US 20190370544A1

(19) **United States**

(12) **Patent Application Publication** (10) **Pub. No.: US 2019/0370544 A1**

Wright, JR. et al. (43) **Pub. Date: Dec. 5, 2019**

(54) **OBJECT INITIATED COMMUNICATION**

(71) Applicant: **PTC Inc.**, Needham, MA (US)

(72) Inventors: **Gerald Wright, JR.**, Cardiff, CA (US); **John James Lechleiter**, Carlsbad, CA (US); **Devender Yamakawa**, Burlingame, CA (US)

(21) Appl. No.: **15/993,550**

(22) Filed: **May 30, 2018**

Related U.S. Application Data

(63) Continuation of application No. PCT/US18/35193, filed on May 30, 2018.

Publication Classification

(51) **Int. Cl.**

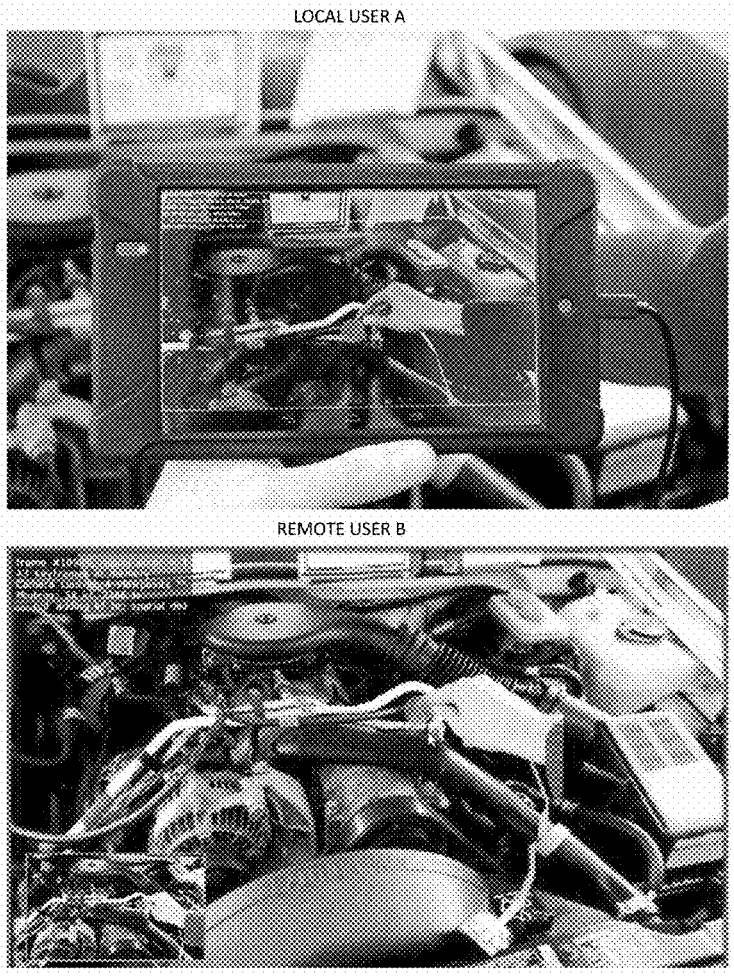
G06K 9/00	(2006.01)
G06K 9/62	(2006.01)
H04N 7/14	(2006.01)
G06T 7/20	(2006.01)
G06F 3/14	(2006.01)
G06K 7/14	(2006.01)

(52) **U.S. Cl.**

CPC ***G06K 9/00671*** (2013.01); ***G06K 9/6201*** (2013.01); ***H04N 7/147*** (2013.01); *G06T 2207/10016* (2013.01); ***G06F 3/1454*** (2013.01); ***G06K 7/1413*** (2013.01); ***G06K 7/1417*** (2013.01); ***G06T 7/20*** (2013.01)

(57) **ABSTRACT**

Communication between two electronic devices can be initiated based on detection of an object in the environment of one of the electronic devices. A method for object initiated communication between a user of a device and a remote individual for assisting the user in interacting with the object includes capturing an image with a camera of the device, detecting the object within the captured image, and locating a record for the detected object within a database of objects. The method can also include locating an identifier for initiating the communication with the remote individual, wherein the identifier is associated with the record of the detected object and identifies an address for initiating the communication. The method can also include initiating the communication between the user and the remote individual based on the identifier.



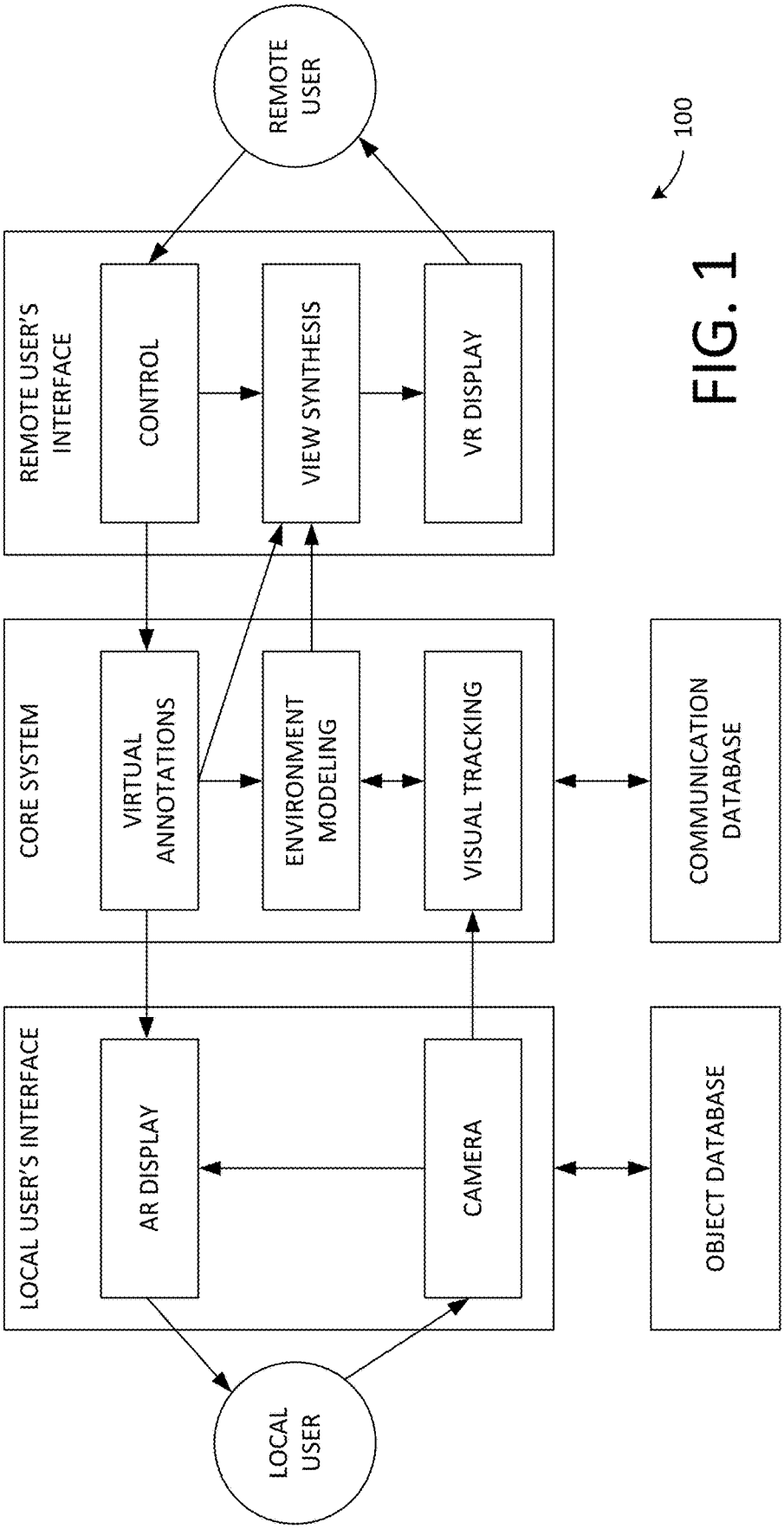

REMOTE USER
100
REMOTE USER'S INTERFACE
CONTROL
VIEW SYNTHESIS
VR DISPLAY
CORE SYSTEM
VIRTUAL ANNOTATIONS
ENVIRONMENT MODELING
VISUAL TRACKING
COMMUNICATION DATABASE
LOCAL USER'S INTERFACE
AR DISPLAY
CAMERA
OBJECT DATABASE
LOCAL USER

FIG. 1

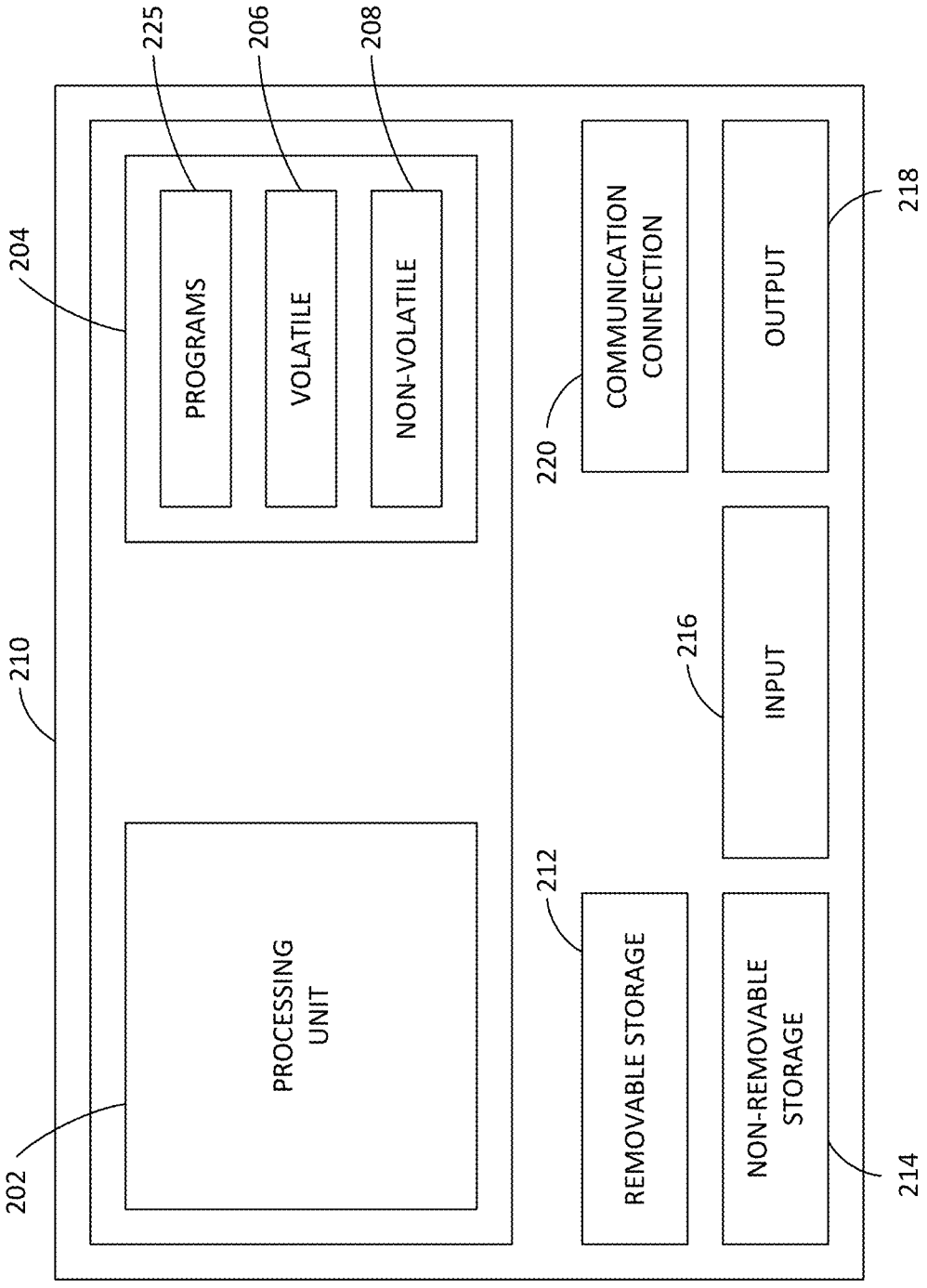
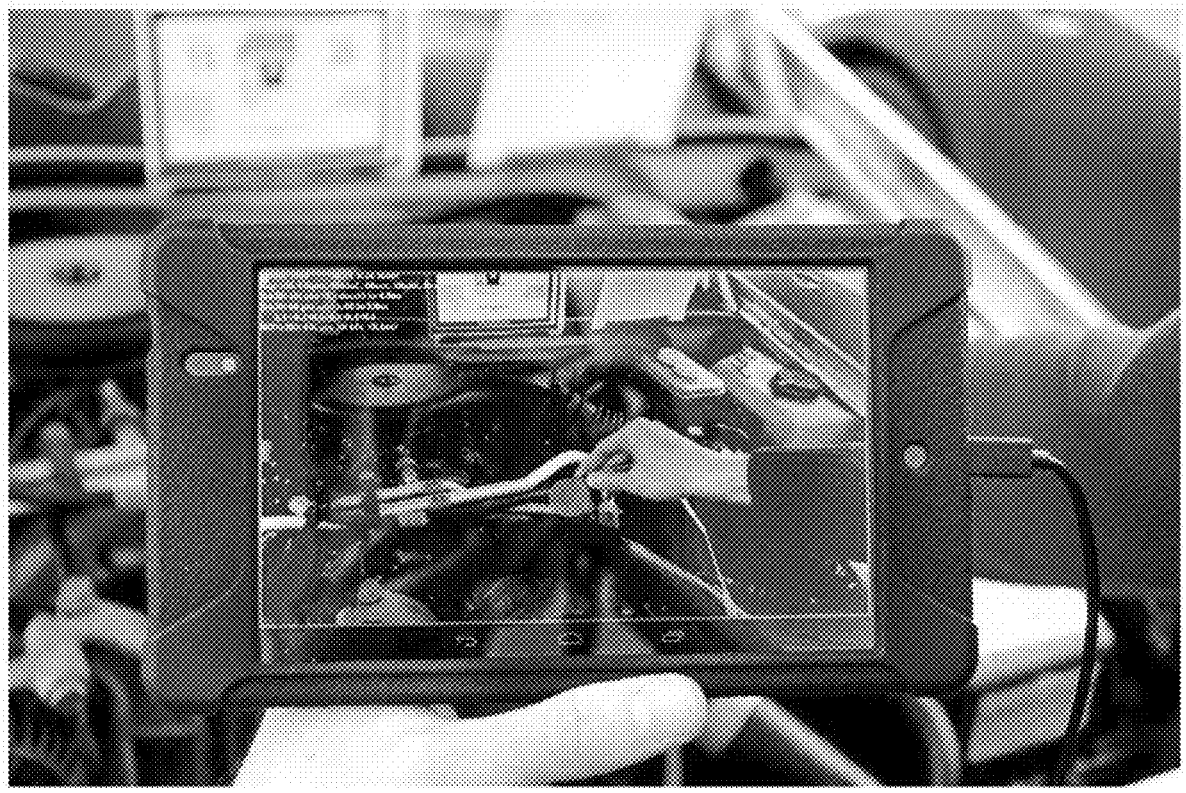

225
206
208
204
PROGRAMS
VOLATILE
NON-VOLATILE
210
PROCESSING UNIT
202
220
COMMUNICATION CONNECTION
218
OUTPUT
216
INPUT
212
REMOVABLE STORAGE
214
NON-REMOVABLE STORAGE

FIG. 2

LOCAL USER A



REMOTE USER B

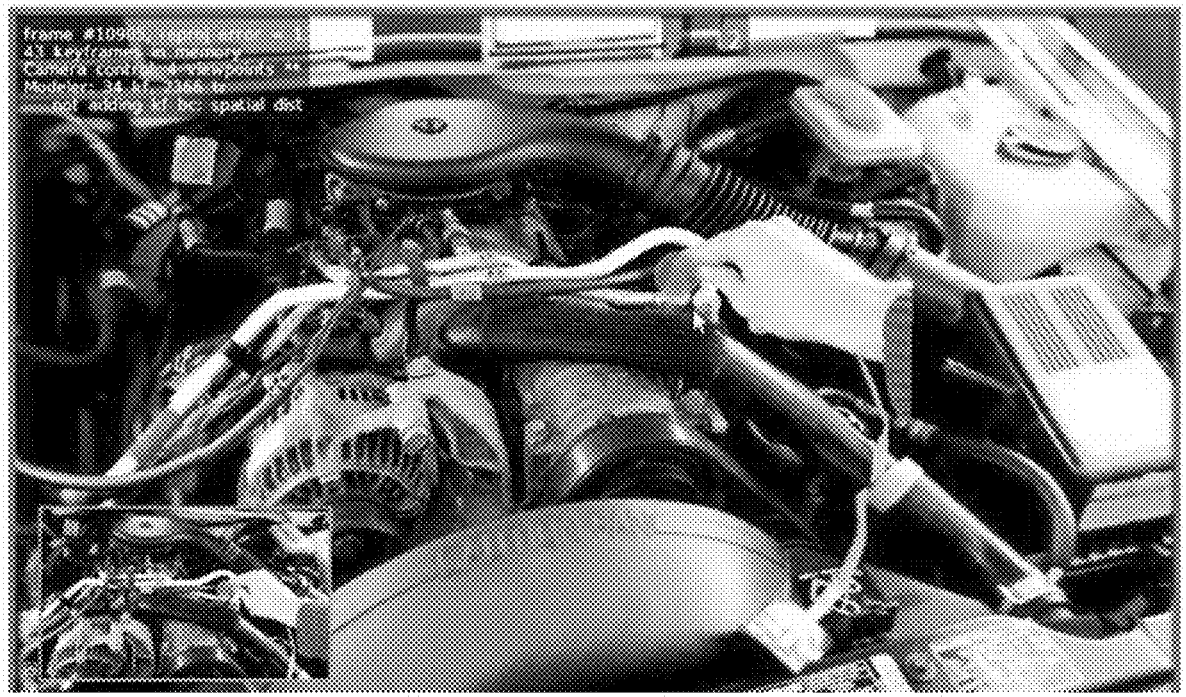


FIG. 3

FIG. 5

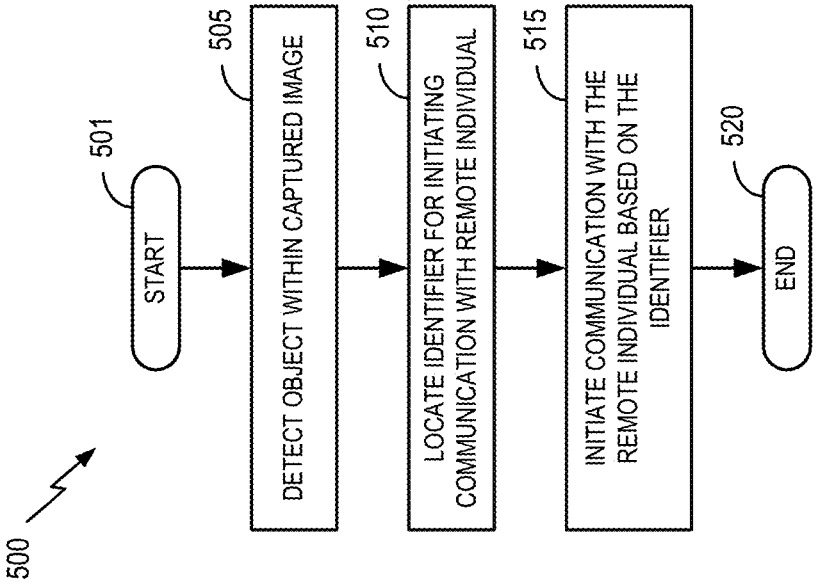

500
501
START
505
DETECT OBJECT WITHIN CAPTURED IMAGE
510
LOCATE IDENTIFIER FOR INITIATING COMMUNICATION WITH REMOTE INDIVIDUAL
515
INITIATE COMMUNICATION WITH THE REMOTE INDIVIDUAL BASED ON THE IDENTIFIER
520
END

