(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
6 September 2013 (06.09.2013)

WIPO | PCT

(10) International Publication Number
WO 2013/127418 A1

(51) International Patent Classification:
H04N 7/14 (2006.01) *G06T 19/00* (2011.01)
H04N 7/15 (2006.01) *G06F 3/01* (2006.01)
G06T 15/20 (2011.01) *G06T 5/00* (2006.01)
G06T 7/00 (2006.01)

(21) International Application Number:
PCT/EP2012/004710

(22) International Filing Date:
13 November 2012 (13.11.2012)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
12001264.6 27 February 2012 (27.02.2012) EP

(71) Applicants: **ETH ZURICH** [CH/CH]; Raemistrasse 101 / ETH transfer, CH-8092 Zurich (CH). **THE TECHNION RESEARCH AND DEVELOPMENT FOUNDATION LTD.**; Senat House, Technion City, Haifa 32000 (IL).

(72) Inventors: **KUSTER, Claudia**; Waffenplatzstrasse 34, CH-8002 Zürich (CH). **POPPA, Tiberiu**; Dolderstrasse 63, CH-8032 Zürich (CH). **BAZIN, Jean-Charles**; Herbstweg 33, CH-8050 Zürich (CH). **GROSS, Markus**; Frohburgstrasse 36, CH-8006 Zürich (CH). **GOTSMAN, Craig**; Clausiusstrasse 35, CH-8006 Zürich (CH).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

(54) **Title**: METHOD AND SYSTEM FOR IMAGE PROCESSING IN VIDEO CONFERENCING FOR GAZE CORRECTION

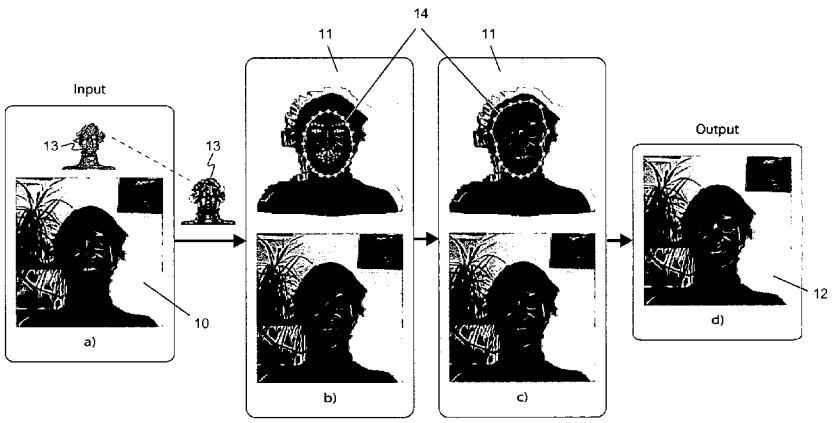

Fig. 2

(57) **Abstract**: A method for image processing in video conferencing, for correcting the gaze of an interlocutor (9) in an image or a sequence of images captured by at least one real camera (1), comprises the steps of • the at least one real camera (1) acquiring an original image of the interlocutor (9); • synthesizing a corrected view of the interlocutor's (9) face as seen by a virtual camera (3), the virtual camera (3) being located on the interlocutor's (9) line of sight and oriented towards the interlocutor (9); • transferring the corrected view of the interlocutor's (9) face from the synthesized view into the original image (10), thereby generating a final image (12); • at least one of displaying the final image (12) and transmitting the final image (12).

WO 2013/127418 A1

METHOD AND SYSTEM FOR IMAGE PROCESSING IN VIDEO CONFERENCING FOR GAZE CORRECTION

The invention relates to the field of video image processing, and in particular to a method and system for image processing in video conferencing as described in the preamble of the corresponding independent claims.

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.

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 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 sceneis 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.

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 Gemmell [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.

It is therefore an object of the invention to create a method and system for image processing in video conferencing of the type mentioned initially, which overcomes the disadvantages mentioned above.

These objects are achieved by a method and system for image processing in video conferencing according to the corresponding independent claims.

The method for image processing in video conferencing, for correcting the gaze of a human interlocutor (or user) in an image or a sequence of images captured by at least one real camera, comprises the steps of

- the at least one real camera acquiring an original image of the interlocutor;
- synthesizing, that is, computing a corrected or novel view of the interlocutor's face as seen by a virtual camera, the virtual camera being located on the interlocutor's line of sight and oriented towards the interlocutor (in particular along the line of sight);
- transferring the corrected view of the interlocutor's face from the synthesized view into the original image, thereby generating a final image;
- at least one of displaying the final image and transmitting the final image, :typically over a data communication network for display at another user's computer.

The gaze correction system is targeted at a peer-to-peer video conferencing model that runs in real-time on average consumer hardware and, in one embodiment, requires only one hybrid depth/color sensor such as the Kinect. One 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 or interlocutor. A main component of the 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 the synthesized image there is no loss of information, the facial expression is preserved as in the original image and the background is also maintained. 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.

In the remainder of the text, the gaze correction system and method shall be explained in terms of a system using only one real video camera (i.e. a color or black and white image camera) in addition to the depth map. It is straightforward to extend the system and method to employ multiple cameras.

In an embodiment, the method comprises the step of acquiring, for each original image, typically at the same time, an associated depth map comprising the face of the interlocutor, and wherein the step of synthesizing the corrected view of the interlocutor's face comprises mapping the original image onto a 3D model of the interlocutor's face based on the depth map, and rendering the 3D model from a virtual camera placed along an estimate the interlocutor's line of sight. If more than one camera is available, their respective images can be blended on the 3D model.

