Aalborg Universitet

AALBORG UNIVERSITY
DENMARK

EyeGaze

Enabling eye contact over video

Kjeldskov, Jesper; Smedegaard, Jacob Haubach; Nielsen, Thomas S.; Skov, Mikael B.; Paay, Jeni

Published in:
Proceedings of the 14th International Conference on Advanced Visual Interfaces, AVI 2014

DOI (link to publication from Publisher):
10.1145/2598153.2598165

Publication date:
2014

Document Version
Early version, also known as pre-print

Link to publication from Aalborg University

Citation for published version (APA):
Kjeldskov, J., Smedegaard, J. H., Nielsen, T. S., Skov, M. B., & Paay, J. (2014). EyeGaze: Enabling eye contact over video. In *Proceedings of the 14th International Conference on Advanced Visual Interfaces, AVI 2014* (pp. 105-112). Association for Computing Machinery. https://doi.org/10.1145/2598153.2598165

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal -

Take down policy
If you believe that this document breaches copyright please contact us at vbn@aub.aau.dk providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from vbn.aau.dk on: April 25, 2024

EyeGaze: Enabling Eye Contact over Video

Jesper Kjeldskov, Jacob H. Smedegård, Thomas S. Nielsen, Mikael B. Skov and Jeni Paay
Aalborg University, Department of Computer Science / Research Centre for Socio+Interactive Design
Selma Lagerlöfs Vej 300, DK-9220 Aalborg East
{jesper, jhaubach}@cs.aau.dk, primogens@gmail.com, {dubois, jeni}@cs.aau.dk

ABSTRACT
Traditional video communication systems offer a very limited experience of eye contact due to the offset between cameras and the screen. In response, we present EyeGaze, which uses multiple Kinect cameras to generate a 3D model of the user, and then renders a virtual camera angle giving the user an experience of eye contact. As a novel approach, we use concepts from KinectFusion, such as a volumetric voxel data representation and GPU accelerated ray tracing for viewpoint rendering. This achieves detail from a noisy source, and allows the real-time video output to be a composite of old and new data. We frame our work in literature on eye contact and previous approaches to supporting it over video. We then describe EyeGaze, and an empirical study comparing it with communication face-to-face or over traditional video. The study shows that while face-to-face is still superior, EyeGaze has added value over traditional video in terms of eye contact, involvement, turn-taking and co-presence.

Categories and Subject Descriptors
H.5.5. Group and Organization Interfaces: Computer-supported cooperative work, Synchronous interaction.

General Terms
Human Factors

Keywords
Eye contact; gaze; virtual camera; Kinect;

1. INTRODUCTION
Eye contact plays an important role in interpersonal face-to-face communication. It is used to provide information, regulate interaction, expressing intimacy, exercising social control, and facilitating service tasks [12]. Contemporary video communication systems, however, offer very limited experience of eye contact with a remote person. This is caused by a fundamental limitation in the way such systems can practically be set up, with cameras placed at the edge of our displays, and not behind or in front of it, which causes an offset between the remote person's actual and apparent viewing angle since one cannot be looking at the screen and straight into the camera at the same time. According to Stokes [28] if this offset is greater than 5 degrees, at conversational distance, then the experience of eye contact is lost. Instead, the remote person appears to be looking slightly upwards or downwards, depending on the position of the camera below or above the screen, when in fact they are looking straight at their communication partner on the screen. Put simply, in order to create an experience of eye contact over video, the camera must be placed where the remote person's eyes are displayed on the screen.

This requirement is obviously challenging because it would be very difficult, if not impossible, to place a physical camera in that position, as it would either occlude, or be occluded by, the screen. What is needed, we suggest, is a way of rendering a virtual camera view from the would-be physical position of the remote conversation partner, behind the screen on which they are displayed. This, we believe, can be achieved by merging the output from a number of cameras placed along the edges of the screen in real time. While others have also explored this idea, and reported promising results, several challenges remain, and no approach has, to our knowledge, been established that renders a live virtual camera view in a usable quality.

With the introduction of Microsoft's Kinect camera it is now much easier and cheaper to capture a physical environment into a 3D model in real time than just a few years ago. Having a physical environment represented as a real time 3D model, we can render views into this environment from any camera perspective desired. We can also move our point of view dynamically, and we can have multiple views from different viewing perspectives. Thereby we can overcome the limitations of physical camera placement, and make this merely a question of coverage, rather than compromise between factors such as placement, focal length and field-of-view, as described in Paay et al. 2011 [22].

In this paper we present such a novel gaze enabling video system called EyeGaze. The system takes live depth data from multiple Kinect cameras and renders a virtual camera view in real time. EyeGaze use advanced 3D reconstruction models and data representations. This allows us to use data from several cameras to continuously refine the 3D model over time. We describe the EyeGaze system and illustrate the value of our approach on the on the quality of the rendering output and systems performance. We then present an empirical study of the experience of using EyeGaze compared to communicating face-to-face and over a traditional video communication system.

1. RELATED WORK
Our research builds on related work on face gaze and eye contact, spatially arranged video conferencing and virtual viewpoints.

1.1 Face Gaze and Eye Contact
Face gaze and eye contact has been subject of research in social psychology since the mid 1960s. In a detailed review [12] Chris Kleinke demonstrates the significance of face gaze and eye contact in human relationships and communication.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
AVI' 14, May 27-29, 2014, Como, Italy
Copyright 2014 ACM 978-1-4503-2775-6/14/05…$15.00.
http://dx.doi.org/10.1145/2598153.2598165

Firstly, people's gazing behaviours have been shown to influence other people's judgements of liking/attraction, attentiveness, competence social skills, credibility, and dominance, and also as an expression of the intensity of their feelings, and of their intimacy. For example, people reportedly like each other more when sharing moderate amounts of gaze and eye contact, over constant or no gaze or eye contact, and use this as a cue in social interactions with each other [1, 10, 26]. People also perceive others as being more sincere and attentive in personal and social conversations, and intelligent or competent in public situations when their mutual gaze and eye contact is high. Conversely, low levels of gaze and eye contact have been found to communicate inattentiveness, non-involvement, lack of credibility, and even suspiciousness and depression.