FIG. 4

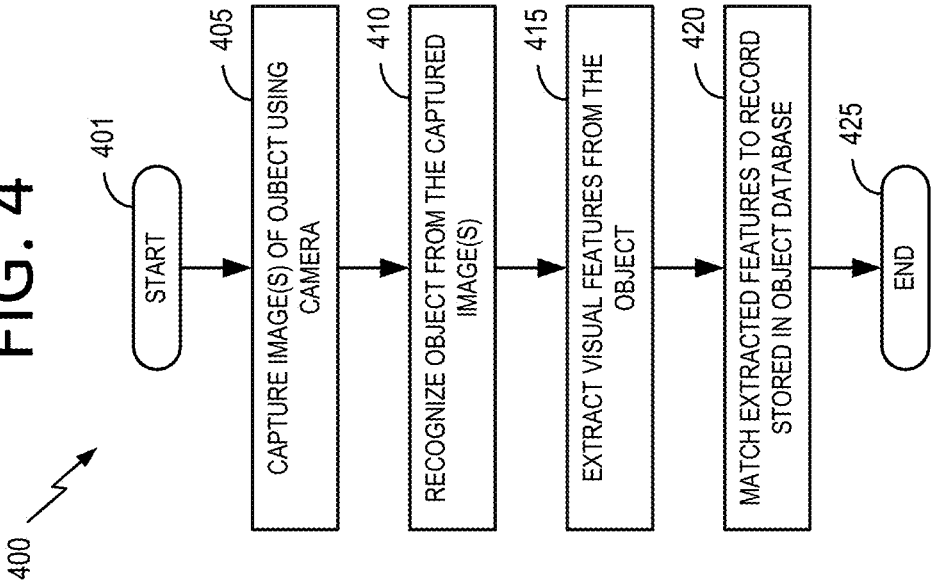

400
401
START
405
CAPTURE IMAGE(S) OF OJBECT USING CAMERA
410
RECOGNIZE OBJECT FROM THE CAPTURED IMAGE(S)
415
EXTRACT VISUAL FEATURES FROM THE OBJECT
420
MATCH EXTRACTED FEATURES TO RECORD STORED IN OBJECT DATABASE
425
END

FIG. 6B

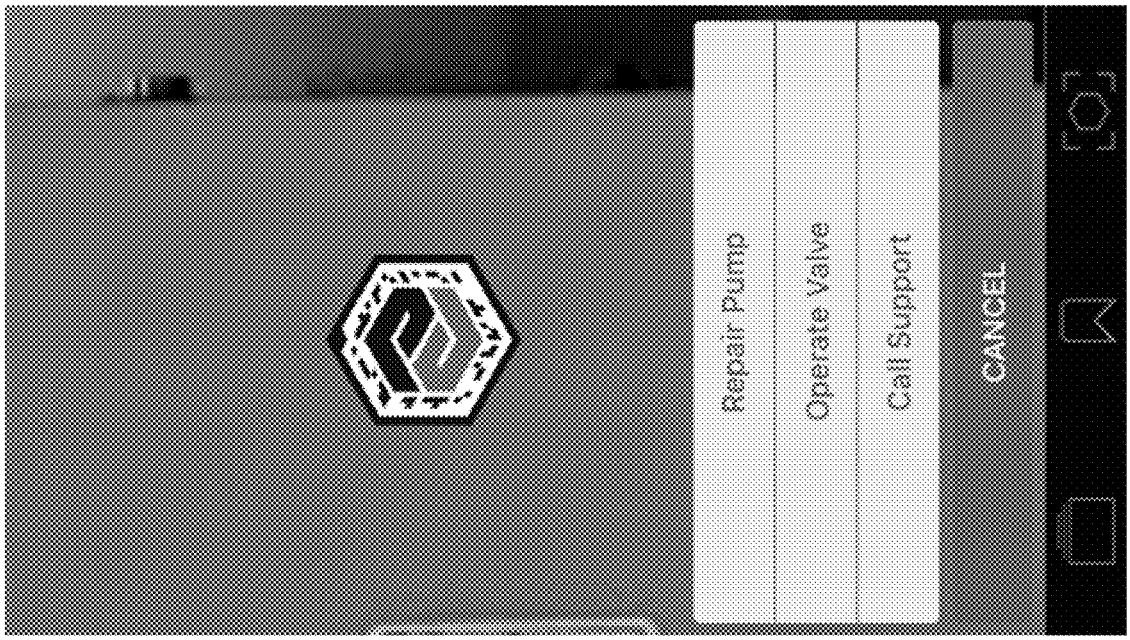
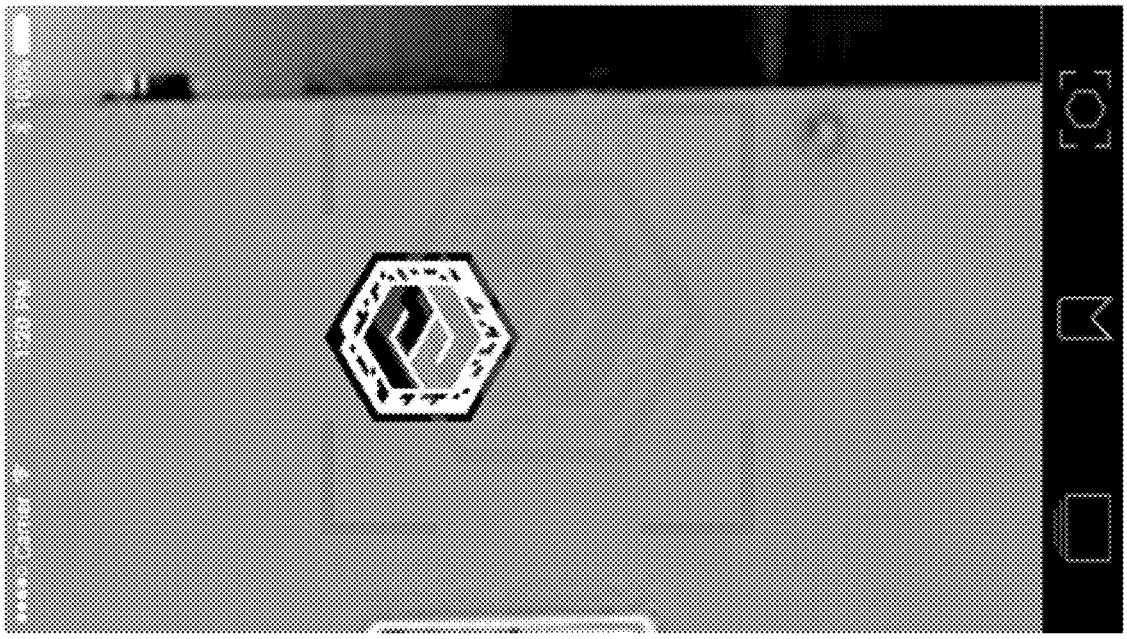


FIG. 6A



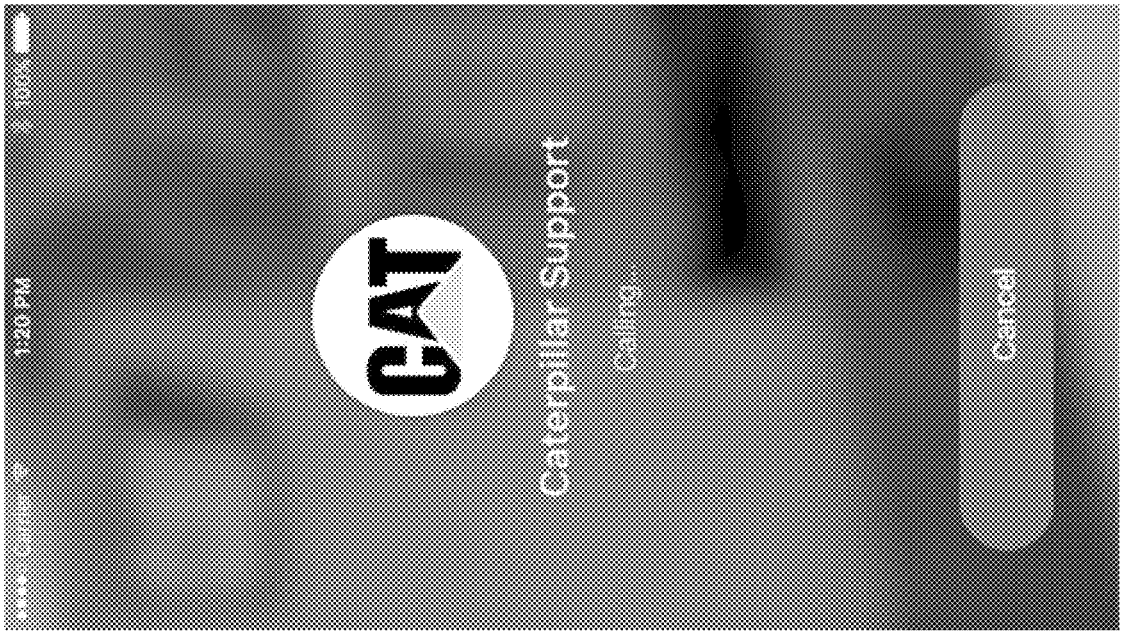

Caterpillar Support
Calling...
Cancel

FIG. 6C

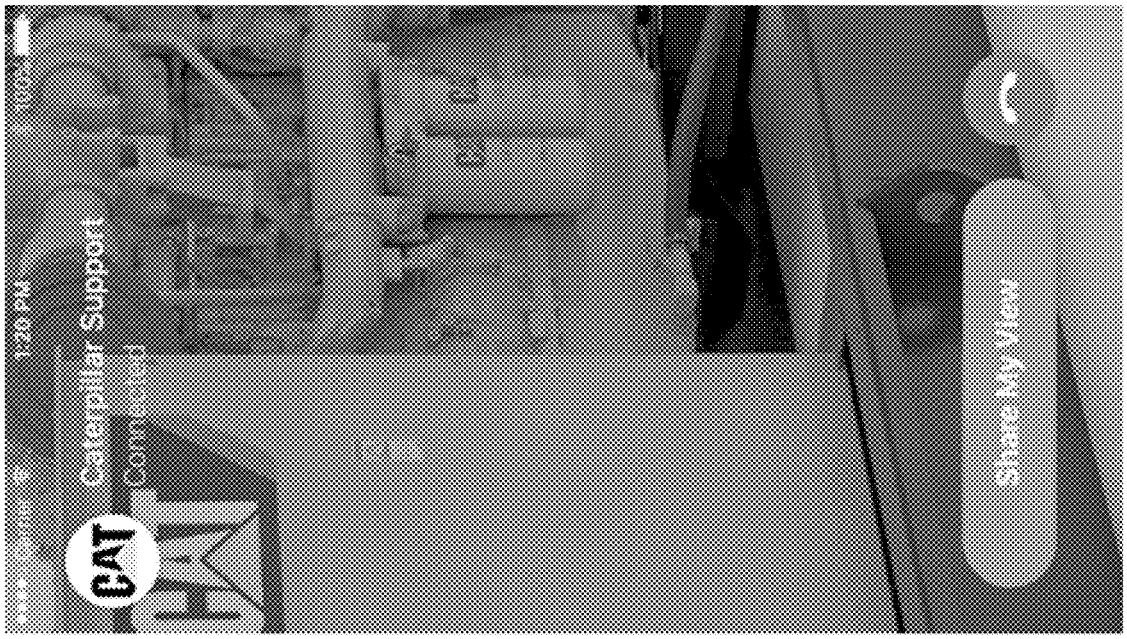
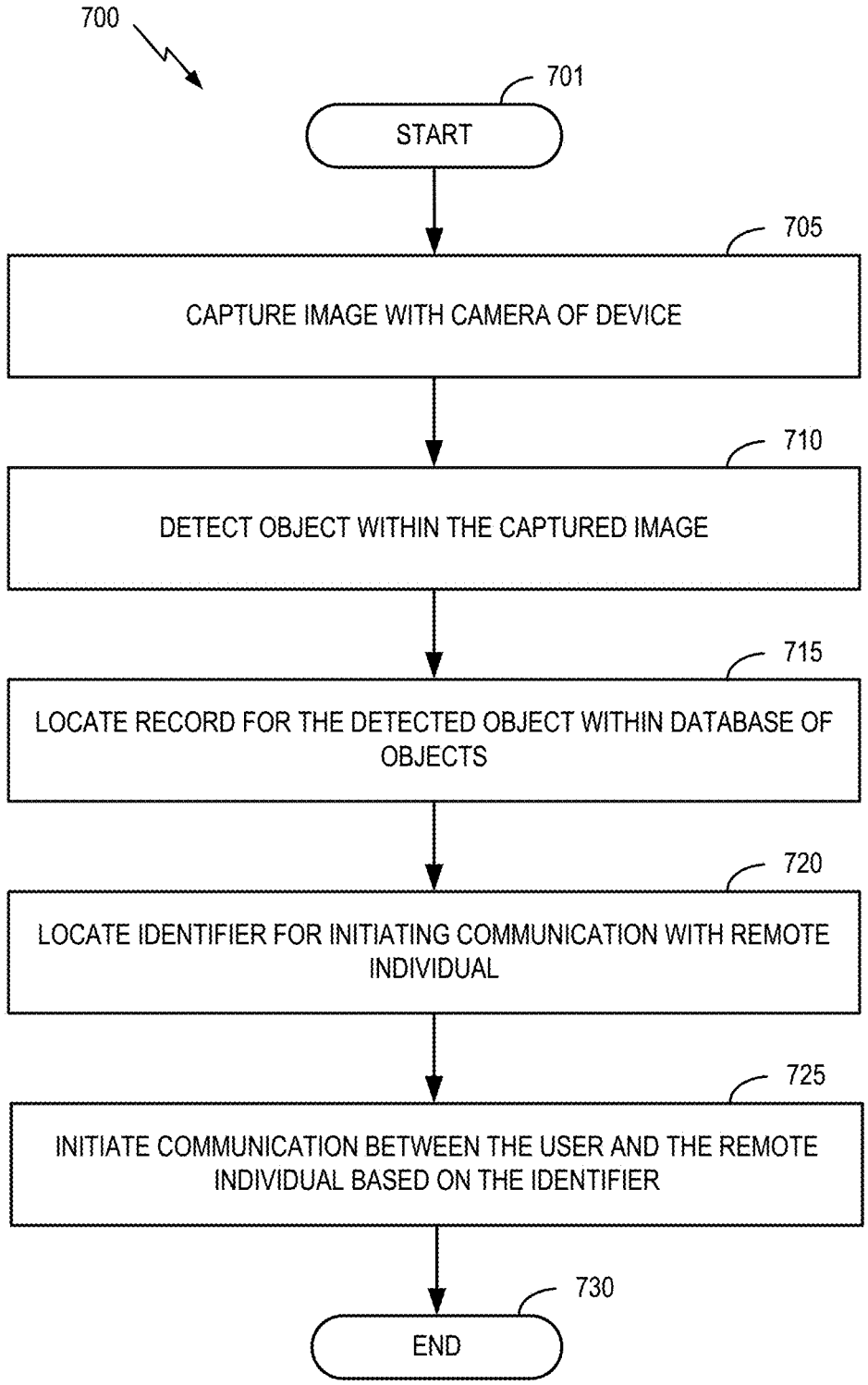

Caterpillar Support
Connected
Share My View

FIG. 6D



FIG. 6E

FIG. 7


700
701
START
705
CAPTURE IMAGE WITH CAMERA OF DEVICE
710
DETECT OBJECT WITHIN THE CAPTURED IMAGE
715
LOCATE RECORD FOR THE DETECTED OBJECT WITHIN DATABASE OF OBJECTS
720
LOCATE IDENTIFIER FOR INITIATING COMMUNICATION WITH REMOTE INDIVIDUAL
725
INITIATE COMMUNICATION BETWEEN THE USER AND THE REMOTE INDIVIDUAL BASED ON THE IDENTIFIER
730
END

