Gaze Correction for Home Video Conferencing

Claudia Kuster[1] Tiberiu Popa[1] Jean-Charles Bazin[1] Craig Gotsman[1,2] Markus Gross[1]
[1]ETH Zurich [2]Technion - Israel Institute of Technology

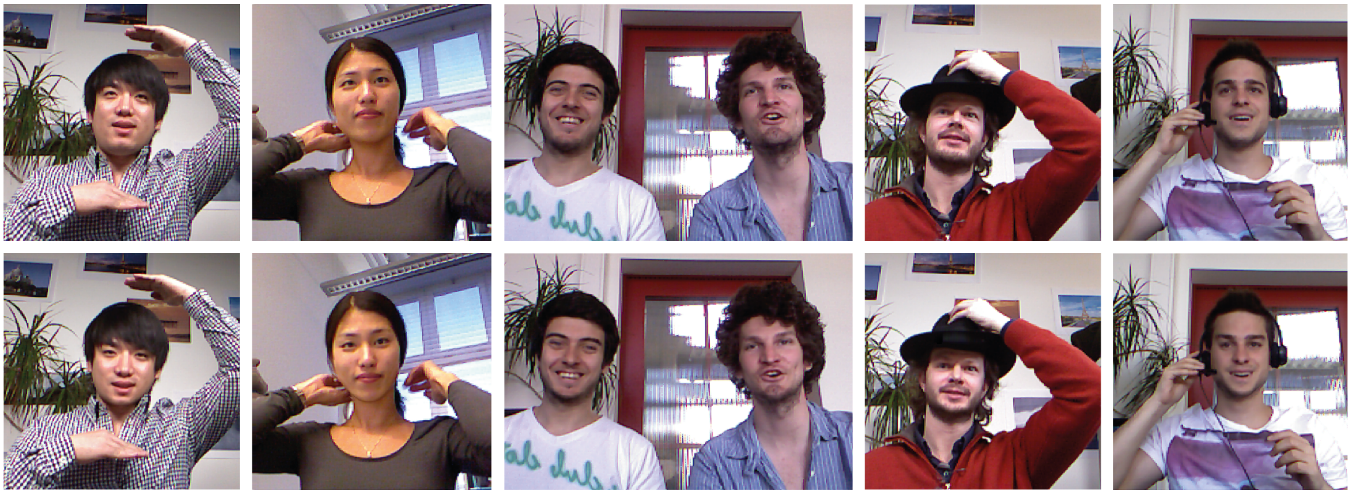

Figure 1: *Top: frames recorded during video-conferencing. Note the lack of eye contact because of the disparity between the locations of the participant and the camera. Bottom: real-time gaze correction with the proposed algorithm.*

Abstract

Effective communication using current video conferencing systems is severely hindered by the lack of eye contact caused by the disparity between the locations of the subject and the camera. While this problem has been partially solved for high-end expensive video conferencing systems, it has not been convincingly solved for consumer-level setups. We present a gaze correction approach based on a single Kinect sensor that preserves both the integrity and expressiveness of the face as well as the fidelity of the scene as a whole, producing nearly artifact-free imagery. Our method is suitable for mainstream home video conferencing: it uses inexpensive consumer hardware, achieves real-time performance and requires just a simple and short setup. Our approach is based on the observation that for our application it is sufficient to synthesize only the corrected face. Thus we render a gaze-corrected 3D model of the scene and, with the aid of a face tracker, transfer the gaze-corrected facial portion in a seamless manner onto the original image.

CR Categories: I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism—Radiosity;

Keywords: video conferencing, depth camera, gaze correction

Links: DL PDF

1 Introduction

It has been firmly established [Argyle and Cook 1976; Chen 2002; Macrae et al. 2002] that mutual gaze awareness (i.e. eye contact) is a critical aspect of human communication, both in person or over an electronic link such as a video conferencing system [Grayson and Monk 2003; Mukawa et al. 2005; Monk and Gale 2002]. Thus, in order to realistically imitate real-world communication patterns in virtual communication, it is critical that the eye contact is preserved. Unfortunately, conventional hardware setups for consumer video conferencing inherently prevent this. During a session we tend to look at the face of the person talking, rendered in a window within the display, and not at the camera, typically located at the top or bottom of the screen. Therefore it is not possible to make eye contact. People who use consumer video conferencing systems, such as Skype, experience this problem frequently. They constantly have the illusion that their conversation partner is looking somewhere above or below them. The lack of eye contact makes communication awkward and unnatural. This problem has been around since the dawn of video conferencing [Stokes 1969] and has not yet been convincingly addressed for consumer-level systems.

