



(51) International Patent Classification:

G06T 5/77 (2024.01) G06V 10/776 (2022.01)

G06T 7/11 (2017.01) G06T 3/18 (2024.01)

G06T 5/50 (2006.01) G02B 27/01 (2006.01)

G06V 10/42 (2022.01)

(21) International Application Number:

PCT/US2024/010312

(22) International Filing Date:

04 January 2024 (04.01.2024)

(25) Filing Language:

English

(26) Publication Language:

English

(71) Applicant: CANON KABUSHIKI KAISHA [JP/JP];
30-2 Shimomaru-ku, 3-chome, Ohta-ku, Tokyo, 146-8501
(JP).

(72) Inventors; and

(71) Applicants: MASEDA, Floyd Albert [US/US]; 15975 Alton
Parkway, Irvine, California 92618 (US). CAO, Xiwu
[US/US]; 15975 Alton Parkway, Irvine, California 92618
(US). DENNEY, Bradley Scott [US/US]; 15975 Alton
Parkway, Irvine, California 92618 (US).

(74) Agent: BUCHOLTZ, Jesse et al.; c/o Canon U.S.A., Inc.,
IP Division, 15975 Alton Parkway, Irvine, California 92618
(US).

(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, CV, CZ, DE, DJ, DK, DM,
DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT,
HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG,
KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY,
MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA,
NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO,
RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH,
TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS,
ZA, ZM, ZW.

(84) Designated States (unless otherwise indicated, for every
kind of regional protection available): ARIPO (BW, CV,
GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST,
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, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE,
SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN,
GQ, GW, KM, ML, MR, NE, SN, TD, TG).

(54) Title: APPARATUS AND METHOD FOR CORRECTION DURING HEAD MOUNT DISPLAY REMOVAL

(57) Abstract: According to the present disclosure an image processing apparatus and image processing method executed by the image processing apparatus are provided. The image processing method includes identifying a region in a live captured image that has been captured by an image capture device to be replaced with a corresponding region from a precaptured image; generating an output image including portions of the live captured image and the precaptured image by blending, in the identified region, pixel information from the precaptured image and the live captured image; and causing display of the output image on a display device.

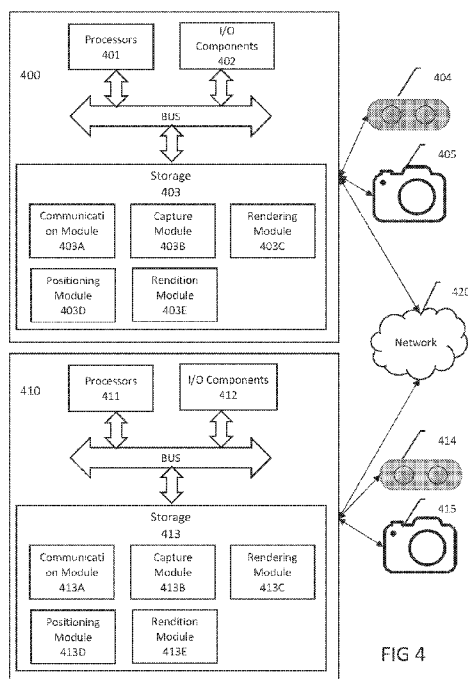


FIG 4

Published:

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

Apparatus and Method for Correction During Head Mount Display Removal

BACKGROUND

Technical Field

[0001] The present disclosure relates generally to video image processing in a virtual reality environment.

Description of Related Art

[0002] Given the progress that has been recently made in mixed reality, it is becoming practical to use a headset or Head Mounted Display (HMD) to join a virtual conference or a get-together meeting and be able to see each other with 3D faces in real-time. The need for these gatherings has been made more important because, in some scenarios such as a pandemic or other disease outbreaks, people cannot meet together in person.

[0003] Headsets are needed so we are able to see the 3D faces of each other using virtual and/or mixed reality. However, with the headset positioned on the face of a user, no one can really see the entire 3D face of others because the upper part of the face will be blocked by the headset. Therefore, to find a way to remove the headset and recover the blocked upper face region from the 3D faces is critical to the overall performance in virtual and/or mixed reality.

[0004] A further issue occurs when there is an attempt to replace the portion of the image containing the headset using a precaptured image. In doing so, it is problematic because simple replacement may not look natural to the visual perception of another user who is viewing that image in mixed or virtual reality.

SUMMARY

[0005] According to the present disclosure an image processing apparatus and image processing method executed by the image processing apparatus are provided. The image processing method includes identifying a region in a live captured image that has been captured by an image capture device to be replaced with a corresponding region from a precaptured image; generating an output image including portions of the live captured image and the precaptured image by blending, in the identified region, pixel information from the precaptured image and the live captured image; and causing display of the output image on a display device.

[0006] According to another embodiment, an image processing apparatus and image processing method includes determining an amount of each of the live captured image and precaptured image to be used for blending the pixel information; and generating the output image using the determined amount of the live captured image and the precaptured image based on amounts of pixel information corresponding to the determined amount.

[0007] According to a further embodiment, an image processing apparatus and image processing method includes determining, from the live captured image, that an predetermined object is present; and selecting, as the identified region, a group of pixels within the live captured image that includes the predetermined object.

[0008] According to another embodiment, an image processing apparatus and image processing method includes extracting, from the live captured image, one or more image characteristics; and selecting a target precapture image to be blended with the live captured image having the extracted one or more image characteristics by searching an image repository of candidate precaptured image each having been processed to include image characteristic information.

[0009] In another embodiment, an image processing apparatus and image processing method includes calculating, for each pixel in the identified region, a confidence score indicating the likelihood that the respective pixel in the identified region represents a portion of a predetermined object; and using the confidence score to determine a percentage of pixel information from each of the precapture image and live capture image when generating the output image.

[0010] Other embodiments provide an image processing apparatus and image processing method that includes generating a pixel mask representative of the all pixels in the identified region; and generating the output image by, for each pixel in the generated pixel mask, using full information for that pixel from the precaptured image when it is determined that the pixel in the identified region is part of a predetermined object; using full pixel information for that pixel from the live captured image when it is determined that the pixel in the identified region is not part of the predetermined object; and using pixel information from each of the live captured image and the precaptured image when determination as to whether that pixel is part of the predetermined object is uncertain.

[0011] In yet another embodiment, an image processing apparatus and image processing method includes that pixel information is color information.

[0012] In another embodiment, an image processing apparatus and image processing method includes performing warping processing to warp the precapture image to the live capture image based on common landmarks detected in each of the live captured image and the precapture image.

[0013] In a further embodiment, an image processing apparatus and image processing method includes performing fade processing to force use of pixel information from either the live captured image or precaptured image based on an orientation error indicating that an object in the identified region has an angle that is outside an acceptable angular threshold range.

[0014] In other embodiments, an image processing apparatus and image processing method includes determining an orientation error between the live captured image and precaptured image representing a sum of the differences between one or more orientation angles in each of the live captured image and the precaptured image.

[0015] In another embodiment, the live captured image includes a user wearing a head mount display device, and the identified region surrounds the head mount display device and the precaptured image is an image of the user captured in the live capture image without the head mount display device which has been captured at a time earlier than the live capturing, wherein each pixel in the identified region includes pixels from the precapture image in response to determining that the pixel is part of the head mount display in the live captured image and pixels from the live capture image in response to determining that the pixel is not part of the head mount display and a combination of pixels from each of the live captured image and precaptured image in response to all other determinations.

[0016] In a further embodiment, an image processing apparatus and image processing method includes displaying the output image on a display device by providing the output image to a user wearing a head mount display device; and displaying the output image on a display screen of the head mount display device.

[0017] In another embodiment, a system is provided and includes a head mount display device configured to be worn by a user; an image capture device configured to capture real time images of the user wearing the head mount display device; and an apparatus configured to execute a method according to any of the embodiments described in the present disclosure.

[0018] These and other objects, features, and advantages of the present disclosure will become apparent upon reading the following detailed description of exemplary embodiments of the present disclosure, when taken in conjunction with the appended drawings, and provided claims.

BRIEF DESCRIPTION OF THE DRAWINGS

[0019] Fig. 1 illustrates a virtual reality capture and display system according to the present disclosure.

[0020] Fig. 2 shows an embodiment of the present disclosure.

[0021] Fig. 3 shows a virtual reality environment as rendered to a user according to the present disclosure.