Secondly, gazing behaviour has been shown to provide an important regulatory function in communication, such as turn taking and synchronization between verbal and kinesic behaviours. For example, Kendon [11] found that speakers tend to end utterances with a prolonged gaze as a turn-yielding cue, and Duncan and Niederehe [4] found that breaking eye contact with a speaker communicated intention to speak. Similarly, Levine and Sutton-Smith [13] found that eye contact aversion before speaking is useful for collecting one's thoughts, while eye contact after speaking communicates that one is listening to feedback.

As a third notable factor, gaze and eye contact has been shown to have an important social control function in relation to acts like persuasion, deception, ingratiation, threat, dominance, avoidance, and compliance. For example, it was found that people tend to increase gaze and eye contact when instructed to be persuasive [18], or when lying [17]. Others have found that in situations with aggression and anger, gaze functions to communicate threat and dominance during conversations and defence of personal space [e.g. 5]. Conversely, avoiding gaze in such situations has been shown to indicate submissiveness and conciliation [e.g. 15].

Finally, research has pointed to a "service-task" function of gazing behaviour in interpersonal communication, applying more to achieving specific goals and outcomes of an interaction rather than to its intimacy or other affective qualities [24]. Service-tasks fall in two categories: seeking information and facilitating communication. In seeking information, gaze and eye contact is used to focus one's attention towards a speaker or a person of interest, for example when seeking feedback on a previous utterance, or establishing contact with a stranger [25]. In facilitating communication, gaze and eye contact enhance people's comfort, with the literature generally reporting preference for face-to-face conversations over the use of video- or telephones. Gaze and eye contact has also been found to enhance teaching and learning by facilitating better participation and satisfaction [12]. Finally, it has been found to foster better cooperation, with people engaging in longer eye contact when cooperating than when competing [6], and being better at negotiating and compromising face-to-face than over the phone [12].

In summary, gaze and eye contact play several important roles in interpersonal communication. When communicating over video, however, these are largely lost.

1.2 Spatially Arranged Video Conferencing

One of the earliest examples of responding to the spatial drawbacks of video communication is the Hydra system from the early 1990s by Sellen, Buxton and Arnott [27]. Hydra supports multiparty meetings preserving the participants' personal space and spatial arrangement. This is done by placing Hydra units around a meeting table, where the other participants would otherwise have been seated. Each Hydra unit has a camera, monitor and speaker, acting in effect as "video surrogates" for the other participants. This conveys conversational acts such as face and eye gaze in a meaningful way to those in the meeting.

Building on the fundamental thinking behind Hydra, more recent research in video conferencing has involved the creation of so-called "blended interaction spaces" where the physical placement of displays, cameras and furniture is used to create the experience of two or more physical locations blending into one. The underlying philosophy of this is that blended spaces may create a more natural video-mediated communication situation, allowing for the use of spatial gestures, such as pointing, and, to some degree, facilitate the use of face and eye gaze. From this perspective, O'Hara et al. [20] analysed a commercial blended space video conferencing system, HP Halo, to understand its essential properties. Halo is designed primarily for business meetings of up to 12 participants distributed, ideally, over two locations. The setup consists of two identical rooms with a long curved table divided into three sections with two seats each. The table is placed in front of a wall with three large widescreen displays mounted side-by-side. On top of each display, a high-resolution video camera feeds into the corresponding display in the other room. When in operation, people in the other room are displayed in life size as if they were sitting physically at the other side of a conference table. While Halo does not support eye contact, due to the offset between cameras and screens, the setup does support a higher level of spatial gestures and even to some extent face gaze due to the identical setups, and the distance to the cameras. This alone reportedly creates an experience closer to being in the same room.

Subsequently, Paay et al. [22] used their experience with the analysis of HP Halo to develop a shared digital workspace facilitating hands-on collaboration with shared applications and digital information across two physical locations, called BIS*i*. In this work they observed that the most important factor in creating the experience of a blended space is the very precise setup of camera views. In addition, they found that the optimal camera view for their setup might be from one or more cameras placed several metres behind the displays. They also speculate that such view, while obviously not possible with a physical camera, might be obtained through the use of virtual camera technology. However, this is not explored in the BIS*i* prototype. Instead, the effect of blending was created through detailed positioning of the cameras, shaping and placing the furniture to match their focal length and field-of-view, and by introducing physical constraints, such as a table leg where not structurally needed, and a false wall behind the chairs, discouraging people from sitting in camera blind spots or where camera views overlap.