While full gaze awareness is a complex psychological phenomenon [Chen 2002; Argyle and Cook 1976], mutual gaze or eye contact has a simple geometric description: the subjects making eye contact must be in the center of their mutual line of sight [Monk and Gale 2002]. Using this simplified model, the gaze problem can be cast as a novel view synthesis problem: render the scene from a virtual camera placed along the line of sight [Chen 2002]. One way to do this is through the use of custom-made hardware setups that change the position of the camera using a system of mirrors [Okada et al. 1994; Ishii and Kobayashi 1992]. These setups are usually too expensive for a consumer-level system.

The alternative is to use software algorithms to synthesize an image from a novel viewpoint different from that of the real camera. Systems that can convincingly do novel view synthesis typically consist

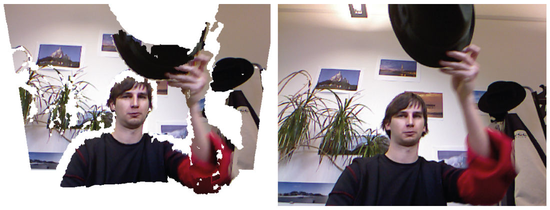

Figure 2: *Comparison between transforming the entire scene (Left) and our approach (Right). The integrity of the scene is well preserved in our approach.*

of multiple camera setups [Matusik et al. 2000; Matusik and Pfister 2004; Zitnick et al. 2004; Petit et al. 2010; Kuster et al. 2011] and proceed in two stages. In the first stage they reconstruct the geometry of the scene and in the second stage, render the geometry from the novel viewpoint. These methods require a number of cameras too large to be practical or affordable for a typical consumer. They have a convoluted setup and are difficult to run in real-time.

With the emergence of consumer-level depth and color cameras such as the Kinect [Microsoft 2010] it is possible to acquire in real-time both color and geometry. This can greatly facilitate solutions to the novel view synthesis problem, as demonstrated by Kuster et al. [2011]. Since already over 15 million Kinect devices have been sold, technology experts predict that soon the depth/color hybrid cameras will be as ubiquitous as webcams and in a few years will even be available on mobile devices. Given the recent overwhelming popularity of such hybrid sensors, we propose a setup consisting of only one such device.

At first glance the solution seems obvious: if the geometry and the appearance of the objects in the scene is known, then all that needs to be done is to render this 3D scene from the correct novel viewpoint. However, some fundamental challenges and limitations should be noted:

- The available geometry is limited to a depth map from a single viewpoint. As such, it is very sensitive to occlusions, and synthesizing the scene from an arbitrary (novel) viewpoint may result in many holes due to the lack of both color and depth information, as illustrated in Fig. 2 (left). It might be possible to fill these holes in a plausible way using texture synthesis methods, but they will not correspond to the true background.
- The depth map tends to be particularly inaccurate along silhouettes and will lead to many flickering artifacts.
- Humans are very sensitive to faces, so small errors in the geometry could lead to distortions that may be small in a geometric sense but very large in a perceptual sense.

In this paper we propose a gaze correction system targeted at a peer-to-peer video conferencing model that runs in real-time on average consumer hardware and requires only one hybrid depth/color sensor such as the Kinect. Our goal is to perform gaze correction without damaging the integrity of the image (i.e., loss of information or visual artifacts) while completely preserving the facial expression of the person. The main component of our system is a face replacement algorithm that synthesizes a novel view of the subject's face in which the gaze is correct and seamlessly transfers it into the original color image. This results in an image with no missing pixels or significant visual artifacts in which the subject makes eye contact. In our synthesized image there is no loss of information, the facial expression is preserved as in the original image and the background is also maintained. Figure 1 shows some examples.

In general, transferring the image of the face from the corrected image to the original may lead to an inconsistency between the vertical parallax of the face and the rest of the body. For large rotations this might lead to perspective aberrations if, for example the face is looking straight and the head is rotated up. A key observation is that in general conferencing applications the transformation required for correcting the gaze is small and it is sufficient to just transform the face, as opposed to the entire body. Figure 3 illustrates this observation. The left column shows two different subjects where the gaze is away from the camera. In the middle column their gaze is corrected by just rotating the geometry. The right column shows our results. Please note that the appearance of the person is similar to just transforming the entire geometry with the advantage that we can preserve the integrity of the scene.