[0022] Fig. 4 illustrates a block diagram of an exemplary system according to the present disclosure.

[0023] Fig. 5 is a flow diagram illustrating an algorithm according to the present disclosure.

[0024] Fig. 6 is a flow diagram illustrating an algorithm according to the present disclosure.

[0025] Figs. 7A – 7D depict output images generated by the algorithms of the present disclosure.

[0026] Throughout the figures, the same reference numerals and characters, unless otherwise stated, are used to denote like features, elements, components or portions of the illustrated embodiments. Moreover, while the subject disclosure will now be described in detail with reference to the figures, it is done so in connection with the illustrative exemplary embodiments. It is intended that changes and modifications can be made to the described exemplary embodiments without departing from the true scope and spirit of the subject disclosure as defined by the appended claims.

DESCRIPTION OF THE EMBODIMENTS

[0027] Exemplary embodiments of the present disclosure will be described in detail below with reference to the accompanying drawings. It is to be noted that the following exemplary embodiment is merely one example for implementing the present disclosure and can be appropriately modified or changed depending on individual constructions and various conditions of apparatuses to which the present disclosure is applied. Thus, the present disclosure is in no way limited to the following exemplary embodiment and, according to the Figures and embodiments described below, embodiments described can be applied/performed in situations other than the situations described below as examples. Further, where more than one embodiment is described, each embodiment can be combined with one another unless explicitly stated otherwise. This includes the ability to substitute various steps and functionality between embodiments as one skilled in the art would see fit.

[0028] **Environment Overview**

[0029] FIG. 1 shows a virtual reality capture and display system 100. The virtual reality capture system comprises a capture device 110. The capture device may be a camera with sensor and optics designed to capture 2D RGB images or video, for example. In one embodiment, the image capture device 110 is a smartphone that has front and rear facing cameras and which can display images captured thereby on a display screen thereof. Some embodiments use specialized optics that capture multiple images from disparate view-points such as a binocular view or a light-field camera. Some embodiments include one or more such cameras. In some embodiments the capture device may include a range sensor that effectively captures RGBD (Red, Green, Blue, Depth) images either directly or via the software/firmware fusion of multiple sensors such as an RGB sensor and a range sensor (e.g., a lidar system, or a point-cloud based depth sensor). The capture device may be connected via a network 160 to a local or remote (e.g., cloud based) system 150 and 140 respectively, hereafter referred to as the server 140. The capture device 110 is configured to communicate via the network connect 160 to the server 140 such that the capture device transmits a sequence of images (e.g., a video stream) to the server 140 for further processing.

[0030] Also, in FIG 1, a user 120 of the system is shown. In the example embodiment the user 120 is wearing a Virtual Reality (VR) device 130 configured to transmit stereo video to the left and right eye of the user 120. As an example, the VR device may be a headset worn by the user. As used herein, the VR device and head mounted display (HMD) device may be used interchangeably. Other examples can include a stereoscopic display panel or any display device that would enable practice of the embodiments described in the present disclosure. The VR device is configured to receive incoming data from the server 140 via a second network 170. In some embodiments the network 170 may be the same physical network as network 160 although the data transmitted from the capture device 110 to the server 140 may be different than the data transmitted between the server 140 and the VR device 130. Some embodiments of the system do not include a VR device 130 as will be explained later. The system may also include a microphone 180 and a speaker/headphone device 190. In some embodiments the microphone and speaker device are part of the VR device 130.

[0031] FIG 2 shows an embodiment of the system 200 with two users 220 and 270 in two respective user environments 205 and 255. In this example embodiment, each user 220 and 270 are equipped with a respective capture devices 210 and 260, respective VR devices 230 and 280, and are connected via respective networks 240 and 270 to a server 250. In some instances, only

one user has a capture device 210 or 260, and the opposite user may only have a VR device. In this case, one user environment may be considered as a transmitter and the other user environment may be considered the receiver in terms of video capture. However, in embodiments with distinct transmitter and receiver roles, audio content may be transmitted and received by only the transmitter and receiver or by both, or even in reversed roles.

[0032] FIG 3 shows a virtual reality environment 300 as rendered to a user. The environment includes a computer graphic model 320 of the virtual world with a computer graphic projection of a captured user 310. For example, the user 220 of FIG 2, may see via the respective VR device 230, the virtual world 320 and a rendition 310 of the second user 270 of FIG 2. In this example, the capture device 260 would capture images of user 270, process them on the server 250 and render them into the virtual reality environment 300.

[0033] In the example of FIG 3, the user rendition 310 of user 270 of FIG 2, shows the user without the respective VR device 280. The present disclosure sets forth a plurality of algorithms that, when executed, cause the display of user 270 to appear without the VR device 280 and as if they were captured naturally without wearing the VR device. Some embodiments show the user with the VR device 280. In other embodiments the user 270 does not use a wearable VR device 280. Furthermore, in some embodiments the captured images of user 270 capture a wearable VR device, but the processing of the user images remove the wearable VR device and replace it with the likeness of the users face.

[0034] Additionally, the addition of the user rendition 310 into the virtual reality environment 300 along with VR content 320 may include a lighting adjustment step to adjust the lighting of the captured and rendered user 310 to better match the VR content 320.

[0035] In the present disclosure, the first user 220 of FIG 2, is shown via the respective VR device 230, the VR rendition 300 of FIG 3. Thus, the first user 220, sees user 270 and the virtual environment content 320. Likewise, in some embodiments, the second user 270 of FIG 2, will see in the same VR environment 320, but from a different view-point, e.g. the view-point of the virtual character rendition of 310 for example.

[0036] In order to achieve the immersive calling as described above, it is important to render each user within the VR environment as if they were not wearing the headset in which they are experiencing the VR content. The following describes the real-time processing performed that

obtains images of a respective user in the real world while wearing a virtual reality device 130 also referred to hereinafter as the head mount display (HMD) device.

[0037] Hardware

[0038] FIG. 4 illustrates an example embodiment of a system for virtual reality immersive calling system. The system includes two user environment systems 400 and 410, which are specially-configured computing devices; two respective virtual reality devices 404 and 414, and two respective image capture devices 405 and 415. In this embodiment, the two user environment systems 400 and 410 communicate via one or more networks 420, which may include a wired network, a wireless network, a LAN, a WAN, a MAN, and a PAN. Also, in some embodiments the devices communicate via other wired or wireless channels.

[0039] The two user environment systems 400 and 410 include one or more respective processors 401 and 411, one or more respective I/O components 402 and 412, and respective storage 403 and 413. Also, the hardware components of the two user environment systems 400 and 410 communicate via one or more buses or other electrical connections. Examples of buses include a universal serial bus (USB), an IEEE 1394 bus, a PCI bus, an Accelerated Graphics Port (AGP) bus, a Serial AT Attachment (SATA) bus, and a Small Computer System Interface (SCSI) bus.

[0040] The one or more processors 401 and 411 include one or more central processing units (CPUs), which may include one or more microprocessors (e.g., a single core microprocessor, a multi-core microprocessor); one or more graphics processing units (GPUs); one or more tensor processing units (TPUs); one or more application-specific integrated circuits (ASICs); one or more field-programmable-gate arrays (FPGAs); one or more digital signal processors (DSPs); or other electronic circuitry (e.g., other integrated circuits). The I/O components 402 and 412 include communication components (e.g., a graphics card, a network-interface controller) that communicate with the respective virtual reality devices 404 and 414, the respective capture devices 405 and 415, the network 420, and other input or output devices (not illustrated), which may include a keyboard, a mouse, a printing device, a touch screen, a light pen, an optical-storage device, a scanner, a microphone, a drive, and a game controller (e.g., a joystick, a gamepad).

[0041] The storages 403 and 413 include one or more computer-readable storage media. As used herein, a computer-readable storage medium includes an article of manufacture, for example a magnetic disk (e.g., a floppy disk, a hard disk), an optical disc (e.g., a CD, a DVD, a Blu-ray), a magneto-optical disk, magnetic tape, and semiconductor memory (e.g., a non-volatile memory

card, flash memory, a solid-state drive, SRAM, DRAM, EPROM, EEPROM). The storages 403 and 413, which may include both ROM and RAM, can store computer-readable data or computer-executable instructions.

