

US Patent & Trademark Office

Patent Public Search | Text View

United States Patent Application Publication

20250252699

Kind Code

A1

Publication Date

August 07, 2025

Inventor(s)

Velasquez; Miguel Andres Granados et al.

CROSS REALITY SYSTEM WITH FAST LOCALIZATION

Abstract

A cross reality system enables any of multiple devices to efficiently and accurately access previously persisted maps, even maps of very large environments, and render virtual content specified in relation to those maps. The cross reality system may quickly process a batch of images acquired with a portable device to determine whether there is sufficient consistency across the batch in the computed localization. Processing on at least one image from the batch may determine a rough localization of the device to the map. This rough localization result may be used in a refined localization process for the image for which it was generated. The rough localization result may also be selectively propagated to a refined localization process for other images in the batch, enabling rough localization processing to be skipped for the other images.

Inventors: Velasquez; Miguel Andres Granados (Thalwil, CH), Gomez Gonzalez; Javier Victorio (Zurich, CH), Hu; Danying (Sunnyvale, CA), Guendelman; Eran (Tel Aviv, IL), Shahrokni; Ali (San Jose, CA), Swaminathan; Ashwin (Dublin, CA), Prasad; Mukta (Uitikon Waldegg, CH)

Applicant: Magic Leap, Inc. (Plantation, FL)

Family ID: 77365305

Assignee: Magic Leap, Inc. (Plantation, FL)

Appl. No.: 19/189479

Filed: April 25, 2025

Related U.S. Application Data

parent US continuation 18077200 20221207 parent-grant-document US 12315097 child US 19189479

parent US continuation 17185741 20210225 parent-grant-document US 11551430 child US 18077200

us-provisional-application US 62981939 20200226

Publication Classification

Int. Cl.: **G06T19/20** (20110101); **G06T7/38** (20170101); **G06T15/00** (20110101); **G06T19/00** (20110101)

U.S. Cl.:

CPC **G06T19/20** (20130101); **G06T7/38** (20170101); **G06T15/005** (20130101); **G06T19/006** (20130101); **G06T2219/2004** (20130101)

Background/Summary

CROSS-REFERENCE TO RELATED APPLICATIONS [0001] This application is a continuation of U.S. patent application Ser. No. 18/077,200, filed on Dec. 7, 2022, entitled “CROSS REALITY SYSTEM WITH FAST LOCALIZATION,” which is a continuation of U.S. patent application Ser. No. 17/185,741, now U.S. Pat. No. 11,551,430, filed on Feb. 25, 2021, entitled “CROSS REALITY SYSTEM WITH FAST LOCALIZATION,” which claims priority to and the benefit under 35 U.S.C. § 119 (e) of U.S. Provisional Patent Application Ser. No. 62/981,939, filed on Feb. 26, 2020, entitled “CROSS REALITY SYSTEM WITH FAST LOCALIZATION.” The contents of these applications are incorporated herein by reference in their entirety.

TECHNICAL FIELD

[0002] This application relates generally to a cross reality system.

BACKGROUND

[0003] Computers may control human user interfaces to create a cross reality (XR) environment in which some or all of the XR environment, as perceived by the user, is generated by the computer. These XR environments may be virtual reality (VR), augmented reality (AR), and mixed reality (MR) environments, in which some or all of an XR environment may be generated by computers using, in part, data that describes the environment. This data may describe, for example, virtual objects that may be rendered in a way that users' sense or perceive as a part of a physical world and can interact with the virtual objects. The user may experience these virtual objects as a result of the data being rendered and presented through a user interface device, such as, for example, a head-mounted display device. The data may be displayed to the user to see, or may control audio that is played for the user to hear, or may control a tactile (or haptic) interface, enabling the user to experience touch sensations that the user senses or perceives as feeling the virtual object.

[0004] XR systems may be useful for many applications, spanning the fields of scientific visualization, medical training, engineering design and prototyping, tele-manipulation and tele-presence, and personal entertainment. AR and MR, in contrast to VR, include one or more virtual objects in relation to real objects of the physical world. The experience of virtual objects interacting with real objects greatly enhances the user's enjoyment in using the XR system, and also opens the door for a variety of applications that present realistic and readily understandable information about how the physical world might be altered.