Also in the area of blended spaces, Nguyen et al. [19] developed MultiView, which is a video conferencing setup designed for multiple participants. MultiView distinguishes itself from Halo and BIS*i* by using a specially designed screen with a retro reflective surface to display different video feeds to different users. This allows MultiView to present people with views into a remote location from their own specific visual perspective, rather than from one or more offset perspectives as in Halo and BIS*i*. To achieve this, MultiView makes use of one camera and projector per participant. While a promising approach to separating video

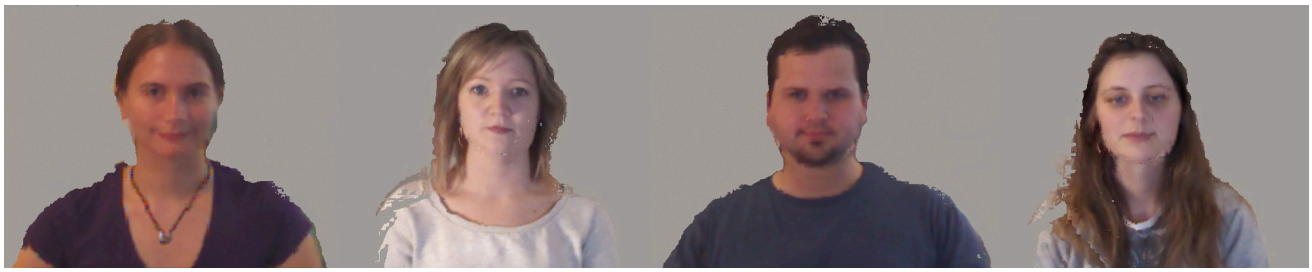


Figure 1: Virtual camera views rendered by EyeGaze for enabling eye contact over video.

output to multiple viewers, which is important for enabling multiple people in the same room to experience eye contact, however, the approach does not solve the offset issue of camera placement on the *input* side, described earlier.

1.3 Virtual Viewpoints

As an alternative to the use of raw camera feeds, research in computer vision and 3D graphics has explored the possibility of rendering live virtual viewpoints. This research includes different uses and combinations of RGB and depth sensing cameras.

An early example of manipulating data from RGB cameras is Ott et al. [21] who calculate a virtual camera view based on stereoscopic analysis. This is done by calculating pixel correspondence between the two views and then rotating one of them according to a disparity map. Later approaches introduce an intermediate step of creating a texturized model of the user, which can be used for output rendering. This typically involves a predefined heavily simplified head model and mapping the texture from one or more cameras to the model. Examples of this approach include Yoon and Lee [30] who use an ellipsoid head model to allow for a computationally fast algorithm. Improving quality and realism Yang and Zhang [29] use a personalised head model. Also using RGB cameras, Gemmell et al. [8] uses a hybrid model combining an avatar and a generic approximation of the user's face and eyes rendered on basis of the captured video feed.

As an example of the use of depth sensing cameras, Zhu et al. [31] combines three RGB cameras with a depth sensor. This makes them able to combine stereo matching with depth data resulting in a 3D point cloud of the user. Data from all RGB cameras are then used for texturing. A very recent contribution is that of Kuster et al. [13] who uses a single Kinect camera for gaze correction of a person. This is done by obtaining an RGB video feed and a 3D capture of the face. Using facial recognition they are then able to isolate the face and reapply a rotated facial texture on top of the video frame, making it seem as if the person is looking straight at the camera. Maimone and Fuchs [16] takes a different approach and present a setup where an entire physical environment is captured in 3D by depth sensors in real time, and virtual viewpoints rendered and displayed on 3D screens. This is done using a number of Kinect cameras for real time merging of the texturized depth input from the cameras, colour matching and eye tracking. Their system uses a frame-by-frame model-generating algorithm where all data obtained is overwritten by the next frame. Hole filling is done on the data from each Kinect without taking data from the other Kinects into account. Missing edges on rendered objects are not recreated or refined over time, but only rendered in frames where data is available.

2. EYEGAZE

Our system, EyeGaze, allows eye contact and face gaze between two people over video by creating visually realistic representations of the users from virtual camera views and rendering it in real time (Figure 1). Building on the work by Maimone and Fuchs [16], we have focussed our technical efforts on using a volumetric data representation. Specifically, we use a voxel data representation, which allows us to merge data from several Microsoft Kinect depth cameras, and improve the quality and completeness of the 3D model over time by only removing data when we are sure it is no longer up to date. In order to achieve high frame rates, we have constructed a GPU accelerated ray tracer that renders the model stored in the voxel grid and texturizes it using the RGB data captured by the Kinect. In the following sections we describe the details of the EyeGaze system and how we have approached some of the technical challenges encountered.

2.1 Volumetric Representation

Using off-the-shelf depth capturing cameras such as the Kinect involves some challenges with data quality. Other studies have shown that the accuracy of depth data obtained from a Kinect camera degrades quadratically with distance [16]. This means that depth accuracy at 2 metres is about 4 cm. Furthermore, the depth data obtained may be incomplete or containing erroneous measures (deviations from the actual surface). This is especially the case for surfaces orthogonal to the camera, or surfaces that are semi-transparent. In order to render a satisfactory image of a person in front of the camera, our system must therefore be able to handle such incomplete and erroneous data, and ensure that it does not seriously affect the quality of the model generated.

Another requirement is that we can move our viewpoint as freely as possible, such that we can support head tracking and exploration. This requirement has several implications for our data structure: *a)* new information should only replace old information if the two are mutually exclusive. *b)* we should faithfully represent the recorded 3D surfaces, such that the spatial arrangement of the scene is intact. *c)* we must be able to handle multiple Kinects, for optimal coverage of the scene.

We take inspiration from Izadi et al. [9] who has designed and implemented a high-resolution 3D scanner, KinectFusion, using a single handheld Kinect Sensor. KinectFusion utilizes a Truncated Signed Distance Function (TSDF), based on the Signed Distance Function presented by Curless et al. [3], as a data structure storing object surfaces as implicit surfaces through surface distances. Izadi et al. assumes that the object in question is static, which allows them to refine the surface to millimetre precision with each new frame of information. The SDF, and its truncated version (the TSDF), are incremental and order-independent when updated. The surface distance is a weighted average over several frames, resulting in an image that is less affected by noise since range uncertainties are averaged out over frames while actual depth discontinuities (such as the jump from background to foreground objects) are preserved. This eliminates the need for separate algorithms for large holes in the rendering where data is scarce or

extremely noisy, since these will be averaged out quickly, at the expense of a slightly imprecise rendering during fast movement.