[0042] The two user environment systems 400 and 410 also include respective communication modules 403A and 413A, respective capture modules 403B and 413B, respective rendering module 403C and 413C, respective positioning module 403D and 413D, and respective user rendition modules 403E and 413E. A module includes logic, computer-readable data, or computer-executable instructions. In the embodiment shown in FIG. 4, the modules are implemented in software (e.g., Assembly, C, C++, C#, Java, BASIC, Perl, Visual Basic, Python, Swift). However, in some embodiments, the modules are implemented in hardware (e.g., customized circuitry) or, alternatively, a combination of software and hardware. When the modules are implemented, at least in part, in software, then the software can be stored in the storage 403 and 413. Also, in some embodiments, the two user environment systems 400 and 410 includes additional or fewer modules, the modules are combined into fewer modules, or the modules are divided into more modules. One environment system may be similar to the other or may be different in terms of the inclusion or organization of the modules.

[0043] The respective capture modules 403B and 413B include operations programed to carry out image capture as shown in 110 of FIG 1, 210 and 260 of FIG 2. The respective rendering module 403C and 413C contain operations programed to carry out the functionality associated with rendering images that are captured to one or more users participating in the VR environment. The respective positioning module 403D and 413D contain operations programmed to carry out the process including identifying and determining position of each respective user in the VR environment. The respective user rendition modules 403E and 413E contains operations programmed to carry out user rendering as illustrated in the following figures described hereinbelow. The prior-training module 403F contains operations programmed to estimate the nature and type of images that were captured prior to participating in the VR environment that are used for the head mount display removal processing. In some embodiments the some modules are stored and executed on an intermediate system such as a cloud server. In other embodiments, the capture devices 405 and 415, respectively, include one or more modules stored in memory thereof that, when executed perform certain of the operations described hereinbelow.

[0044] Fading HMD Removal Processing

[0045] As noted above, in view of the progress made in augmented virtual reality, it is becoming more common to enter into an immersive communication session in a VR environment where each user is in their own location wearing a headset or Head Mounted Display (HMD) to join together in virtual reality. However, the HMD device has blocked the capability of achieving better user experience if HMD removal is not applied since you won't see the full face of others while in VR and others are unable to see your full face.

[0046] It is therefore desirable to provide a system and method that advantageously removes, from a 2D face image of a user that is wearing the HMD and participating in a VR environment. Removing the HMD from a 2D image of a user's face and not from the 3D object is practical. The reason is that humans can perceive a 3D effect from a 2D human image by inserting the 2D image into a 3D environment.

[0047] Through this virtual reality environment provided inside HMD, we are able to create some interaction between different users or meet different users in the same environment. The users can also see each other if a camera is placed in front of them to capture their live-time images and transfer to others' HMD device. However, seeing others wearing an HMD device in the view will downgrade the user experience if one person can only see an HMD image of a person instead of their full face. It is therefore desirable to replace or fill the HMD region of a face image with some precaptured or generated face. But, in doing so, it is not as simple to merely identify a region of a live capture and replace a portion of that image with a different image that does not include the HMD device. By merely replacing the portion of an image with a precaptured image, there is a negative impact on quality because there may be noticeable boundaries in the area from the image being used for replacement and the original, live captured image. This results in a disjointed composition which, during the course of a live, virtual reality communication session between users where one user wearing the HMD is viewing the VR environment sees an unnatural depiction of the other user with whom they are communicating. Therefore, the present algorithm resolves this problem by providing a blending algorithm that advantageously considers a plurality of image characteristics to perform blend processing which smooths out the composite image such that the replacement portion (e.g. upper face region) seamlessly connects with the live captured image resulting in the user being depicted, in a natural state, in the VR environment.

[0048] In exemplary HMD removal processing, the VR communication application executing on a server accesses a repository containing a plurality of precaptured images representing a face of

the user. The precapture processing is performed by an application executing on an information processing device that includes image capture hardware. In one embodiment, the information processing device that performs the precapture processing is a mobile phone. The precapture application includes a series of instructions that are visible to a user illustrated on a display device of the information processing device (e.g. mobile phone screen) that prompts the user to move their head and face in a plurality of different directions and orientations so that a series of image frames of the user are captured and stored in association with a user's profile. These images represent a set of precaptured images of the user which are stored in the repository and are accessible by the VR communication application which may use one or more of these precaptured images as part of the HMD removal processing.

[0049] As part of the precapture processing performed, each image frame captured is processed to extract image characteristic information identifying one or more image characteristics associated with the image. The extracted image characteristic information will be used in the blend processing described hereinbelow. Image characteristic information that is extracted from each of the precapture image frames captured during precapture processing includes (a) a set of facial landmarks identifying the location in pixel coordinates of a set of predetermined features of the human face, such as eyes, nose, mouth, etc.; (b) orientation information of the user's face/head relative to the optical axis (e.g. pitch, yaw, and roll) that is extracted from the above set of facial landmarks and/or from the image itself; and (c) feature information such as facial expressions, whether or not a user is blinking, eye gaze direction, etc. In one embodiment, the extraction of landmarks may include using an image or set of facial landmarks of a canonical or reference face, such as provided by Mediapipe, which is known to be looking directly along the optical axis to extract the PYR values from that rotation matrix. In another embodiment, the facial landmarks may be identified by providing the image to a trained machine learning model (e.g. neural network or CNN) on the image or landmarks directly. The extracted image characteristic information is associated with each of the precaptured image frames and stored in a data structure that allows for the VR communication application to query and retrieve one or more of the images based on image characteristic information. As will be described below, the basis for the query performed is information associated with a live captured image of the user which is then inserted, by the VR communication application, into the VR environment whereby the upper region of the user's face which is occluded by the HMD device is replaced with a portion of one of the precaptured images

having similar image characteristic information thereby resulting in a seamless combination so that the user appears in VR as they do in real-space without an HMD on their face.

[0050] In the context of the removal of a head mounted device (HMD) from a frame of video of a person wearing one (hereafter the “live image”), via selecting the upper facial region from a pre-captured image of a person’s face and inserting that region into the corresponding region of the live image, we describe herein a method of blending the two images together providing for a seamless appearance. Fig. 5 is a flow diagram detailing an algorithm for blend processing that determines, for each pixel in an image to be output for display in a VR environment, a percentage of the pixel value from either or both of a live captured image and a precaptured image that will be used to generate the image being displayed. This processing is performed on each frame of the live captured image. The steps in Fig. 5 represent instructions stored in memory and executed by one or more processors of a server to configure the server (or other computing device) to perform the described operations.

[0051] In step 502, an application executing on a server (or other computing device) obtains a live captured image of a human body wearing an HMD device. In one embodiment, the live captured image can be captured by an image capturing application executing on an information processing device (e.g. mobile phone) whereby the image capturing application captures and provides, in real-time live images on a frame by frame basis. In this embodiment, the server is in communication with the image capturing application and receives these live images as input images which are processed by a VR communication application such as described above with respect to Figs. 1 – 4 and which executes the processing on the live image to effectively identify and replace a portion of the live captured image with a portion of a precaptured image as described herein so that a final image is output for display in the VR environment. In another embodiment, the live captured images are obtained by an image capture apparatus (e.g. camera) that is in communication (either direct wired communication or through a communication network such as the internet or LAN) with the server.

[0052] In step 504, one or more image characteristics are extracted from the live captured image. The one or more image characteristics are at least the same as those described above with respect to the precaptured images. In one embodiment, extracting image characteristics from the live image corresponding to orientation information is performed based on one or more sensors in the HMD device 130 shown in Fig. 1 that is also in communication with the VR application executing

on the server. For example, the HMD device 130 in Fig. 1 contains an Inertial Measurement Unit (IMU) sensor that will continuously generate 3D orientations of an HMD device being worn by a user. It is preferred that the 3D orientations of the IMU sensor represent the 3D orientations (pitch, yaw and roll) of the head of the user wearing the HMD in the captured live image. In some embodiments an alignment process may be undergone to align the IMU coordinate axes with the camera's optical axis. Other image characteristics such as facial landmarks, expressions, etc., may be determined by various trained machine learning models, or by additional sensors in the HMD device such as eye/face tracking sensors.