[0005] To realistically render virtual content, an XR system may build a representation of the physical world around a user of the system. This representation, for example, may be constructed by processing images acquired with sensors on a wearable device that forms a part of the XR system. In such a system, a user might perform an initialization routine by looking around a room or other physical environment in which the user intends to use the XR system until the system acquires sufficient information to construct a representation of that environment. As the system operates and the user moves around the environment or to other environments, the sensors on the

wearable devices might acquire additional information to expand or update the representation of the physical world.

BRIEF SUMMARY

[0006] Aspects of the present application relate to methods and apparatus for providing cross reality (XR) scenes. Techniques as described herein may be used together, separately, or in any suitable combination.

[0007] In some aspects, a system that supports specification of a position of virtual content relative to one or more persisted maps in a database of persisted maps is provided. The system comprises a localization service comprising a communication component configured to receive from a portable electronic device information about a plurality of collections of features in images of a three-dimensional (3D) environment of the portable electronic device; and at least one processor configured to execute computer-executable instructions. The computer-executable instructions implement a localization component, connected to the communication component, the localization component configured to: compute a plurality of candidate localizations for the plurality of collections of features by, for a first collection of features of the plurality of collections of features, perform rough localization of the first collection of features with respect to one or more persisted maps in the database of persisted maps, wherein performing rough localization comprises computing a first rough transformation of the portable electronic device with respect to a persisted map in the database of persisted maps. For the first collection of features, refined localization of the first collection of features with respect to the persisted map is performed, wherein performing refined localization of the first collection of features comprises computing a first candidate localization of the portable electronic device based on the first rough transformation. For a second collection of features of the plurality of collections of features refined localization of the second collection of features with respect to the persisted map is performed, wherein performing refined localization of the second collection of features comprises computing a second candidate localization of the portable electronic device based on the first rough transformation. A localization based on consensus among the plurality of candidate localizations, including the first candidate localization and the second candidate localization, is selected.

[0008] In some embodiments, performing rough localization further comprises identifying a first set of features among the first collection of features, identifying a second set of features in the persisted map, wherein features of the first set of features correspond to respective features of the second set of features based on similarity of at least one characteristic.

[0009] In some embodiments, the at least one characteristic of a feature comprises a feature descriptor computed from pixels in an image depicting the feature.

[0010] In some embodiments, the feature descriptor comprises a deep sparse feature (DSF).

[0011] In some embodiments, performing rough localization further comprises, for a plurality of repetitions, such that a plurality of candidate rough transformations and associated measures of error are computed, selecting a first subset of the first set of features and a second subset of the second set of features and determining a candidate rough transformation between the first subset of features and the second subset of features based on a measure of alignment error between the first subset transformed with the candidate transformation and the second subset, selecting a selected candidate rough transformation from among the plurality of candidate rough transformations based on associated measures of alignment error, and providing the first rough transformation based on the selected candidate rough transformation.

[0012] In some embodiments, the selected localization comprises a localization transformation between a local coordinate frame of the portable electronic device and a coordinate frame of the persisted map, and the localization component is further configured to send the localization transformation to the portable electronic device.

[0013] In some embodiments, the plurality of collections of features each comprises a posed feature rig.

[0014] In some embodiments, the posed feature rig comprises a plurality of feature descriptors and a pose associated with each of the plurality of features descriptors.

[0015] In some embodiments, the computer-executable instructions implementing the localization component further comprise instructions for determining whether to perform refined localization of the second collection of features based on satisfaction of one or more criteria.

[0016] In some embodiments, the computer-executable instructions implementing the localization component further comprise instructions for performing refined localization of the second collection of features based on a second rough transformation computed from the second collection of features when the one or more criteria are not satisfied.

[0017] In some embodiments, the one or more criteria comprise parallax above a threshold amount between an image from which the first collection of features was extracted and an image from which the second collection of features was extracted.

