task performance of conveying a cue which we believe is hardly redundantly coded by multiparty task performance if we would leave out such motion video cues? Photographs showing the direction in which people look effectively remove such cues, yet are still able to convey physical appearance, gaze direction, and, to some extent, eye contact. We called this type of video still-video, and stated the following hypothesis about it.

2.2.1. Hypothesis 1
In a standardized situation of group communication, still-video which conveys gaze direction will not result in:
1. lower performance on a collaborative task; or
2. a lower subjective quality of the communication, compared to full-motion video which includes gaze direction.

2.2.2. Hypothesis 2
In a standardized situation of group communication, full-motion video which conveys gaze direction will result in:
1. better performance on a collaborative task; and
2. a higher subjective quality of the communication, compared to full-motion video which does not include gaze direction.

3. Materials and methods
We tested the hypotheses with an experiment, in which people communicated by means of a video mediated system. To avoid confounding variables, we used conversations between one subject and two stooges, one male and one female.

3.1. Conditions
We use three separate groups of experimental subjects. To each group, we applied one of three conditions:
A. the subject saw full-motion images of the stooges. By turning their heads, the stooges indicated the direction of their gaze frontal to the subject (as in Fig. 1a), sideways to the other stooge (as in Fig. 1b), or slightly down to the screen (as in Fig. 1c);
B. the subject saw full-motion images of the stooges. The stooges did not show the direction of their gaze; and
C. the subject saw still images of the stooges. By pressing keys on a keyboard, the stooges indicated their gaze direction, and the corresponding still-video images were shown to the subject. These images are shown in Fig. 1.

To give the subjects some sense of eye contact with the stooges, we had the stooges look into the camera from time to time. When looking at the subject, they spent some time looking at their screens, and some time looking into the
camera. Since this might be a confounding variable, we tried to keep the ratio between the time spent looking at the screen and the amount of time spent looking into the camera equal for all conditions.

3.2. Task

The subjects and the stooges were given a group problem solving task, consisting of a number of problems. In each problem, each participant was presented a different fragment of a sentence. Together, they had to construct as many meaningful and syntactically correct sentences of these fragments as possible. After they had given all correct solutions to a problem, the next problem would be presented. The objective of this task was that a more successful interaction between the subject and the stooges resulted in a better performance (more completed problems within 15 minutes).

For the experimental subject each problem appeared on a computer screen, but the stooges had a sheet of paper on which all problems were given. However, they acted as if they had a computer screen just like the subject. To prevent a learning effect of the stooges, they knew all correct solutions to each problem. Their sheets also indicated how many correct solutions they were allowed to give away, and when to give an incorrect answer. This was done to increase the interaction with the subject and to prevent the subject from bearing suspicion.

In this task, each participant had a specific role. The experimental subject’s role was to officially submit each solution they collectively agreed upon to be correct. Then stooge 1 would enter the sentence on a computer keyboard, and stooge 2 would report whether this solution was correct or not, pretending this was indicated on a computer screen. In addition to these roles, the stooges were to ensure active participation of each subject. Following is a typical example of this task.

3.2.1. An example of a typical problem solving process during the experimental task

Suppose the subject sees the following fragment on his screen: ‘‘from now on’’. Stooge 1 tells him he has: ‘‘there will be’’ and stooge 2 reads aloud: ‘‘no more complaints’’. First, they enter the sentence ‘‘from now on there will be no more complaints’’ because this is the most obvious. Stooge 2 tells that this sentence is correct, but nothing else happens, so they decide to try another one. They come up with ‘‘will there be no more complaints from now on’’, but it appears to be wrong, since ‘‘will there be’’ is a manipulation of ‘‘there will be’’, which is not allowed. Because they do not know which sentence they have had, they enter ‘‘from now on there will be no more complaints’’ again. According to Stooge 2 the computer answers this one is correct, but still nothing happens. Only when they enter ‘‘there will be no more complaints from now on’’ is the problem is completed and the subject gets another fragment on his screen.

At the start of the experiment, the subject was given a simple task to perform with the stooges to get used to them, as well as to the system and their roles. Their task was to give as many names of mammals as they could, starting with a ‘‘D’’. Their roles were the same as in the actual experimental task. This task took 1 minute.

3.3. Experimental subjects

In a meeting prior to the experimental sessions, we tested our subjects on their language competence and spatial ability by means of a pen-and-paper test (in Dutch [8], based on [3]). The subjects’ command of (the Dutch) language was likely to be related to their performance on the task since it involved the creation of (Dutch) sentences. We tested their spatial ability, which is related to their intelligence, because this variable might also influence task performance [13,19,20]. To prevent undesirable effects on task performance due to subjects not being able to read from a screen, we also performed an eye test on each subject.