[0053] In step 506, a continuous segmentation mask is generated based on the obtained live captured image, e.g. from a trained machine learning model devoted to this task. The segmentation mask can be interpreted as describing the confidence level that each pixel of the live captured image includes the HMD or does not include the HMD whereby pixels having a value of 1 represent a pixel that is 100% certain to contain the HMD and pixels having a value of 0 represent pixels that are 100% certain not to have the HMD. Typically pixels far away from the HMD will have values close to zero while those in the center of the HMD will have values close to one. Pixels on the border of the HMD will typically have less confident values (e.g. around 0.5) indicating an uncertainty whether the respective pixel is definitely inside the HMD region.

[0054] In step 508, a repository of candidate precaptured images of the same user being captured in the live capture image is queried to obtain a set of candidate precapture images having image characteristics corresponding to the extracted image characteristics from the live captured image. In this processing the query is performed using data representing the extracted image characteristics to return at least one image having substantial similarity in orientation and/or In one embodiment, characteristics representing the orientation are used as primary factors in the query such that images returned have similar orientation angles as the live captured. In another embodiment, facial expression information similarity is used alone or in combination with orientation information such that if the live captured image is determined to include a "happy" expression (e.g. presence of a smile is detected), an image in the repository tagged as "happy" or "with smile" will be returned by the search. The landmarks from the live image are extracted from an inpainting step using a CAD model of the HMD and the precapture image itself to allow mediapipe to detect landmarks there. Because the eye/nose landmarks in the live image are "fake", we ignore them during the search, only determining orientation from the lower face landmarks.

[0055] In step 510, using the retrieved precaptured image that best corresponds to the user in the current frame of the live capture image, a final output image is generated. In generating the final output image, a determination is made as to what pixel or percentage of the pixel from either the live capture image frame or the precaptured image frame will be used. In exemplary operation using the landmarks in the live and pre-captured images, the pre-captured image can be warped, e.g. by a projective transformation, so that the landmarks in the pre-captured image align with those of the live image. The HMD segmentation mask can then serve as a pixelwise blending parameter between the warped pre-captured image and the live image, according to the following equation:

$$\text{Final\_image} = (1 - \text{HMD\_mask}) * \text{live\_image} + \text{HMD\_mask} * \text{precaptured\_image} \quad (1)$$

This processing is performed on a pixel-by-pixel basis and therefore “Final_Image” refers to a respective one of the plurality of pixels used to generate the final output image in step 510. For example, for pixels where the HMD segmentation mask has a value being within a predetermined range of “1”, the algorithm selects pixels from the pre-captured image to be used in the final output image. Where the HMD segmentation mask has a value within a predetermined range of “0”, the algorithm selects a pixel from the live captured image to be used in the final output image. For HMD segmentation mask values between 0 and 1, a blend of the color values for the given pixel from each of the live captured image and the precaptured will be selected. For example if for a specific pixel the HMD segmentation mask has a value of 0.3, that pixel in the final image will be a blend of a color value 30% of the pre-captured image pixel and 70% of the live image pixel. As noted, this is performed for every pixel in a given image frame and also performed on each image frame in a series of image frames captured by the information processing device. In other embodiments, a cutoff value for either the precapture or live capture image can be set such that if the pixel is determined to exceed the cutoff value, then either the precapture or live capture image is used by itself.

[0056] In a case where it is determined that the candidate images from the repository does not include a target image having the desired image characteristics (e.g. orientation), some small differences in orientation between the live image and the selected pre-captured image may exist. In this case, step 510, a warping process is applied to the pre-captured image to align its landmarks

to the live image thereby compensating for small orientation differences. In case where the difference is too large, e.g. more than a few degrees along any axis, the upper face in the warped pre-captured image may appear distorted or otherwise not match up seamlessly with the rest of the face in the live image. In these cases, it may be desirable for the pre-captured image to be gradually faded out revealing the HMD even in the final image.

[0057] In an embodiment, a fade processing algorithm is provided to operate in conjunction with the blend processing algorithm in order to determine which pixels from a set of pixels in a live captured image and precaptured image are used to generate the output image. As used herein, fading processing represents a decision to, regardless of the pixel values determined above in Fig. 5, force the final image to always use a pixel from either the live captured image or the precaptured image. If the live captured image is in the zone of confidence, the algorithm causes the precapture image to be displayed and, if the error value exceeds an upper bound of an error range, the live capture image including the HMD is caused to be displayed. Accordingly, in some instances, this may result in a portion of the HMD being shown in the final output because without it would negatively impact the perception of the user viewing the final output image. This is particularly important, as will be discussed below when orientation error between the live capture image and the precaptured image is above a threshold error value.

[0058] Fig. 6 is a flow diagram detailing the processing steps associated with fade determination processing that is performed according to the present disclosure. The steps in Fig. 6 represent instructions stored in memory and executed by one or more processors of a server (or other computing device) to configure the server to perform the described operations.

[0059] In step 602, a determination is made as whether orientation information associated with the live image which is obtained based on readings from the IMU of the HMD and provided with live capture image frame is within a zone of confidence indicating that the head of the user wearing the HMD in the live image is within a predetermined orientation range and thus no fading is to be performed. In one embodiment, the zone of confidence represents a range of a respective type of orientation angles. For example, the orientation angle may be a yaw angle and the range may be $+50^\circ$ and -50° may not allow fading. But, if the orientation angle as determined from the IMU is outside that range, predetermined fading processing is performed when generating the final output image in step 510 in Fig. 5. If the result of the determination in step 602 indicates that orientation angle is within the predetermined zone of confidence, then no fading processing is performed and

blending processing as described in Fig. 5 is used as shown in step 604. An exemplary final output image 700 generated after step 604 is shown in Fig. 7A. In the final output image, when it has been determined that the orientation angle is within the zone of confidence, lower face region 702 comprising pixels from the live view image are combined with upper face region 704 comprising pixels from the precaptured image selected, in accordance with the processing and blending described above with respect to Fig. 5. In this exemplary final output image, the HMD represented by the dotted line labeled 706 is not visible in the final output image.

[0060] If the determination in step 602 is negative indicating that the orientation of the live captured image is outside of the zone of confidence indicating that one or more of the orientation angles are outside a predetermined threshold range, a determination is made in step 603 which determines whether an orientation error is below a predetermined orientation error threshold. If the result of that determination is positive indicating that the orientation error is lower than the threshold, then processing proceeds to step 604 as discussed above. If the result of the determination in step 604 is positive processing proceeds to step 605.

[0061] In exemplary operation, the orientation error determination is performed in the following manner. Given orientation information, e.g. in the form of pitch, yaw, and roll, for the pre-captured image and for the live image, a total orientation error is defined among the angles in any normed fashion. For example, the L1-norm over all three angles is shown in Equation 2 below:

$$\text{error} = |\text{pitch}_p - \text{pitch}_l| + |\text{yaw}_p - \text{yaw}_l| + |\text{roll}_p - \text{roll}_l| \quad (2)$$

where “pitch_p”, “yaw_p” and “roll_p” are orientation angles obtained from the precaptured image and “pitch_l”, “yaw_l”, “roll_l” are orientation angles from the live image. In Equation 3, the orientation error over the orientation angles is the combination of absolute value of each difference between orientations for each type of orientation angle. Because a difference in roll is more easily accounted for by many warping functions like projective transforms, in some embodiments the error in this direction may be weighted less severely compared to pitch and yaw. In others, a different norm such as the L2 norm (or a weighted version thereof) may be used. Once a sufficient error metric has been chosen, the magnitude of the error may then be used to inform the mixture in equation (1), e.g. by artificially lowering the value of the HMD segmentation mask when error is large. This will have the effect of putting an upper bound on the percentage of pre-capture color values to use in the live image. A projective transform can include a rotation in image space, which

is largely functionally equivalent to a user rolling their head along the camera's optical axis. A head tilt of 20 degrees can be achieved by rotating the image 20 degrees CW and the eyes/nose will look substantially the same. If on the other hand I change my pitch by 20 degrees (i.e. I look up), there is no transform that can be applied to mimic this since new parts of the face (e.g. below my chin) will be visible and others (e.g. top of head) will no longer be visible. Since the different angles have different effects on the image, their weights in equation (2) may be modified to matter more or less. For example, an A can be placed in front of pitch, B in front of yaw, and C in front of roll. In the default L1 norm, $A=B=C=1$, so all angles are weighted equally but this can vary so that, for example, $A=B=1$ and $C=0.01$, then roll will matter 100x less than pitch and yaw.