[0018] In some embodiments, the one or more criteria comprise computing a confidence in the first candidate localization that is above a first threshold.

[0019] In some embodiments, computing the first candidate localization comprises rejecting as the first candidate localization localizations with a confidence less than a second threshold, wherein the first threshold is higher than the second threshold.

[0020] In some embodiments, the computer-executable instructions to select the localization based on consensus among the plurality of candidate localizations computed for the plurality of collections of features determine whether there is consensus above a first threshold among the plurality of candidate localizations computed for the plurality of collections of features, including the second candidate localization, and based on a determined consensus below the first threshold, re-computing candidate localizations for collections of features of the plurality of collections of features for which the candidate localization was computed based on a rough transformation from a different collection of features.

[0021] In some embodiments, the computer-executable instructions to select the localization based on consensus among the plurality of candidate localizations computed for the plurality of collections of features determine whether there is consensus above a second threshold among the plurality of candidate localizations computed for the plurality of collections of features, wherein the plurality of candidate localizations comprises the re-computed candidate localizations, wherein the first threshold is higher than the second threshold.

[0022] In some embodiments, determining whether there is consensus above the first threshold comprises comparing a mode of the plurality of candidate localizations to the first threshold.

[0023] In some embodiments, performing rough localization for the first collection of features further comprises selecting a sub-region the persistent map based on location metadata associated with the first collection of features and the selected sub-region of the persistent map.

[0024] In some embodiments, in combination with a portable electronic device configured to operate within a cross reality system, the portable electronic device having a device coordinate frame, wherein the portable electronic device comprises one or more sensors configured to capture information about the 3D environment, the captured information comprising a plurality of images, and at least one second processor configured to execute second computer executable instructions, wherein the second computer executable instructions comprise instructions for generating the information about the plurality of collections of features by for each image of the plurality of images, extracting a plurality of features and associating with each of the plurality of features pose information in the device coordinate frame, sending the information about the plurality of collections of features over a network to a localization service, comprising sending information about the plurality of extracted features and the associated pose information, and receiving from the localization service a transformation relating the device coordinate frame to a second coordinate frame.

[0025] In some embodiments, the portable electronic device further comprises a buffer, and the

computer executable instructions of the portable electronic device further comprise instructions for storing in the buffer, for each of the plurality of images, information about the extracted collection of features and the associated pose information and sending the information about the plurality of collections of features over the network, comprising sending contents of the buffer.

[0026] In some embodiments, the portable electronic device comprises a display, and the computer-executable instructions of the portable electronic device further comprise instructions for rendering virtual content having a location specified in the second coordinate frame on the display in a position computed based, at least in part, on the received transformation.

[0027] In some embodiments, the information about the plurality of extracted features comprises descriptors for individual features.

[0028] In some aspects, a method of computing a pose between a first component and a map, wherein a pose of the first component is represented by a plurality of collections of features posed with respect to a coordinate frame of the first component is performed. The method comprises computing a plurality of candidate localizations based on respective collections of features of the plurality of collections of features; and determining a localization based on the plurality of candidate localizations. Computing the plurality of candidate localizations comprises, for a first portion of the plurality of collections of features, computing a first rough localization for a collection of features in the first portion of the plurality of collections of features; and computing, as a first candidate localization of the plurality of candidate localizations, a first refined localization based on the first rough localization. For a second portion of the plurality of collections of features, one or more criteria are evaluated to determine whether to compute, as a second candidate localization of the plurality of candidate localizations, a second refined localization based on the first rough localization.

[0029] In some embodiments, the method further comprises, for a collection of features in the second portion of the plurality of collections of features for which it is determined to compute the second refined localization based on the first rough localization, computing, as the second candidate localization of the plurality of candidate localizations, the second refined localization based on the first rough localization.

[0030] In some embodiments, the method further comprises, for a collection of features in the second portion of the plurality of collections of features for which it is determined not to compute the second refined localization based on the first rough localization, computing a second rough localization for the collection of features in the second portion of the plurality of collections of features for which it is determined not to compute the second refined localization based on the first rough localization, and computing, as the second candidate localization of the plurality of candidate localizations, the second refined localization based on the second rough localization.