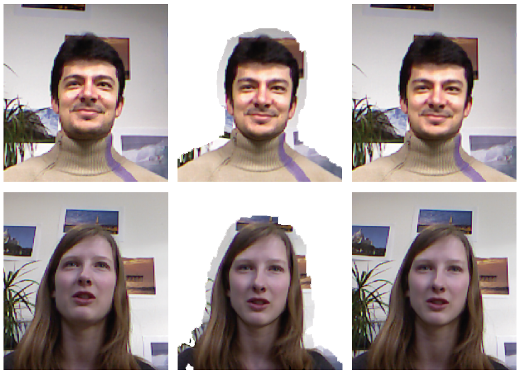

Figure 3: *Left: original image from the color camera (i.e. without correct gaze). Middle: eye contact achieved by rotating the entire scene. Right: our method. Transferring the rotated face onto the original image does not lead to perspective aberrations.*

2 Related Work

Gaze correction is a very important issue for teleconferencing and many experimental and commercial systems support it [Jones et al. 2009; Nguyen and Canny 2005; Gross et al. 2003; Okada et al. 1994]. However, these systems often use expensive custom-made hardware devices that are not suitable for mainstream home use. Conceptually, the gaze correction problem is closely related to the real-time novel-view synthesis problem [Matusik et al. 2000; Matusik and Pfister 2004; Zitnick et al. 2004; Petit et al. 2010; Kuster et al. 2011]. Indeed if a scene could be rendered from an arbitrary viewpoint then a virtual camera could be placed along the line of sight of the subject and this would achieve eye contact. Novel view synthesis using simple video cameras has been studied for the last 15 years, but unless a large number of video cameras are used, it is difficult to obtain high-quality results. Such setups are not suitable for our application model that targets real-time processing and inexpensive hardware.

There are several techniques designed specially for gaze correction that are more suitable for an inexpensive setup. Some systems only require two cameras [Criminisi et al. 2003; Yang and Zhang 2002] to synthesize a gaze-corrected image of the face. They accomplish this by performing a smart blending of the two images. This setup constrains the position of the virtual camera to the path between the two real cameras. More importantly, the setup requires careful calibration and is sensitive to light conditions which makes it impractical for mainstream use.

Several methods use only one color camera to perform gaze correction. Some of these [Cham et al. 2002] work purely in image space, trying to find an optimal warp of the image, and are able to obtain reasonable results only for very small corrections. This is because without some prior knowledge about the shape of the face it is difficult to synthesize a convincing image. Thus other methods use a proxy geometry to synthesize the gaze-corrected image. Yip et al. [2003] uses an elliptical model for the head and Gem-