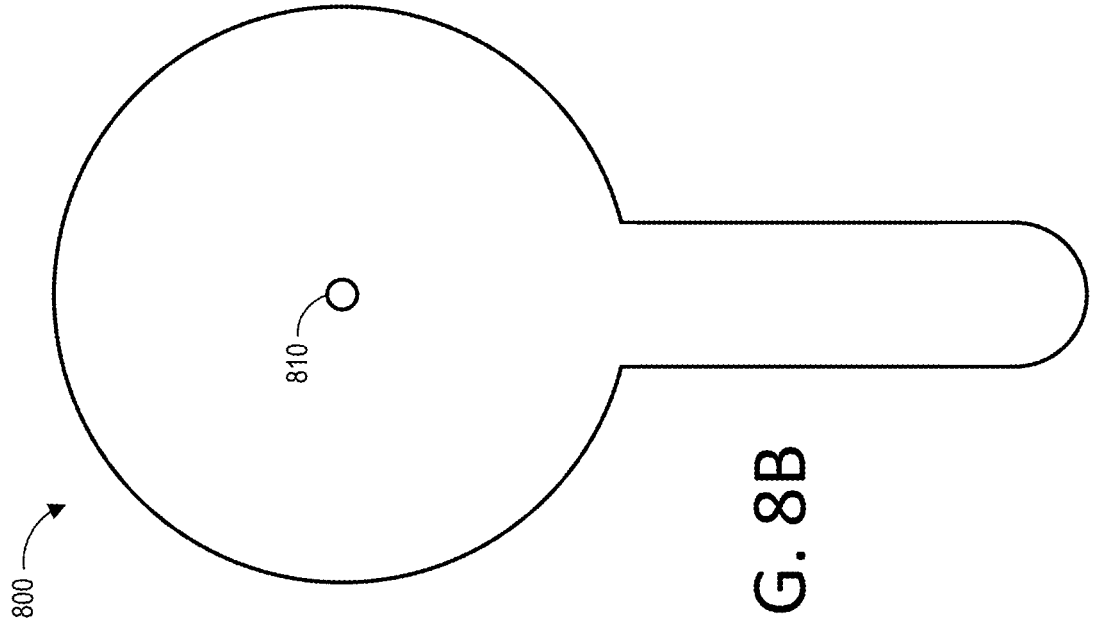


FIG. 8B

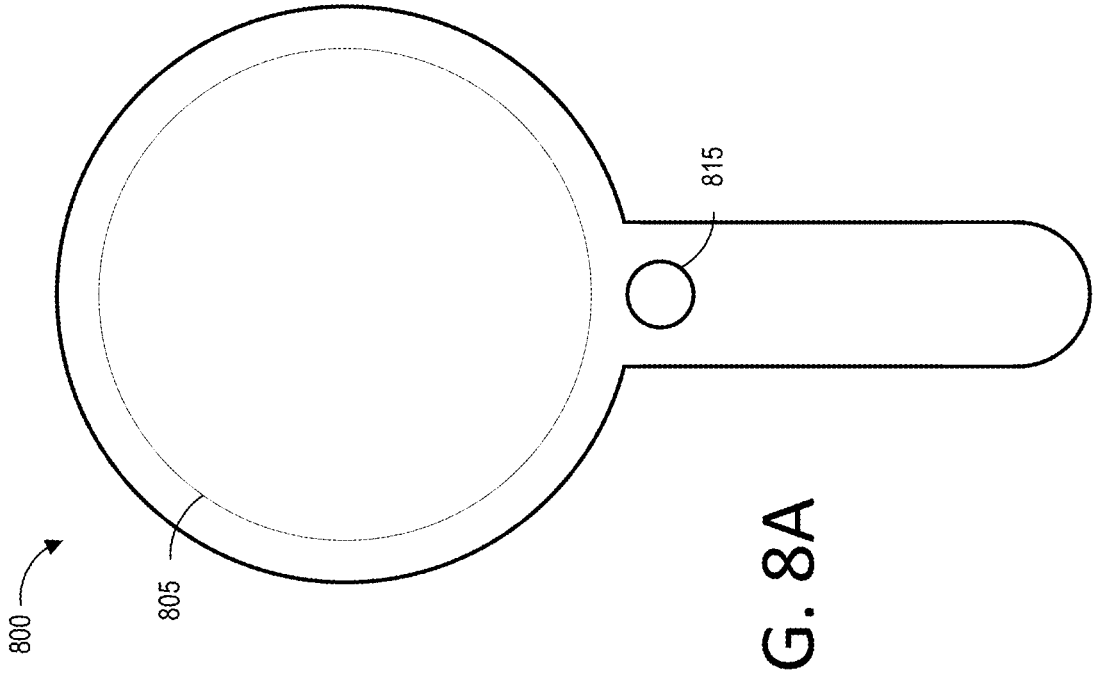


FIG. 8A

OBJECT INITIATED COMMUNICATION

RELATED APPLICATIONS

[0001] The subject matter of this application is related to International (PCT) Application No. PCT/US18/35193, filed on 2018 May 30 and to U.S. Provisional Application No. 62/512,696, filed on 2017 May 30, which applications are hereby incorporated by reference in their entireties.

BACKGROUND OF THE INVENTION

[0002] A user may require and/or desire technical or other assistance in interacting with physical object(s) in his/her environment. Traditionally, a user would solve technical issues by consulting written documentation describing the typical issues that may arise from malfunction of the object along with the typical indicators associated with the object/malfunction and solutions to those issues. However, there may be limitations in consulting written documentation to solve these types of issues. For example, rare issues may not be included in the written documentation, leaving the user unable to diagnose and/or solve these issues using the written documentation alone. Additionally, the written documentation may be outdated, providing solutions to the issues which are no longer the most up-to-date solutions to the issues. Further, when the user does not have a sophisticated understanding of the object and its potential issues, the user may not be able to fully understand the written documentation due to assumptions in knowledge in the documentation and/or the use of technical jargon.

[0003] The user may also be able to call a third-party user (also referred to hereinafter as a remote user or remote individual) and/or an automated system for assistance in interacting with the object. A phone call to a third-party may solve some of the problems associated with written documentation by providing recently updated diagnosis/solution information related to the object. However, the user may not have readily available access to the contact information for the third-party. That is, the user may not be able to determine, from inspecting the object alone, which third-party to contact to obtain assistance with the object. Thus, the user may be required to determine which third-party is available and able to assist the user in interacting with the object and find the contact information for the third-party. These additional steps may require a non-trivial amount of time for the user.

[0004] Moreover, a voice call may have certain limitations in the type of assistance available to the user. For example, the local user may not be able to accurately describe the object and/or the issues that the user is experiencing with the object. The local user may also not understand the terminology used by the remote user in describing actions to take to address the issues. Since the remote user is unable to see the object when assisting via a voice phone call, it may be difficult for the user and the remote user to efficiently communicate in a manner which is understandable by both parties.

SUMMARY OF THE INVENTION

[0005] In one aspect, there is provided a method for initiating communication between a user of a device and a remote individual or party for assisting the user in interacting with an object. The method may involve capturing an image with a camera of the device; detecting the object within the captured image; locating a record for the detected object within a database of objects; locating an identifier for initiating the communication with the remote individual, wherein the identifier is associated with the record of the detected object and identifies an address for initiating the communication; and initiating the communication between the user and the remote individual based on the identifier.

[0006] In another aspect, there is provided a method for initiating a communication session. The method may involve capturing a first image of at least a portion of an object with a first device; accessing a first database having a plurality of records storing characteristic data for identifying objects based on captured images; locating a first matching record within the first database based on the first image; accessing a second database having a plurality of records storing identifiers for initiating communication sessions; using the first matching record within the first database to locate a second matching record in the second database; obtaining an identifier for initiating a communication session from the second matching record; and using the identifier for initiating a communication session to initiate a communication session between the first device and a second device.

[0007] In yet another aspect, there is provided a method for initiating a person-to-person communication session. The method may involve initiating an augmented reality (AR) video session using a camera and a display on a first device operated by a first user; identifying an object within a field of view of the camera; performing motion tracking using the identified object within the field of view of the camera during the AR video session; accessing a database having a plurality of records storing identifiers for initiating communication sessions; using the identification of the object to locate a matching record in the database; obtaining an identifier for initiating a communication session from the matching record; based on the identifier for initiating a communication session, presenting to the first user, during the AR video session, an option to initiate a person-to-person communication session; and in response to selection of the option by the user, using the identifier for initiating a communication session to initiate a person-to-person communication session between the first user operating the first device and a second user operating a second device.

BRIEF DESCRIPTION OF THE DRAWINGS

[0008] FIG. **1** illustrates an example shared AR system platform **100** that can be used as platforms to support various embodiments.

[0009] FIG. **2** illustrates an example computing device that can be used to support various systems and methods disclosed.

[0010] FIG. **3** illustrates an example operation of an AR-based remote collaboration system.

[0011] FIG. **4** illustrates a method for identifying an object and matching the detected object to a record stored in an object database.

[0012] FIG. **5** illustrates a method for initiating communication with a remote user based on object detection.

[0013] FIGS. **6**A to **6**E illustrate a local user interface displayed via a display of the electronic device in accordance with one embodiment.

[0014] FIG. **7** illustrates a method for initiating communication between a local user and a remote user based on the detection of an object in the local user's environment.

[0015] FIGS. **8**A and **8**B provide two views of an example shared augmented reality device.

DETAILED DESCRIPTION

[0016] In the following description, references are made to various embodiments in accordance with which the disclosed subject matter can be practiced. Some embodiments may be described using the expressions one/an/another embodiment or the like, multiple instances of which do not necessarily refer to the same embodiment. Particular features, structures or characteristics associated with such instances can be combined in any suitable manner in various embodiments unless otherwise noted.

[0017] Shared Augmented Reality System Platform Augmented reality systems may enable a more natural and efficient communication system between two parties. A shared augmented reality (shared AR) system implementation may support sharing images and/or video captured by the local user with a remote user or remote individual. At least one of the local user and the remote user may be able to add annotations (e.g., markings, notes, drawings, etc.) to certain objects within the environment captured within the images and/or video. These annotations and the shared images and/or video may improve the communication between the local user and the remote user by, for example, allowing the local user and the remote user to visually identify specific objects in the local user's environment. Some example shared AR systems are described in U.S. Patent Application Publications 2015/0125045 and 2016/0358383.

[0018] FIG. **1** illustrates an example shared AR system platform **100** that can be used as platforms to support various embodiments disclosed herein. In particular, FIG. **1** illustrates an example system for a particular example application that enables a remote user to explore a physical environment via live imagery from a camera that the local user holds or wears, such as a camera of a mobile device or a wearable computing device, which may be or include a networked, wearable camera. The remote user is able to interact with a model fused from images captured from the surroundings of the local user and create and add virtual annotations in it or transfer live imagery (e.g., of gestures) back. The example system is an example of a system on which various embodiments may be implemented. Further, the various method and portions thereof, may be performed on one or more of the various computing devices, on a cloud, network-based server, on a local or remote user's computing device, and various combinations thereof. The example system platform illustrated in FIG. **1** is just one of many possible system platforms on which the various AR embodiments herein may be implemented.

[0019] The system illustrated in FIG. **1** can support live mobile tele-collaboration. The system can include a tracking and modeling core, which enables the system to synthesize views of the environment and thus decouple the remote user's viewpoint from that of the local user, giving the remote user some control over the viewpoint, and to register virtual annotations to their real world referents. Note, however that this is a particular example application, and not all embodiments include virtual annotation, model viewing, or model navigation functionality.

[0020] This system is compatible with hardware devices and systems that are already ubiquitous (e.g., smartphones), but also scales to more advanced high-end systems, including augmented reality, mixed reality and virtual reality (VR) devices. Further, some embodiments are compatible with various types of displays for the local user, including eye-worn, head-worn and/or projector-based displays.

[0021] In the illustrated embodiment of the FIG. **1**, the local user may hold or wear a device that integrates a camera and a display system (e.g., hand-held tablet, mobile device, digital eyewear, or other hardware with a camera), which is used to both sense the environment and display visual/spatial feedback from the remote user correctly registered to the real world. In the case of a hand-held device, the handheld device acts as sort of a "magic lens" (i.e., showing the live camera feed and virtual annotations, when the embodiment includes AR). Since a collaboration system typically aids the user an actual task being performed rather than distracts from it, an interface which is simple and easy to comprehend is typically provided such as to facilitate an active user who may be looking at and working in multiple areas.

[0022] A device of the remote user may also be a mobile device, such as a handheld computer, a tablet, a smartphone, and the like. However, the device of the remote user may also or alternatively be any computing device, such as a personal computer, a wearable computing device, and the like. The remote user, in some embodiments, is presented with a view into the local user's environment, rendered from images obtained by the local user's camera. The remote user, in augmented reality embodiments, can place annotations that will be displayed to both users, correctly registered to their real-world referents from their respective points of view. Annotations may include point-based markers, more complex three-dimensional annotations, drawings, or live imagery, such as hand gestures.

[0023] In a simple embodiment, the remote user's viewpoint may be restricted to being identical to the local user's current camera view. In such embodiments, little image synthesis is needed. However, the remote user may be permitted be able to decouple a presented viewpoint and control the viewpoint independently, as far as supported by the available imagery of the environment. In such embodiments where the system allows for decoupled views, only the viewpoint is decoupled; the video is still synthesized and updated from live images to enable consistent communication.

[0024] FIG. **2** illustrates an example computing device that can be used to support various systems and methods disclosed. The computing device of FIG. **2** may be implemented as one or more of the computing devices of the local user's interface, core system, and remote user's interface as illustrated and described with regard to FIG. **1**.

[0025] In one embodiment, multiple such computer systems are utilized in a distributed network to implement multiple components in a transaction-based environment. An object-oriented, service-oriented, or other architecture may be used to implement such functions and communicate between the multiple systems and components. One example computing device in the form of a computer **210** may include a processing unit **202**, memory **204**, removable storage **212**, and non-removable storage **214**. Although the example computing device is illustrated and described as computer **210**, the computing device may be in different forms in different embodiments. For example, the computing device may instead be a smartphone, a tablet, smartwatch, another wearable computing device type, or other

computing device including the same, similar, fewer, or more elements than illustrated and described with regard to FIG. **2**. Devices such as smartphones, tablets, and smartwatches are generally collectively referred to as mobile devices. Further, although the various data storage elements are illustrated as part of the computer **210**, the storage may also or alternatively include cloud-based storage accessible via a network, such as the Internet.

[0026] Returning to the computer **210**, memory **204** may include volatile memory **206** and non-volatile memory **208**. Computer **210** may include or have access to a computing environment that includes a variety of computer-readable media, such as volatile memory **206** and non-volatile memory **208**, removable storage **212** and non-removable storage **214**. Computer storage includes random access memory (RAM), read only memory (ROM), erasable programmable read-only memory (EPROM) and electrically erasable programmable read-only memory (EEPROM), flash memory or other memory technologies, compact disc read-only memory (CD ROM), Digital Versatile Disks (DVD) or other optical disk storage, magnetic cassettes, magnetic tape, magnetic disk storage or other magnetic storage devices, or any other medium capable of storing computer-readable instructions.

[0027] Computer **210** may include or have access to a computing environment that includes input **216**, output **218**, and a communication connection **220**. The input **216** may include one or more of a touchscreen, touchpad, one or more cameras, mouse, keyboard, one or more device-specific buttons, one or more sensors integrated within or coupled via wired or wireless data connections to the computer **210**, and other input devices. The input **216** may further include inertial measurement unit (IMU), which may include an accelerometer and/or a gyroscope. As discussed below, the computer **210** may use the camera and/or the IMU to determine whether the computer has been put down (e.g., placed on a surface such as a table) or placed in the local user's pocket or a bag. The computer **210** may operate in a networked environment using a communication connection **220** to connect to one or more remote computers, such as database servers, web servers, and other computing device. An example remote computer may include a personal computer (PC), server, router, network PC, a peer device or other common network node, or the like. The communication connection **220** may be a network interface device such as one or both of an Ethernet card and a wireless card or circuit that may be connected to a network. The network may include one or more of a Local Area Network (LAN), a Wide Area Network (WAN), the Internet, and other networks. In some embodiments, the communication connection **220** may also or alternatively include a transceiver device, such as a Bluetooth device that enables the computer **210** to wirelessly receive data from and transmit data to other Bluetooth devices.

[0028] Computer-readable instructions stored on a computer-readable medium are executable by the processing unit **202** of the computer **210**. A hard drive (magnetic disk or solid state), CD-ROM, and RAM are some examples of articles including a non-transitory computer-readable medium. For example, various computer programs **225** or apps, such as one or more applications and modules implementing one or more of the methods illustrated and described herein or an app or application that executes on a mobile device or is accessible via a web browser, may be stored on a non-transitory computer-readable medium.

[0029] FIG. **3** illustrates an example operation of an AR-based remote collaboration system. A local user is located in physical location A, while a remote user is located in physical location B. FIG. **3** shows a physical view of the local user in location A in front of a car engine, identifying a particular element with his hand. FIG. **3** also shows a screen view of the remote user in location B onto the scene in physical location A, which (at this moment) shows more surrounding context than the more limited view on the screen of the user in location A. The local user's view is shown as an inset on the bottom left of the remote user's view as well as being projected onto the model as viewed by the remote user. The user in location B can browse (e.g. pan, tilt, zoom) this environment independently of the current pose of the image capture device in physical location A, and can set annotations, which are immediately visible to the user in physical location A in AR. In this case, the remote user in location B has virtually marked the element identified by the local user in location A in the screen view of the scene using a virtual circular marker that is also visible to the local user A.

[0030] The above-describe shared AR system platform supports an augmented shared visual space for live mobile remote collaboration on physical tasks. The user in physical location B can explore the scene independently of the current pose of the image capture device in physical location A and can communicate via spatial annotations that are immediately visible to user in in physical location A in AR. This system can operate on off-the-shelf hardware and uses real-time visual tracking and modeling, thus not requiring any preparation or instrumentation of the environment. It can create a synergy between video conferencing and remote scene exploration under a unique coherent interface.

[0031] Object Initiated Communication

[0032] A user of an electronic device (also referred to herein as a "local user") may require and/or desire assistance in interactive with object(s) in the local user's environment. Certain embodiments of this disclosure relate to systems and techniques for initiating communication between two electronic devices or between one electronic device and a remote party based on object detection performed by one of the devices. This assistance may take the form of a phone call to a help desk/automated system, a text-based chat session, a shared augmented reality communication session, etc. The user may be able to communicate with a third-party (e.g., a remote user and/or automated electronic system) to receive assistance with any issues the user may be experiencing with the object(s) to diagnose the source of the issues and/or provide instructions to the individual to aid in the interaction with the object. The object(s) and/or type of communication may vary depending on the user's circumstances. For example, the user may be performing maintenance on an engine of a vehicle and may encounter object(s) within the engine having issues that the user is unable to diagnose and/or address.

[0033] The described technology is not limited to a specific type of object for interaction with the user. That is, similar limitations may exist for user interaction with object (s) other than the engine of a vehicle. A non-exhaustive list of such objects include: assembly of furniture and/or electronic devices, performing maintenance on machinery, diagnosing and/or troubleshooting mechanical errors, etc.

[0034] Augmented reality systems can be programmed to automatically locate and identify objects from within the view of the images captured by a camera of an electronic device (e.g., a mobile phone, table computer, "virtual looking glass", etc.). There are a number of different methods for object detection.

[0035] FIG. **4** illustrates a method **400** for identifying an object and matching the detected object to a record stored in an object database. As used herein, an object may refer to an entire physical object or a portion of a physical object which is visually identifiable based on an image of the portion. In certain embodiments, the object can include a machine-readable code (such as a graphical marking), which may be printed on the object to facilitate identification of the object. One or more of the steps illustrated and described in connection with the method **400** may be omitted and/or performed in a different order to the described order without departing from this disclosure.