[0062] In one embodiment a piecewise function of orientation error is defined with a lower threshold below which no fading will occur, i.e. the HMD segmentation mask will be allowed to max out at 1 and therefore the final image may have pixels that are taken from the pre-captured image in their entirety. and an upper threshold above which the pre-capture image will be faded *entirely*, i.e. the HMD segmentation mask is forced to all zeros meaning the final image will contain only pixels taken directly from the live image. In mathematical form, this is shown in Equation 3 as follows:

$$\text{max\_opacity} = \text{clip}((\text{error} - \text{upper}) / (\text{lower} - \text{upper}), 0, 1) \quad (3)$$

where $\text{clip}(x, \text{min}, \text{max})$ returns min if $x < \text{min}$, max if $x > \text{max}$, and x if $\text{min} < x < \text{max}$. Note that $\text{max\_opacity} = 0$ if $\text{error} \geq \text{upper}$, $\text{max\_opacity} = 1$ if $\text{error} \leq \text{lower}$, and $0 < \text{max\_opacity} < 1$ if $\text{lower} < \text{error} < \text{upper}$. The mixture equation (1) may then be modified by replacing HMD_mask with $\text{HMD\_mask} * \text{max\_opacity}$. In another embodiment, the max_opacity definition can be represented in a different form such as an exponentially decaying function, a sigmoid-type function, or any suitable function $f(\text{error})$ such that $f(0) = 1$ and $f(x)$ tends to 0 as x tends to infinity.

[0063] Turning back to Fig. 6, in response to the determination in step 605 which determines if the orientation error is above an upper bound of orientation errors, processing proceeds to step 606 where fading processing is performed in step 606 such that the final output image will have the HMD fully visible because the pixels selected for the final output image will be entirely from the live captured image. An example of the final output image 710 generated after step 606 is shown in Fig. 7B. Therein, the heavy fading results in pixels from the live captured image that contain

the HMD device 716 are shown more prominently with lower face region pixels 712 from the live captured image and the pixel values from the precaptured image 714 are deemphasized.

[0064] If the determination in step 605 indicates that the orientation error does not exceed an upper bound, processing provides to step 608 whereby the max opacity value described above is reduced such that HMD will be partially visible. This final output image 720 generated in accordance with step 608 is shown in Fig. 7C.

[0065] In another embodiment, rather than (or in addition to) considering orientation error, determination as to the fade amount may be done by considering one or more of the orientation angles directly, starting the fade only when one or more angles is outside of a predetermined range. For example to fade at large pitch, similar to the above, one can set a lower and upper threshold for fading and use the pitch from the live or pre-captured image directly instead of error in equation (2).

[0066] In another embodiment, fade processing is based on values contained in an alpha channel that is transmitted with and associated with each of the precapture images and the live capture images. In some embodiments where pre-captured and live images contain an alpha channel indicating transparency, it may be desired to keep the alpha channel e.g. of the pre-captured image even if the RGB (or other suitable color space) channels will be faded. In this case, the max_opacity parameter may only be multiplied to the HMD segmentation mask for the blending of the color channels while the alpha channel may use the original unfaded HMD segmentation mask. In one embodiment, this is used to generate a visual anchor to allow the HMD to be visible outside of the user's face in case of a misalignment between the face in the pre-captured image and that in the live image. In this case, the HMD segmentation mask can be binarized above some threshold, and the alpha channel of the final image may be forced to 1 in this region (e.g. all pixels where $HMD\_mask > 0.5$ have alpha channel forced to 1). This is illustrated in Fig. 7D which shows the final output image 730 generated in accordance with fade processing using the alpha channel. As shown in Fig. 7D, the final output image includes the border of the HMD 736 to be force displayed over the precapture upper face region 734 and the lower face region from the live captured image 732.

[0067] After the processing described hereinabove the generated final output image is provided as input to the VR communication application which transmits, during a communication session, the image of the user that includes the blended face region determined in accordance with the

disclosure herein so that the user can appear, in the virtual reality environment to another user with whom they are communicating such that the user is able to see a live captured image of the user with the HMD removed in a manner that is seamless and visually acceptable for the communication session. These images are able to exist in the VR environment and the display thereof is maintained throughout the communication session whereby the blended face regions are continually updated on a frame by frame basis throughout the VR communication session.

[0068] At least some of the above-described devices, systems, and methods can be implemented, at least in part, by providing one or more computer-readable media that contain computer-executable instructions for realizing the above-described operations to one or more computing devices that are configured to read and execute the computer-executable instructions. The systems or devices perform the operations of the above-described embodiments when executing the computer-executable instructions. Also, an operating system on the one or more systems or devices may implement at least some of the operations of the above-described embodiments.

[0069] Furthermore, some embodiments use one or more functional units to implement the above-described devices, systems, and methods. The functional units may be implemented in only hardware (e.g., customized circuitry) or in a combination of software and hardware (e.g., a microprocessor that executes software).

[0070] Additionally, some embodiments of the devices, systems, and methods combine features from two or more of the embodiments that are described herein. Also, as used herein, the conjunction “or” generally refers to an inclusive “or,” though “or” may refer to an exclusive “or” if expressly indicated or if the context indicates that the “or” must be an exclusive “or.”

[0071] While the present disclosure has been described with reference to exemplary embodiments, it is to be understood that the invention is not limited to the disclosed exemplary embodiments.

CLAIMS

We Claim,

1. An image processing method comprising:
identifying a region in a live captured image that has been captured by an image capture device to be replaced with a corresponding region from a precaptured image;
generating an output image including portions of the live captured image and the precaptured image by blending, in the identified region, pixel information from the precaptured image and the live captured image; and
causing display of the output image on a display device.
2. The image processing method according to claim 1, further comprising:
determining an amount of each of the live captured image and precaptured image to be used for blending the pixel information; and
generating the output image using the determined amount of the live captured image and the precaptured image based on amounts of pixel information corresponding to the determined amount.
3. The image processing method according to claim 1, further comprising:
determining, from the live captured image, that an predetermined object is present;
selecting, as the identified region, a group of pixels within the live captured image that includes the predetermined object.
4. The image processing method according to claim 1, further comprising:
extracting, from the live captured image, one or more image characteristics; and
selecting a target precapture image to be blended with the live captured image having the extracted one or more image characteristics by searching an image repository of candidate precaptured image each having been processed to include image characteristic information.
5. The image processing method according to claim 1, further comprising:

calculating, for each pixel in the identified region, a confidence score indicating the likelihood that the respective pixel in the identified region represents a portion of a predetermined object; and

using the confidence score to determine a percentage of pixel information from each of the precapture image and live capture image when generating the output image.

6. The image processing method according to claim 1, further comprises:
generating a pixel mask representative of the all pixels in the identified region; and
generating the output image by, for each pixel in the generated pixel mask,
using full information for that pixel from the precaptured image when it is determined that the pixel in the identified region is part of a predetermined object;
using full pixel information for that pixel from the live captured image when it is determined that the pixel in the identified region is not part of the predetermined object; and
using pixel information from each of the live captured image and the precaptured image when determination as to whether that pixel is part of the predetermined object is uncertain.

7. The image processing method according to claim 1, wherein pixel information is color information.

8. The image processing method according to claim 1, further comprising:
performing warping processing to warp the precapture image to the live capture image based on common landmarks detected in each of the live captured image and the precapture image.

9. The image processing method according to claim 1, further comprising
performing fade processing to force use of pixel information from either the live captured image or precaptured image based on an orientation error indicating that an object in the identified region has an angle that is outside an acceptable angular threshold range.

10. The image processing method according to claim 1 further comprising:

determining an orientation error between the live captured image and precaptured image representing a sum of the differences between one or more orientation angles in each of the live captured image and the precaptured image.

11. The image processing method according to claim 1, wherein the live captured image includes a user wearing a head mount display device, and the identified region surrounds the head mount display device.

12. The image processing method according to claim 11, wherein the precaptured image is an image of the user captured in the live capture image without the head mount display device which has been captured at a time earlier than the live capturing.

13. The image processing method according to claim 12, wherein each pixel in the identified region includes pixels from the precapture image in response to determining that the pixel is part of the head mount display in the live captured image and pixels from the live capture image in response to determining that the pixel is not part of the head mount display and a combination of pixels from each of the live captured image and precaptured image in response to all other determinations.

14. The image processing method according to claim 1, wherein displaying the output image on a display device further comprises:

providing the output image to a user wearing a head mount display device; and
displaying the output image on a display screen of the head mount display device.