Alternatively, in an embodiment, it is also possible to estimate a 3D face model from one or more images alone, for example by adapting a generic 3D model to the facial features recognized in the image. Also, a generic 3D face model can be used without adapting it to the image.

The gaze correction approach can be based on a depth map acquired with a single depth scanner such as a Microsoft Kinect sensor, and 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. The 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. The approach is based on the observation that for such an 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.

In an embodiment, the step of transferring the corrected view of the interlocutor's face from the synthesized view into the original image comprises determining an optimal seam line between the corrected view and the original image that minimizes a sum of differences between the corrected view and the original image along the seam line. In an embodiment, the differences are intensity differences. The intensity considered here can be the grey value or gray level, or a combination of intensity differences from different color channels.

In an embodiment, determining the optimal seam line comprises starting with

- either an ellipsoidal polygon fitted to chin points determined by a face tracker
- or with a polygon determined for a previous image of a sequence of images

and adapting vertex points of the polygon in order to minimize the sum of differences.

In an embodiment, only vertices of an upper part of the polygon, in particular of an upper half of the polygon, corresponding to an upper part of the interlocutor's face, are adapted.

In an embodiment, the method comprises the step of temporal smoothing of the 3D position of face tracking vertices over a sequence of depth maps by estimating the 3D position of the interlocutor's head in each depth map and combining the 3D position of the vertices as observed in a current depth map with a prediction of their position computed from their position in at least one preceding depth map and from the change in the head's 3D position and orientation, in particular :from the preceding depth map to the current depth map.

In an embodiment, the method comprises the steps of, in a calibration phase, determining transformation parameters for a geometrical transform relating the position and orientation of the real camera and the virtual camera by: displaying the final image to the interlocutor and accepting user input from the interlocutor to adapt the transformation parameters until the final image is satisfactory. This can be done by the interlocutor entering a corresponding user input such as clicking on an "OK" button in a graphical user interface element displayed on the screen.

In an embodiment, the method comprises the steps of, in a calibration phase, determining transformation parameters for a geometrical transform relating the position and orientation of the real camera and the virtual camera by:

- acquiring a first depth map of at least the interlocutor's face when the interlocutor is looking at the real camera;
- acquiring a second depth map of at least the interlocutor's face when the interlocutor is looking at a display screen that is placed to display an image of an interlocutor's conference partner; and
- computing the transformation parameters from a transform that relates the first and second depth maps.

In an embodiment, the method comprises the steps of, in a calibration phase, adapting a 2D translation vector for positioning the corrected view of the interlocutor's face in the original image by: displaying the final image to the

interlocutor and accepting user input from the interlocutor to adapt the 2D translation vector until the final image is satisfactory. In a variant of this embodiment, adapting the 2D translation vector is done in the same step as determining the transformation parameters mentioned above.

In an embodiment, the method comprises the step of identifying the 3D location of the interlocutor's eyeballs using a face tracker, approximating the shape of the eyeballs by a sphere and using, in the depth map at the location of the eyes, this approximation in place of the acquired depth map information.

In an embodiment, the method comprises the step of smoothing the depth map comprising the face of the interlocutor, in particular by Laplacian smoothing.

In an embodiment, the method comprises the step of artificially extending the depth map comprising the face of the interlocutor. In addition or alternatively, the step of filling holes within the depth map can be performed. .

According to one aspect of the invention, the image and geometry processing method serves to correct the gaze of an interlocutor recorded, requires only one (depth and color) camera, where the interlocutor's line of sight is not aligned with that camera, and comprises the following steps:

a. localizing the perimeter of the head of the interlocutor by using an appropriate algorithm and
b. generating a novel view by applying a transformation to the image inside the identified perimeter so that the line of sight aligns with the camera, while leaving the image outside of the identified perimeter of the head essentially unchanged.

In an embodiment of this image and geometry processing method according to this aspect, smoothing is applied to the depth map, preferably Laplacian smoothing.

In an embodiment of this image and geometry processing method according to this aspect, the geometry around the identified perimeter through the discontinuities in the depth map is artificially enlarged to take into account the low resolution of the depth map.

In an embodiment of this image and geometry processing method according to this aspect, the transformed image inside the identified perimeter is pasted back onto the original image along an optimized seam that has as little changes as possible when comparing the transformed image with the original image.

In an embodiment of this image and geometry processing method according to this aspect, the method in addition comprises a calibration step to set up the transformation that needs to be carried out to the image inside the identified perimeter, defining the relative position of the line of sight with respect to the camera.

In an embodiment, a computer program or a computer program product for image processing in video conferencing is loadable into an internal memory of a digital computer or a computer system, and comprises computer-executable instructions to cause one or more processors of the computer or computer system execute the method for image processing in video conferencing. In another embodiment, the computer program product comprises a computer readable medium having the computer-executable instructions recorded thereon. The computer readable medium preferably is non-transitory; that is, tangible. In still another embodiment, the computer program is embodied as a reproducible computer-readable signal, and thus can be transmitted in the form of such a signal.

A method of manufacturing a non-transitory computer readable medium comprises the step of storing, on the computer readable medium, computer-executable instructions which when executed by a processor of a computing system, cause the

computing system to perform the method steps for image processing in video conferencing as described in the present document.

Further embodiments are evident from the dependent patent claims. Features of the method claims may be combined with features of the device claims and vice versa.