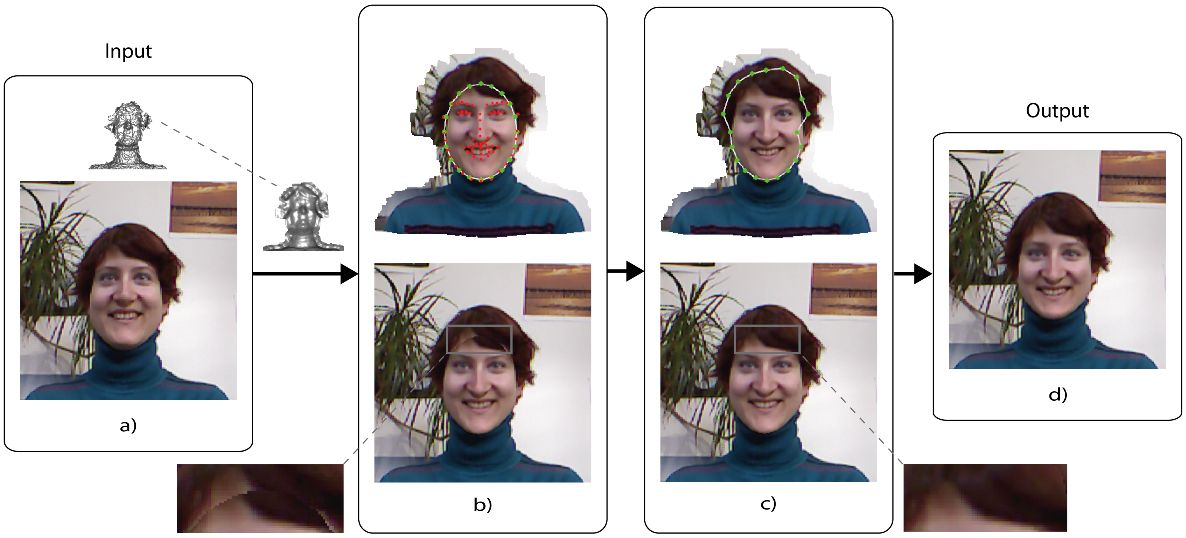

Figure 4: *System overview: a) Input: color and depth images from the Kinect b) Synthesize an image of the subject with the gaze corrected (by performing an appropriate 3D transformation of the head geometry). Top: The subject overlayed with the face tracker (red points) and an ellipse fitted to the chin points of the face tracker (green). Bottom: Use the ellipse as a stencil to copy the gaze-corrected rendering and paste it into the original image. Seam artifacts are visible. c) Optimize the seam. Top: The subject overlayed with the new seam (green). Much fewer visible artifacts. d) Blend the images along the seam edges to obtain the final result.*

mell [2000] uses an ad-hoc model based on the face features. However, templates are static and faces are dynamic. So a single static template will typically fail to do a good job when confronted with a large variety of different facial expressions.

Since the main focus of many of these methods is reconstructing the underlying geometry of the head or face, the emergence of consumer-level depth/color sensors such as the Kinect, giving easy access to real-time geometry and color information, is an important technological breakthrough that can be harnessed to solve the problem. Zhu et al. [2011] proposed a setup containing one depth camera and three color cameras and combined the depth map with a stereo reconstruction from the color cameras. However this setup only reconstructs the foreground image and still is not inexpensive.

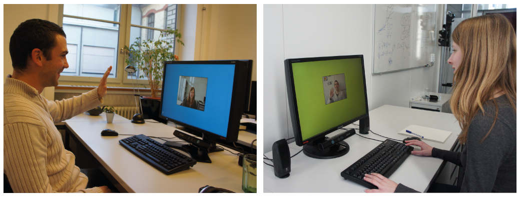

Figure 5: *Our setup is composed of a single Kinect device. The angle between the Kinect and the screen window is typically between 19 and 25 degrees.*

3 System Overview

The physical layout of our system is simple. The only device required is a single hybrid depth/color sensor such as the Kinect (Figure 5). Although webcams are usually mounted on the top of the screen, the current hybrid sensor devices are typically quite bulky and it is more natural to place them at the bottom of the screen. Our gaze correction system first synthesizes a novel view where the subject makes eye contact using the geometry from the depth camera (Figure 4b). The resulting image has holes and artifacts around the silhouettes due to occlusions and depth errors. To construct a complete image that preserves both the integrity of the background and foreground, the facial expression of the subject as well as the eye contact, we transfer only the face from the synthesized view seamlessly into the original image. This allows us to completely preserve both the spatial and temporal integrity of the image without any loss of information and achieve eye contact while simultaneously preserving the facial expression of the subject.

An overview of the system is shown in Figure 4. The steps of our algorithm are as follows:

1. Smooth and fill holes on the Kinect depth map (Figure 4a thumbnails) using Laplacian smoothing. In practice, to improve performance, we do this only on the foreground objects that are obtained using a simple depth threshold. Moreover, the silhouette from the Kinect is very inaccurate and it is possible that chunks of the face geometry can be missing. Therefore, we extend the geometry artificially by around 25 pixels.

2. Generate a novel view where the gaze is corrected (Figure 4b) This is accomplished by applying a transformation to the geometry to place the subject in the coordinate frame of the virtual camera. The parameters of this transformation are computed only once during the calibration stage (cf. section 3.1). The face now has correct gaze, but the image is no longer complete and consistent.

3. Extract the face from the corrected image and seamlessly transfer it into the original image. We use a state-of-the-art face tracker [Saragih et al. 2011] to track facial feature points in the original color image. The tracker computes 66 feature points along the chin, nose, eyes and eyebrows. We compute an optimal stencil to cut the face from the transformed image (Figure 4c). The optimal stencil has to ensure the spatial consistency of the image (cf. section 3.2) as well as the temporal consistency of the sequence (cf. section 3.3). The face is transferred by blending the two images on a narrow $5-10$ pixel wide band along the boundary (Figure 4d).