[0036] Method **400** may be performed by a processor, such as processor unit **202**. The method **400** begins at block **401**. At block **405**, one or more image(s) including at least a portion of an object are captured using a camera. At block **410**, the processor recognizes the object from the captured image(s). That is, the processor may recognize that the captured image includes an object which can be identified from an object database. In one implementation, this may include highlighting or otherwise visually identifying the recognized object. When more than one object is present in the image(s), each of the objects may be visually identified. The user may select, via a user interface, one or more of the objects to be matched with a record in the object database. Alternatively, the processor may identify an object without visually identifying the object via a display. In certain implementations, the processor may automatically select one of the objects to be matched with a record in the object database based on at least one of: the object's size, the object's position within the image, etc.

[0037] The identification of an object may also include identifying a plurality of objects to be matched with a record in the object database. For example, a plurality of objects may be related in some way such that they may be considered a set of objects that could match a single record in the object database. In one example, the user may wish to receive assistance with assembling a piece of furniture. However, prior to assembly, the furniture may include a plurality of different objects that are to be attached to one another in order to assemble the furniture. In this case, the processor may identify the plurality of objects as related so as to match the plurality of objects together with a single record stored in the object database.

[0038] At block **415**, the processor may extract visual features from the identified object. The visual features may be any aspect of the object that can be extracted based on the information in the image(s) captured by the camera. This may include: the two-dimensional (2D) and/or three-dimensional (3D) shape of the object, a graphical marking (e.g., machine-readable code) on the object which encodes data, image recognition (e.g., an image printed on the object or the image of the object itself), CAD-model based data (e.g., shapes and/or features of the object that may be matched to a CAD-model of the object), geo-location data (e.g., when the object is located at a unique location), etc.

[0039] In certain embodiments, the extracted visual features can be characteristic data that can be used to identify the object based on matching the characteristic data to a record of an object stored in an object database. The object database may be stored locally on the electronic device or may be stored on a remote server accessed over a network such as the Internet (also referred to as being stored in the "cloud"). The processor may extract one or more of the visual features in order to use more than one type of visual feature in matching the object to a record in the object database or to search more than one object database.

[0040] In certain embodiments, the extracted visual features can include a machine-readable code. In some embodiments, the machine-readable code can be capable of being algorithmically decoded without a database lookup, such as a traditional barcode or a QR code. In some embodiments, the machine-readable code may require a database lookup in order to be decoded, such as a Vuforia™ VuMark™, available from applicant, PTC Inc. The database can include records defining formats for identifying and decoding various types of machine-readable codes.

[0041] At block **420**, the processor can attempt to match the extracted visual features to records of objects stored in the object database. This may include the processor accessing the object database and locating a record in the database that matches the extracted visual features. In response to the visual features matching one of the records stored in the object database, the processor may highlight the object on the display or otherwise indicate to the user that an object has been detected and identified.

[0042] In some implementations, the extracted visual features may not be sufficient for matching a record within the database or may only broadly identify the object but not have specific information about the object. For example, when matching a record in the object database based on the shape of the object, the model of the object may be identified. However, the identified object model may have been manufactured over a number of years in different batches and using different methods and/or materials, some of which may affect how the user interacts with the object. In order to gather more specific information, the user may be able to scan a further graphical marking or machine-readable code printed on the object to obtain more specific information.

[0043] In one embodiment, the electronic device may prompt the user to scan the graphical marking or machine-readable code printed on the object. This may include displaying a location on the object where the graphical marking may be found. For example, if the object is lying on the ground such that the graphical marking is obscured from the user's view, the user may not be aware that a graphical marking is present on the object. Thus, the display of the graphical marking's location may prompt the user to rotate the object and/or move the electronic device such that the graphical marking may be found. The graphical marking may encode information about the object, such as the object's serial number, manufacture date, batch number, etc. Based on this additional information, the object database may be able to access information that can assist the user in interacting with the object. For example, the manufacture date and/or batch number of the object may indicate that the object was manufactured during a batch that was later identified as having manufacturing errors, which may indicate that the object is faulty and required replacement. After decoding the graphical marking, the electronic device may send the decoded information to the remote user's computer

system. Thus, the remote user may be able to use this information in assisting the user.

[0044] After an object has been detected by the user's electronic device, the user may be able to interact with the object in a number of ways. In one implementation, the object may be highlighted on a display associated with the electronic device and the user may be able to select the object to bring up a menu of options to interaction with the object within an AR environment. The menu may be able to display information associated with the object, such as the name, model, part number, owner, serial number, or a unique identifier, etc. The menu may also include an option to display stored history associated with the object, such as past communications sessions for help or service, operating specifications/parameters, and other real-time diagnostic data. The historical data may be stored on a remote server in the cloud.

[0045] The menu may also display a list of options for the user to obtain assistance for the identified object. These options may include option for: a phone call with a corresponding phone number, a shared AR session, a text-based chat session, or another type of person-to-person communication. The assistance options may also include options for obtaining assistance instructions and/or troubleshooting techniques for interacting with the object. The user may follow the steps associated with the assistance instructions and/or troubleshooting techniques prior to initiating a communication session with a remote individual. Another option which may be displayed by the menu is link(s) to any previously recorded communication sessions relevant to the identified object that can be replayed. These recorded sessions may be stored locally and/or in the cloud.

[0046] FIG. **5** illustrates a method **500** for initiating communication with a remote user based on object detection. Method **500** may be performed by a processor, such as processor unit **202**. The method **500** begins at block **501**. At block **505**, the processor detects an object within a captured image. This may be performed by the local user's electronic device. Once the object has been detected by the electronic device, the device can provide an indication to the user that the object has been detected, such as by presenting an augmented reality overlay corresponding to the object as shown on a display of the device. The device can also present to the user a number of options for interacting with the detected object, where those options can include initiating a live communication session with a remote user or party. At block **510**, the processor may locate contact information (also referred to as an identifier for initiating a communication session with a remote user) by searching a second database (hereinafter also referred to as a communication database). The identifier for initiating a communication session may identify an address for initiating the communication session. For example, the address may be an IP address, a phone number, etc. Similar to the object database, the communication database may be stored locally on the electronic device or may be stored in the cloud. Once the identifier is located, the device may either automatically initiate communication with the remote individual or may prompt the user to request input regarding whether the user would like to initiate the communication.

[0047] A key for searching the communication database may be retrieved from the record associated with the identified object stored in the object database. The key may be identifier of the object (e.g., a unique object identifier) that can be used to locate the corresponding record in the communication database. Thus, the processor may access the communication database using information from the matching record retrieved from the object database. Once the record for the object has been located in the communication database, the electronic device may obtain an identifier for initiation a communication session from the record. At block **515**, the electronic device may then initiate communication with the remote individual based on the retrieved identifier. The method ends at block **520**.

[0048] The communication session may be any type of real-time communication between the user and the remote individual. For example, the user may initiate a shared AR session with the remote individual in which a live video stream or sequence of images captured by a camera of the user's electronic device is transmitted to a remote user's interface (e.g., a remote computer system or other device capable of displaying the video stream or sequence of images). As discussed in detail above, one or more of the local user and the remote user may communicate verbally and/or by annotating objects within the environment which are displayed by either the local user's electronic device or the remote user's interface. The shared AR session may involve motion tracking so that the core system is able to track the environment as the electronic device is moved by the local user. In certain implementations, the motion tracking is performed using the object identified by the processor as a known point about which motion may be tracked.

[0049] In other implementations, the communication session may be a phone call, a VOIP call, a text-based chat session, etc. These types of communication sessions may also be integrated to an AR session which is only displayed to the local user (e.g., the video stream is not transmitted to the remote user). In certain situations, the user may not be connected to a data network that has sufficient bandwidth to transmit data, and thus, may not be able to establish an AR session with the remote user. Thus, the user may establish a local AR session that is not viewable by the remote user. The local AR session may be established prior to initiating the communication session and/or prior to identifying an object.

[0050] As discussed above, the user may follow the steps associated with the assistance instructions and/or troubleshooting techniques prior to initiating a communication session. Information relating to the user's progress through the steps may be transmitted to the remote user upon establishing the communication session. This may aid the remote user in determining what actions the local user has already taken prior to initiating the communication session, thereby enabling the remote individual to understand the specific issues facing the local user without requiring the remote user go through all of these steps again with the local user. In one implementation, the remote user may follow a predetermined set of instructions to guide the local user through a troubleshooting process. In response to receiving the steps that the user has already performed, the remote user's computer may skip these steps, resulting in a more efficient use of time for both the local and remote users.

[0051] The remote user may also receive other information from the local user's electronic device in response to the initiation of the communication session that is helpful in assisting the local user. This information may include information associated with the object, such as the name, model, part number, owner, serial number, or other unique ID, etc.; and historical information associated with the object, such as

past communications sessions for help or service, operating specifications/parameters, and other real-time diagnostic data.

[0052] FIGS. **6**A to **6**E illustrate a local user interface displayed via a display of the electronic device in accordance with one embodiment. FIG. **6**A illustrates the user preparing to line up a graphical marking on an object with the camera of the electronic device. After the graphical marking has been scanned, a portion of the graphical marking may be highlighted, as shown in FIG. **6**B, to provide the user with a visual confirmation that the graphical marking has been scanned. The display may also provide the user with a list of options for interacting with the object via AR. The option to initiate a support call, based on the detected object, is also shown in FIG. **6**B.

[0053] FIG. **6**C illustrates a display after the user has selected the option to call support. Once the call to support has been established, as shown in FIG. **6**D, the user may be given the option to share the user's view, thereby transitioning the call to an AR communication session. FIG. **6**E illustrates an example annotation that may be anchored to an object in the environment. The annotation may be added by either the local user or the remote user and displayed on the electronic device.

[0054] In the illustrations of FIG. **6**A to **6**E, the graphical marking shown is a VuMark™ machine-readable code. VuMark™ machine-readable codes and their associated programs, applications and systems, were released in August 2016 under the Vuforia™ brand by the Applicant, PTC Inc. More generally, a user-designed, machine-readable code, also referred to herein as a target, is printed or otherwise disposed on an object in order to support one or both of machine-reading of encoded data from the code and machine-determination of the position and orientation of the object. The user-designed, machine-readable code can deliver a unique augmented reality experience on any object and/or enable the machine-reading of encoded data while allowing design freedom for a custom look and feel of the code itself. The form and visual representation of the user-designed code can be highly flexible, with the ability to support visually appealing graphics, logos and designs. The user-designed code overcomes the limitations of existing bar code solutions that do not support augmented reality experiences and can detract from a product's appearance. The user-designed code supports encoding varying amounts and kinds of data such as URLs or a product serial numbers. The user-designed code can be designed to distinguish any number of objects or products, since the range of the encoded data can be a user-determined design parameter. A single template for a user-designed code can be used to encode multiple instances of the code where each instance encodes different data. The user-designed code can be detected and tracked, for example, with a mobile device with a camera. Position and orientation within a camera image can be determined automatically, so the user-designed code can be used for tracking an object, such as within an augmented reality application.

[0055] Although in the illustrations of FIGS. **6**A to **6**E, the graphical marking is user-designed, machine readable code, any appropriate machine-readable code can be used to identify the object or provide a link to a database to support initiation of communication. Certain machine-readable codes can also support object identification and/or object tracking.

[0056] FIG. **7** illustrates a method **700** operable by an electronic device, or component(s) thereof, for initiating communication between a local user and a remote user based on the detection of an object in the local user's environment. For example, the steps of method **700** illustrated in FIG. **7** may be performed by a processor, e.g., processing unit **202**, of the electronic device. For convenience, the method **700** is described as performed by a processor of a local user's electronic device.

[0057] The method **700** begins at block **701**. At block **705**, the processor captures an image with a camera of the device. At block **710**, the processor detects an object within the captured image. At block **715**, the processor locates a record for the detected object within a database of objects. At block **720**, the processor locates an identifier for initiating communication with a remote individual. The identifier may be associated with the record of the detected object. The identifier may also identify an address (e.g., an IP address) for initiating the communication. At block **725**, the processor initiates the communication between the user and the remote individual based on the identifier. The method **700** ends at block **730**.

[0058] Shared AR Device

[0059] Aspects of this disclosure may be performed using a number of different electronic devices. One such electronic device that can be used by a local user for a shared augmented reality session is a shared augmented reality device that can be embodied in the form of a "virtual looking glass". FIGS. **8**A and **8**B provide two views of an example shared augmented reality device. Specifically, FIG. **8**A illustrates a front view and FIG. **8**B illustrates a rear view of the electronic device **800** of this embodiment.

[0060] Getting assistance from a remote individual may present certain challenges, especially for elderly persons, disabled persons, and/or children. Thus, there is a need for a device and a system that makes requesting assistance as easy as possible. The device **800** illustrated in FIGS. **8**A and **8**B has the form factor of a large looking glass. However, in place of the mirror in a looking glass, this embodiment of the electronic device **800** comprises a circular screen **805** and a back facing camera **810** mounted in the center of the back side of the screen. The device **800** may further include a processing unit (not illustrated), a communication unit (not illustrated) and an audio system (not illustrated) to enable the device **800** to perform aspects of this disclosure related to object initiated communication. For example, the device **800** may be configured to trigger a remote assistance session with a helper (e.g., a remote individual) that is preconfigured in a database. The remote helper may be able to see through the camera **810**, provide instructions using his/her voice and draw augmented reality annotations that appear to stick to objects within the local user's environment when looking at the objects through the "virtual looking glass" **800**. In one exemplary application, the local user may be an elderly individual who frequently has trouble with his/her medication. The device **800** may be able to recognize when the local user requires assistance with his/her medication (e.g., by recognizing the medication container) and can initiate a help session with an individual designated to assist the local user with the medication (e.g., a family member).

[0061] Remote help sessions can be run on an existing mobile devices (e.g., on smart phones and tablets). Unfortunately, the cameras of these devices are typically placed off center with respect to the main body of the device

(sometimes even on a corner of the device). This placement of the camera can disturbs the "look through" experience of the device. For example, objects closer to the camera (such as a user's hand when interacting with an object in the environment) may not be displayed by the device's display in a position that directly corresponds to the user's hand since the camera is off-center. This can lead to situations where the user may experience disorientation (especially for elderly persons and/or children who may have a hard time determining how to point the camera at a certain area within the environment when they are close to a desired object). Another limitation of traditional electronic devices is that holding the device with one hand may be difficult especially for larger factor devices such as a tablet.

[0062] A remote communication help session may require at least two devices. The first device **800**, used by a local user (e.g., the helpee), may have the form factor of a looking glass and may include a back facing camera **810** located approximately in the center of the looking glass, a circular screen **805** which may be configured to accept touch input, audio input and output devices (not illustrated), a processor unit (not illustrated), one or two buttons **815**, a communication interface (not illustrated) configured to connect with another electronic device via a network (such as the Internet). Although the screen **805** is described an illustrated as being substantially circular, this disclosure is not so limited and the device **800** and/or the screen **805** may have a different shape such as a rectangular or square shape in other embodiments.

[0063] The device **800** may also be configured to access a database, which may be stored locally on a memory (not illustrated) or remotely in the cloud. The database can be configured to store object descriptions and associated identifiers (e.g., IP address) for initiating a remote help session with a remote individual.

[0064] The second device (not illustrated) can be used by a remote individual, who may be associated with at least one of the identifiers stored in the database. The second device may be embodied as, for example, a desktop computer, a mobile device (e.g., a mobile phone), a tablet, a head mounted display, etc. An exemplary embodiment of the second device is shown in FIG. **1** and discussed in detail above. The second device may further include a screen configured to accept touch input, a processing unit, a communication interface configured to be connected to the local user via the network, and audio input and output devices.

[0065] In one implementation, the local user can initiate a call to the remote individual. For example, when the local user requires help, the local user may press one of the buttons **815** of the device **800** and/or point the camera **810** of the device **800** at an object. The device **800** may then recognize the object and initiate communication with the remote individual via at least one of the object initiated communication methods described herein. For example, the device **800** may perform object recognition and retrieve an identifier from a preconfigured database. In another embodiment, the local user may say the name of the remote user to initiate the communication session. In this embodiment, a processor of the device **800** or a remote processor may perform voice recognition to identify the spoken name with the identifier for contacting the remote individual. This process may include matching the spoken name with a record stored in a database.

[0066] During a communication session between the local user and the remote individual, the remote individual is able to provide help by talking with the remote user over an audio link and by drawing annotations on objects within the environment captured by the camera **810**. In addition to that the local user is able to hold the device **800** in one hand and reach into the environment while looking through the "virtual looking glass". In this way the remote individual is able to see what the local user is doing and provide assistance while the user is interacting with an object. For example, the remote user can provide instructions to the local user such as: "Turn the knob a bit further, yes, just that much. Now stop!" The specific design of the camera **810** (positioned at the center of the back side of the screen **805** and having a focal length that provides a natural "look through" feeling in an arm reach distance) makes the experience natural and easy to use even for elderly persons of children. The help session can be ended by pressing the button **815** on the handle of the device **800**.

[0067] The specific form factor of the "virtual looking glass" device may have certain advantages over other form factors. For example, the device **800** may have an improved ease of use, allowing the local user to point the device **800** at an object, press a button, and/or say "Call . . . " in order to receive assistance immediately. The specific form factor allows the local user to hold the device **800** with one hand and use the other hand to reach into the view and interact with an object. The centrally positioned camera **810** generates a true "looking through feeling" and does not generate any confusion while reaching into the view. This may be especially important when getting help on manual tasks that need continuous feedback.

[0068] User Interface Sharing Between Multiple Local Devices

[0069] In a remote assistance scenario, the local user (e.g., the individual receiving assistance from a remote user) may have multiple tasks to perform. These tasks may include administrative tasks such as receiving his work order, reading documentation, calling for assistance, talking to a remote individual providing the assistance and receiving instructions while performing the manual repair work. While the use of a head mounted display (HMD) may be desirable for use when performing physical repair work (e.g., enabling the local user to go hands free for a given task), an HMD may not provide a convenient interface to wear all the time or to use while performing certain tasks, such as administrative tasks. For these tasks, a handheld mobile device may be more suited to the required tasks. Thus, aspects of this disclosure also relate to the use of a combination of a handheld device and a head worn device (such as an HMD) that enables the user to share a user interface (UI) across these devices, thereby streamlining the user experience.

[0070] Remote assistance systems may use a mobile phone or tablet to share the back-facing camera view of the local user with a remote individual. In certain embodiments, the remote individual is able to draw annotations directly on the camera view provided by the back-facing camera and share those annotations back to the local user. In these embodiments, the user experience may be acceptable as long as the user is doing the administrative tasks (e.g., receiving a work order, accepting the work order, reading documentation, calling for assistance, etc.). However, once the local user is connected to the remote individual in order to receive instructions while simultaneously performing manual tasks,

the local user may be required to split time between performing the required tasks and interacting with the remote individual. This may require the local individual to put away the handheld electronic device while performing the manual tasks and to stop doing the manual tasks in order to interact with the electronic device for communication or other tasks which require the device.