[0031] In some embodiments, evaluating the one or more criteria to determine whether to compute the second refined localization based on the first rough localization comprises determining whether to compute the second refined localization based on the first rough localization based on parallax between an image from which the collection of features in the first portion was extracted and an image from which the collection of features in the second portion was extracted being above a threshold.

[0032] In some embodiments, evaluating the one or more criteria to determine whether to compute the second refined localization based on the first rough localization comprises determining whether to compute the second refined localization based on the first rough localization based on a measure of error for the first refined localization for the collection of features in the first portion being below a threshold.

[0033] In some embodiments, determining the localization based on the plurality of candidate localizations comprises evaluating the plurality of candidate localizations for consensus based on consensus criteria dependent on whether the second candidate localization for the collection of features in the second portion was computed as the second refined localization based on the first

rough localization.

[0034] In some embodiments, the plurality of collections of features comprise a batch of posed feature rigs received from a portable electronic device.

[0035] In some embodiments, the plurality of collections of features comprise posed feature rigs in a tracking map formed by a portable electronic device.

[0036] In some embodiments, selecting the map from a database of stored maps based on location metadata included with the posed feature rigs of the batch.

[0037] In some embodiments, computing the first rough localization for a collection of features in the first portion comprises identifying a set of features in the collection of features in the first portion that correspond to features in the map, repetitively forming subsets of the set of corresponding features and determining a candidate rough transformation for each subset and an associated accuracy, and selecting a rough transformation from among the repetitively formed candidate rough transformations based on a determined accuracy of the selected rough transformation.

[0038] In some embodiments, computing the first refined localization based on the first rough localization comprises searching for a transformation that aligns the set of features identified in the collection of features in the first portion and the corresponding features in the map, wherein the searching is constrained around the rough transformation computed for the first rough localization.

[0039] In some aspects, a non-transitory computer-readable medium comprising computer-executable instructions configured to, when executed by at least one processor, perform a method of computing a pose between a first component and a map is provided. A pose of the first component is represented by a plurality of collections of features posed with respect to a coordinate frame of the first component. The method comprises computing a plurality of candidate localizations based on respective collections of features of the plurality of collections of features; and determining a localization based on the plurality of candidate localizations. Computing the plurality of candidate localizations comprises, for a first portion of the plurality of collections of features, computing a first rough localization for a collection of features in the first portion of the plurality of collections of features; and computing, as a first candidate localization of the plurality of candidate localizations, a first refined localization based on the first rough localization. For a second portion of the plurality of collections of features, one or more criteria are evaluated to determine whether to compute, as a second candidate localization of the plurality of candidate localizations, a second refined localization based on the first rough localization.

[0040] The foregoing summary is provided by way of illustration and is not intended to be limiting.

Description

BRIEF DESCRIPTION OF THE DRAWINGS

[0041] The accompanying drawings are not intended to be drawn to scale. In the drawings, each identical or nearly identical component that is illustrated in various figures is represented by a like numeral. For purposes of clarity, not every component may be labeled in every drawing. In the drawings:

[0042] FIG. 1 is a sketch illustrating an example of a simplified augmented reality (AR) scene, according to some embodiments;

[0043] FIG. 2 is a sketch of an exemplary simplified AR scene, showing exemplary use cases of an XR system, according to some embodiments;

[0044] FIG. 3 is a schematic diagram illustrating data flow for a single user in an AR system configured to provide an experience to the user of AR content interacting with a physical world, according to some embodiments;

[0045] FIG. 4 is a schematic diagram illustrating an exemplary AR display system, displaying

virtual content for a single user, according to some embodiments;

[0046] FIG. 5A is a schematic diagram illustrating a user wearing an AR display system rendering AR content as the user moves through a physical world environment, according to some embodiments;

[0047] FIG. 5B is a schematic diagram illustrating a viewing optics assembly and attendant components, according to some embodiments;

[0048] FIG. 6A is a schematic diagram illustrating an AR system using a world reconstruction system, according to some embodiments;