3.1 Initial Calibration

A few parameters of our system depend on the specific configuration and face characteristics that are unique to any given user. For instance, the position of the virtual camera depends on the location of the application window on the screen as well as the height of the person and the location of the sensor. These parameters are set by the user only once at the beginning of a session using a simple and intuitive interface. The calibration process typically takes less than 30 seconds. After that the system runs in a fully automatic way.

The first parameter that needs to be set is the position of the virtual camera. This is equivalent to finding a rigid transformation that, when applied to the geometry, results in an image that makes eye contact. We provide two mechanisms for that. In the first one, we allow the user, using a trackball-like interface, to find the optimal transformation by him/herself. We provide visual feedback by rendering the corrected geometry onto the window where the user is looking. This way, the user has complete control over the point at which to make eye contact. The second one is a semi-automatic technique where two snapshots are taken from the Kinect: one while the user is looking straight at the Kinect and one while the user is looking straight at the video conference window. From these two depth images we can compute the rigid transformation that maps one into the other. This is accomplished by matching the eye-tracker points in the two corresponding color/depth images.

When rigidly pasting the face from the gaze corrected to the original image we still have two degrees of freedom: a 2D translation vector that positions the corrected face in the original image. To determine this second parameter, we automatically align the facial features such that the eye and mouth positions of the two faces coincide. If necessary, we allow the user to refine the results with a simple and interactive interface with direct visual feedback. The translation is then used throughout the sequence.

3.2 Seam Optimization

In order to transfer the face from the corrected view to the original image, a seam that minimizes the visual artifacts has to be found in every frame. To accomplish this we compute a polygonal seam S that is as similar as possible in the source image and the corrected image. When blended together, the seam will appear smooth. We minimize the following energy, similar to [Dale et al. 2011]:

$$E_{TOTAL} = \sum E(p_i) \forall p_i \in S \qquad (1)$$

$$\text{where } E(p) = \sum \|I_s(q_i) - I_o(q_i)\|_2^2 \forall q_i \in B(p)$$

where I_o and I_s are the pixel intensities in the original and synthesized images and $B(p)$ is a 5×5 block of pixels around p.

Due to performance constraints we chose a local optimization technique. While this does not lead to a globally optimal solution, our experiments show that it typically leads to a solution without visible artifacts. First an ellipse is fitted to the chin points of the face tracker and offset according to the calibration (Figure 4b).

Each vertex on the upper half of the ellipse is iteratively optimized by moving it along the ray connecting the vertex to the ellipse center. We construct ellipses that have 20 to 30 points in total and our scheme converges in about four iterations. This procedure is very efficient because each vertex moves only in one dimension (the final solution will always be a simple star-shaped polygon), yet results in an artifact-free seam.We optimize only the top half of the ellipse because, unlike the forehead, the chin seam corresponds to a true depth discontinuity on the face. Therefore, we expect to see a discontinuity that makes the chin distinctive. Imposing a smooth seam along the chin will lead to unnatural visual artifacts. To further speed up the process, the optimization takes advantage of temporal coherence, and in each frame starts with the polygon from the previous frame as an initial guess.

3.3 Temporal Stabilization

Large temporal discontinuities from the Kinect geometry can lead to disturbing flickering artifacts as illustrated in the accompanying video. Although the 2D face tracking points are fairly stable in the original color image, when projected onto the geometry, their 3D positions are unreliable, particularly near depth discontinuities such as the silhouettes (see Figure 6). As this error is most predominant in the z-direction of the initial view, we fix the problem by optimizing the face tracker vertices along the respective projective rays depicted as black arrows in Figure 6 (left). A naïve averaging of the z-values over several frames would stabilize the stencil, but would create strobing artifacts when the person moves back and fourth. Instead, we first estimate the translational 3D motion of the head using the tracked points around the eyes. These points are more reliable because they are not located near a depth discontinuity. Using this information we perform temporal smoothing of the 3D face tracking vertices by averaging the z-values over several frames, subtracting the global translation between frames. This stabilization technique comes at nearly no penalty in computing resources and successfully provides a temporally consistent gaze correction even when the subject performs a wide range of motions as illustrated in the accompanying video.