A feature of the TSDF is that of selective updating of distances. Using this feature, we can preserve information in the 3D model, which has become obstructed, such as surface information for backgrounds or for the upper body of people while obscured by their arms. Finally, the use of the SDF carries no restriction on the type of model encoded. There is no assumption that the model is a person, an object or a room. Using an SDF, or TSDF, has thus several advantages over treating each camera and image frame as a separate textured model as done by, for example, Maimone & Fuchs [16]. Rendering each frame and camera independently, while faster, would mean that old information is not preserved. In order to capture the background, dedicated cameras must be used.

2.2 Merging camera inputs into the voxel grid

The graphics pipeline of EyeGaze has two parts: 1) a merging algorithm storing data from cameras in the voxel grid, and 2) an engine for ray tracing the scene from a remote user's perspective.

The merging algorithm relies on known static pose (position and orientation) of each Kinect sensor. This can easily be computed using various Computer Vision techniques. For calibration we used a checkerboard pattern with 117 corner points, captured by each Kinect sensor. These corner points were transformed into point clouds and transformation matrices between a known pose of a single Kinect and each unknown Kinect was computed through the use of an Iterative Closest Point algorithm. The algorithm captures depth frames from all Kinects, and then for each Kinect transform the voxels into camera space and project them into 2D positions on the Kinect camera's depth frame. It then calculates the distance between the voxel and the surface observed by the Kinect camera, and finally average the surface distances obtained from the depth frames with the weighted distance obtained from the voxel's truncated SDF value, and store it in the grid.

2.2.1 GPU Accelerated Voxel Representation

In order to achieve higher frame rates, all operations performed on the voxel grid are done on the GPU, with the voxel grid stored persistently on the GPU. Doing so means that only depth and RGB frames from each Kinect camera needs to be transferred, resulting in much less data than if we were transferring the entire voxel grid. The voxel grid is stored efficiently as a single dimensional array on the GPU, similar to KinectFusion.

2.2.2 Truncated Signed Distance Function (TSDF)

We capture and merge depth frames into the SDF to enrich the detail and update moving objects. Like Izadi et al. [9], we *truncate* the SDF such that only voxels within a truncation distance from the surface are encoded with a distance. Each voxel is transformed into camera coordinates, and then perspective-projected such that depth value can be extracted from the depth frame. It is then updated by calculating the weighted average between the existing distance value, and the new distance value estimated from the Kinect depth frame. Since we are not dealing with a static scene, objects and people can move and thus their old data should be deleted. We are interested in retaining as much information as possible, which means that all voxels between the surface and the camera are updated with their distance value and truncated to a max value. Voxels within the truncation distance behind the surface are also updated, however voxels farther away is left with their old values, thus preserving their data.

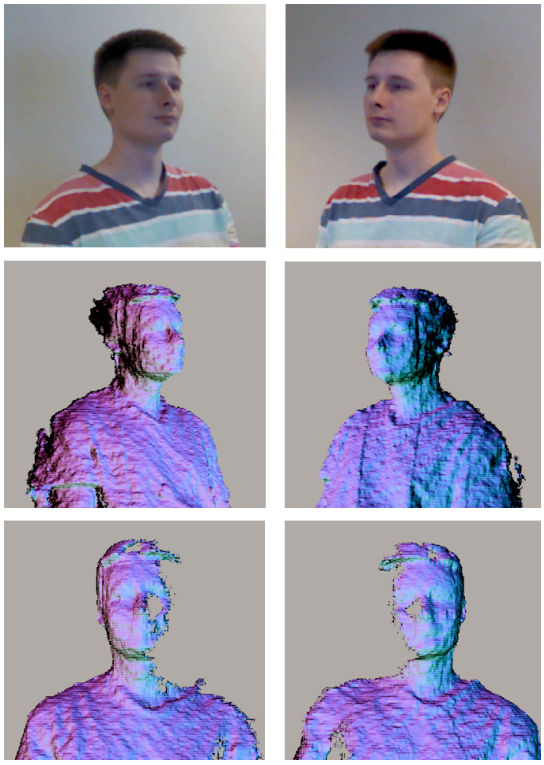


Figure 2: Output from the RGB camera (top) with corresponding 3D models generated from the depth camera, each rendered from two different perspectives. The perspective rendered in the bottom show that the models are both incomplete, but complementary and partly overlapping.

Figure 2 illustrates the data obtained from two Kinect cameras placed either side of a person. The top row shows the output from the RGB camera. The middle row shows the corresponding 3D models generated from the depth camera. In the bottom row we have rotated the 3D models to view the face straight on. This illustrates that they are both incomplete, but also that they are complementary and partly overlapping, with each missing data that is contained in the other. Figure 5b shows the merged 3D model created from the two depth cameras and stored in the SDF. The merged model is then ray traced and texturised with colours obtained from the Kinects' RGB cameras.

2.3 Ray Tracing

In order to obtain live video output we have constructed a GPU accelerated ray tracer that renders the 3D model stored in the voxel grid by interpolating the surface. We use trilinear interpolation to estimate the surface zero crossing and to generate the surface normal for texturing. We also adopt an interactive ray tracing method for isosurface rendering [23] as well as quick approximation of the ray length to the zero crossing [9] to reduce the number of interpolations needed. Because natural light is a part of the texture, we do not need virtual lighting as in a traditional ray tracer. This has great advantages for performance.