The subject matter of the invention will be explained in more detail in the following text with reference to exemplary embodiments which are illustrated in the attached drawings, in which:

Figure 1	shows a setup of the system's user interface hardware;
Figure 2	shows images as they are processed and combined by the system;
Figure 3	a comparison of different approaches;
Figure 4	examples of images generated by the system;
Figure 5	implementation of a 2D offset correction; and
Figure 6	correction of z-values at the boundary of pasted image segments.

The reference symbols used in the drawings, and their meanings, are listed in summary form in the list of reference symbols. In principle, identical parts are provided with the same reference symbols in the figures.

The system and method described below represent possible embodiments of the claimed invention. They can be realized by using a real camera 1 combined with a depth scanner 2, as shown in a schematic fashion in **Figure 1**. Further elements of a videoconferencing system comprise a microphone 4, a display 5, a loudspeaker 6, all typically connected to a general purpose computer 7 or dedicated video conferencing device comprising a data processing unit and programmed to perform the method as described herein. A combination of such elements and in particular of a camera and a depth scanner is embodied, for example, in a Microsoft Kinect device capable of acquiring a hybrid depth/color image, wherein the color image, in this context the

"original image" 10 is obtained using a regular camera and a depth map 13 is obtained from reflection patterns of a structured infrared image. In general, the color and depth information can be retrieved by any other adequate way. Instead of a color image 10, a black and white image 10 can be used just as well.

In an embodiment, only device required is a single hybrid depth/color sensor such as the Kinect. 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. As shown in an exemplary manner in **Figure 1**, the setup comprises a single Kinect device comprising both a real camera 1 and a depth scanner 2 viewing a person 9 or subject or interlocutor in a video conference. The angle between the Kinect and the screen window 5 is typically between 19 and 25 degrees. The gaze correction system first synthesizes a novel view where the subject 9 makes eye contact using the geometry from the depth camera (**Figure 2**b). 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 9 as well as the eye contact, we transfer only the face from the synthesized view seamlessly into the original image 10. This allows 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 9.

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 (small dots around eyes, nose, mouth and chin) and a seam line 14 in the shape of an ellipse fitted to the chin points of the face tracker (large dots).

- Bottom: Use the ellipse as a stencil to copy the gaze-corrected rendering and paste it into the original image 10. Seam artifacts are visible.

c) Optimize the seam line 14.

- Top: The subject overlayed with the new seam line 14 (large dots). Much fewer visible artifacts.

d) Blend the images along the seam line defined by edges linking the vertices to obtain the final result.

The steps of the algorithm are as follows:

1. Smooth and fill holes on the Kinect depth map 13 (**Figure 2a**) 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, especially at the edges around the face. Therefore, we extend the geometry artificially by around 25 pixels and/or fill any holes (i.e. points without depth information) in the depth map.
2. Generate a novel view or corrected view 11 where the gaze is corrected (**Figure 2b**, top). This is accomplished by applying a transformation to the geometry to place the subject in the coordinate frame of the virtual camera 3, oriented so as to look at the virtual camera 3, i.e. with the interlocutor's line of sight directed towards the virtual camera 3. 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. The line of sight is determined by the 3D head position detection, for example, as a line orthogonal to the head, i.e. in the direction of the head orientation, and passing through one of the eyes or through a point in the middle between the two eyes.
3. Extract the face from the corrected image 11 and seamlessly transfer it into the original image 10 (**Figure 2b**, bottom). 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 e.g. 66 feature points along the chin, nose, eyes and eyebrows (see also **Figure 6**) . We compute an optimal stencil or seam line 14 to cut the face from the transformed, i.e. corrected image (**Figure 2c**, top and bottom). The optimal stencil preferably ensures the spatial consistency of the image as well as the temporal consistency of the sequence. The face, that is, a segment 15 of the corrected image, is transferred by blending the two images on a narrow 5 - 10 pixel wide band along the boundary (**Figure 2d**).

Extending the geometry artificially and/or filling any holes in the depth map can be done with the following steps:

- select a kernel, i.e. a window, for example of size K=25,
- for each pixel of unknown depth, position the kernel such that its center is at this pixel and compute the average value of the known depth values inside this kernel (if no known values, then skip, i.e. ignore pixels that have no known depth value) and assign this value to the studied center pixel.

This will fill the holes whose size is smaller than K. The sequence of pixels being processed typically starts with pixels that lie next to pixels of known depth.

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 original images 10 of two different subjects where the gaze is away from the camera. In the middle column their gaze is corrected by just rotating the geometry, giving corrected views 11. The right column shows the results, i.e. the final images 12. Transferring the rotated face onto the original image 10 does not lead to perspective aberrations. Please note that the appearance of the person 9 is similar to just transforming the entire geometry with the advantage that we can preserve the integrity of the scene. **Figure 4** shows further examples of the application of the

system or method, with the original images 10 in the top row and final images 12 (or output images) with corrected gaze in the bottom row.

Initial Calibration

A few parameters of the system depend on the specific configuration and face characteristics that are unique to any given user. For instance, the position of the virtual camera 3 depends on the location of the videoconferencing application window on the display screen 5 as well as the height of the person 9 and the location of the depth sensor 2. 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 3. 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 9 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's camera 1: one while the user is looking straight at the Kinect's camera 1 and one while the user is looking straight at the video conference window on the display 5. From these two depth images we can compute the rigid transformation that maps one into the other. This can be accomplished by matching the eye-tracker points in the two corresponding color/depth images.