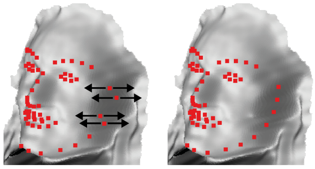

Figure 6: *3D positions of the tracked facial feature points. Left: without stabilization. The points near depth discontinuities (from the perspective of the camera) can slide arbitrarily along the z-direction, depicted as black arrows. Right: with the proposed stabilization. Points are stable in 3D even near depth discontinuities.*

4 Results and Discussion

To demonstrate and validate our system we ran it on 36 subjects. We calibrated the system for each user and let the user talk in a video conference setup for a minute. Depending on the subject, the rotation of the transformation applied for the geometry varies from 19 to 25 degrees. For this type of application, seeing the results in action is critical to evaluate the method, so we invite the reader to view the accompanying video. To keep the video to a reasonable size, we selected 11 subjects showing an average of 20 seconds each. The calibration process is very short (i.e., around 30 seconds) and the results are convincing for a variety of face types, hair-styles, ethnicities, etc. In Fig. 7 we selected a subset of challenging and interesting situations. Fig. 7a,b) illustrate how the expressiveness of the subject is preserved, in terms of both facial expression and gestures. This is crucial in video-conferencing since the meaning of non-verbal communication must not be altered. In Fig. 7b), our system rectifies the gaze of two persons simultaneously. This is done by dividing the window and applying our algorithm on each face individually. Fig. 7c) illustrates how our system is robust against lighting conditions (dimmed light and overexposure) and illumination changes. This would cause problems for a stereo-based method. Fig. 7c,d) illustrate how our method is robust to appearance changes. When the subjects pull their hair back or change their hair style, the gaze is still correctly preserved and the dynamic seam does not show any artifacts. The accompanying video contains additional results to show the temporal consistency and robustness of our method against partial occlusion, exaggerated facial expressions, pose, lighting and dynamic background.

The system runs at about 20 Hz on a consumer computer. The convincing results obtained with our method and the simplicity of use motivated the development of a Skype plugin. Users can download it from the authors' website and install it on their own computer in a few clicks. Our plugin seamlessly integrates in Skype and is very intuitive to use: a simple on/off button enables/disables our algorithm. The plugin brings real-time and automatic gaze correction to the millions of Skype users all over the world.

Limitations When the face of the subject is mostly occluded, the tracker tends to fail [Saragih et al. 2011]. This can be detected automatically and the original footage from the camera is displayed. Examples are shown in the video. Although our system is robust to many accessories that a person might wear, reflective surfaces like glasses cannot be well reconstructed resulting in visual artifacts. Since our method performs a multi-perspective rendering, the face proportions might be altered especially when the rotation is large.

5 Conclusion

Our system accomplishes two important goals in the context of video-conferencing. First and foremost, it corrects the gaze in a convincing manner while maintaining the integrity and the information of the image for both foreground and background objects, leading to artifact-free results in terms of visual appearance and communication. Second, the calibration is short and trivial and the method uses inexpensive and available equipment that will be as ubiquitous as the webcam in the near future. Given the quality of the results and its simplicity of use, our system is ideal for home video-conferencing. Finally, our intuitive Skype plugin brings gaze correction to the mainstream and consumer level.

Acknowledgements

This research, which is carried out at BeingThere Centre, is supported in part by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office. This work was also supported through the ETH Zurich Research Grant Program. Claudia Kuster is supported by the Swiss National Science Foundation (grant No. 140549). The authors thank the persons appearing in the figures and video. Thanks to Michal Nir, Saar Gross and Omri Azencot of Technion for developing the Skype plugin.

References

ARGYLE, M., AND COOK, M. 1976. *Gaze and mutual gaze*. Cambridge University Press.

CHAM, T.-J., KRISHNAMOORTHY, S., AND JONES, M. 2002. Analogous view transfer for gaze correction in video sequences. In *ICARCV*, vol. 3, 1415–1420.

CHEN, M. 2002. Leveraging the asymmetric sensitivity of eye contact for videoconference. In *CHI*, 49–56.

CRIMINISI, A., SHOTTON, J., BLAKE, A., AND TORR, P. H. S. 2003. Gaze manipulation for one-to-one teleconferencing. In *ICCV*, 191–198.