15. An information processing apparatus comprising:

One or more memories storing instructions; and

One or more processors that, upon execution of the stored instructions, are configured to execute a method according to any of claims 1 – 14.

16. A non-transitory computer readable storage medium storing instructions that, when executed by one or more processors of an apparatus, configures the apparatus to perform a method according to any of claims 1 - 14.

17. A system comprising:

- a head mount display device configured to be worn by a user;
- an image capture device configured to capture real time images of the user wearing the head mount display device; and
- an apparatus configured to execute a method according to any of claims 1 – 14.

1/7

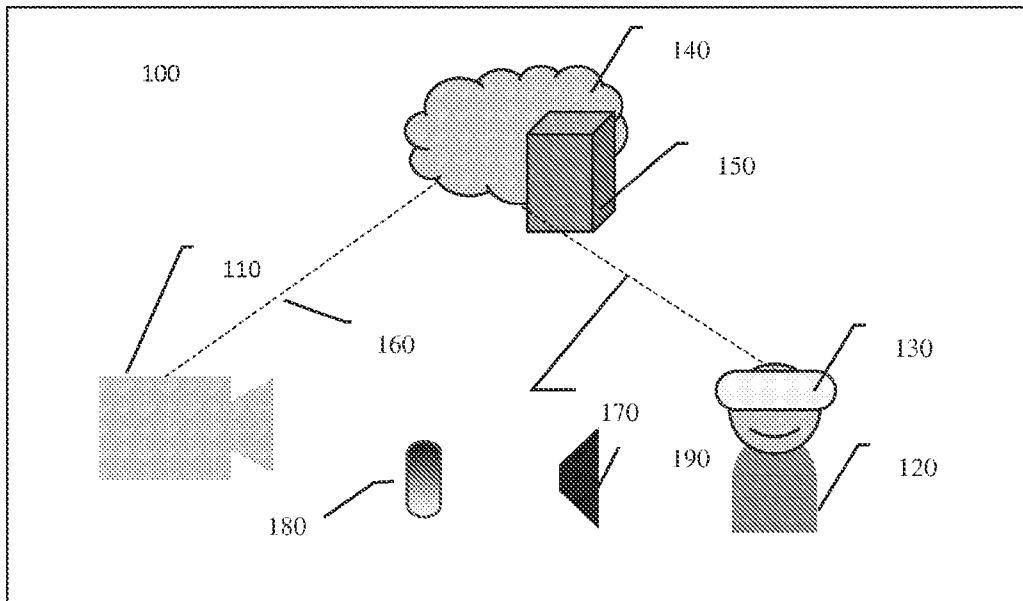


FIG 1

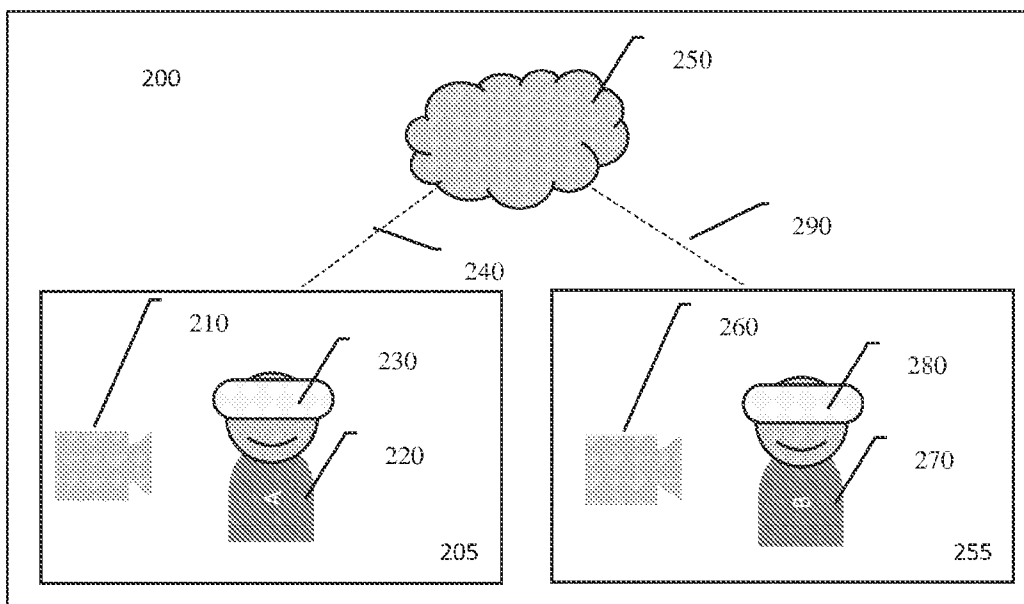


FIG 2

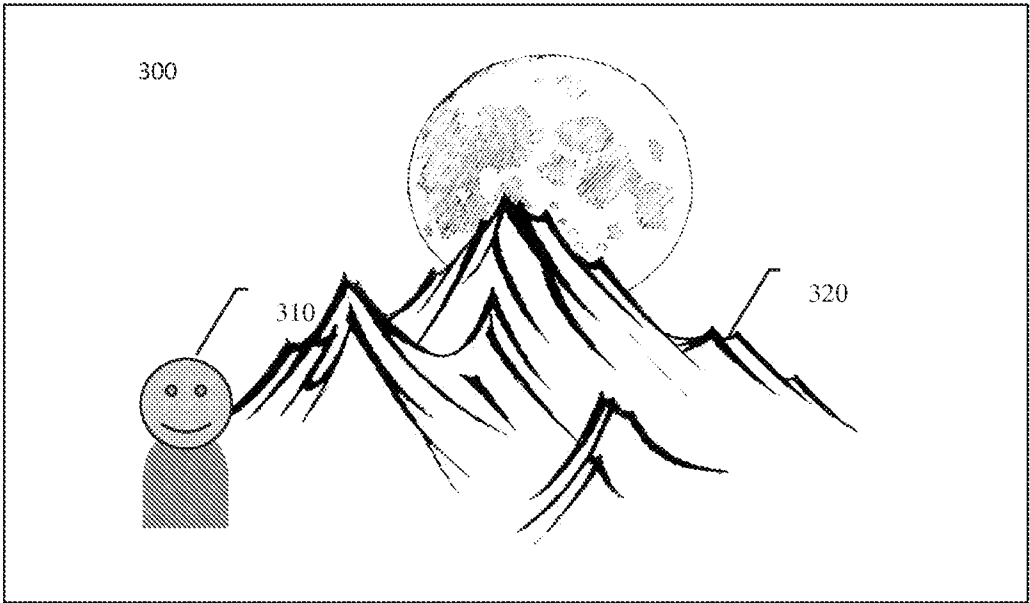


FIG 3

3/7

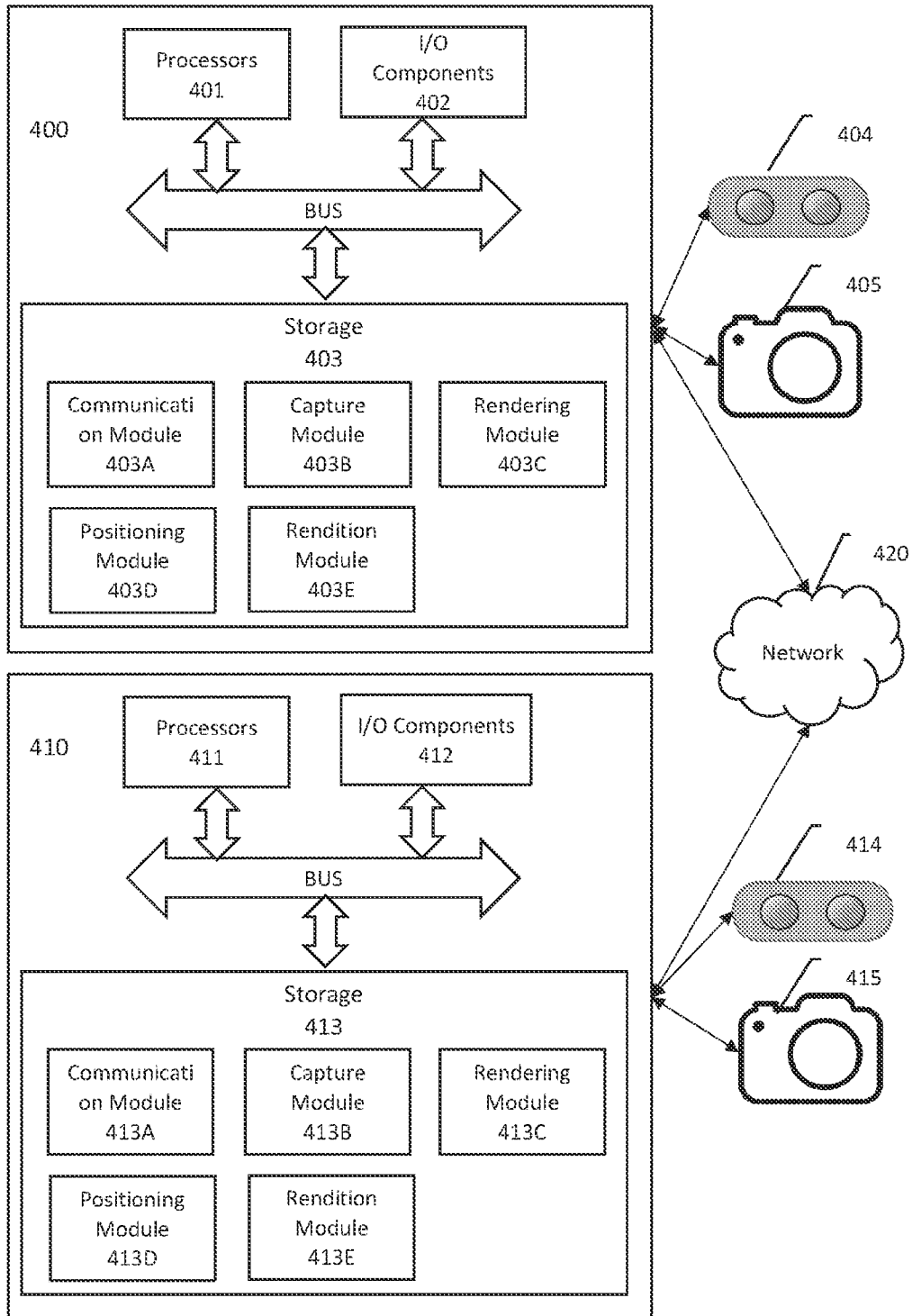


FIG 4

4/7

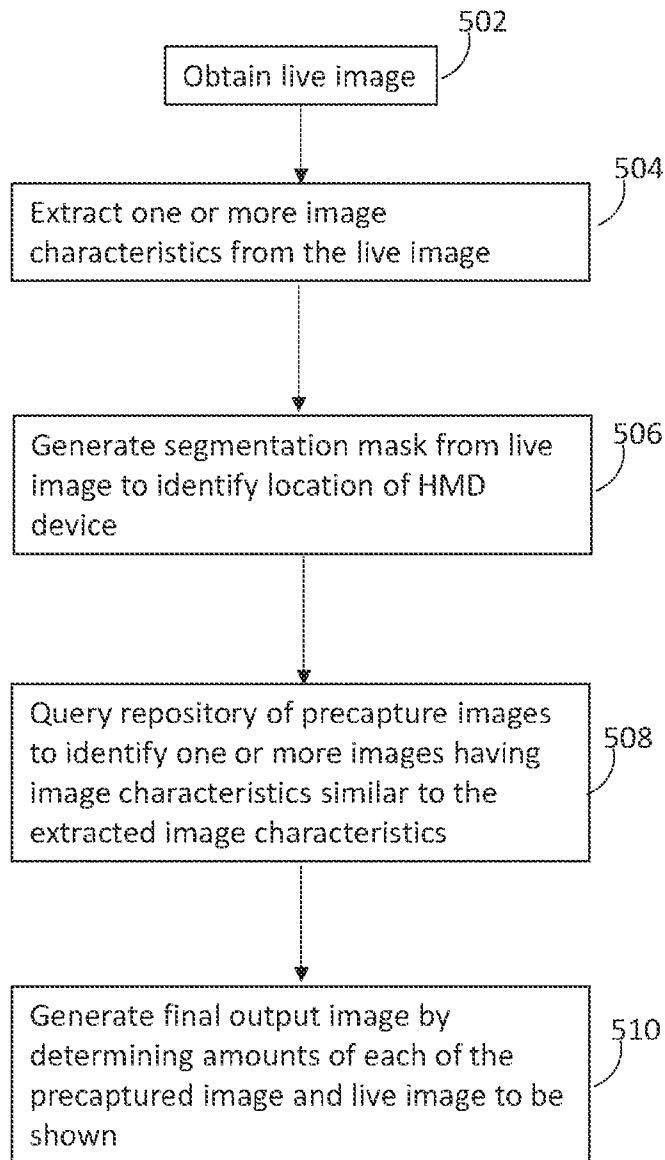


FIG 5

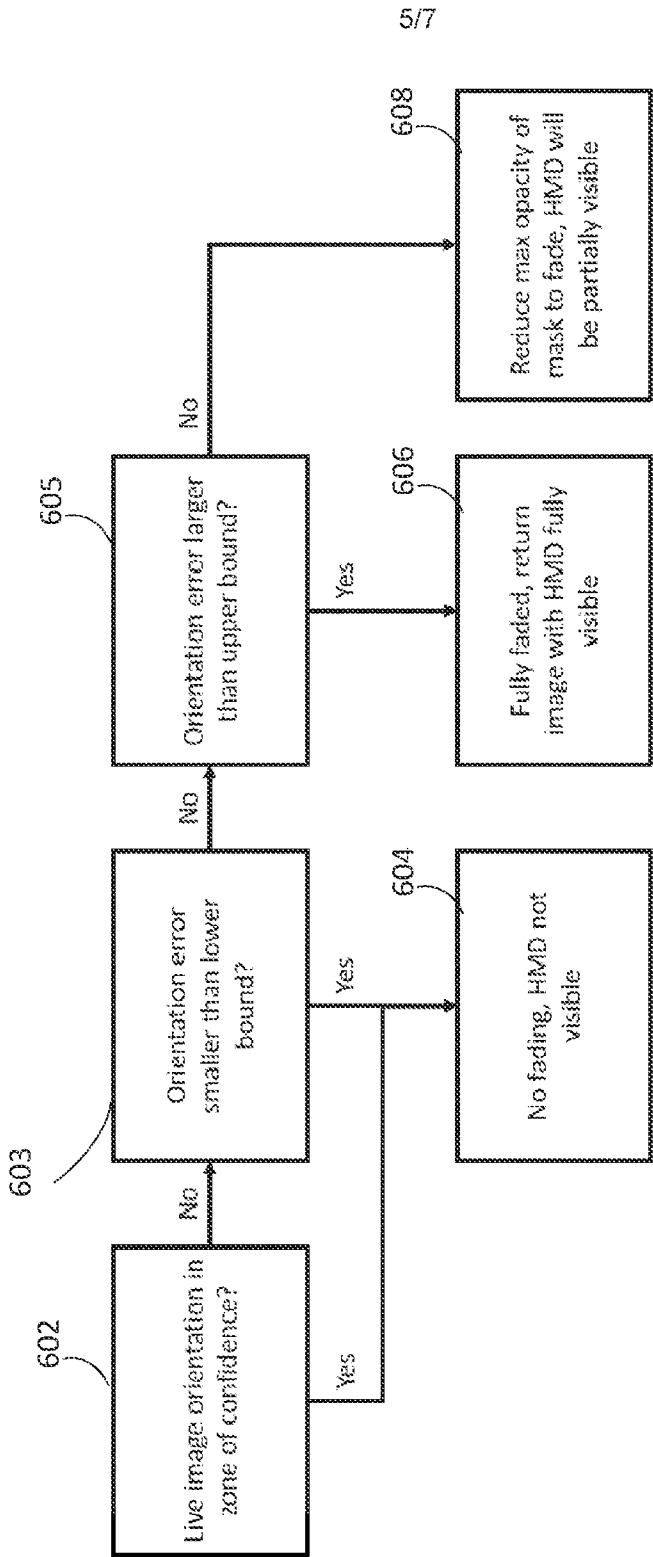


FIG 6

6/7

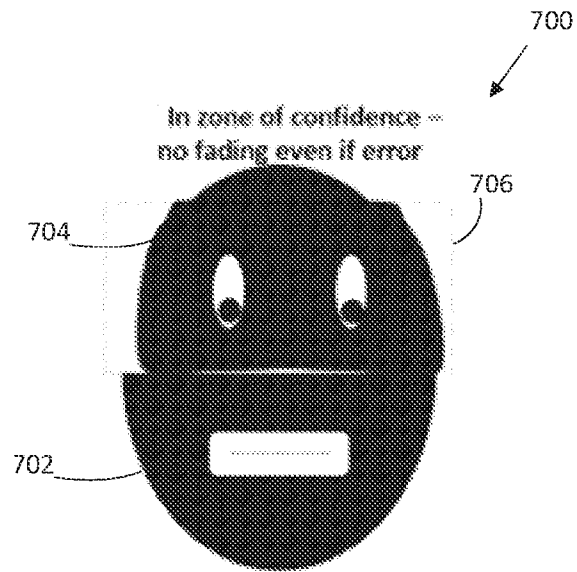


FIG 7A

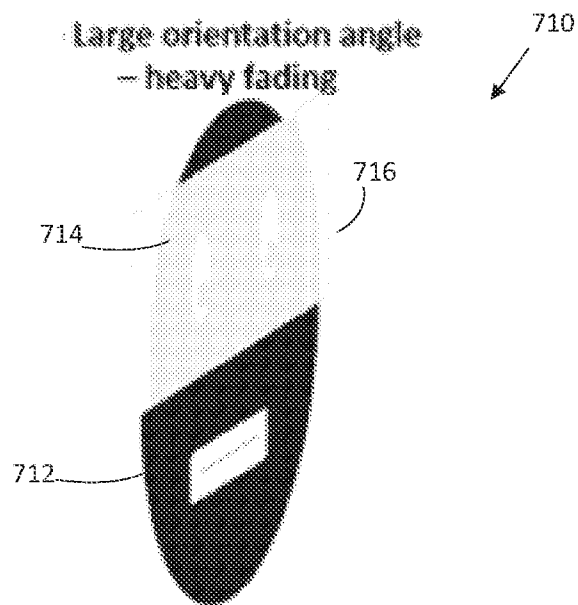


FIG 7B

7/7

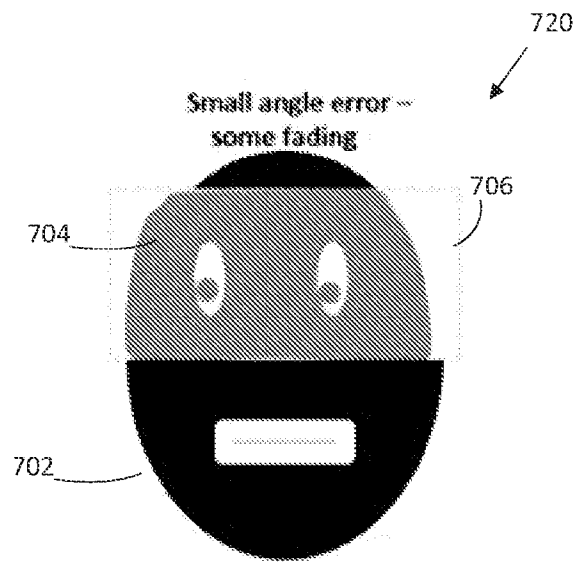


FIG 7C

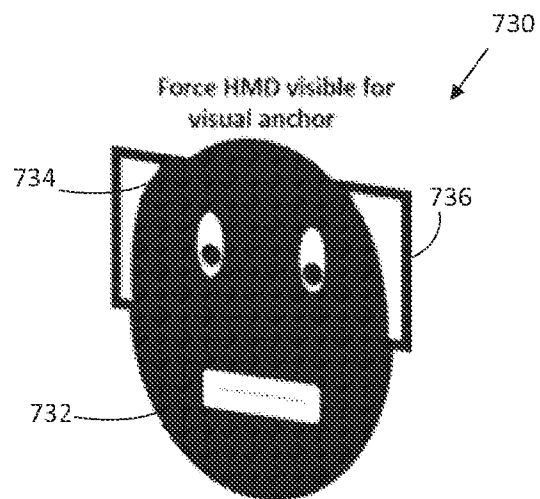


FIG 7D

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US2024/010312

A. CLASSIFICATION OF SUBJECT MATTER G06T 5/77(2024.01)i; G06T 7/11(2017.01)i; G06T 5/50(2006.01)i; G06V 10/42(2022.01)i; G06V 10/776(2022.01)i; G06T 3/18(2024.01)i; G02B 27/01(2006.01)i 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) G06T 5/77(2024.01); G06F 18/00(2023.01); G06K 9/00(2006.01); G06T 19/00(2011.01); G06T 