When rigidly pasting the face from the gaze corrected image onto the original image 10 we still have two degrees of freedom: a 2D translation vector that positions the corrected face in the original image 10. Thus a second parameter that requires an initial setting is the pasted face offset. The simplest way to determine this parameter

is to automatically align the facial features (**Figure 5** left) such that the eye and mouth positions will roughly coincide. Unfortunately, this does not lead to appealing results because the proportions of the face are still unnatural. For example, the forehead will look unnaturally shrunk. This is because in reality when we rotate the head, the face features, such as the eyes and the mouth, are moved lower in a perspective transformation. By slightly moving the decal down (**Figure 5** right) the proportions of the face can be restored. The same problem does not appear along the chin because the chin belongs to the actual silhouette of the face where there is a depth discontinuity from both the real camera and from the virtual camera 3. The user can drag the decal of the face, that is, the segment 15 of the corrected image that is transferred to the original image 11, in one frame downward until the proportions look natural. This offset can then be used throughout the process.

Seam Optimization

In order to transfer the face from the corrected view 11 to the original image 10, a seam that minimizes the visual artifacts has to be found in every frame. To accomplish this we compute a polygonal seam line 14 or 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 x 5 block of pixels around p. The pixel intensities are, for example, grey values or a combination of color intensities.

Due to performance constraints, a local optimization technique can be chosen. While this does not lead to a globally optimal solution, the 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 2**b).

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 the 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 line 14. 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.

Eye Geometry Correction

One important challenge in synthesizing human faces stems from the fact that human perception is very sensitive to faces, especially the eyes. Relatively small changes in the geometry of the face can lead to large perceptual distortions.

The geometry from the depth sensor is very coarse and as a result artifacts can appear. The most sensitive part of the face is near the eyes where the geometry is unreliable due to the reflectance properties of the eyeballs. Therefore, the eyes may look unnatural. Fortunately the eyes are a feature with relatively little geometric variation from person to person, and can be approximated well by a sphere having a radius of approximately 2.5 cm. They can be added artificially to the depth map 13, replacing the depth values provided by the depth sensor, by identifying the eyes position using the face tracker.

Temporal Stabilization

Large temporal discontinuities from the Kinect or depth map geometry can lead to disturbing flickering artifacts. Although the 2D face tracking points (facial feature 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. **Figure 6** shows 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, i.e. respective projective rays of the camera, depicted as arrows. 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. A naïve averaging of the z-values (in the camera coordinate system) over several frames would stabilize the stencil, but would create strobing artifacts when the person 9 moves back and forth. 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. The resulting corrected facial feature points are seen in Figure 6, right. 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.

Results and Discussion

To demonstrate and validate the 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. 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. 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. The system can rectify the gaze of two persons simultaneously. This is done by dividing the window and applying the method tothe each face individually. The system is robust against lighting conditions (dimmed light and overexposure) and illumination changes. This would cause

problems for a stereo-based method. The 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

The system runs at about 20 Hz on a consumer computer. The convincing results obtained with the 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. The plugin seamlessly integrates in Skype and is very intuitive to use: a simple on/off button enables/disables the 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. Although the system is robust to many accessories that a person 9 might wear, reflective surfaces like glasses cannot be well reconstructed resulting in visual artifacts. Since the method performs a multi-perspective rendering, the face proportions might be altered especially when the rotation is large.

Conclusion

The 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, the system is ideal for home video-conferencing. Finally, the intuitive Skype plugin brings gaze correction to the mainstream and consumer level.

While the invention has been described in present embodiments, it is distinctly understood that the invention is not limited thereto, but may be otherwise variously embodied and practised within the scope of the claims.

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: Asoftware 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., W¨URMLIN, 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. Multicamera 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.

PATENT CLAIMS

1. A **method** for image processing in video conferencing, for correcting the gaze of an interlocutor (9) in an image or a sequence of images captured by at least one real camera (1), comprising the steps of
   - the at least one real camera (1) acquiring an original image of the interlocutor (9);
   - synthesizing a corrected view of the interlocutor's (9) face as seen by a virtual camera (3), the virtual camera (3) being located on the interlocutor's (9) line of sight and oriented towards the interlocutor (9);
   - transferring the corrected view of the interlocutor's (9) face from the synthesized view into the original image (10), thereby generating a final image (12);
   - at least one of displaying the final image (12) and transmitting the final image (12).

2. The method of claim 1, comprising the step of acquiring, for each original image (10), an associated depth map (13) comprising the face of the interlocutor (9), and wherein the step of synthesizing the corrected view of the interlocutor's (9) face comprises mapping the original image (10) onto a 3D model of the interlocutor's (9) face based on the depth map (13), and rendering the 3D model from a virtual camera (3) placed along an estimate the interlocutor's (9) line of sight

3. The method of claim 1 or claim 2, wherein the step of transferring the corrected view of the interlocutor's (9) face from the synthesized view into the original image (10) comprises determining an optimal seam line (14) between the corrected view and the original image (10) that minimizes a sum of differences between the corrected view and the original image (10) along the seam line (14).

4. The method of claim 3, wherein the differences are intensity differences.

5. The method of claim 3 or claim 4, wherein determining the optimal seam line (14) comprises starting with

- either an ellipsoidal polygon fitted to chin points determined by a face tracker
- or with a polygon determined for a previous image of a sequence of images

and adapting vertex points of the polygon in order to minimize the sum of differences.

6. The method of claim 5, comprising adapting only vertices of an upper part of the polygon, in particular of an upper half of the polygon, corresponding to an upper part of the interlocutor's (9) face.

7. The method of one of the preceding claims, comprising the step of temporal smoothing of the 3D position of face tracking vertices over a sequence of depth maps (13) by estimating the 3D position of the interlocutor's (9) head in each depth map (13) and combining the 3D position of the vertices as observed in a current depth map (13) with a prediction of their position computed from their position in at least one preceding depth map (13) and from the change in the head's 3D position and orientation.