DALE, K., SUNKAVALLI, K., JOHNSON, M. K., VLASIC, D., MATUSIK, W., AND PFISTER, H. 2011. Video face replacement. In *SIGGRAPH Asia*, 1–10.

GEMMELL, J., TOYAMA, K., ZITNICK, C. L., KANG, T., AND SEITZ, S. 2000. Gaze awareness for video-conferencing: A software approach. *IEEE MultiMedia 7*, 26–35.

GRAYSON, D. M., AND MONK, A. F. 2003. Are you looking at me? eye contact and desktop video conferencing. *ACM Trans. Comput.-Hum. Interact. 10*, 221–243.

GROSS, M., WÜRMLIN, S., NAEF, M., LAMBORAY, E., SPAGNO, C., KUNZ, A., KOLLER-MEIER, E., SVOBODA, T., VAN GOOL, L., LANG, S., STREHLKE, K., MOERE, A. V., AND STAADT, O. 2003. Blue-c: a spatially immersive display and 3D video portal for telepresence. In *SIGGRAPH*, 819–827.

ISHII, H., AND KOBAYASHI, M. 1992. Clearboard: a seamless medium for shared drawing and conversation with eye contact. In *CHI*, 525–532.

JONES, A., LANG, M., FYFFE, G., YU, X., BUSCH, J., MCDOWALL, I., BOLAS, M., AND DEBEVEC, P. 2009. Achieving eye contact in a one-to-many 3D video teleconferencing system. In *SIGGRAPH*, 64:1–64:8.

KUSTER, C., POPA, T., ZACH, C., GOTSMAN, C., AND GROSS, M. 2011. FreeCam: a hybrid camera system for interactive free-viewpoint video. In *VMV*, 17–24.

MACRAE, C. N., HOOD, B., MILNE, A. B., ROWE, A. C., AND MASON, M. F. 2002. Are you looking at me? eye gaze and person perception. In *Psychological Science*, 460–464.

MATUSIK, W., AND PFISTER, H. 2004. 3D TV: a scalable system for real-time acquisition, transmission, and autostereoscopic display of dynamic scenes. In *SIGGRAPH*, 814–824.

MATUSIK, W., BUEHLER, C., RASKAR, R., GORTLER, S. J., AND MCMILLAN, L. 2000. Image-based visual hulls. In *SIGGRAPH*, 369–374.

MICROSOFT, 2010. http://www.xbox.com/en-US/kinect.

MONK, A. F., AND GALE, C. 2002. A look is worth a thousand words: Full gaze awareness in video-mediated conversation. *Discourse Processes 33*, 3, 257–278.

MUKAWA, N., OKA, T., ARAI, K., AND YUASA, M. 2005. What is connected by mutual gaze?: user's behavior in video-mediated communication. In *CHI*, 1677–1680.

NGUYEN, D., AND CANNY, J. 2005. Multiview: spatially faithful group video conferencing. In *CHI*, 799–808.

OKADA, K.-I., MAEDA, F., ICHIKAWAA, Y., AND MATSUSHITA, Y. 1994. Multiparty videoconferencing at virtual social distance: Majic design. In *Proc. Conference on Computer supported cooperative work (CSW)*, 385–393.

PETIT, B., LESAGE, J.-D., MENIER, C., ALLARD, J., FRANCO, J.-S., RAFFIN, B., BOYER, E., AND FAURE, F. 2010. Multi-camera real-time 3D modeling for telepresence and remote collaboration. *Intern. Journ. of Digital Multi. Broadcasting*.

SARAGIH, J., LUCEY, S., AND COHN, J. 2011. Deformable model fitting by regularized landmark mean-shift. *IJCV 91*, 200–215.

STOKES, R. 1969. Human factors and appearance design considerations of the mod II picturephone station set. *IEEE Transactions on Communication Technology 17*, 2, 318–323.

YANG, R., AND ZHANG, Z. 2002. Eye gaze correction with stereovision for video-teleconferencing. In *ECCV*, 479–494.

YIP, B., AND JIN, J. S. 2003. Face re-orientation in video conference using ellipsoid model. In *OZCHI*, 167–173.

ZHU, J., YANG, R., AND XIANG, X. 2011. Eye contact in video conference via fusion of time-of-flight depth sensor and stereo. *3D Research 2*, 1–10.