[0071] In certain embodiments, a remote assistance system includes an HMD worn by the local user. In these embodiments, the local user experience may be improved during the remote communication session via, for example, freeing the user's hand for interacting with objects in the environment. The local user can thus use his/her hands to do the required manual tasks while receiving vocal instructions from the remote individual. The local user is also able to see what the remote individual is doing (e.g., adding annotations or other instructions) and is able to immediately respond to the feedback from the remote individual as soon as the remote individual receives the feedback. However, performing administrative tasks on an HMD may be tedious and inconvenient due to the restricted user interface associated with HMDs. For example, the lack of a physical keyboard and/or touch screen makes the input of text difficult when using an HMD.

[0072] Although aspects of this disclosure have been described in connection with a remote assistance setting, this disclosure is not limited thereto and may also relate to embodiments where the local user is not in communication with a remote individual. For example, the local user may be required to perform manual tasks while receiving information about the local environment. The tasks can relate to service and maintenance of an object/machinery, where the local user receives step-by-step instructions from a database.

[0073] Embodiments relating to sharing a user interface may be applied by a system that comprises at least two mobile devices (e.g., a HMD and a handheld device such as a mobile phone or tablet). The handheld device may include a processing unit, a communication interface that allows the device to communicate with a second device. In certain embodiments, the communication with the second device may include communicating with a remote individual over the Internet. The handheld device may further include a touch screen configured to receive touch input from the local user. The HMD may include a front facing camera, a processing unit, a communication interface, a display, an audio input device (e.g., a microphone) and an audio output device (e.g., speakers or earphones).

[0074] The camera of the HMD may be configured to capture images of the field of view of the user. The processing unit may be configured to perform tracking and in some cases 3D object recognition and tracking of the environment based on the images captured by the camera. The communication interface may be configured to communicate with the first device (e.g., via Wi-Fi, Bluetooth, etc.). The display may be a stereoscopic display configured to generate 3D image for the user. The display may also provide annotations which can be "anchored" to objects within the environment.

[0075] One application of the above-described system is a repair scenario. The user may perform repair of an object using both the handheld device and the HMD. For example, the user may use the handheld mobile device to run a specific application (e.g., a maintenance and repair application) on the handheld device. This application may allow the user to receive a work order (e.g., using the touch screen). The work order may be bundled with "AR experiences" (e.g., a set of computer-executable instructions which can be run on the HMD) for each repair sequence. For example, an AR experience may include tracking a target as well as providing predefined functionality to the user (e.g., step-by-step instructions may be provided to the user via the display and/or the audio output device).

[0076] The user may choose to read instructions and/or a manual of the object for repair using the touch screen of the handheld device. The user may then select and start a specific repair sequence using the touch screen. In response to this selection, the application run on the handheld device may send the associated "AR experience" to the second device (HMD). Thereafter, the user can put on the HMD and run the associated "AR experience". The HMD may include a specific application which can be run on the HMD and is configured to receive the "AR experience" from the handheld device for execution.

[0077] This application on the HMD can receive the specific AR experience and execute the same. Part of the "AR experience" may include the application running on the HMD starting the camera and an AR engine to detect and track objects within the local user's environment. The object tracked by the HMD may be defined within the received "AR experience". Once one of the target(s) is detected, the application may display certain content (e.g., annotations) over the target.

[0078] The application run on the HMD may also be configured to respond to certain voice commands received from the user. For example, the user may be able to say certain keywords, such as "next" or "repeat" in order to navigate the "AR experience". Alternatively, the user may be able to navigate the "AR experience" via gesture input detected using the camera of the HMD or via motion detection (e.g., the user may be able to move his/her head as input to the HMD). Accordingly, the local user can move through the "AR experience" while being hands free, enabling the user to perform manual tasks while simultaneously receiving instructions from the "AR experience" of the HMD.

[0079] Once the AR experience has completed execution, the application run on the HMD may signal back to the handheld device that the "AR experience" has been completed. Thereafter, the first application running on the handheld device is informed of the status of the "AR experience" and may then continue displaying information and receiving input for communicating with the local user.

[0080] In certain implementations, the second device (e.g., the HMD), is configured to function as a slave to a master application run on the handheld device. However, in other implementations, the HMD may be configured to run standalone software that communicates with the handheld device and can be run independent from the handheld device. In certain implementations, while the AR session is running, the handheld device may be configured to perform certain tasks, such as communicate with a backend system to fetch and receive additional data over a network (e.g., a cellular network, the Internet, etc.), provide computational services to the HMD to offload a portion of the processing from the HMD to the handheld device (e.g., performing object recognition on images received by an camera of the HMD) to balance the computational load between the HMD and the handheld device.

[0081] Another application is a remote assistance communication situation. In this embodiment, the user may interact with the handheld device to run an application (e.g., a remote assistance application) on the handheld device. This application may enable the user to perform one or more of the following actions: select a remote individual from a phone book stored on the handheld device or in the cloud, request a help session from the selected remote individual, and receive a call from the remote individual. Once a call with the remote individual has been initiated, the application running on the handheld device may send an "AR experience" to the HMD which is configured to run a remote assistance session with the remote individual.

[0082] The user may then put on the HMD, which may have an application running thereon so as to receive the "AR experience" from the handheld device and execute the "AR experience". The received "AR experience" may include: starting the camera on the HMD, starting an AR engine to track and reconstruct the environment based on image received from the camera, sharing the video and audio feed received by the HMD with the remote individual, and sharing a 3D reconstruction of the environment with the remote individual. The "AR experience" may further include receiving audio and annotations from the remote individual, and displaying the annotations and playing audio via the speakers of the HMD anchored to objects in the environment tracked by the HMD.

[0083] Once the AR experience has completed executing (e.g., the communication session with the remote individual has ended), the application running on the HMD may signal back to the handheld device that the "AR experience" has completed. Accordingly, the application running on the handheld device may be informed of the status of the "AR experience". The handheld device may then be configured to receive further input from the local user (e.g., to perform other administrative tasks).

[0084] Low Power Mode and Suspending Tracking During an AR Session

[0085] Vision based AR applications may consume a large portion of the available computing resources in order to track the device within the environment and to recognize new objects that come into sight. For example, it may be required for the AR application to know its position within the environment to correctly anchor annotations or other graphical elements to objects within the environment. However, there are situations where the local user temporarily puts away the electronic device in which the tracking and recognizing of objects is no longer necessary. For example, in a remote assistance communication application, the local user may follow certain instructions from a remote user and/or from a set of predefined instructions. In this case, the local user may be required to perform a manual task that required both of the user's hands, preventing the user from holding onto the electronic device.

[0086] Accordingly, if the electronic device is able to detect situations in which the user is not actively using the device for tracking and/or recognition of objects (e.g., viewing the environment through the electronic device), the computing resources, and thus battery power, of the electronic device can be conserved by turning these algorithms off. The electronic device may be able to detect these situations by analyzing the images received from the camera and an IMU to switch off camera input and computer vision tasks. The IMU input may be periodically checked to determine when the full computer vision algorithms are required to be reactivated (e.g., when the signals output from the IMU are indicative of the device being picked up by the user or moved around).

[0087] Many AR applications are vision based and thus use the camera of the electronic device for various computer vision algorithms such as device tracking (determining the position of the device within the environment), object detection and tracking (determining the position of known objects relative to the device), and/or 3D reconstruction (creating a digital 3D representation of the environment). All those algorithms require a relatively large amount of computing power compared to other algorithms run within the AR application. As discussed above, there may be situations in which a user does not require the computer vision algorithms to be run, e.g., when the user puts away the device to perform some other task.

[0088] In one implementation, in a remote assistance communication situation, the remote individual may provide instructions to the local user to perform a manual task. In this case, the local user may put away the device, perform the task with one or more of the user's hands and later pick up the device to show the results of the task to the remote expert. In another implementation, for example, when the AR application is providing step-by-step instructions, the user may be required to perform s manual task based on the instructions. When the user is not using an HMD, the local user may be required to put the device away and pick the electronic device back up after the task is performed.

[0089] While the device is put away, there may not be any relevant computer vision algorithms for the electronic device to perform. When the device is placed on a flat surface or in the user's pocket, there may not be any visual information on which to perform the computer vision algorithms. Thus, all the known objects which were previously being tracked may have disappeared from the view of the camera (e.g., since the camera is either pointing down onto the surface where the device is lying now or up seeing the sky or the ceiling). If the computer vision algorithms are not turned off in these cases, the electronic device may continuously attempt to locate known objects in the received images, thereby expending resources on computer vision tasks that are not beneficial.

[0090] Accordingly, in at least one embodiment, a combination of the images received from the camera and the signal received from the IMU are used to determine if the device has been put away or is not currently being used for AR. The electronic device may determine that the device is not currently being used in response to at least one of: i) the camera image being nearly completely black (e.g., the camera is covered) and the IMU indicating that the device is not experiencing changes in acceleration (e.g., changes to the signal are within a threshold of IMU noise and drift); ii) camera image does not change significantly from the previous image (e.g., the sum of absolute differences (SAD) between the two images is less than a threshold value) and the acceleration values from the IMU are within a threshold of zero (as discussed above).

[0091] In each of the above two situations, the electronic device may temporarily turn off the camera input and suspend all computer vision algorithms running in the AR application. However, the IMU signal may be periodically samples to determine when the computer vision algorithms should be restarted. Once the signal received from the IMU

includes acceleration values that are larger than the above-discussed threshold, the camera may be restarted and the computer vision algorithms may also be restarted. In addition to automatically turning off tracking, either user can be given the ability to manually turn off tracking (e.g. a freeze button) and the display can then revert to previously captured key frames so that either user can browse through or manipulate a view into the previously captured environment while the camera is not capturing a scene of interest.

[0092] Virtual Navigation and Digital Visualization of a Remote Individual

[0093] Virtual Navigation may be performed in the context of interacting with a live streamed, augmented scene, and in particular, when a remote individual is assisting a local user of an electronic device. Virtual navigation may generally refer to decoupling of the remote user's (or the local user's) view from the live streamed video and locating the remote user within the tracked environment "virtually" by displaying a reconstruction of the remote user's position within the environment. As an example, for key frame-based virtual navigation, the remote user may enter virtual navigation by selecting one of the key frame views.

[0094] In certain remote assistance systems, a remote individual (e.g., a remote expert) is helping a local user using voice and video communication as well as drawn annotations. In situations where the environment the local user needs to operate is relatively large (e.g., the environment may be distributed over more than one room and/or simply very large machine) the remote user may be allowed to select his/her own viewpoint from a reconstructed version of the local user's 3D environment. In order to avoid confusion caused by the disconnect between what the local user sees and hears, the location of the remote individual may be visualized to provide the local user with the context of the a simulated location of the remote user. For example, the remote individual may be visualized in the form of an avatar comprising one or more of a head, hand(s) and voice, which may be rendered within the 3D environment of the user as augmentations to the environment.

[0095] In the situations where the environment in which the local worker is required to work is relatively large, it may be desirable for the remote individual to navigate the space of the environment independently of the local user to enable the remote individual to provide annotations to the portions of the environment in which the user is not currently located and/or to guide the user to a desired location within the environment. In one implementation, the remote individual is able to capture snapshots (e.g., images) from the screen during the session and annotate the captured snapshots. The localization of these annotations can be determined by the system based on the location of the objects within the snapshot to be anchored to objects in the local user's environment.

[0096] However, the ability of the remote user to annotate any given snapshot may create a disconnect between remote individual and the local user. That is, if the remote individual is not looking at the same part of the environment as the local user, the voice instructions given by the remote individual may not be related to what the local user is seeing. This may cause confusion to the local user, and may lead to errors due to miscommunication.

[0097] In order to solve this problem, the system may employ a 3D reconstruction technique that allows the local user to share a rough representation of the environment with the remote individual, thereby allowing the remote individual to navigate freely in the virtual representation and annotate anywhere within that environment. In one embodiment, the shared representation of the environment may include a 3D reconstruction of the environment as well as so-called key frames (also referred to as virtual navigation "anchors") including the pose of the camera that has taken those key frames. In certain implementations, a key frame may be a screen shot taken from the images captured by the local user's camera while the local user is navigating the environment. A key frame may also include information relating to the pose, e.g., the relative position and orientation of the camera in the environment when the image of the key frame was captured.

[0098] Certain actions by the local user may indicate that a particular area or perspective is of interest to him/her and/or is relevant for the task at hand. These actions may include: making an annotation, "pausing" a current action being performed by the user (e.g., looking at the "same" view for longer than a threshold period of time). These actions can be detected automatically. For detecting a pausing action, the system may be required to distinguish between a pause in which the user looks at an object/scene of interest from a pause in which the electronic device is put aside.

[0099] Since these actions may be indicative of the captured area and/or perspective as being of interest, the system may and add a virtual navigation anchor for those views (e.g., mark the associated image and/or perspective as a key frame).

[0100] The system's virtual navigation features may be supported by an environment tracking and modeling system (for example, a visual or visual-inertial Simultaneous Localization and Mapping (SLAM) system, a scene- or object-specific tracker, etc.). This system can provide a spatial frame of reference and understanding of the environment's spatial layout. Given this data as well as data about the local user's actions—e.g., where the local user looked and/or drew annotations—certain embodiments may automatically select a set of views (e.g., a combination of an image and corresponding perspective of the camera at the time the image was taken) that provide coverage of an object with which the local user is interacting. In some embodiments, the system may also weigh areas in which the local user has shown an interest in (e.g., as automatically identified above) more strongly than other areas when selecting the views. This weighting can be accomplished by a variety of different techniques, including multi-graph-cut algorithms, and others.

[0101] As described above, this freedom on the remote individual's side may cause confusion for the local user since the explanations and instructions via voice may be related to what the remote individual is seeing but not what the local user is seeing. The confusion may be related to the fact that the local user is not aware of the portion of the environment that the remote individual is viewing. For example, if the remote individual was in the same room as the local user, the local user would be able to hear the voice of the remote individual as coming from a specific direction and would also be able to look in the specific direction and see what the remote individual is looking at.

[0102] One aspect of this disclosure addresses this problem by introducing a digital representation of the remote individual and displaying this representation in the local

user's environment. The representation (e.g., an avatar) may include a representation of a head and/or a hand of the remote individual. In certain implementations, the head and hands may be displayed as semi-transparent to prevent obstructing relevant parts of the environment from the local user. Additionally, a location of the voice of the remote individual can be simulated using 3D sound to signal to the local user the position of the remote individual.

[0103] In certain embodiments, a representation of the remote individual's hand may be displayed only when the remote individual is drawing an annotation. Similarly, the head of the avatar may only be displayed when the remote individual has navigated to a different point of view than that of the local user.

[0104] In one embodiment, the position and orientation of the avatar's head and/or hand(s) can be determined using the following technique. When the remote individual selects a view that is different from the view of the local user, the remote individual may be virtually navigating the digital representation of the environment received from the electronic device. There may be two different ways for the remote individual to navigate the environment: i) the remote individual may select one of the captured key frames directly and ii) the remote individual may select a position that is on a transition between two or three key frames. Limiting the navigation of the remote individual to these two options may have the two advantages: i) this technique provides a simple method for the remote individual to navigate the environment and ii) the pose of the virtual camera rendering the view for the remote expert may be determined by interpolation of the poses of the camera in capturing the used key frames.

[0105] When the local user is navigating the environment using the electronic device (e.g., in order to see all the annotations in the environment), the pose of his/her electronic device relative to the environment can be determined. Similarly, the pose of the remote individual can be determined based on the navigation of the remote individual as discussed above. From the poses of both parties, the camera position and orientation of the remote individual relative to the local user can be determined and rendered. In a similar way, the voice of the remote individual can be rendered using the same relative positioning, so that the local user is able to see and hear the remote individual through the avatar representation of the remote individual in the environment.

[0106] Whenever the remote individual is drawing an annotation, a hand holding a pencil may be displayed to indicate to the local user both the position of the remote individual and the fact that the remote individual is adding a new annotation. To place an annotation, the remote individual may select one or more of the key frames as discussed above. The remote individual may then draw a sketch using either a mouse or a touch interface or any other 2D input device. In this way, the remote individual can specify a sequence of 2D locations (describing lines or points) on the screen. Based on the 2D pose of the current input, the 3D pose of the key frame selected by the remote individual and the local user's pose in the environment, the pose (e.g., the position and orientation) of the remote individual's hand can be determined relative to the local user's viewing position, thereby enabling the electronic device to render the hand accordingly.

[0107] As described above, the remote user may enter virtual navigation by selecting one of the key frame views. However, in certain embodiments, virtual navigation may also be entered automatically by the system. This may be performed based on the system state (e.g., whether a remote assistance session is active, whether there is a live stream of video, etc.), environment signals (e.g., the location of the local user, etc.), and context (e.g., the actions previously performed by the remote user, etc.).

[0108] In certain situations, the live video stream may not provide useful information to the remote and/or local users. For example, when the local user puts the electronic device down (or places the device in his/her pocket)—for example, to perform a task with two hands, or to perform another, parallel task in the meantime. The system may be able to detect this state automatically, for example, based on signals such as: (lack of) electronic device movement (inferred to via inertial and/or visual measurements), electronic device orientation (inferred to via inertial and/or visual measurements), and the camera image. If the state of electronic device has been determined to be put down, the system may automatically send the remote user into virtual navigation. Putting the remote user into virtual navigation may include selecting an appropriate alternative view (e.g., key frame view), for example an overview view and/or the last view used by the remote user.

[0109] In certain embodiments, automatically initiating virtual navigation can be applied not only to the remote user(s), but also to the local user. For example, as the local user puts down the electronic device, the local user can be presented with the same view selected by the remote user and/or the last view that the remote user drew annotations on.

[0110] Shared AR Session Recording and Playback

[0111] Augmented reality or virtual reality (VR) systems described herein may be configured to record images, video, and/or annotations for concurrent communication to a remote system for display or for subsequent access and viewing by the remote system. In some embodiments, the AR or VR system can be programmed to communicate concurrently while also creating a recording for later review. Accordingly, in some embodiments, an AR or VR session may be generated based on a "live" or current issue being faced by the user and recorded for later review by the user or other users or recorded for later review by the users and other users.

[0112] In some embodiments, an offline AR or VR instruction session may be automatically recorded from an online session between the user and, for example, a remote expert. Accordingly, both the remote expert and the user may be using devices that support AR or VR. Furthermore, at least one of the remote expert's and the user's device (or an external device or system) may be further configured to record the AR or VR session. Either device may also replay the recorded session. Accordingly, AR or VR may provide instructions in an offline mode.