8. The method of one of claims 1 to 7, comprising the steps of, in a calibration phase, determining transformation parameters for a geometrical transform relating the position and orientation of the real camera (1) and the virtual camera (3) by: displaying the final image (12) to the interlocutor (9) and accepting user input from the interlocutor (9) to adapt the transformation parameters until the final image (12) is satisfactory.

9. The method of one of claims 1 to 7, comprising the steps of, in a calibration phase, determining transformation parameters for a geometrical transform relating the position and orientation of the real camera (1) and the virtual camera (3) by:
   - acquiring a first depth map (13) of at least the interlocutor's (9) face when the interlocutor (9) is looking at the real camera (1);
   - acquiring a second depth map (13) of at least the interlocutor's (9) face when the interlocutor (9) is looking at a display screen that is placed to display an image of an interlocutor's (9) conference partner; and
   - computing the transformation parameters from a transform that relates the first and second depth maps (13).

10. The method of one of the preceding claims, comprising the steps of, in a calibration phase, adapting a 2D translation vector for positioning the corrected view of the interlocutor's (9) face in the original image (10) by: displaying the final image (12) to the interlocutor (9) and accepting user input from the interlocutor (9) to adapt the (2D) translation vector until the final image (12) is satisfactory.

11. The method of one of the preceding claims, comprising the step of identifying the 3D location of the interlocutor's (9) eyeballs using a face tracker, approximating the shape of the eyeballs by a sphere and using, in the depth map (13) at the location of the eyes, this approximation in place of the acquired depth map (13) information.

12. The method of one of claims 2 to 11, comprising the step of smoothing the depth map (13) comprising the face of the interlocutor (9), in particular by Laplacian smoothing.

13. The method of one of claims 2 to 12, comprising the step of artificially extending the depth map (13) comprising the face of the interlocutor (9) at its boundaries, and/or filling holes within the depth map (13).

14 A **data processing system** for image processing in video conferencing, for correcting the gaze of an interlocutor (9) in an image or a sequence of images captured by at least one real camera (1), comprising at least one real camera (1) for acquiring an image of the interlocutor (9), the system being programmed to execute the method according to one of claims 1 to 13.

15. A **non-transitory computer readable medium** comprising computer readable program code encoding a computer program that, when loaded and executed on a computer, causes the computer to execute the method according to one of claims 1 to 13.