[0049] FIG. 6B is a schematic diagram illustrating components of an AR system that maintain a model of a passable world, according to some embodiments;

[0050] FIG. 7 is a schematic illustration of a tracking map formed by a device traversing a path through a physical world;

[0051] FIG. 8 is a schematic diagram illustrating a user of a cross reality (XR) system, perceiving virtual content, according to some embodiments;

[0052] FIG. 9 is a block diagram of components of a first XR device of the XR system of FIG. 8 that transform between coordinate systems, according to some embodiments;

[0053] FIG. 10 is a schematic diagram illustrating an exemplary transformation of origin coordinate frames into destination coordinate frames in order to correctly render local XR content, according to some embodiments;

[0054] FIG. 11 is a top plan view illustrating pupil-based coordinate frames, according to some embodiments;

[0055] FIG. 12 is a top plan view illustrating a camera coordinate frame that includes all pupil positions, according to some embodiments;

[0056] FIG. 13 is a schematic diagram of the display system of FIG. 9, according to some embodiments;

[0057] FIG. 14 is a block diagram illustrating the creation of a persistent coordinate frame (PCF) and the attachment of XR content to the PCF, according to some embodiments;

[0058] FIG. 15 is a flow chart illustrating a method of establishing and using a PCF, according to some embodiments;

[0059] FIG. 16 is a block diagram of the XR system of FIG. 8, including a second XR device, according to some embodiments;

[0060] FIG. 17 is a schematic diagram illustrating a room and key frames that are established for various areas in the room, according to some embodiments;

[0061] FIG. 18 is a schematic diagram illustrating the establishment of persistent poses based on the key frames, according to some embodiments;

[0062] FIG. 19 is a schematic diagram illustrating the establishment of a persistent coordinate frame (PCF) based on the persistent poses, according to some embodiments;

[0063] FIGS. 20A to 20C are schematic diagrams illustrating an example of creating PCFs, according to some embodiments;

[0064] FIG. 21 is a block diagram illustrating a system for generating global descriptors for individual images and/or maps, according to some embodiments;

[0065] FIG. 22 is a flow chart illustrating a method of computing an image descriptor, according to some embodiments;

[0066] FIG. 23 is a flow chart illustrating a method of localization using image descriptors, according to some embodiments;

[0067] FIG. 24 is a flow chart illustrating a method of training a neural network, according to some embodiments;

[0068] FIG. 25 is a block diagram illustrating a method of training a neural network, according to some embodiments;

[0069] FIG. 26 is a schematic diagram illustrating an AR system configured to rank and merge a

plurality of environment maps, according to some embodiments;

[0070] FIG. **27** is a simplified block diagram illustrating a plurality of canonical maps stored on a remote storage medium, according to some embodiments;

[0071] FIG. **28** is a schematic diagram illustrating a method of selecting canonical maps to, for example, localize a new tracking map in one or more canonical maps and/or obtain PCFs from the canonical maps, according to some embodiments;

[0072] FIG. **29** is flow chart illustrating a method of selecting a plurality of ranked environment maps, according to some embodiments;

[0073] FIG. **30** is a schematic diagram illustrating an exemplary map rank portion of the AR system of FIG. **26**, according to some embodiments;

[0074] FIG. **31A** is a schematic diagram illustrating an example of area attributes of a tracking map (TM) and environment maps in a database, according to some embodiments;

[0075] FIG. **31B** is a schematic diagram illustrating an example of determining a geographic location of a tracking map (TM) for geolocation filtering of FIG. **29**, according to some embodiments;

[0076] FIG. **32** is a schematic diagram illustrating an example of geolocation filtering of FIG. **29**, according to some embodiments;

[0077] FIG. **33** is a schematic diagram illustrating an example of Wi-Fi BSSID filtering of FIG. **29**, according to some embodiments;

[0078] FIG. **34** is a schematic diagram illustrating an example of use of localization of FIG. **29**, according to some embodiments;

[0079] FIGS. **35** and **36** are block diagrams of an XR system configured to rank and merge a plurality of environment maps, according to some embodiments.