3/00(2006.01); G06T 5/00(2006.01); G06T 5/50(2006.01); G06T 7/30(2017.01) Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched Korean utility models and applications for utility models Japanese utility models and applications for utility models Electronic data base consulted during the international search (name of data base and, where practicable, search terms used) eKOMPASS(KIPO internal) & Keywords: virtual reality, HMD (head mounted display), occlusion, precaptured image, synthesizing, warping		
C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X Y	KR 10-2017-0085477 A (FACEBOOK, INC.) 24 July 2017 (2017-07-24) paragraphs [0004], [0039]; claims 1-2, 4, 9-10; and figure 2	1,11-12,14-17 2-10,13
Y	US 2022-0398705 A1 (GOOGLE LLC) 15 December 2022 (2022-12-15) paragraphs [0041], [0077]-[0078], [0108], [0125]; and claim 2	2-7
Y	KR 10-2020-0074780 A (SAMSUNG ELECTRONICS CO., LTD.) 25 June 2020 (2020-06-25) paragraphs [0067], [0087]; and claims 1, 3, 5	8-10
Y	KR 10-2016-0135652 A (SAMSUNG ELECTRONICS CO., LTD.) 28 November 2016 (2016-11-28) paragraph [0051]	13
A	US 2018-0158246 A1 (INTEL IP CORPORATION) 07 June 2018 (2018-06-07) paragraphs [0025]-[0159]; and figures 1A-15	1-17
☐ Further documents are listed in the continuation of Box C. ☒ 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 “D” document cited by the applicant in the international application “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 20 September 2024		Date of mailing of the international search report 20 September 2024
Name and mailing address of the ISA/KR Korean Intellectual Property Office 189 Cheongsa-ro, Seo-gu, Daejeon 35208, Republic of Korea Facsimile No. +82-42-481-8578		Authorized officer YANG, Jeong Rok Telephone No. +82-42-481-5709