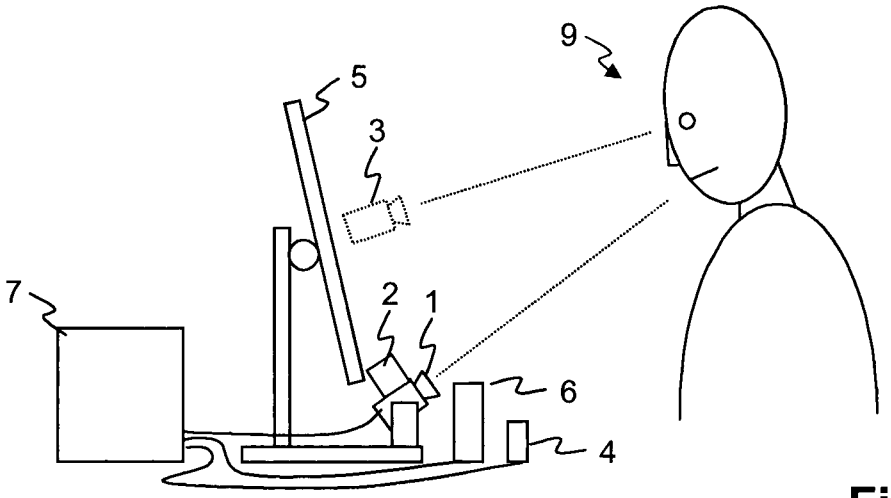

Fig. 1

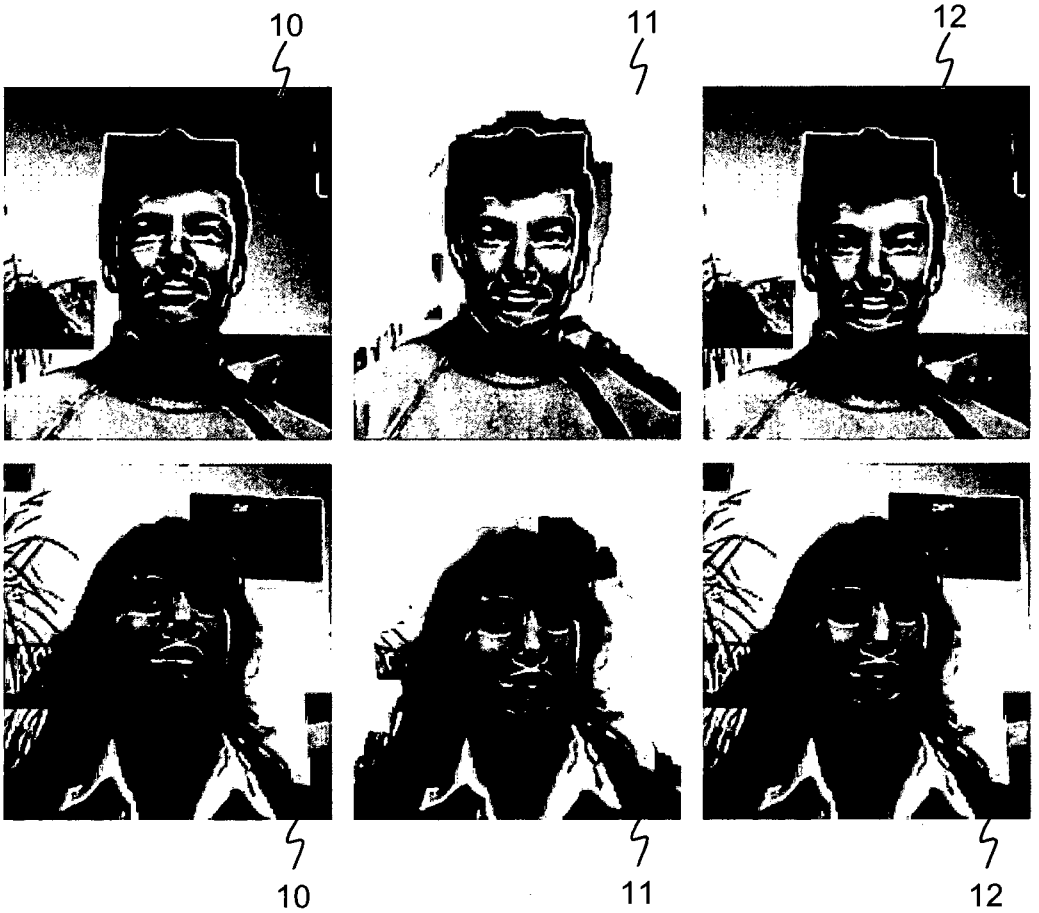

Fig. 3

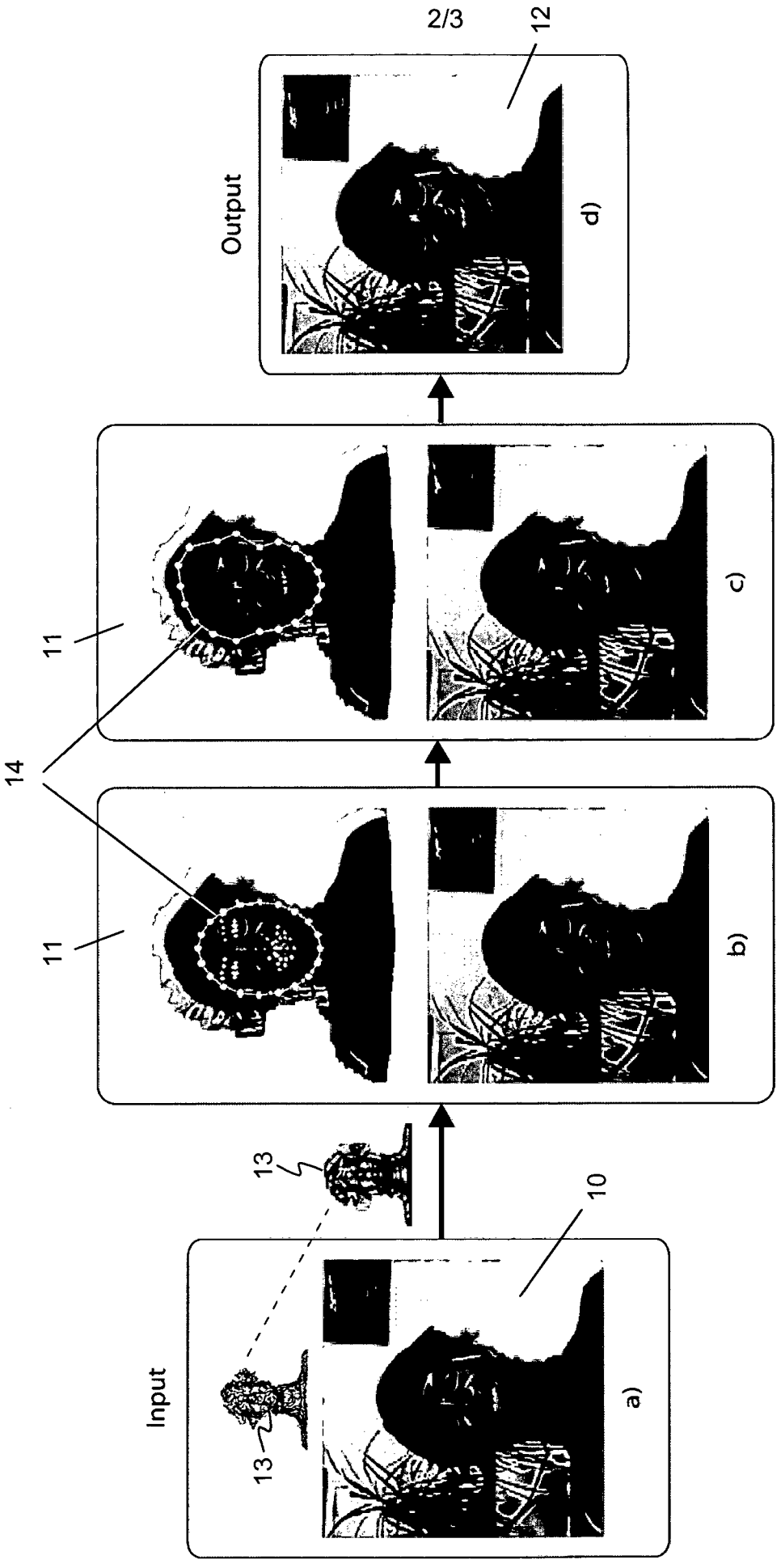

Fig. 2

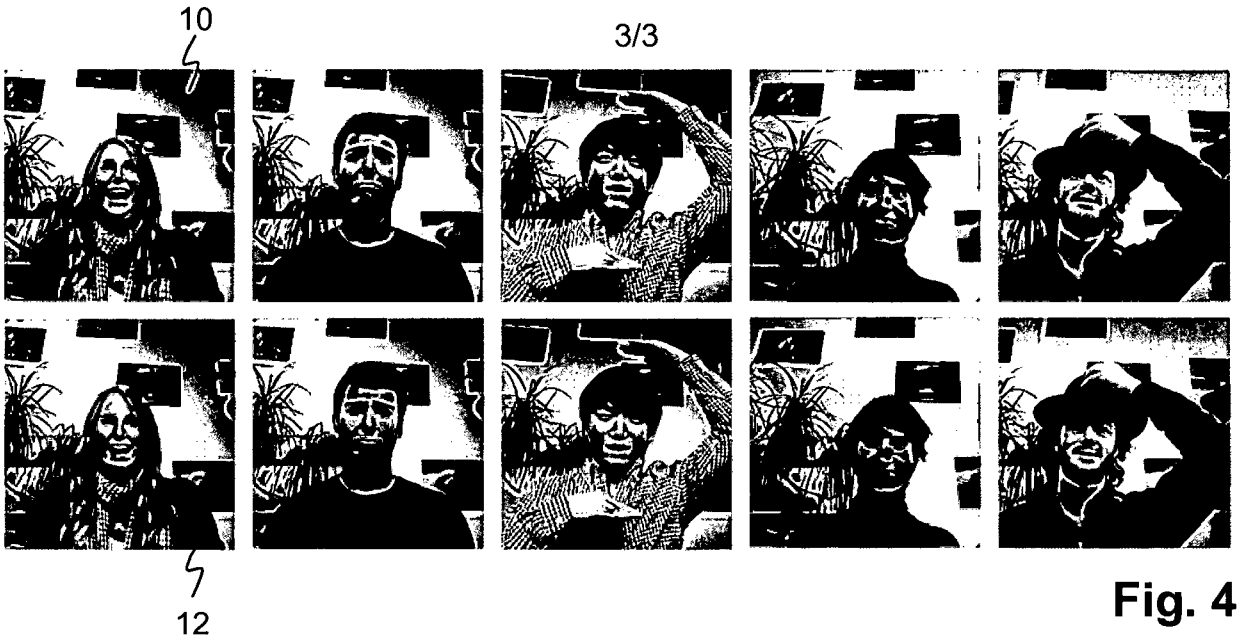

Fig. 4

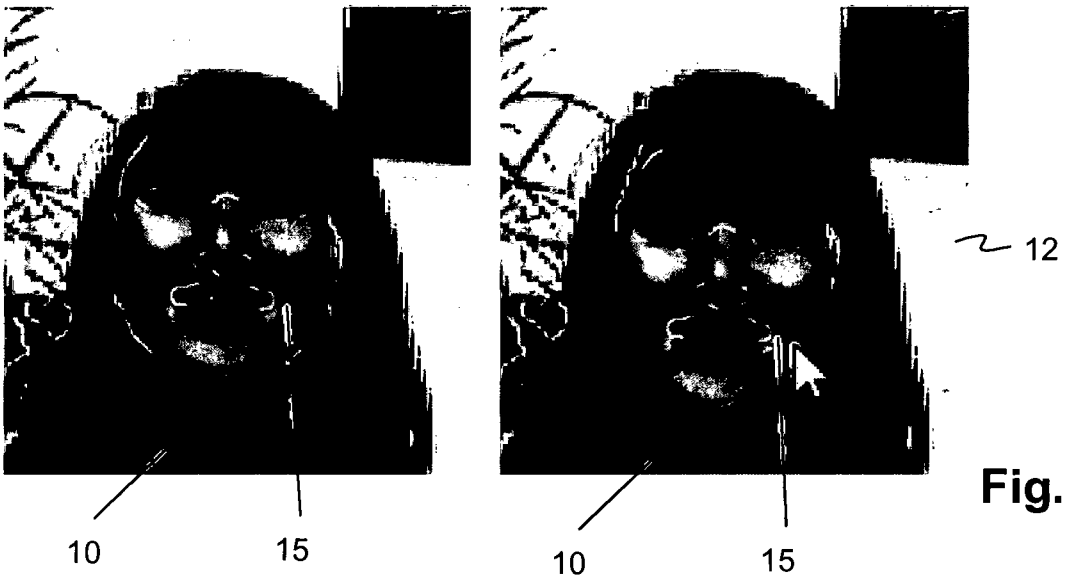

Fig. 5

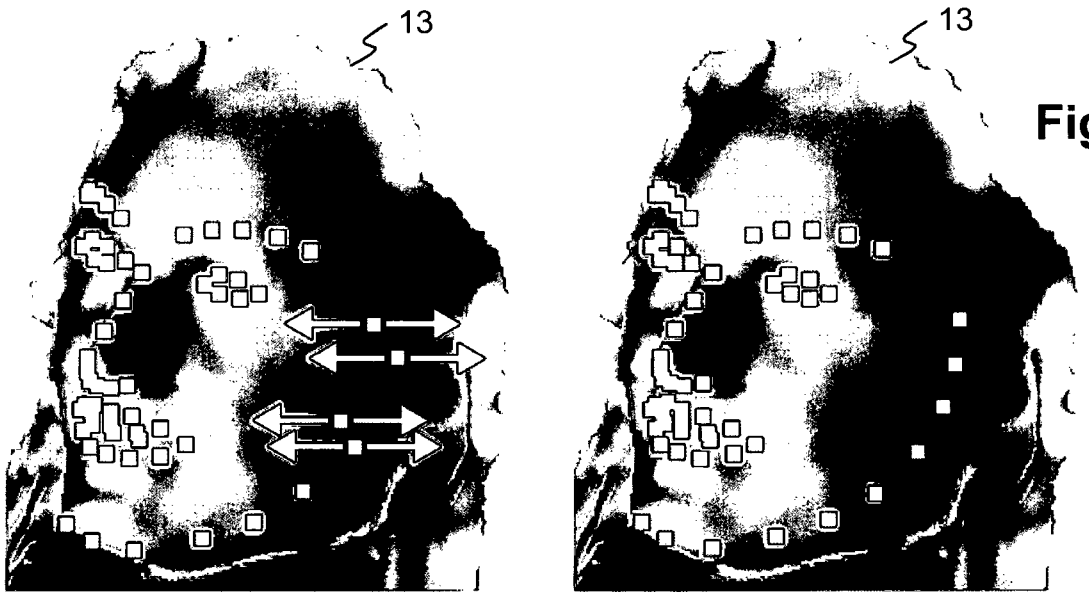

Fig. 6

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2012/004710

A. CLASSIFICATION OF SUBJECT MATTER
INV. H04N7/14 H04N7/15 G06T15/20 G06T7/00 G06T19/00
G06F3/01 G06T5/00
ADD.
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
H04N G06T G06F G06K

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2011/267348 A1 (LIN DENNIS [US] ET AL) 3 November 2011 (2011-11-03) paragraph [0022] - paragraph [0074]; figure 1 -----	1,2, 7-10, 12-15
X	WOLFGANG WAIZENEGGER ET AL: "Patch-sweeping with robust prior for high precision depth estimation in real-time systems", IMAGE PROCESSING (ICIP), 2011 18TH IEEE INTERNATIONAL CONFERENCE ON, IEEE, 11 September 2011 (2011-09-11), pages 881-884, XP032080635, DOI: 10.1109/ICIP.2011.6116699 ISBN: 978-1-4577-1304-0 the whole document -----	1,2,7, 12-15
	-/--	

[X] Further documents are listed in the continuation of Box C.

[X] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
11 February 2013	29/04/2013

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	McGrath, Simon

2

Form PCT/ISA/210 (second sheet) (April 2005)