[0113] In some embodiments, the user's device may correspond to the computing device of FIG. **2** or the local user interface of FIG. **1**. The user's device may allow the user to communicate with the expert user and get help using AR or VR. The communication between the user and the expert user may consist of an audio communication (unidirectional or bidirectional), a video communication from the user to the expert user (allowing the expert user to see, in real time, the same environment as the user), and a data communication

(via which content overlaid over the video communication, such as annotations, may be displayed for both the expert user and the user).

[0114] In some embodiments, the user and the expert device may include one or more components for tracking the respective device in its environment and localizing the respective device within its environment. Accordingly, either device may perform simultaneous localization and mapping of itself to track a perspective or location of the device in relation to one or more critical anchor points. The processing unit **202** or other component of either device may further include a sparse environment reconstruction method that allows the devices to share a "rough" or "coarse" 3D representation of the user's environment with the other user. Additionally, or alternatively, the devices may include one or more components that allow the users to provide or exchange instructions and/or comments in several ways. For example, the expert's device may include audio inputs allowing the expert to provide verbal instructions and or annotations. In some embodiments, verbal annotations may include any verbal statements or instructions provided by the expert. In some embodiments, the expert's device may include visual inputs that allow the expert to draw or type annotations into the video feed being received from the user. The processing unit **202** and/or the component for tracking and localizing of the expert device may localize the visual annotations within the environment of the video feed using the 3D reconstruction. Assuming the user is viewing the expert's video in real time, the expert's annotations would, accordingly, be communicated to the user in real-time as they are drawn by the expert and will appear to the user to "stick" within the user's environment. Sticking within the user's environment means that the annotations may remain positioned at their original location in the user's environment regardless of the position of the user's device in the user's environment. Accordingly, the expert's annotations may be augmented or added to the environment as viewed by the user while allowing the user to freely choose his/her viewpoint to look at the user's environment as well as the annotations. By allowing the expert to include both verbal and visual annotations, the expert may provide the user with all necessary instructions for a very specific repair. As the AR session may be recorded, the expert's annotations may be provided for use by various users, and the "sticking" annotations can then be used to overlay or annotate or overlay over different video captured by the devices of those various users. Further, this may allow the various users to be at any perspective or position in their respective environments without losing the effect or message of the expert's annotations.

[0115] The AR and VR sessions (e.g., the video feed from the user along with the annotations by the expert and the localization information for the expert's annotations) may be captured and/or stored in a very specific form as help sessions for distribution to those having similar problems in similar environments. In some embodiments, these help sessions may be stored on a server or device and allow users to access the help sessions as needed for replay at future times (e.g., "on demand"). Since the help session was an AR or VR session and not merely a video, the user may replay the sound, video, and annotations in the user's 3D environment. Therefore, independent of the viewing position of the user viewing the help session in the environment, the annotations, sounds, and video provided by the expert are displayed at the right position in the environment and will follow the same sequence and timing as when recorded.

[0116] When help sessions are recorded and stored, each stored help session may include a 3D reconstruction of the environment. In some embodiments, each 3D reconstruction of the environment may include one or more significant feature points that allow the user viewing or replaying the help session to identify and/or recognize the environment and track the environment and any movements or annotations within the environment. Accordingly, a user viewing or replaying any help session may use their device to recognize the environment and track the 3D location and timing of any expert annotation that has been drawn along with the voices of both the expert user and any other participants, including the spatial location of any user or expert device relative to the environment (e.g., the spatial location of recording position of the user device and the spatial location of the viewing position of the expert device). Given the 3D reconstruction of the environment and the spatial locations of the original user and the expert, the user that is viewing or replaying the help session may view or replay the help session in any environment that is similar to the original user's environment or that at least includes the one or more significant feature points (e.g., a front of a vehicle being worked on, etc.).

[0117] In some embodiments, recorded sessions can be used in conjunction with object recognition and/or SLAM to overlay recorded annotations over video captured from a new environment. Thus, many users in varying environments but dealing with similar issues with the same or similar object may be able to utilize the pre-recorded or stored help session without being in exactly the same environment and dealing with exactly the same object. As such, the user(s) needing assistance may obtain the needed assistance without relying on the expert to be available at that exact moment, instead accessing the stored help session (e.g., in a help session database accessible by either of the user devices or any device).

[0118] By recording and saving the 3D representations, the help sessions may be viewed or replayed in various ways using various devices and technologies and environments with different captured video. For example, a recorded or saved help session (either VR or AR) may be replayed on the same object in the same environment. For example, another user with access to the original object and environment and the original user device may replay a recorded or stored help session to review and/or understand what actions were performed on the object. In this case, the one or more significant feature points of the environment may be used to recognize and track from the previously recorded actions and/or annotations (verbal and visual) with proper positioning in the 3D environment.

[0119] In another or the same embodiment, the same or another user may replay the recorded or stored help session (either VR or AR) on a similar looking, but different object in either the same or a different environment. For example, if the user or other user is a mechanic at an automotive repair shop, the environment may always be the same but repairs may be on similar vehicles. As noted herein, the stored feature points may be used to "anchor" the help session content to the object and allow the reuse of the help sessions on objects and/or environments.

[0120] In another or the same embodiment, the recorded or stored help session (e.g., an AR help session) may be

replayed in a virtual environment. For example, since a 3D representation of the environment (including significant feature points) of the help session is stored as part of the help session, another user may be able to view or replay the help session in VR application or environment. For example, the VR application or environment may be able to recreate the environment in a virtual manner based on the stored 3D representation. Furthermore, as discussed herein, the user may be able to track movements, objects, annotations, etc., based on the stored significant feature points. Accordingly, the user viewing or replaying the help session may be immersed into the reconstructed environment and may experience the annotations (verbal and visual) and witness the issue and solution without having access to the real physical environment. In some embodiments, the recorded or stored help session may be used for training purposes.

[0121] In some embodiments, the help session may be viewed or replayed on a non-VR or non-AR computer or device by rendering the 3D representation of the environment directly on the screen. Accordingly, the stored significant feature points may be displayed using the screen of the computer or device in proper relation to the environment and object shown on the screen and the visual annotations may be displayed on the screen in conjunction with the displayed environment and object. Additionally, audio components of the computer or device may replay the verbal annotations stored as part of the help session. The user viewing or replaying the help session may be able to navigate the environment freely, which may be useful in educational and/or training environments. In some embodiments, the viewed or replayed help session may be reviewed without navigation capabilities and may instead may reviewed or replayed from the perspective of either the original user or the expert.

[0122] In some embodiments, a recorded help session may include long periods of time where nothing relevant to the object or the environment occurs or when no useful annotations (verbal or visual) are provided. Accordingly, these long periods of time may be automatically detected and deleted from the help session before the help session is played back. In some embodiments, these periods may simply be skipped over or fast-forwarded through.

[0123] In some embodiments, one or more subsequent users may be able to add annotations to the stored help session. For example, a supervisor may review an employee's repair, etc., and provide comments or the work performed or provide suggestions and/or feedback. In some embodiments, the system storing the help session may detect indicators provided by one of the user or the expert that indicate when critical or important periods of the help session begin and/or end, and these indicators may be used to automatically detect and delete periods that are not useful. In some embodiments, simple mechanisms may be used to identify and remove sections of the recording that don't contain any information. For example, periods of no verbal and/or visual annotations may be detected and deleted. In some embodiments, the periods may be detected and/or deleted based on a specified minimum length threshold (e.g., where the detected and/or deleted period must be greater than a minimum length). In some embodiments, the detected periods may not be deleted but rather skipped over during playback. Accordingly, a compact representation of the help session can be generated and/or replayed.

[0124] In some embodiments, each person (e.g., user, expert, other user) that adds or provides information for the help session may add a new layer to the help session. For example, the original recording of the environment and/or object provided by the user may comprise a single layer or portion in the help session. The expert's verbal and/or visual annotations may comprise a second layer or portion in the help session. A later other user may provide feedback and/or additional comments as a third layer or portion of the help session. In some embodiments, the other user (or the expert or user) may include another help session as part of the help session. Accordingly, help sessions may be embedded in other help sessions during recording and storage. In this way, multiple versions of help sessions with multiple layers of information will be generated and stored for later retrieval and use to support future help sessions or an offline use as described.

[0125] Offline use may comprise viewing and/or replay of the help session by a user that is not connected or in communication with the expert at the time that the help session is reviewed. Alternatively, or additionally, the offline use may comprise the viewing or replay of the help session when not connected to a communication medium. In some embodiments, the expert or another user may use a tool (e.g., an offline or online editing tool) to edit stored help sessions, for example generating step-by-step instruction from the help sessions. Additionally, one or more users (e.g., the original user, the expert user, or other users) may use a tool that allows for selection of short sequences from a complete help session and defining of the selected short sequences as single steps in a step-by-step instruction. Such step-by-step instructions can then be replayed in any AR or VR capable device.

[0126] Recording instruction sessions using video is well known and widely used. However, since every environment might look slightly differently, especially from different viewpoints, such instructions might cause more confusion than they would help. With our innovation the instructions are played back in any similar environment directly on top of the environment and allow the user to be consumed by choosing a free viewing angle. This is especially important when the user is wearing a head mounted display to be able to use his/her hands to follow the instructions. It is also important in situation where the user is not able to look from the same viewing angle as the recording was done (maybe because there is some other object(s) in his environment that blocks him/her form going there).

[0127] Examples of advantages of being able to record and replay VR and AR help sessions may include that such help sessions may be used to automatically generate instructions (e.g., step-by-step instructions) based on the stored help session. Additionally, when the stored help sessions include the tracking information for the recording user and the various objects, the help session and/or instructions can be replayed in similar environments where similar tracking information or points can be identified in the replay environment. Accordingly, the help session or instructions can be replayed while freely choosing the point of view of the user. In some implementations, the help session or the instructions can be compacted so that they can be stored, communicated, and replayed in a compressed format.

[0128] In some embodiments, the user may initiate the AR or VR session by capturing a video of the object in the user's environment using the user device (e.g., camera). The video

may be shared with the expert, who may add and save annotations to one or more objects in the video. Each annotation may be saved with reference to a spatial relationship to the object and a temporal relationship to a time within the session. These annotations may be stored with the session so that the user or any other user may review the session and see the annotations in relation to the object in question.

[0129] In some embodiments, the expert may initiate a second augmented reality session, capturing an object in his/her environment and superimposing a new session and/or objects on the saved annotations, wherein the each annotation is superimposed in spatial relationship to the second object based on the saved spatial relationship and in temporal relationship to a time within the second augmented reality session based on the saved temporal relationship. For example, this may allow the expert to create a session that the user can replay in a local environment without first sending the environment to the expert.

[0130] In one embodiment, a method can include: initiating a first augmented reality session where a first camera captures video of a first object within the camera's field of view; saving augmented reality annotations created during the first augmented realty session, where each annotation is saved: in spatial relationship to the first object, and in temporal relationship to a time within the first augmented reality session; initiating a second augmented reality session where a second camera captures video of a second object with the camera's field of view; and superimposing the saved augmented reality annotations over the video captured by the second camera during the second augmented reality session, wherein the each annotation is superimposed: in spatial relationship to the second object based on the saved spatial relationship, and in temporal relationship to a time within the second augmented reality session based on the saved temporal relationship.

[0131] Shared AR Session Bandwidth Adjustment

[0132] Some remote assistant systems employ audio/video (A/V) feeds or conferencing techniques to allow the expert to see the user's environment and conditions and provide help accordingly. In some embodiments, an audio medium or channel may provide bidirectional voice communications and a video medium or channel may provide bidirectional video communications. Such systems may utilize large data transfers to communicate the A/V information. Such large data transfers may necessitate high bandwidth and/or speed connections between the expert and user. In general, the transfer of A/V data utilize high bandwidth communications to ensure complete transfer of actions, comments, etc., in real time or with minimal delay, allowing all viewers or recipients of data to receive the conveyed recorded or live data in a useful manner. Accordingly, such systems may have reduced functionality in environments where communication coverage is reduced, such as power plants, mines, cellars, equipment rooms, industrial facilities, etc. In A/V conference systems, the quality of the A/V feed or conference may be reduced or the video may be turned off completely. In situations where the video feed was relied upon for conveying important information, the reduction to only video may result in communication errors and subsequently higher error rates by the user. For example, when A/V data is generally transmitted in low bandwidth conditions, the A/V data may be received by a viewer in short bursts or with larger periods of buffering as compared with high bandwidth conditions. These shorts bursts or increased buffering periods may increase frustration or make it difficult to follow what is being shown or conveyed in the A/V data feed.

[0133] In some embodiments, when there is no high bandwidth communication available, the A/V feeds or conference techniques may revert to audio transmissions alone. In reverting to audio transmissions alone, much longer communication times may be needed to relay the necessary information between the user and the expert and subsequent increases in miscommunication and errors are seen.

[0134] However, in AR and VR systems, high speed and high bandwidth connections may not be required for the duration of the communication. For example, once the initial environment is communicated as a 3D digital representation, information such as verbal or visual annotations may be provided at much lower bandwidths. Accordingly, AR and VR systems may share the 3D representation over a high bandwidth connection and maintain visual and verbal annotations communication should the connection between the user and the expert degrade to levels where video/audio conference systems would be forced to drop to audio only connections.

[0135] In some embodiments, the AR or VR system of FIG. **1** may determine that the communications between the user and the expert have degraded or are insufficient to support both video and audio communications. In some embodiments, the determination of the sufficiency of the communications may be made by at least one of the processing unit **202** or the communication connection **220**. If the determination is made that the communications link still exists but conditions have degraded below a specified threshold, the AR or VR system may determine that no video will be shared while voice and data are shared.

[0136] In some embodiments, this determination may be made before or after the initial 3D representation of the user's environment is conveyed from the user device to the expert device. Accordingly, the AR or VR system may reduce communications exchanged between the devices to audio and data only. Since the expert device already has the 3D representation of the user's environment, only minimal information needs to be conveyed from the user to the expert. This minimal information may include data regarding 3D representations of particular objects, etc., that the user is actively viewing or manipulating. The 3D representation of the object may be communicated as data and may be reconstructed by the expert device the AR or VR environment. The 3D representation of the object may be conveyed as data using less bandwidth than video information of the same object. Accordingly, by conveying the 3D representation of the object, movements, annotations, or other details of the object may be conveyed even in low bandwidth or speed conditions.

[0137] In operation, after detecting a low bandwidth or speed condition, the user device and the expert device need not share video, but only voice and some data. While the user is viewing his/her environment, the user device generates the 3D reconstruction of the environment and object(s) being viewed and shares that 3D reconstruction with the expert. The expert device receives the 3D reconstruction of the user's environment and generates a virtual or augmented display of the user's environment. The expert may use the virtual or augmented view of the user's environment generated based on the 3D reconstruction and may provide

verbal or visual annotations. The expert device may convey the visual annotations as 3D reconstructions to the user device along with any verbal annotations from the expert. The user device may receive the verbal and visual annotations from the expert device and provide the visual annotations in the proper location in the environment and play the verbal annotations from the expert.

[0138] By reducing the communications between the user device and expert device to the 3D representations instead of video, the AR or VR system is able to reduce the data that needs to be communicated between the devices. Since a video stream contains a lot of redundant information (e.g., information regarding objects, etc., that do not change between frames of the video), and the 3D representations do not, the communications between the user and the expert are "compressed".

[0139] In some embodiments, the user device of FIG. **2** may have no connectivity to the expert device in the environment of the object at issue for the user. Accordingly, there may not be any video, audio, or data communication for the user device. In embodiments where no connectivity existed, no information regarding the user's environment was provided to the expert. In such an instance, the user may scan his/her environment with the user device and create a 3D reconstruction of the environment plus a tracking map. In some embodiments, the tracking map may provide the significant feature points according to or relative to which movements, annotations, and other actions may be tracked. Once the user device creates the 3D reconstruction information, the user may move to a location where communication is possible and the generated 3D reconstruction information and the tracking map is communicated to the expert. The expert may receive the 3D representation and tracking map and may provide verbal and/or visual annotations. These verbal and/or visual annotations may be communicated as video, audio, and/or data. For example, as described herein, the verbal and/or visual annotations may be communicated as 3D representations. Once the user receives the annotations from the expert, the user may move back to the original environment without communication and replays or views the annotations from the expert via the user device in AR or VR. Thus, the user may obtain the assistance needed regardless of the communication capabilities in the environment where assistance is needed.

[0140] In such embodiments, the request for assistance may be split into two parts. The first part allows the user to submit information to the expert and receive the expert's instructions and comments and the second part allows the user to replay the expert's instructions and comments to guide him/her through the issue. Since only the first part needs a connection between the user and the expert and could be performed with a pre-captured model of the user's environment, this two part technique may help provide a solution in situations where there is no connectivity at all at the actual environment.

[0141] Associating Annotations with Recognized Objects

[0142] When using augmented reality or virtual reality in an environment accessible to or by multiple users, sometimes one or more of the users may annotate an object in the environment so that others are able to see the annotation. One common annotation may be a rough outline drawn around the object to highlight or indicate that the object is selected or identified by the user creating the annotation. However, such an outline may be dependent on various factors, including a viewing direction from which the outline was drawn, a viewing direction from which the outline is viewed, a shape of the object being outlined, an amount of the object that is exposed to view, among others. While the outline as drawn by the user might perfectly fit the object from the drawing direction, the outline might not even be close to the object's outline when viewed from another angle.

[0143] For example, in many VR or AR systems, the user device may include a flat or two-dimensional touchscreen. Accordingly, use the touchscreen to display an augmented view of the object or environment and may draw or directly interact with the object or environment to insert visual annotations directly onto the augmented view. While viewing the object or environment via the user device, the object or environment may be 2-dimensional, and accordingly, any added visual annotations may also be 2-dimensional and based on the 2-dimensional view of the object. Thus, even if the 2-dimensional annotation is perfectly displayed in relation to the object (e.g., a perfect outline of the 2-dimensional view of the object), the 2-dimensional annotation may be improperly shown or perceived from other viewing directions. This deficiency may be further complicated when the object has a complex shape.

[0144] In some embodiments, when the object is viewed and annotated by the user device, the user device may generate a 3D model or reconstruction of the object. This 3D model or reconstruction may utilize information from a depth sensor to identify the 3D shape data of the object and, thus, render the 3D model or reconstruction. However, such use of a depth sensor may require additional hardware in the user device, which may not be desirable.

[0145] Alternatively, or additionally, 3D object recognition and automatic outline generation may help ensure that the outline drawn by the user is properly shown and/or tracked regardless of the perspective of other users viewing the object. Accordingly, the other users may view the object as it was originally selected or identified by the user regardless of the other user's perspective of the object. The 3D object recognition and automatic outline generation may allow for the varying of viewing viewpoints independent of the creation viewpoint regardless of when the annotations are created and viewed. Additionally, the 3D object recognition and automatic outline generation may allow for the varying of viewing viewpoints independent of the creation viewpoint regardless of whether the annotations are created and viewed in the same environment with the same or similar object or in different environments with the same or similar object.