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/US2024/010312

Patent document cited in search report				Publication date (day/month/year)		Patent family member(s)		Publication date (day/month/year)	
KR	10-2017-0085477	A	24 July 2017	CN	107004296	A	01 August 2017		
				CN	107004296	B	29 December 2020		
				EP	3178066	A1	14 June 2017		
				EP	3178066	A4	17 January 2018		
				JP	2017-534096	A	16 November 2017		
				JP	6654625	B2	26 February 2020		
				KR	10-2422469	B1	20 July 2022		
				US	10445863	B2	15 October 2019		
				US	10593025	B2	17 March 2020		
				US	11238568	B2	01 February 2022		
				US	2017-0243334	A1	24 August 2017		
				US	2019-0392559	A1	26 December 2019		
				US	2020-0175655	A1	04 June 2020		
				WO	2016-020921	A1	11 February 2016		
				US	2022-0398705	A1	15 December 2022	CN	115529835
EP	4091141	A1	23 November 2022						
JP	2023-524326	A	12 June 2023						
KR	10-2022-0140402	A	18 October 2022						
KR	10-2612529	B1	11 December 2023						
WO	2022-216333	A1	13 October 2022						
KR	10-2020-0074780	A	25 June 2020	KR	10-2607789	B1	30 November 2023		
				WO	2020-130579	A1	25 June 2020		
KR	10-2016-0135652	A	28 November 2016	CN	106170083	A	30 November 2016		
				CN	106170083	B	07 August 2020		
				EP	3096208	A1	23 November 2016		
				EP	3096208	B1	13 March 2019		
				KR	10-2581453	B1	21 September 2023		
				US	10527846	B2	07 January 2020		
				US	10684467	B2	16 June 2020		
				US	2016-0341959	A1	24 November 2016		
				US	2018-0143429	A1	24 May 2018		
				US	2018-0149864	A1	31 May 2018		
				US	9910275	B2	06 March 2018		
				US	2018-0158246	A1	07 June 2018	None	