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	SEITZ S M ET AL: "VIEW MORPHING", COMPUTER GRAPHICS PROCEEDINGS 1996 (SIGGRAPH). NEW ORLEANS, AUG. 4 - 9, 1996; [COMPUTER GRAPHICS PROCEEDINGS (SIGGRAPH)], NEW YORK, NY : ACM, US, 4 August 1996 (1996-08-04), pages 21-30, XP000682718,	1,2,14, 15
A	the whole document	8-10

X	US 2003/197779 A1 (ZHANG ZHENGYOU [US] ET AL) 23 October 2003 (2003-10-23)	1,14,15
A	the whole document	8-10

X	US 2006/017804 A1 (LEE MI-SUEN [US] ET AL) 26 January 2006 (2006-01-26) the whole document	1,14,15

A	DONG TIAN ET AL: "View synthesis techniques for 3D video", APPLICATIONS OF DIGITAL IMAGE PROCESSING XXXII, 2 September 2009 (2009-09-02), page 74430T, XP055041884, San Diego, CA DOI: 10.1117/12.829372 the whole document	1,2, 7-10, 12-15

A	EP 2 150 065 A2 (ST MICROELECTRONICS SRL [IT]) 3 February 2010 (2010-02-03) the whole document	12,13

2

INTERNATIONAL SEARCH REPORT

International application No.
PCT/EP2012/004710

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

see additional sheet

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying an additional fees, this Authority did not invite payment of additional fees.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

1, 2, 7-10, 12-15

Remark on Protest

☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.

☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

☐ No protest accompanied the payment of additional search fees.

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