[0146] While the example annotation described herein relates to outlines of objects, the same issues of properly tracking and/or indicating annotations in relation to or corresponding to a target object from various perspectives different from the original user may exist for any visual annotation.

[0147] When the user adds an annotation indicating or identifying an object via the touchscreen of the user device (e.g., in AR), the user may draw a rough outline around the object or may draw an arrow pointing to the object. When the indication (either outline or arrow) is rough or not very specific, the user device may automatically identify which object is being indicated. For example, if the circle drawn by the user surrounds approximately 80% of the object and 20% of another object, the user device may determine that

the user intends to identify the object (e.g., based on a comparison of what is included in the drawn circle). Similarly, when the user draws an arrow, the user device may determine the object that is mostly likely being identified by the drawn arrow. In some embodiments, since the user indicates an area where the object exists when the circle or arrow is drawn, the user device may use that area to identify a region in which object recognition is performed (e.g., via an object recognition algorithm). By focusing the object recognition only on or near the area indicated by the user, efficiency of the object detection may be improved.

[0148] The object recognition algorithm may utilize a database of all objects in the environment that users might annotate. The database may include 3D geometry of each of the objects as well as various parameters of the object (e.g., name, label, previous annotations, etc.). Once the object recognition algorithm identifies the object from the database, the user device may determine a pose of the object (e.g., a 6DoF pose including (x,y,z) coordinates of position and (x,y,z) coordinates of orientation, such as an indication of three (3) rotations around a main axis of the object relative to a reference coordinate frame). Accordingly, the user device may create the outline or arrow identifying the object as it is seen from this specific point of view.

[0149] When the annotation is communicated from the user to another user, the annotation may include information about the annotation (e.g., 3D object tracking information and the annotation identifier) and/or information about the object annotated. For example, the annotation information may include information relating the annotation to a tracking or reference point in the environment to track or localize the object and/or annotation. The annotation information may also identify the type of annotation (e.g., outline, arrow, etc.). By tracking the object being annotated, the annotation can be located appropriately in relation to the object regardless of the viewpoint of the other user viewing the annotation. Accordingly, regardless of the pose of the object in the view of the other user, the annotation may appropriately identify the object (e.g., the arrow will be pointed at the right object or the outline will be applied to the current viewed pose of the object). For example, the outline of the object created by the user may be aligned with the object as viewed by another user by identifying the pose of the object as viewed by the other user. The identified pose may then be used in rendering the outline based on the geometric representation of the object as stored in the database.

[0150] In some embodiments, 3D object tracking may update a new pose for each frame of the annotation and/or object. This new pose allows for rendering the outline of the object from the geometric representation of the object. The object recognition as described herein may be performed using various methods, for example feature based methods, template matching, or recognizing objects using a trained neural network. Additionally, or alternatively, object tracking can be performed using various techniques, including, but not limited to, feature tracking, edge tracking, tracking by detection (meaning the detection methods described herein are used for every frame).

[0151] In another version, the user creating the annotation just touches the object that is displayed on the screen. The input from the user's touch is used to perform automated object recognition using machine based recognition based on a database of objects to be recognized. Additionally, the database of objects may include computer aided drafting ("CAD") representations for the objects in a hierarchical structure of assemblies, sub-assemblies, parts and subparts. Accordingly, the user device may utilize the CAD representations of the hierarchical structure to draw an outline around or select the object. For example, the user device may draw an outline around or select the smallest identified part under or near the user's finger. As the user continues to maintain the touch (e.g., keeping the finger contacting the touchscreen), the user device may expand the outline or selection to now select the next higher assembly in the product hierarchy. This expansion of the outline may continue until the user discontinues the touch (meaning the user is satisfied with the object outline or selection). In some embodiments, the user might move a finger around on the touchscreen and each part of the assembly or environment being touched may be included in the selection and the outline may be drawn around all of the selected parts.

[0152] Alternatively, or additionally, the CAD model of the object as stored in the database of object may be used to highlight selectable features of the object as detected in the AR field of view. In some embodiments, the selectable features or parts can be internal to the object (based on the CAD model) and might not necessarily be visible in the camera's actual field of view. Accordingly, such features or parts may be displayed in a "window" or other frame of the object that indicates that they are internal to the object. The user may then point and/or click to select various highlighted parts (visible or obscured) of the object. Additional functionality or features can then be presented to the user based on the selected part. Accordingly, the user may be able to select a portion or part of an identified object and keep that portion or part highlighted. The CAD model could then be continually referenced to update the perspective of the highlighting of the selected part consistent with the camera angle/perspective.

[0153] Additional techniques and methods that can be used in various embodiments for selecting parts of recognized objects for annotation are described in Appendix 3 "DISPLAYING CONTENT IN AN AUGMENTED REALITY SYSTEM" of priority U.S. Provisional Application No. 62/512,696, referenced above. For example, a CAD model of an object can be used to highlight selectable features of the object in the AR field of view. The selectable features or parts can be internal to the object (based on the CAD model) and might not necessarily be visible in the camera's actual field of view. The user can then point and/or click to select various highlighted parts (visible or obscured) of the object. Additional functionality or features can then be presented to the user based on the selected part. The part selection functionality can be used to present the user the ability to select a portion or part of an identified object and keep that part highlighted. Further, the selected part might be one that is obscured or not visible within the camera's field of view. The CAD model can then be continually referenced to update the perspective of the highlighting of the selected part consistent with the camera angle/perspective.

[0154] In some embodiments, the object identification or recognition algorithm described above and implemented by the processing unit may be configured to identify the object, and the processing unit may further identify details regarding the identified object in a database. For example, the processing unit, after identifying an object as being a sink faucet, may access a CAD file of the sink faucet and use the CAD file to identify hidden components (e.g., components

that are not readily visible) of the sink faucet. In some embodiments, this may be shown as a semi-transparent or transparent overlay view or layer. This overlay view may show all or some selectable parts. In some embodiments, the overlay view or the CAD file may include a menu or list of parts, where selection of an item in the menu or list by the user will highlight the part in the overlay view. In some embodiments, there may not be an overlay view and the parts will be selectable or highlighted on the main view itself.

[0155] Accordingly, as described herein, annotation may be created using prior knowledge on the environment (e.g., stored in the database). This prior knowledge may be used to augment the annotations to better identify objects and draw the annotations in proper relation to the object independent of the viewing direction. This allows the viewing user to be positioned at any position or location around the object and still see the object annotation. This may be an improvement over systems where the annotation sticks to the position where it was placed and therefore is hardly visible from other viewing direction or does not change shape according to the perspective changes of the object. In some embodiments, the database of objects may be utilized in conjunction with a depth sensor.

[0156] Automatic or Assisted Management of Annotations Augmented reality (AR), mixed reality, and virtual reality technologies may facilitate communication between participating users. For example, one user may annotate a specific object or object part in a scene shared between the users to convey a message regarding the object or object part. One or more of the users may wish to label one or more of the objects or object parts so that the other users are able to visually see the name of the object or object part in addition to hearing the name. However, as discussed herein, placing a label on an object in the 3D environment may be complicated, requiring identification of the object, locating of the object in 3D, and annotation of the object with label text. Placing the label on the object may be further complicated when the environment of the object is not pre-known and when the users are using mobile devices (e.g., head mounted displays) where text typing is difficult and/or tedious.

[0157] In some embodiments, simultaneous localization and mapping (SLAM) may be used to generate 3D representations of the objects and annotations and track the target object and annotations. In some embodiments, the user may draw abstract circles/outlines and provide verbal comments, such as "do something to this object" while identifying an approximate location or drawing a rough shape of the object. The verbal comments may include nouns and verbs, and the AR system (e.g., the user device or another component of the AR or VR system) may use the nouns to identify and label the object drawn by the user. In some embodiments, the system may identify an anchor point based on the object that is drawn by the user. In some embodiments, the system may automatically generate labels for annotations of objects. In some embodiments, the system may automatically update annotations based on verbal comments of the user or users. In some embodiments, the verbs of the verbal comments may be indicated with icons or animations. For example, the verbal statement "open this panel" may result in the panel object being automatically identified when generally circled or indicated by the user and a button or animation for "opening" may be displayed in relation to the identified panel object. In some embodiments, the various actions to be performed may be presented in a particular sequence or may be displayed in a delayed fashion as actions are completed. In some embodiments, a combination of SLAM with user annotations and/or other input may be used to track a position of the object in the AR/VR world.

[0158] In some embodiments, annotations and icons generated based on verbal comments may be conveyed to a deep learning system, which may supplement the creation of the annotation labels (e.g., if the user's annotation is not enough, deep learning based on the instructions may identify the proper target object). Additionally, or alternatively, deep learning may be used predictively to assist with generating labels and/or annotations based on the objects in the view and the verbal commands/statements to simply user interaction in generating the annotations. For example, if the user is typing annotations, deep learning may provide words of objects/commands associated with a particular view.

[0159] In some embodiments, annotations and/or labels may automatically be removed as actions are completed or objects addressed. For example, if annotations or instructions involve removing a screw, the annotation area may be examined by the system to determine if a screw exists in the view that needs to be removed. If not, then the annotation/label/actions may be removed. If so, then the annotation/label/actions may remain until the screw is removed. Accordingly, deep learning may be used to more easily create labels and also to remove labels as actions are completed or the objects disappear. Accordingly, the system may continue monitoring the annotation area to determine when to remove labels and proceed to subsequent steps. In some embodiments, the user performing the actions or instructions may indicate completion of each action or command, which may reduce computation cost by reducing scanning of the annotation area.

[0160] Furthermore, in some embodiments, the system may determine to pause video, transmissions, recording, deep learning, etc., while no sound or annotation is being added to the environment. Such automatic pausing may save or reduce processing by the system. Alternatively, or additionally, the system may await a command or "okay" from the user.

[0161] In some embodiments, the AR or VR system may include a device for each user (e.g., a mobile device like a tablet or mobile phone, a head worn device like Google Glass or any other head mounted device, desktop PC, or laptop). The user device may comprise a camera that allows the user to capture a stream of images from the user's environment and a processing unit that analyzes images captured by the camera and determines relative position of the device within the environment (e.g., performs device tracking). The user device also comprises a processing unit (might be the same processing unit described above) that reconstructs a rough digital model from the images captured by the camera and the associated poses of the object (e.g., creates a 3D reconstruction). The user device further comprises an audio input/output component to communicate with the other users. The user device also comprises a touchscreen to be able to view and interact with (e.g., draw or sketch over) the environment captured by the camera and shown on the touchscreen. A communication unit of the user device allows multiple users to communicate with each other and share data such as a video stream that the camera is delivering, a relative position of the user device (as computed by the tracking algorithm running on the process-

ing unit), and a 3D reconstruction of the environment of the user device so that it can be represented on a remote location. The user device is coupled to or configured to couple to a server network (such as the internet), that allows the user devices to connected to each other and to share the described data.

[0162] The system or user device described above may create and share a digital representation of the environment of the user device. In some embodiments, SLAM may be used to calculate a relative pose of the user device in respect to the environment and to generate a rough representation of the visible surfaces of the environment in 3D. As soon as the user is moving around in the environment with the camera of the user device activated, the user device or system may start building an internal 3D model of the environment. This 3D model may be used to calculate for each frame where the camera is relative to the environment (e.g., calculate the pose of the camera). The system or user device may also create an annotation for an object in the environment. The users may be able to draw annotations on the touchscreen while they are looking "through" the touchscreen onto the environment. In many cases, the users may draw outlines around the object, arrows pointing towards the object, or crosses on top of the object to identify it. In some embodiments, the users may automatically verbally identify a name of the object and/or an action associated with the object. For example, the user may say "look here, this is a . . . !" or "please turn that red knob!" or "please press this button!" or "you need to open this screw!" While the users are providing verbal annotations, they may also draw the annotations to indicate which object(s) is really meant with "this." The system or user device may use both inputs and timely relationship between the two to identify what object was really being indicated.

[0163] The system or user device may perform multiple steps when creating annotations. For example, the system or user device may identify commonly used shapes or symbols (e.g., circle, arrow, cross) by using a classifier on the drawn strokes or annotations. Additionally, the system or user device may identify the point of reference from the shapes or symbols (i.e. the tip of the arrow, the center of the circle, the crossing point of the cross, or, if no shapes or symbols were identified, the center of mass of the drawing). Furthermore, the system or the user device may run a voice recognition method on the sound input with a window (e.g., of a few seconds before and after the shape or symbol was drawn) to identify nouns and verbs in a verbal comment. The system and user device may use the nouns to generate annotations displayed as a label text. In some embodiments, the annotations may include a leader line starting at the reference point leading to the label text. The system and user device may use the verbs to display little icons that visualize what actions are to be performed close to the reference point of the drawn shape or symbol.

[0164] In some situations, the system and/or user device may use a trained neural network associated with the label text to segment out image areas that fit to the label text and highlight the one area that contains the reference point. Accordingly, the object will be highlighted. Where the user has drawn an outline, the system or user device may identify edges of the object that are closest to the drawn outline and replace the drawn outline with those edges. The combination of the "new" drawn outline together with the label text may be provided as additional training input to improve the neural network.

[0165] The system and/or user device may track annotation within subsequent image frames using patch tracking of the segmented area. Alternatively, or additionally, the device or system may project the annotation onto the 3D reconstruction of the environment and/or share a 3D representation of the annotation with other participating users, so that all users are able to visualize the annotations within their own environments or fields of view even if the users are looking at the annotations from different viewpoints. In situations when the viewpoint changed more than a predefined threshold from the viewpoint the annotation was created, the system or device may update the 3D representation of the highlight area by again using the neural network to classify the camera image from the new user's position.

[0166] In some embodiments, the system and user device may determine how/when to maintain annotations. In some instances, the user may change the environment during a session. For example, in a maintenance or repair session, the user may remove parts or opening compartments, or reposition items. If an annotation was provided for a part that was removed in a subsequent step, the user device and/or system may be configured to remove the annotation once the step is performed. The system or user device may perform this removal automatically by scanning the environment or wait for a trigger from the user. For example, at regular intervals (e.g., every second), the system or user device may perform (for each annotation visible in the view of any user) a search request in an associated neural network to obtain a new segmentation for the current camera image. If there is no positive segmentation area provided under the reference point of the annotation over a sequence of subsequent frames, we can safely assume that the object has been removed and the associated annotation should not be displayed any more. Alternatively, or additionally, if the user is prompted with a request about removing a part or completing a step, the user's response may be used to remove the annotation. Alternatively, or additionally, the user may provide feedback as steps are completed to indicate to the system or user device that an annotation may be removed.

[0167] Environment-Indexed Database of Shared AR Sessions

[0168] A user of a user device in an environment may wish to use the user device to view objects in the environment in augmented reality (AR). For example, the user may hold the user device up to view a street corner while traveling and obtain a translation of street signs and restaurant names. Additionally, or alternatively, the user may be performing maintenance on a product and use the user device to identify individual parts of the product for ordering and replacement by pointing the user device at the product. In such instances, the touchscreen of the user device may show the street corner or product as viewed by the camera of the user device and include annotations in the way of translations or part labels, etc. These products or objects associated with these annotations may be identified using visual object recognition.

[0169] Visual object recognition may be difficult to perform quickly and efficiently. Methods utilizing Deep Neural Networks (DNNs) may more efficiently perform visual object recognition. However, DNNs may utilize labeled input images as training data, where accuracy of the visual

object recognition is dependent at least in part on an amount of training data provided. Training the DNN may be a tedious and labor intensive process, requiring the obtaining of and providing of the labeled input images. The labeled input images may comprise an image having one or more objects displayed in the image labeled with a label that is to be used by the DNN to identify that object. For example, to train the DNN to visually identify a red tricycle in an image, the DNN will need to be provided with an imaging have a red tricycle that is labeled in the image. Accordingly, much effort is expended obtaining labeled pictures to input into the DNN for training.

[0170] In some embodiments, a user may annotate various objects with labels before sharing the annotations with other users. These annotations may be used as the labeled images for input into the DNN. For example, when the user annotates an object in the interior of a car as a steering wheel and communications the annotation to the other user, one or more of the annotation, the target object, and the environment may be input into the DNN for training. Accordingly, the DNN may be trained to identify steering wheels in images.

[0171] In some embodiments, the remote assistant application described herein may allow users to help each other and at the same time. In user, the application may automatically generate such training images and labels as a byproduct, as described in the example above. Since the application may be used in many different environments and by many different people around the world, the application may generate various training images for deep learning by the DNN.

[0172] Once the DNN is trained, the DNN may be used to perform object recognition in an unrestricted environment. The DNN may compare images with/without objects to identify when the object exists in one picture but not in the other picture. Accordingly, DNNs may provide a method for performing visual object recognition.

[0173] Alternatively, or additionally, visual object recognition may be performed via database searching where the database is generated based on images received with pre-identified labels. For example, when the user is on a street corner in Paris, there may exist a database or listing of all objects at that particular street corner. In some embodiments, the location of the user may be determined by GPS or some other positioning method. The user's position or location may be used to search for a corresponding database for that position or location. When the database is found or identified, the user device may display annotations and/or labels for all objects viewed via the user device, assuming annotations and/or labels exist for all the objects.

[0174] In some embodiments, the user of the remote assistant application may annotate various objects with labels before sharing the annotations with other users. In use, the application may automatically generate a database of such images, objects, labels, and the tracking map. This information may allow for the generation of a database of locations and objects so that object recognition can be performed based on the stored labels, images, and locations. In some embodiments, the DNN visual object recognition may be combined with the location database.

[0175] In some embodiments, one or both of the DNN image input and the location database may be further benefited when a user provides some verbal instructions. For example, when the user is capturing an image or video, the sound sequence associated with the captured image or video may provide additional annotations or labels. For example, the user may identify one or more objects in the image or video verbally while they are annotating (e.g., the user may instruct to "open that compartment" or "turn that handle"). Accordingly, voice-to-text programs or algorithms may be used to analyze the sound sequence captured in association with the image or video and identify nouns that can be attached as labels to the image or video. Additionally, or alternatively, geolocation of the images or video may also be captured based on the user's position or location. The labels and objects being labeled may be associated with that geolocation. In some embodiments, the geolocation may be used in combination with object recognition from DNNs to filter results based on the location of the user. For example, there exist several Eifel Towers (and replicas) in the world. Based on visual object recognition alone, the DNN may not be able to distinguish from the Eifel Tower in Paris and the Eifel Tower in Las Vegas. However, the two Eifel Tower's may be easily distinguished based on their locations.

[0176] In some embodiments, simultaneous localization and mapping (SLAM) may provide tracking of the user device relative to the environment. This localization and mapping may be used to generate a rough 3D reconstruction of the environment. Accordingly, a rough 3D model of the environment, including many visual features describing the environment uniquely, may be generated. Such 3D reconstructions or models can be stored in a geolocation indexed database, for example using the geolocation of the user as an index.