Once an intersection has been found, the challenge is to decide which Kinect RGB camera should provide the texturing colour for this point. Each point in the scene may be captured by more than one Kinect, and thus we need to identify what Kinect has the best view of it. We base this decision on the angle between the surface normal of the intersection point and the vectors of the Kinect

cameras, as illustrated in Figure 3. Knowing the surface normal (***n***) of the point *p*, and the vector of each Kinect (***a*** and ***b***) we can identify what Kinect has the smallest angle of disparity, and hence the best view of *p*.

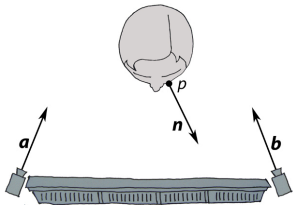


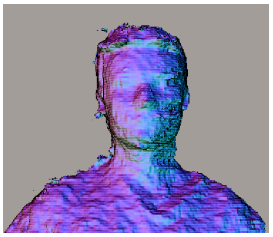
Figure 3: Using vectors to decide what camera provides the best RGB data for texturing.

In this example, the angle between ***n*** and ***b*** is smaller than the angle between ***n*** and ***a***, telling us that the camera on the right side of the figure has the best data for texturing. The position of *p* is then transformed into the camera space for that Kinect, and perspective-projected, so that we acquire the pixel coordinates (*x, y*) to lookup data in the RGB image captured by it.

3. QUALITY AND PERFORMANCE

In the following we discuss the quality and performance of EyeGaze. For the purpose of comparison, the images in Figure 4 show the output from a traditional webcam placed on top of the screen (a), the merged 3D model stored in the SDF (b), and rendered video from EyeGaze (c and d).

(a) Camera view from a webcam on top of the screen



(b) The merged 3D model stored in the SDF

(c) EyeGaze rendering - with texture from one Kinect

(d) EyeGaze rendering - with texture from two Kinects

Figure 4: Video output from webcam on top of the screen (a), a simple virtual camera approach (b), and EyeGaze (c and d).

As Figure 4c and 4d show, EyeGaze is capable of producing a high quality virtual camera view of the user's head and torso in good detail and enabling the experience of eye contact and face gaze. The difference between using one or two cameras for texturing is that two cameras provide more detail in "shaded" areas such as under the chin and in the eye sockets (Figure 4d), while using only one camera results in those areas being blurred (Figure 4c). As can be seen in Figure 4d, our current texturing from two cameras does, however, create a bit of colouring-noise, requiring some correctional filtering, as in [16].

EyeGaze is able to capture, model and render with an average time lag of less than 40ms. This means that the system is experienced as working in real time. The rendered video also has minimal flickering along borders of objects, unlike [16], which contributes to a smoother video stream. Using multiple Kinect cameras, EyeGaze can reliably capture enough perspective of the user's face and torso to render a realistic face-on view with very little missing data in each video frame. Due to volumetric data representation, EyeGaze "remembers" objects that are temporarily occluded, which minimizes frames with missing data.

3.1 Model quality

Using a volumetric approach for storing information about the users and the background allows us to achieve a high quality image from a noisy input source. This is illustrated in Figure 4c/d where we see that the model has very little missing data due to the volumetric approach allows us to compensate for the lack of data in one frame by reusing information from a previous one.

3.2 Frame rate

For our current prototype implementation the merging and ray tracing implementations are alternating in execution, limiting the output video to at or below the 30 FPS of the Kinect cameras which are polled for frames in serial. Running EyeGaze on a computer with a single Nvidia GeForce GTX 770 graphics card at 1080p and a voxel resolution of 512^3 voxels currently allows the video to be rendered at 25 FPS, with some frames from the Kinect sensor being skipped due to timing constraints. While the frame rate does not match the performance in [16] they are still promising given the computation required for merging data from two depth cameras into a joint 3D model and then rendering a video of it at those resolutions. The cause of limited frame rate is thus to be found in the specific hardware and pipeline used for rendering, rather than in our underlying approach, we are certain about being able to achieve higher performance with a more asynchronous rendering pipeline. Another step that could be taken is to more sparsely update the voxel grid, such that truncated voxels are not re-truncated on each frame.

4. EMPIRICAL STUDY

We have compared the experience of EyeGaze to face-to-face communication and the use of Skype through a within-subject laboratory experiment with 30 participants (8 females) with randomized order of the three conditions. The participants were grouped in pairs of two and given a selection of discussion topics to engage with each other in. They then spent fifteen minutes talking to each other, equally distributed on the three conditions. For the face-to-face condition the participants were seated across from each other at a meeting table. For the Skype and EyeGaze conditions they were seated in front of a display (Figure 5).

Afterwards the participants each filled out a questionnaire. This was based on the one used by Garau et al. [7] to measure the impact of eye gaze on communication. It contained 20 statements, like "I had a real sense of personal contact with my conversation-partner". For each statement participants were asked to respond on a 9-point Likert scale from "strongly agree" to "strongly disagree" for each condition. They were also asked to express for each statement which of the two video-mediated systems provided the best experience, on a 9-point scale from "Skype" to "EyeGaze". In our analysis we treated the Likert items as interval scales, allowing us to use descriptive and inferential statistics [2].

Figure 5: EyeGaze setup for the empirical study

5. FINDINGS

In the following we present the findings from the empirical study. Overall, the findings show that the face-to-face condition always provided the best communication experience. This is followed by a general tendency towards EyeGaze providing a better experience than Skype, albeit in some areas only slightly. In reporting the findings we begin with the relative ratings of Skype and EyeGaze against each other. We then compare the mean ratings for all 20 questions across conditions. Lastly we go into details with four measures; eye contact, involvement, turn-taking, and co-presence, for which we found significant and otherwise notable differences.