This International Searching Authority found multiple (groups of) inventions in this international application, as follows:

1. claims: 1, 2, 7-10, 12-15

   Methods of gaze correction for video conferencing using a depth map and a 3D model to synthesize the corrected view of an interlocutor's face.

   ---

2. claims: 1, 3-6

   Methods of gaze correction for video conferencing which synthesize the corrected view of an interlocutor's face. Corrected view of face to pasted into "original image" (ie background) by selecting seam line that minimizes the sum of differences between the corrected view and the original image.

   ---

3. claims: 1, 11

   Methods of gaze correction for video conferencing which synthesize the corrected view of an interlocutor's face, and which using a face tracker approximate (in an unclear way) the shape of the eyeballs by a sphere, and using the sphere instead of a depth map.

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/EP2012/004710

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2011267348 A1	03-11-2011	NONE	
US 2003197779 A1	23-10-2003	NONE	
US 2006017804 A1	26-01-2006	AU 2003286303 A1	30-06-2004
		CN 1774726 A	17-05-2006
		EP 1573674 A2	14-09-2005
		JP 4568607 B2	27-10-2010
		JP 2006510081 A	23-03-2006
		KR 20050084263 A	26-08-2005
		US 2006017804 A1	26-01-2006
		WO 2004053795 A2	24-06-2004
EP 2150065 A2	03-02-2010	EP 2150065 A2	03-02-2010
		JP 2010045776 A	25-02-2010
		US 2010026712 A1	04-02-2010