The experimental subjects were recruited amongst university students of both technical and social disciplines. We allocated them to three groups, matched to language competence and spatial ability, as well as to age, sex and field of study. Group A contained 19 subjects (13 male, 6 female, mean age: 21.7), group B 22 (15 male, 7 female, mean age: 21.3) and group C 17 (11 male, 6 female, mean age: 22.2).

The actual experimental sessions were performed with one subject at a time. After each session we tested the subject’s field dependency. Typically, field dependent...
persons tend to be distracted by all sorts of coincidental or seemingly irrelevant visual details. Field independent persons, on the other hand, abstract what they consider to be relevant [25]. This variable could influence task performance because conditions differed in the amount and relevance of details provided (e.g. facial expressions, head movements, etc.). We used a test based on recognizing simple figures, some of which contained irrelevant elements [21,24]. With the results of this test we could statistically control for the effect on task performance of this variable.

3.4. Questionnaire

After the experimental subjects completed their experimental session, they filled in a questionnaire. It contained both open questions and multiple choice questions. Each multiple choice question was answered by selecting one of five options, ranging from strongly positive to strongly negative statements (e.g. “strongly agree”, “agree”, “undecided”, “disagree” and “strongly disagree”). We arranged the order of the options from strongly positive to strongly negative for about half the questions, and in the other direction for the other half. By doing so, we reduced response set.

3.5. Configuration of the subject’s equipment

In terms of equipment, we attempted to construct the minimal setup required to simulate our conditions. The subject’s equipment setup consisted of two video monitors, a computer screen, two cameras with built-in microphones and two loudspeakers. Fig. 2 shows the subject’s equipment configuration in front view (a) and top view (b).

On each video monitor, the subject saw one of the stooges. Because both the full-motion and the still-video images were displayed on these screens by computers, they were of exactly the same size and quality for each condition. The cameras on top of the video monitors provided each stooge with the experimental subject’s sound and image. The stooges always saw one full-motion image of the subject. Through each of the loudspeakers next to the video monitors the subject heard a stooge’s voice. In the full-motion condition without gaze direction (B), the subject only perceived the stooges looking in his direction, and had to think they also saw a similar image of him. To make this more credible, a dummy camera was placed right in front of him, on top of his computer screen.

On the computer screen, the subject was given the information he needed for the experimental task. This information was displayed by a computer controlled by the experimenter who lead the sessions. During a session, this experimenter pressed a key on a keyboard when the subject and the stooges completed a problem. The computer then presented the subject the information for the next problem, and generated a beep to inform the stooges. This computer also registered the number of problems solved by each subject.

The experimenter had a microphone connected to a loudspeaker which was mounted on the back of the computer screen, barely visible from the subject’s perspective. By this speaker the experimenter was able to guide the subject through the experimental session. The video monitors were placed in an angle of 30° with the edge of the table. Both monitors were tiled backwards 5°, and the computer screen was tilted backwards 10°.

The size of the video images was 25 × 18 cm on a 14 inch video monitor in all conditions. There was no delay between the video and audio. The characters of the problem text on the subject’s computer screen were displayed in white on black and were 0.5 cm high (height of the letter “x”). All subjects were seated on a chair with the same fixed position. The seat of this chair was at 47 cm above the floor, and the height of the table was 75 cm. The chair’s back was 37 cm from the edge of the table. The distance from the subject’s head to each monitor was approximately 60 cm.

4. Results

After the experiments, we first checked the data on the field dependency. Analysis of variance (two-tailed at the 0.05 level) showed that this variable did not differ significantly amongst all three conditions. Plots showed that it was normally distributed. Since there was no correlation between subjects’ sequence number and task performance, there was no learning effect for the stooges.

4.1. Task performance

During the experiments the computer registered the
number of problems each experimental subject completed. From this data, we computed scores representing the total number of correct solutions each subject gave. Per condition the means of these scores are presented in Table 1, together with their standard deviations. Analysis of variance (two-tailed at the 0.05 level) showed no significant differences \[ F(2,55) = 0.87, p = 0.42 \].

4.2. Questionnaire

We tested the subjective quality of the interactions by means of a questionnaire with both multiple choice and open questions. We rated the answer to the multiple choice questions from the questionnaire from 1 to 5, in which a score of 1 corresponds to the most negative option and a score of 5 to the most positive option. In four cases a subject chose two adjacent options instead of one. These were replaced by the options with the lowest score. The scores to each question are shown in Table 2. We performed a \( \chi^2 \) test on the scores of each multiple choice question (two-tailed at 0.05). The only question which reached significance was the question whether it was always clear to the experimental subject whom the stooges were talking to \( \[ \chi^2(2) = 10.9, p = 0.004 \] \). In the condition with full-motion video and gaze direction (A) 47% of the subjects answered this question with “clear” or “very clear”. In the full-motion condition without gaze direction (B) this was 45%, but in the still-video condition (C) it was 88%.