5.1 Relative experience of Skype and EyeGaze

When asked to rate which of the two video-mediated conditions provided the best experience, the users expressed a preference for EyeGaze in 18 out of 20 questions. The two exemptions, where Skype rated higher than EyeGaze were Question 7, which asks if the conversation felt natural, and Question 12, which asks if the participants were easily distracted from the conversation. For these two questions Skype was and EyeGaze were rated almost equally. Performing a one-way ANOVA test on the averaged ratings for Skype and EyeGaze shows that this overall preference for EyeGaze is significant, $F(1, 19) = 33.17$, $p<0.01$. One-way ANOVA tests on the ratings for each individual question furthermore show a significant preference for EyeGaze over Skype on four specific questions. Firstly, participants rated EyeGaze significantly higher than Skype when asked if they had a good sense of eye contact (question 1), $F(1, 29) = 6.64$, $p<0.05$. This is an important finding as it shows that the participants experienced a notable difference in terms of eye contact in the two video-mediated systems. Secondly, participants rated EyeGaze significantly higher when asked if they felt absorbed in the conversation (question 11), $F(1, 29) = 10.76$, $p<0.01$. Thirdly, EyeGaze was rated significantly higher when participants were asked about their experienced awareness of their conversation partner (question 14), $F(1, 29) = 6.8$, $p<0.01$, and finally when asked if their conversation partner seemed attentive (question 19), $F(1, 29) = 4.28$, $p<0.05$. For the remaining 14 questions with a preference towards EyeGaze this was not statistically significant.

5.2 Face-to-face, Skype and EyeGaze

Figure 6 gives an overview of the responses to the questionnaire across the three conditions of Face-to-face, Skype and EyeGaze. It shows the mean values for the participants' ratings of their experience with the three conditions in response to each of the 20 questions. Overall, looking at this figure shows that face-to-face was always rated higher than any of the two video-mediated conditions. This is to be expected as none of the two video-mediated conditions are close to replicating the actual experience of being together in the same room. It is also evident that there is a general tendency for EyeGaze to rate slightly higher than Skype, and that Skype never rated higher than EyeGaze. This is also to be expected, as EyeGaze actually facilitates the experience of eye contact, while Skype does not. Especially for question 1, which specifically asks about the experience of eye contact, EyeGaze is rated a lot higher than Skype. Looking at the values for questions 15, 16 and 17 it is notable that these are very close to each other across conditions. These questions were all in the theme of "partner evaluation", showing that the perception of one's communication partner was not affected greatly by the ability to have eye contact. This is an unexpected finding as the literature tells us that eye contact typically leads to more positive perceptions of communication partners.

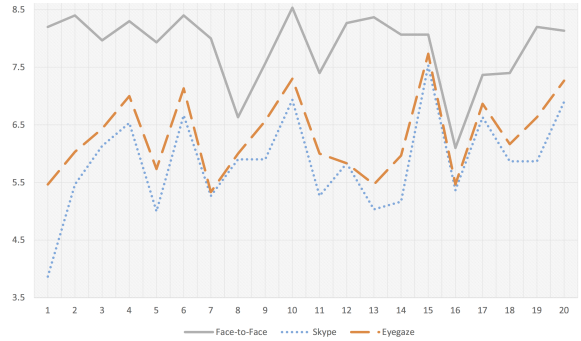


Figure 6: Mean ratings for all 20 questions across conditions

5.3 Eye Contact

Figure 7 show mean values for responses to question 1 where the participants were asked if they "had a good sense of eye contact with [their] conversation partner". In response to this face-to-face was rated highest and EyeGaze was rated higher than Skype.

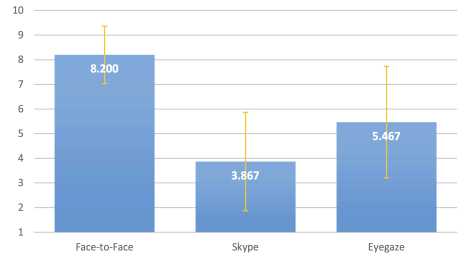


Figure 7: Sense of eye contact

Running a one-way ANOVA test on these responses showed that these differences between the three conditions are significant, $F(2, 58) = 42.84$, $p<0.01$. Furthermore, a Tukey HSD post hoc test shows that there was a significant difference between Skype and EyeGaze ($p<0.01$). This shows that participants experienced a better sense of eye contact when using EyeGaze than when using Skype, and confirms that the essential capability of our prototype system to facilitate eye contact was actually experienced by the participants. We also found significant differences between face-to-face and Skype ($p<0.01$), and face-to-face and EyeGaze ($p<0.01$). This shows that neither of the video-mediated conditions provided as good an experience of eye contact as face-to-face.

5.4 Involvement

We measured the participants' experience of involvement in the conversations through two questions derived from Garau et al. [7] inquiring into the participants' ability to keep track of and be absorbed in the conversation. Figure 8 show the mean values for

responses to question 11 where the participants were asked if they "felt completely absorbed in the conversation". This shows, again, highest ratings for face-to-face, followed by EyeGaze, followed by Skype, and running a one-way ANOVA test show that these differences are significant $F(2,58) = 25.93$, $p<0.01$.

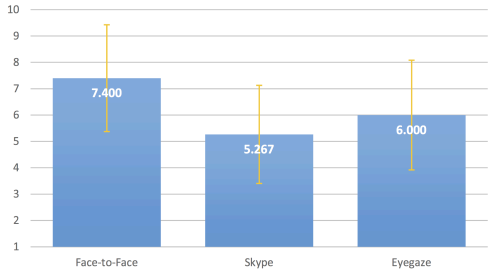


Figure 8: Feeling absorbed in the conversation

Furthermore, a Tukey HSD post hoc test shows that the difference between Skype and EyeGaze is also significant ($p<0.05$). This points to a stronger feeling of involvement in a conversation, in terms of feeling absorbed, when able to establish eye contact (EyeGaze) than when not (Skype). However, Tukey HSD post hoc tests also showed that the differences were significant between face-to-face and Skype ($p<0.01$) and face-to-face and EyeGaze ($p<0.01$), which shows that face-to-face is still superior to the two video-mediated conditions. The second question in relation to involvement, asking if the participants "found it easy to keep track of the conversation" showed no significant difference between EyeGaze and Skype, but only significant differences between face-to-face and the other two condition ($p<0.01$).