ZITNICK, C. L., KANG, S. B., UYTTENDAELE, M., WINDER, S., AND SZELISKI, R. 2004. High-quality video view interpolation using a layered representation. *SIGGRAPH 23*, 600–608.

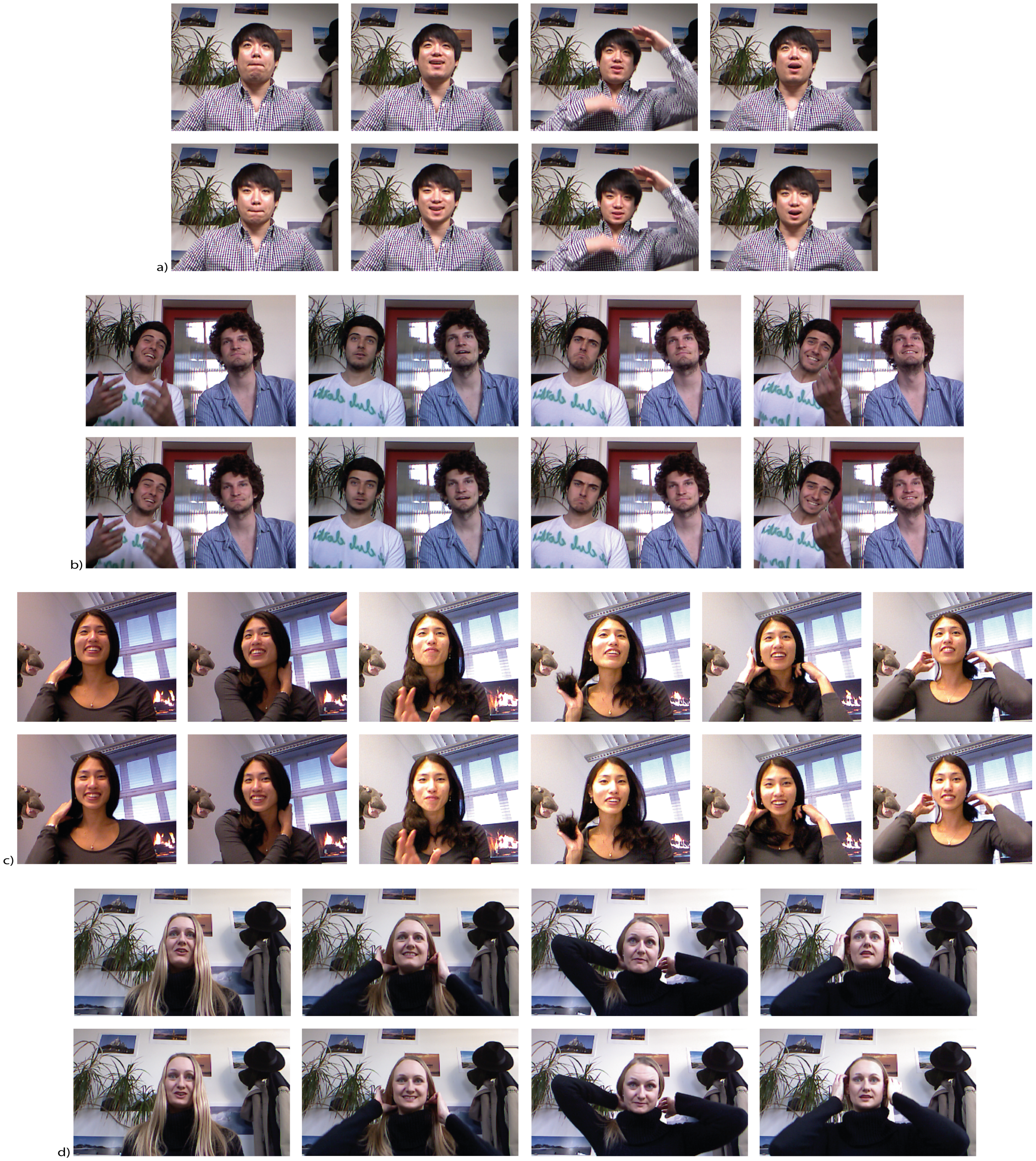

Figure 7: *A representative selection from our results. Top rows: original images from the real camera. Bottom: gaze corrected images obtained with our system. Please refer to the accompanying video for the complete sequences and additional results.*