The differences in responses to the question whether the subject found the collaboration with the two stooges pleasant were close to significance \( \[ \chi^2(2) = 5.53, p = 0.063 \] \). For the condition with full-motion and gaze direction (A), 84% of the subjects said it was pleasant or very pleasant. For the full-motion condition without gaze direction (B) this was 82%, and for the still-video condition (C) 100%.

5. Discussion

With our experiment, we tested two hypotheses on the influence of video on both the performance on a collaborative task and the subjective quality of the communication. Based on the first hypotheses we expected no negative effect of conveying still-video instead of full-motion video (both

<table>
<thead>
<tr>
<th>Question</th>
<th>Significance</th>
<th>Condition A</th>
<th>Condition B</th>
<th>Condition C</th>
</tr>
</thead>
<tbody>
<tr>
<td>It was hard to create those sentences with the three of us.</td>
<td>Not significant</td>
<td>3.26 (0.872)</td>
<td>3.30 (0.630)</td>
<td>3.53 (0.624)</td>
</tr>
<tr>
<td>The collaboration with the two partners was pleasant.</td>
<td>Not significant ( p = 0.063 )</td>
<td>4.21 (0.713)</td>
<td>3.86 (0.467)</td>
<td>4.25 (0.437)</td>
</tr>
<tr>
<td>This way of communication is pleasant.</td>
<td>Not significant</td>
<td>3.63 (0.761)</td>
<td>3.55 (0.596)</td>
<td>3.50 (0.707)</td>
</tr>
<tr>
<td>This communication system is easy to work with.</td>
<td>Not significant</td>
<td>3.95 (0.911)</td>
<td>4.05 (0.576)</td>
<td>4.06 (0.659)</td>
</tr>
<tr>
<td>It was always clear whom my partners were talking to.</td>
<td>( p = 0.004 )</td>
<td>3.37 (0.684)</td>
<td>3.23 (0.973)</td>
<td>4.09 (0.906)</td>
</tr>
<tr>
<td>I could easily see what my partners were looking at.</td>
<td>Not significant</td>
<td>3.16 (0.900)</td>
<td>3.36 (1.09)</td>
<td>2.97 (1.08)</td>
</tr>
</tbody>
</table>

Table 1
The mean scores on the task for each condition; standard deviations are shown in brackets

<table>
<thead>
<tr>
<th>Condition A</th>
<th>Condition B</th>
<th>Condition C</th>
</tr>
</thead>
<tbody>
<tr>
<td>(full-motion video with gaze direction)</td>
<td>(full-motion video without gaze direction)</td>
<td>(still-video with gaze direction)</td>
</tr>
<tr>
<td>30.9</td>
<td>31.5</td>
<td>28.8</td>
</tr>
<tr>
<td>(7.26)</td>
<td>(6.42)</td>
<td>(5.49)</td>
</tr>
</tbody>
</table>

Table 2
The mean scores for the questionnaire; significances are tested two-tailed at the 0.05 level; standard deviations are in brackets
conditions conveying gaze direction). Based on our second hypothesis, we expected adding gaze directional cues to full-motion video to have a positive effect.

5.1. Task performance

As we expected, there was no significant difference in task performance between the full-motion and still-video conditions. This seems to confirm our hypothesis that the use of full-motion video does not necessarily improve performance in this collaborative task. However, we did expect to find differences in task performance between conditions with and without gaze directional cues. Apparently, the information provided by gaze directional cues was redundant with regard to completing this joint task in this situation.

We assume that gaze direction cues may have been redundant because of the small size of the groups, and the clearly defined roles within these groups.

In four-person conversations, situations can occur in which two pairs of people engage in two different side conversations, introducing two different focuses of attention [15]. Since we used groups of three persons, we eliminated the possibility of side conversations. As a result, it may have been less important for the subject to know whether the speaker was addressing him or the other stooge.

Because each participant had a clearly defined role, it may have been relatively easy for the subjects to derive information about whom the stooges addressed from the internal context of communication. In the questionnaire, subjects mentioned that the role organization helped them avoid problems in turn-taking behaviour.