5.5 Turn-taking

We measured the effect on turn taking by asking if the participants felt they often interrupted their conversation partner (question 8). The mean value responses to this question are shown in Figure 9.

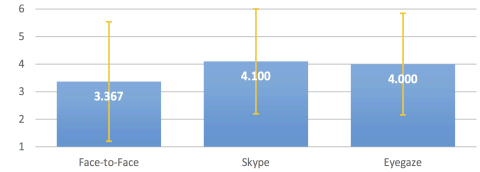


Figure 9: Feeling interruptive in the conversation

While at first glance Skype and EyeGaze appear to be performing equally well, a one-way ANOVA test shows a significant difference between the three conditions, $F(2, 58) = 3.76$, $p<0.05$. Furthermore, Tukey HSD post hoc tests show that the difference between face-to-face and EyeGaze are not significant, while the difference between face-to-face and Skype is ($p< 0.05$). This means that EyeGaze performed more similar to face-to-face than Skype did in terms of facilitating turn-taking.

5.6 Co-presence

We measured the participants' experience of co-presence through two questions derived from Garau et al. [7] asking the participants were asked if they "had a real sense of personal contact with [their] conversation partner" and "was very aware of [their] conversation partner" (question 13 and 14). Figure 10 show the mean values for responses to the Likert scale measurement summing up the results from this set of Likert items. This continues the trend towards higher ratings of the face-to-face condition, followed by EyeGaze, and with Skype being rated lowest, with the difference being significant when in a one-way ANOVA test $F(2, 58) = 48.25$, $p<0.01$.

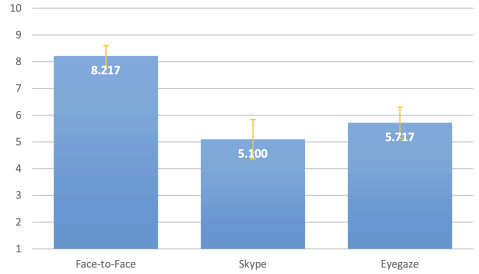


Figure 10: Feeling co-present

However, although the mean value rating for EyeGaze was higher than for Skype, and the standard deviation lower, the difference between the two was significant in a Tukey HSD post hoc test. There was, however, significant difference between face-to-face and Skype ($p<0.01$) and face-to-face and EyeGaze ($p<0.01$). This confirms that the feeling of co-presence is very difficult to achieve over a video connection but also indicate some added value to this by facilitating the experience of eye contact.

6. CONCLUSIONS AND FURTHER WORK

We have presented EyeGaze, a novel gaze enabling video system that takes live depth data from two Microsoft Kinect cameras and renders a virtual camera view in real time. Our goal was to construct a two-way video link enabling the experience of eye contact and face gaze by means of virtual camera view renderings. We have shown that using two Kinects, volumetric voxel data representations, concepts from KinectFusion [9], and exploring the power of GPUs for ray tracing, EyeGaze can create a visually convincing representation of a remote person from a virtual camera perspective enabling the experience of eye contact. EyeGaze renders this representation of the remote person in real time with low lag and promising frame rates.

We have conducted an empirical study comparing the user experience of EyeGaze to face-to-face communication and the use of Skype. As expected neither of the video-based systems performed as well as face-to-face in terms of the participants' subjective experience of the communication situation. This confirms the limitations of such technology reported in the literature, and fuels motivation for exploring ways of improving video-mediated communication further. Looking at the relative user experience ratings of the two video-mediated conditions, with and without eye contact, we found that EyeGaze was rated statistically significantly higher overall and on the four measures of sense of *eye contact, feeling absorbed in the conversation, awareness,* and *experienced attentiveness of the communication partner.* Looking at the experience ratings across face-to-face and video-mediated conditions we found that EyeGaze was rated statistically significantly higher on the three measures of *eye contact, involvement, and co-presence.* These are all factors of great importance in interpersonal communication and are known to suffer from mediating technology. Although one might perhaps have expected to see larger differences, finding any advances to these show that it is possible to improve the experience of video-mediated interpersonal communication by enabling face gaze and eye contact like EyeGaze does.

Our work with EyeGaze opens several avenues for further work. Firstly, we are exploring the effect of rendering the virtual camera view from a dynamic set of coordinates matching the exact

location of the viewer's eyes. We currently do this using the Kinects' built-in skeletal tracking, with promising results. Related to this we are also exploring the rendering of a stereoscopic image, and displaying this to the remote user on a 3D screen. Another potential is to render multiple perspectives for multiple remote viewers. This could be used to facilitate eye contact in meetings with several participants in a spatial arrangement similar to that of Hydra. If combined with multi-perspective screens [19] this could also be used to facilitate eye contact in meetings between *groups* of people, in a spatial arrangement similar to HP Halo and BIS*i*.

Finally, technologies that enable eye contact over video should be studied "in the wild". Here we would like to see further comparisons with face-to-face communication and traditional video conferencing in real world contexts and over time.

ACKNOWLEDGEMENTS

This research is supported by the Obel Family Foundation and Aalborg University's Faculty of Engineering and Science. We thank Erik Frøkjær for comments on previous drafts. The empirical study was done with Anne K. Jensen.