[0177] In operation, the user device may be used for object recognition in AR contexts using a DNN or by accessing a geolocated database with labels for all objects at that location. The DNN based object recognition may utilize the labeled training images provided by the remote assistant application to identify labels of objects being viewed by the user device. The geolocation database may store additional information about the object (e.g., indexed by the object label), including geolocation information for the object. In some embodiments, the database may be filtered according to geolocation to identify objects at that geolocation, which are then compared to an image captured by the user device camera to determine what objects are shown that are also in the database. In some embodiments, the database may store environment maps and may be filtered or searched by the geolocation to retrieve an environment map of that particular geolocation. Once the environment map is retrieved, it is loaded the user device compares features of objects identified in an image captured by the camera of the user device with information from the environment map. When there is a match of objects in the image to information in the environment map, find the geometric relationship (pose) of the device relative to the environment. The user device then displays information from the environment map for objects that are visible within the view of the current camera image.

Aspects of Various Embodiments

[0178] A method can be performed by a portable computing device having a processing unit, a memory, and a camera. The method can include: capturing an image with the camera; detecting an object within the captured image based on a database record for the object within a database of objects; based on the database record, retrieving an identifier for initiating a live communication with a remote

party; in response to the retrieving the identifier, presenting, in association with an indication that the object has been detected, a user option to initiate the live communication with the remote party; and in response to receiving user input, initiating the live communication between the device and the remote party based on the identifier. The live communication can include one or more of: voice, video, and text messaging. The live communication can include a live video stream captured by the camera of the device. The method can further include: transmitting first augmented reality (AR) graphics generated by the device to the remote party along with the live video stream to be displayed by a remote computer system; receiving second AR graphics from the remote party via the remote computer system; and displaying the second AR graphics overlaid on the live video stream captured by the camera of the device on a display of the device. The detecting the object can include: determining a shape of the object based on the captured image; and matching the shape of the object to a stored shape associated with the record for the object. The detecting the object can include: scanning a machine-readable code disposed on the object within the image; and dereferencing the machine-readable code using the database. The method can further include: based on the database record, displaying on a display of the device an expected location of a graphical marking on the object as augmented reality graphics overlaid on a live video stream captured by the camera of the device; scanning the graphical marking with the camera, wherein the graphical marking encodes additional information about the object; and transmitting the additional information to the remote party. The database can further include user instructions for interacting with the object associated with the record for the detected object. The method can further include displaying the user instructions via the display. The database can be located on a remote server. The database can be located on the device.

[0179] A method can be performed by a system including a first portable computing device having a processing unit, a memory, and a camera. The method can include: the first device capturing a first image of at least a portion of an object; accessing a first database having a plurality of records storing characteristic data for identifying objects based on captured images; locating a first matching record within the first database based on the first image; accessing a second database having a plurality of records storing identifiers for initiating communication sessions; using the first matching record within the first database to locate a second matching record in the second database; retrieving an identifier for initiating a communication session from the second matching record; and using the identifier for initiating a communication session to initiate a live communication session between the first device and a second device. The first matching record can include an identifier of the object. The second matching record can be located based on the identifier of the object. The object can include a machine-readable code captured in the first image. The machine-readable code can have a format defined in the first matching record. The method can further include decoding the machine-readable code using the format. The communication session can include the first device sending a live video stream, captured by the first device, to the second device. The method can further include: capturing a second image of at least a portion of the object with the first device, wherein the second image includes a machine-readable code; decoding the machine-readable code; and the first device transmitting the decoded machine-readable code to the second device within the communication session.

[0180] A method can include: initiating an AR video session using a camera and a display on a first device operated by a first user; identifying an object within a field of view of the camera; performing motion tracking using the identified object within the field of view of the camera during the AR video session; accessing a database having a plurality of records storing identifiers for initiating communication sessions; using the identification of the object to locate a matching record in the database; obtaining an identifier for initiating a communication session from the matching record; based on the identifier for initiating a communication session, presenting to the first user, during the AR video session, an option to initiate a person-to-person communication session; and in response to selection of the option by the user, using the identifier for initiating a communication session to initiate a person-to-person communication session between the first user operating the first device and a second user operating a second device. The communication session can include the first device sending to the second device a live video stream representing the AR video session. The communication session can include the second device sending AR graphics, specified by the second user during the communication session, to the first device, wherein the AR graphics are displayed on the first device in motion relative to the object in the field of view. The object can include a machine-readable code, and the identifying the object can be performed using the machine-readable code. The motion tracking can be performed using the machine-readable code.

[0181] A method of sharing a user interface between first and second mobile devices can include: receiving a selection of an instructional sequence from a user with the first mobile device; transmitting data indicative of the selected instructional sequence from the first mobile device to the second mobile device; in response to receiving the data indicative of the selected instructional sequence: capturing a first sequence of images with a camera of the second mobile device, detecting an object within the first sequence of images, the detected object being identified by the selected instructional sequence, tracking the object within the first sequence of images, and displaying the instructional sequence overlaid on the first sequence of images with a display of the second mobile device. The first mobile device can be or include a handheld mobile device and the second mobile device can be or include a head mounted display. The first mobile device can include a touch screen configured to receive input from the user, and the second mobile device can include an auxiliary input device, different from the touch screen, configured to receive input from the user. The instructional sequence can include a set of repair instructions for repairing the object. The method can further include receiving a voice command from the user to navigate the instructional sequence with a microphone of the second mobile device. The method can further include: transmitting data indicative of completion of the instructional sequence from the second mobile device to the first mobile device; and receiving user input related to the completed instructional sequence from the user with the first mobile device. The method can further include: capturing a second sequence of images with a camera of the first mobile device; detecting the object within the second sequence of images; and

displaying a plurality of instructional sequences associated with the detected object with a display of the first mobile device, wherein the selection of the instructional sequence received from the user is in response to displaying the plurality of instructional sequences associated with the detected object. The displaying the instructional sequence can include: establishing a connection between the second mobile device and a remote individual; receiving, at the second mobile device, at least a portion of the instructional sequence via the connection with the remote individual; and displaying the received portion of the instructional sequence overlaid on the first sequence of images with the display of the second mobile device. The received portion of the instructional sequence can include an annotation anchored to the tracked object. The method can further include receiving administrative data from the user associated with the selected instructional sequence with the first mobile device prior to transmitting the data indicative of the selected instructional sequence to the second mobile device.

[0182] A head mounted display (HMD) can include: a display configured to display images; a camera; a processor; and a memory in communication with the processor and having stored thereon computer-executable instructions to cause the processor to: receive data indicative of a selected instructional sequence from a mobile device, the instructional sequence being received at the mobile device based on a user selection, capture a first sequence of images with the camera in response to receiving the data indicative of the selected instructional sequence, detect an object within the first sequence of images, the detected object being identified by the selected instructional sequence, track the object within the first sequence of images, and display the instructional sequence overlaid on the first sequence of images with the display. The mobile device can include a touch screen configured to receive input from the user, and the HMD can include an auxiliary input device, different from the touch screen, configured to receive input from the user. The HMD can include a semi-transparent display screen configured to display the images in a semi-transparent manner. The HMD can further include a microphone, wherein the memory further has stored thereon computer-executable instructions to cause the processor to receive a voice command from the user to navigate the instructional sequence with the microphone. The memory can further have stored thereon computer-executable instructions to cause the processor to: establish a connection with a remote individual, receive at least a portion of the instructional sequence via the connection with the remote individual, and display the received portion of the instructional sequence overlaid on the first sequence of images with the display.

[0183] A non-transitory computer readable storage medium can have stored thereon instructions that, when executed by a processor, cause a HMD to: receive data indicative of a selected instructional sequence from a mobile device, the instructional sequence being received at the mobile device based on a user selection; capture a first sequence of images with a camera of the HMD in response to receiving the data indicative of the selected instructional sequence; detect an object within the first sequence of images, the detected object being identified by the selected instructional sequence; track the object within the first sequence of images; and display the instructional sequence overlaid on the first sequence of images with a display of the HMD. The mobile device can include a touch screen configured to receive input from the user and the HMD can include an auxiliary input device, different from the touch screen, configured to receive input from the user. The instructional sequence can include a set of repair instructions for repairing the tracked object. The non-transitory computer readable storage medium can further have stored thereon instructions that, when executed by the processor, cause the HMD to receive a voice command from the user to navigate the instructional sequence with a microphone of the HMD. The non-transitory computer readable storage can further have stored thereon instructions that, when executed by the processor, cause the HMD to: establish a connection with a remote individual; receive at least a portion of the instructional sequence via the connection with the remote individual; and display the received portion of the instructional sequence overlaid on the first sequence of images with the display.

[0184] A method of person-to-person communication session playback can include: initiating a person-to-person communication session between a first device operated by a first user and a second device operated by a second user; capturing a live video stream with a camera of the first device during the communication session; sending at least a portion of the live video stream from the first device to the second device; receiving with the first device at least one audio-visual communication specified by the second user during the communication session; providing the audio-visual communication to the first user via an output device of the first device; and storing the audio-visual communication for playback to a third user of a third device outside of a person-to-person communication session. The audio-visual communication can be stored independent of a viewing position of the first device. The method can further include: generating a three-dimensional (3D) reconstruction of a first environment of the first device from the live video stream during the communication session; and storing the 3D reconstruction of the first environment for the playback to the third user. The 3D reconstruction of the first environment can include a plurality of feature points derived from the live video stream, the feature points enabling the tracking of a second environment during playback of the audio-visual communication with the third device. The method can further include: detecting a first object within the live video stream; and anchoring the audio-visual communication to the detected first object, wherein the storing of the audio-visual communication includes storing an indication of the anchoring of the audio-visual communication to the detected first object independent of the live video stream. The method can further include: initiating a playback session for the third user outside of the person-to-person communication session; capturing an image with a camera of the third device; detecting a second object with the captured image, the second object being the same as or similar to the first object; and displaying the audio-visual communication anchored to the detected second object. The audio-visual communication can include at least one of: a verbal instructions and a graphical annotation. The method can further include initiating a playback session for the third user in a virtual environment outside of the person-to-person communication session. The person-to-person communication session can include an augmented reality video session. The method can further include: detecting a period of time during the communication session in which the first device does not receive any audio-visual communication specified by the second

user; and marking the period of time such that the period of time is not included in playback of the communication session.

[0185] A first device for person-to-person communication session playback can include: a camera; an output device; a processor; and a memory in communication with the processor and having stored thereon computer-executable instructions to cause the processor to: initiate a person-to-person communication session between the first device operated by a first user and a second device operated by a second user; capture a live video stream with the camera during the communication session; send at least a portion of the live video stream to the second device; receive at least one audio-visual communication specified by the second user during the communication session; provide the audio-visual communication to the first user via the output device; and store the audio-visual communication in the memory for playback to a third user of a third device outside of a person-to-person communication session. The audio-visual communication can be stored independent of a viewing position of the first device. The memory can further have stored thereon computer-executable instructions to cause the processor to: generate a three-dimensional (3D) reconstruction of a first environment of the first device from the live video stream during the communication session; and store the 3D reconstruction of the first environment in the memory for the playback to the third user. The 3D reconstruction of the first environment can include a plurality of feature points derived from the live video stream, the feature points enabling the tracking of a second environment during playback of the audio-visual communication with the third device. The memory can further have stored thereon computer-executable instructions to cause the processor to: detect a first object within the live video stream; and anchor the audio-visual communication to the detected first object, wherein the storing of the audio-visual communication includes storing an indication of the anchoring of the audio-visual communication to the detected first object independent of the live video stream.

[0186] A non-transitory computer readable storage medium can have stored thereon instructions that, when executed by a processor, cause a first device to: initiate a person-to-person communication session between the first device operated by a first user and a second device operated by a second user; capture a live video stream with a camera of the first device during the communication session; send at least a portion of the live video stream from the first device to the second device; receive at least one audio-visual communication specified by the second user during the communication session; provide the audio-visual communication to the first user via an output device of the first device; and store the audio-visual communication in the memory for playback to a third user of a third device outside of a person-to-person communication session. The audio-visual communication can be stored independent of a viewing position of the first device. The non-transitory computer readable storage medium can further have stored thereon instructions that, when executed by the processor, cause the first device to: generate a three-dimensional (3D) reconstruction of a first environment of the first device from the live video stream during the communication session; and store the 3D reconstruction of the first environment in the memory for the playback to the third user. The 3D reconstruction of the first environment can include a plurality of feature points derived from the live video stream, the feature points enabling the tracking of a second environment during playback of the audio-visual communication with the third device. The non-transitory computer readable storage medium can further have stored thereon instructions that, when executed by the processor, cause the first device to: detect a first object within the live video stream; and anchor the audio-visual communication to the detected first object, wherein the storing of the audio-visual communication includes storing an indication of the anchoring of the audio-visual communication to the detected first object independent of the live video stream.

CONCLUDING COMMENTS

[0187] Implementations disclosed herein provide systems, methods and apparatus for initiating communication between two electronic devices based on the detection of an object as well as for other disclosed purposes.

[0188] The methods disclosed herein comprise one or more steps or actions for achieving described methods. Method steps and/or actions may be interchanged with one another without departing from the scope of the claims. In other words, unless a specific order of steps or actions is required for proper operation of the method that is being described, the order and/or use of specific steps and/or actions may be modified without departing from the scope of the claims.

[0189] As used herein, the term "plurality" denotes two or more. For example, a plurality of components indicates two or more components. The term "determining" encompasses a wide variety of actions and, therefore, "determining" can include calculating, computing, processing, deriving, investigating, looking up (e.g., looking up in a table, a database or another data structure), ascertaining and the like. Also, "determining" can include receiving (e.g., receiving information), accessing (e.g., accessing data in a memory) and the like. Also, "determining" can include resolving, selecting, choosing, establishing and the like. The phrase "based on" does not mean "based only on," unless expressly specified otherwise. In other words, the phrase "based on" describes both "based only on" and "based at least on."

[0190] Although the subject matter has been described in terms of certain embodiments, other embodiments, including embodiments which may or may not provide various features and advantages set forth herein will be apparent to those of ordinary skill in the art in view of the foregoing disclosure. The specific embodiments described above are disclosed as examples only, and the scope of the patented subject matter is defined by the claims that follow.

[0191] In the claims, the terms "based upon" or "based on" shall include situations in which a factor is taken into account directly and/or indirectly, and possibly in conjunction with other factors, in producing a result or effect. In the claims, a portion shall include greater than none and up to the whole of a thing.

1. A method performed by a portable computing device having a processing unit, a memory, and a camera, the method comprising:

capturing an image with the camera;

detecting an object within the captured image based on a database record for the object within a database of objects;

based on the database record, retrieving an identifier for initiating a live communication with a remote party;

in response to the retrieving the identifier, presenting, in association with an indication that the object has been detected, a user option to initiate the live communication with the remote party; and

in response to receiving user input, initiating the live communication between the device and the remote party based on the identifier.

2. The method of claim **1**, wherein the live communication comprises one or more of:

voice, video, and text messaging.

3. The method of claim **1**, wherein the live communication comprises a live video stream captured by the camera of the device, the method further comprising:

transmitting first augmented reality (AR) graphics generated by the device to the remote party along with the live video stream to be displayed by a remote computer system;

receiving second AR graphics from the remote party via the remote computer system; and

displaying the second AR graphics overlaid on the live video stream captured by the camera of the device on a display of the device.

4. The method of claim **1**, wherein detecting the object comprises:

determining a shape of the object based on the captured image; and

matching the shape of the object to a stored shape associated with the record for the object.

5. The method of claim **1**, wherein detecting the object comprises:

scanning a machine-readable code disposed on the object within the image; and

dereferencing the machine-readable code using the database.

6. The method of claim **1**, further comprising:

based on the database record, displaying on a display of the device an expected location of a graphical marking on the object as augmented reality graphics overlaid on a live video stream captured by the camera of the device;

scanning the graphical marking with the camera, wherein the graphical marking encodes additional information about the object; and

transmitting the additional information to the remote party.

7. The method of claim **1**, wherein the database further comprises user instructions for interacting with the object associated with the record for the detected object, the method further comprising displaying the user instructions via the display.

8. The method of claim **1**, wherein the database is located on a remote server.

9. The method of claim **1**, wherein the database is located on the device.

10. A method performed by a system including a first portable computing device having a processing unit, a memory, and a camera, the method comprising:

the first device capturing a first image of at least a portion of an object;

accessing a first database having a plurality of records storing characteristic data for identifying objects based on captured images;

locating a first matching record within the first database based on the first image;

accessing a second database having a plurality of records storing identifiers for initiating communication sessions;

using the first matching record within the first database to locate a second matching record in the second database;

retrieving an identifier for initiating a communication session from the second matching record; and

using the identifier for initiating a communication session to initiate a live communication session between the first device and a second device.

11. The method of claim **10**, wherein the first matching record comprises an identifier of the object, and wherein the second matching record is located based on the identifier of the object.

12. The method of claim **10**, wherein the object comprises a machine-readable code captured in the first image.

13. The method of claim **12**, wherein the machine-readable code has a format defined in the first matching record, the method further comprising decoding the machine-readable code using the format.

14. The method of claim **10**, wherein the communication session comprises the first device sending a live video stream, captured by the first device, to the second device.

15. The method of claim **14**, further comprising:

capturing a second image of at least a portion of the object with the first device, wherein the second image comprises a machine-readable code;

decoding the machine-readable code; and

the first device transmitting the decoded machine-readable code to the second device within the communication session.

16. A method comprising:

initiating an augmented reality (AR) video session using a camera and a display on a first device operated by a first user;

identifying an object within a field of view of the camera;

performing motion tracking using the identified object within the field of view of the camera during the AR video session;

accessing a database having a plurality of records storing identifiers for initiating communication sessions;

using the identification of the object to locate a matching record in the database;

obtaining an identifier for initiating a communication session from the matching record;

based on the identifier for initiating a communication session, presenting to the first user, during the AR video session, an option to initiate a person-to-person communication session; and

in response to selection of the option by the user, using the identifier for initiating a communication session to initiate a person-to-person communication session between the first user operating the first device and a second user operating a second device.

17. The method of claim **16**, wherein the communication session comprises the first device sending to the second device a live video stream representing the AR video session.

18. The method of claim **17**, wherein the communication session comprises the second device sending AR graphics, specified by the second user during the communication session, to the first device, wherein the AR graphics are displayed on the first device in motion relative to the object in the field of view.

19. The method of claim **16**, wherein the object comprises a machine-readable code, and wherein the identifying the object is performed using the machine-readable code.

20. The method of claim **19**, wherein the motion tracking is performed using the machine-readable code.

* * * * *