Since neither gaze direction nor motion video cues had any impact on task performance, it is conceivable this task situation would have resulted in similar performance in an audio-only condition. The question therefore arises as to whether our task was appropriate to gauge the differences between conditions. With this experiment, however, we did not attempt to prove the existence of differences as such. Instead, we assert that from the functional perspective of a relatively well-organized triadic office communication task, gaze directional information seems to be largely redundant, and the need for full-motion video can be questioned. There are tasks conceivable which may result in different conclusions. Typically, we believe, such task situations should involve conflict, more than three participants, and numerous use of deictic references to visual information. We leave it to the readers to decide to what extent our joint task definition correlates with communication tasks typical to their office environment.

5.2. Questionnaire

Subjects reported that it was clearer whom the stooges were addressing in the still-video condition (C) than in the full-motion conditions (A and B). Apparently, the still-video images seemed to supply more information about the stooges’ focus of attention than the full-motion images. This effect may be explained by the difference in the amount of information supplied in both conditions. The full-motion images contained more information than the still-video images. As a result, in the still-video images the information that was present (the gaze direction cues) became more obvious.

Sellen [15] gave a similar explanation for the observation that gaze direction did not seem to influence the turn-taking process much in her experiment. She attributed the lack of difference between conditions with and without gaze direction to the fact that video mediation of such cues decreases their effectiveness. When they appear on a screen, they lose their power to attract attention.

The differences in the responses to the question whether the subject found the collaboration pleasant were remarkably near significance. This might indicate that the subjects from the still-video condition tended to enjoy the collaboration more than those from the condition with full-motion video. Perhaps the still-video condition was more anonymous, making the subject feel more at ease. It might also be that the collaboration was more pleasant because of the fact that it was clearer to the subject whom the stooges were talking to.

The questionnaire also provided some information about how the subjects experienced the different conditions. The subjects from the still-video condition complained about the stammering movements of the images, and some of them said they missed their participants’ facial expressions. However, of these, some judged the images to contain all relevant information.

By instructing the stooges to look into the camera from time to time, we tried to give the subjects some sense of eye contact with the stooges. In the still-video condition, we used images in which the stooges always looked into the camera when they looked at the subject. In the full-motion conditions, they looked both at their screen and into the camera for some time when looking at the subject. Nevertheless, in all conditions some subjects complained about a lack of eye contact. This was probably due to asymmetry. When the subject tried to establish eye contact (by looking at his screen), the stooges did not always respond to this by looking in to the camera.

For all conditions the subjects generally regarded video mediated communication as a supplement, rather than a substitute to face-to-face communication. Apparently in VMC, even when conveying full-motion video and gaze direction, people still miss other cues that play a role in face-to-face communication, resulting in a lack of social presence [1,5]. For example, since people are not together in one room, they cannot influence each other physically, and know the others cannot really influence them in this way.
6. Conclusions

Designing a mediated system is a question of conveying the least redundant information first. If bandwidth is cheap, one can simply choose to convey as much information as possible. If, however, bandwidth is at a premium, the question remains what type of cues should be conveyed. We regarded this problem from a functional point of view: how redundant are cues with regard to successful completion of a joint task?

The experiment we carried out confirmed our hypotheses partly. It showed that the use of full-motion images does not necessarily result in better task performance than the use of still-video. Triads are able to perform our abstract version of an office task with still-video just as well as with full-motion video. Subjects found still-video to be clearer with regard to knowing who is talking to whom. Because still-video contains far less information than full-motion video, the gaze directional cues conveyed by this type of video are more easily perceived. Although this aspect may have been a factor, we have no ready explanations for the surprising near significance of the subjects’ preference to the still-video condition with regard to pleasantness of collaboration.

However, with regard to task performance of triads with clearly defined roles, gaze direction appeared more redundant than we expected. Although it may have provided extra regulative information, such information did not increase task performance. However, in larger, unstructured groups such extra regulative information might in fact increase task performance.

Although we did convey gaze direction in our experiments, we did not offer full eye contact between the communicating partners. The fact that our experimental subjects complained about a lack of eye contact might indicate that eye contact is more important than absolute gaze direction as such. However, our experiment is not conclusive in this respect.

Subjects rated VMC as an inferior alternative to face-to-face communication. Even when conveying gaze direction, lip movements, physical appearance, facial expressions and head movements, VMC does not seem to offer the same sense of presence as face-to-face communication. The extent to which conversational partners think they can affect each other’s behaviour seems to be inherently limited by the fact that their communication is mediated. This makes it unlikely that VMC will ever be a complete substitute for face-to-face communication, no matter how realistically it models it. We should therefore continue to research the possibilities of mediated communication as a supplement rather than a substitute for face-to-face communication.

References