REFERENCES

1. Argyle, M., Lefebvre, L.M. and Cook, M. The meaning of five patterns of gaze. *European Journal of Social Psychology 4,* 2 (1974), 125-136.
2. Carfio, J. and Perla, R. Ten Common misunderstandings, misconceptions, Persistent Myths and Urban Legends about Likert Scales and Likert Response Formats and their Antidotes. *Journal of Social Sciences 3,* 3 (2007), 106-116.
3. Curless, B. and Levoy, M. A volumetric method for building complex models from range images. In *Proc. SIGGRAPH 96,* ACM (1996), 303-312.
4. Duncan, S.D. and Niederehe, G. On signalling that it's your turn to speak. *Journal of Experimental Social Psychology 10,* 3 (1974), 234-247.
5. Exline, R.V., Ellyson, S.L. and Long, B. Visual behaviour as an aspect of power role relationships. In *Nonverbal communication of aggression.* Plenum Press (1975), 21-52.
6. Foddy, M. Patterns of gaze in cooperative and competitive negotiation. *Human Relations 31,* 11 (1978), 925-938
7. Garau, M., Slater, M., Bee, S., and Sasse, M.A. The impact of eye gaze on communication using humanoid avatars. In *Proc. CHI 2001,* ACM (2001), 309–316.
8. Gemmell, J., Toyama, K., Zitnick, C. L., Kang, T. and Seitz, S. Gaze Awareness for Video-Conferencing. *IEEE MultiMedia 7,* 4 (2000), 26-35.
9. Izadi, S. Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., Shotton, J., Hodges, S. Freeman, D., Davison, A. and Fitzgibbon, A. KinectFusion: real-time 3D reconstruction and interaction using a moving depth camera. In *Proc. UIST 2011,* ACM (2011), 559-568.
10. Kendon, A. and Ferber, A. A description of some human greetings. In R.P. Michael and J.H. Crook, (eds.) *Comparative Behaviour and Ecology of Primates.* Academic Press (1973), 591-668.
11. Kendon, A. Some functions of gaze-direction in social interaction. *Acta Psychologica 26,* 1 (1967), 22-63.
12. Kleinke, C.L. Gaze and Eye Contact: A Research Review. *Psychological Bulletin 100,* 1 (1986), 78-100.
13. Kuster, C., Popa, T., Bazin, J.C., Gotsman, C., Gross, M. Gaze Correction for Home Video Conferencing. In *Proc. ACM SIGGRAPH Asia (2012)*
14. Levine, M.H., and Sutton-Smith, B. Effects of age, sex, and task on visual behaviour during dyadic interaction. *Developmental Psychology 9,* 3 (1973). 400-405.
15. Lochman, J.E. and Allen, G. Nonverbal communication of couples in conflict. *Journal of Research in Personality 15,* 2 (1981), 253-269.
16. Maimone, A. and Fuchs, H. Encumbrance-free telepresence system with real-time 3D capture and display using commodity depth cameras. In *Proc. ISMAR 2011,* ACM (2011), 237-146.
17. Mehrabian, A. Nonverbal betrayal of feeling. *Experi-mental Research in Personality 5,* 1 (1971), 64-73
18. Mehrabian, A., Williams, M. Nonverbal concomitants of perceived and intended persuasiveness. *Journal of Personality and Social Psychology 13,* 1 (1969), 37-58.
19. Nguyen, D. and Canny, J. MultiView: spatially faithful group video conferencing. *Proc. CHI 2005,* ACM (2005), 799-808.
20. O'Hara, K., Kjeldskov, J. and Paay, J. Blended interaction spaces for distributed team collaboration. *Transactions on Computer-Human Interaction 18,* 1 (2011), Article No. 3.
21. Ott, M., Lewis, J. P. and Cox, I. Teleconferencing eye contract using a virtual camera. In *Proc. INTERACT'93 and CHI'93,* ACM (1993), 109-110.
22. Paay, J., Kjeldskov, J. and O'Hara, K. BISi: a blended interaction space. *Ext. Abstracts CHI 2011,* ACM (2011), 185-200.
23. Parker, S., Shirley, P., Livnat, Y., Hansen, C. and Sloan, P. Interactive ray tracing for isosurface rendering. In *Proc. VIS '98,* ACM (1998), 233-238.
24. Patterson, M.L. A sequential functional model of nonverbal exchange. *Psychological Review 89*, 3 (1982), 231-249.
25. Rutter, D.R. and Stephenson, G.M. The functions of looking: Effects of friendship on gaze. *British Journal of Social and Clinical Psychology 18* (1979), 203-205.
26. Scherer, S.E. Influence of Eye Contact on Impression Formation. *Perceptual and Motor Skills 38,* 2 (1974)
27. Sellen, A.J., Buxton, W., and Arnott, J. 1992. Using spatial cues to improve videoconferencing. *In Proc. CHI 1992,* ACM (1992), 651-652.
28. Stokes, R. Human Factors and Appearance Design Considerations of the Mod II PICTUREPHONE & Station Set. *IEEE Transactions on Communication Technology 17,* 2 (1969), 318-323.
29. Yang, R. and Zhang, Z. Eye gaze correction with stereovision for video-teleconferencing. *IEEE Transactions on Pattern Analysis and Machine Intelligence 26,* 7 (2004), 956-960.
30. Yoon, N. and Lee, B. Viewpoint Interpolation Using an Ellipsoid Head Model for Video Teleconferencing. In *Proc. Advances in Visual Computing 2005,* Springer LNCS (2005), 287-293.
31. Zhu, J., Yang, R. and Xiang, X. Eye contact in video conference via fusion of time-of-flight depth sensor and stereo. *3D Research 2,* 3 (2011), 1-10