[0080] FIG. **37** is a block diagram illustrating a method of creating environment maps of a physical world, in a canonical form, according to some embodiments;

[0081] FIGS. **38A** and **38B** are schematic diagrams illustrating an environment map created in a canonical form by updating the tracking map of FIG. **7** with a new tracking map, according to some embodiments.

[0082] FIGS. **39A** to **39F** are schematic diagrams illustrating an example of merging maps, according to some embodiments;

[0083] FIG. **40** is a two-dimensional representation of a three-dimensional first local tracking map (Map 1), which may be generated by the first XR device of FIG. **9**, according to some embodiments;

[0084] FIG. **41** is a block diagram illustrating uploading Map 1 from the first XR device to the server of FIG. **9**, according to some embodiments;

[0085] FIG. **42** is a schematic diagram illustrating the XR system of FIG. **16**, showing the second user has initiated a second session using a second XR device of the XR system after the first user has terminated a first session, according to some embodiments;

[0086] FIG. **43A** is a block diagram illustrating a new session for the second XR device of FIG. **42**, according to some embodiments;

[0087] FIG. **43B** is a block diagram illustrating the creation of a tracking map for the second XR device of FIG. **42**, according to some embodiments;

[0088] FIG. **43C** is a block diagram illustrating downloading a canonical map from the server to the second XR device of FIG. **42**, according to some embodiments;

[0089] FIG. **44** is a schematic diagram illustrating a localization attempt to localize to a canonical map a second tracking map (Map 2), which may be generated by the second XR device of FIG. **42**, according to some embodiments;

[0090] FIG. **45** is a schematic diagram illustrating a localization attempt to localize to a canonical map the second tracking map (Map 2) of FIG. **44**, which may be further developed and with XR content associated with PCFs of Map 2, according to some embodiments;

[0091] FIGS. **46A-46B** are a schematic diagram illustrating a successful localization of Map 2 of FIG. **45** to the canonical map, according to some embodiments;

[0092] FIG. **47** is a schematic diagram illustrating a canonical map generated by including one or more PCFs from the canonical map of FIG. **46A** into Map 2 of FIG. **45**, according to some embodiments;

[0093] FIG. **48** is a schematic diagram illustrating the canonical map of FIG. **47** with further expansion of Map 2 on the second XR device, according to some embodiments;

[0094] FIG. **49** is a block diagram illustrating uploading Map 2 from the second XR device to the server, according to some embodiments;

[0095] FIG. **50** is a block diagram illustrating merging Map 2 with the canonical map, according to some embodiments;

[0096] FIG. **51** is a block diagram illustrating transmission of a new canonical map from the server to the first and second XR devices, according to some embodiments;

[0097] FIG. **52** is a block diagram illustrating a two-dimensional representation of Map 2 and a head coordinate frame of the second XR device that is referenced to Map 2, according to some embodiments;

[0098] FIG. **53** is a block diagram illustrating, in two-dimensions, adjustment of the head coordinate frame which can occur in six degrees of freedom, according to some embodiments;

[0099] FIG. **54** is a block diagram illustrating a canonical map on the second XR device wherein sound is localized relative to PCFs of Map 2, according to some embodiments;

[0100] FIGS. **55** and **56** are a perspective view and a block diagram illustrating use of the XR system when the first user has terminated a first session and the first user has initiated a second session using the XR system, according to some embodiments;

[0101] FIGS. **57** and **58** are a perspective view and a block diagram illustrating use of the XR system when three users are simultaneously using the XR system in the same session, according to some embodiments;

[0102] FIG. **59** is a flow chart illustrating a method of recovering and resetting a headpose, according to some embodiments;

[0103] FIG. **60** is a block diagram of a machine in the form of a computer that can find application in the present invention system, according to some embodiments;

[0104] FIG. **61** is a schematic diagram of an example XR system in which any of multiple devices may access a localization service, according to some embodiments;

[0105] FIG. **62** is an example process flow for operation of a portable device as part of an XR system that provides cloud-based localization, according to some embodiments;

[0106] FIGS. **63A, B, and C** are an example process flow for cloud-based localization, according to some embodiments;

[0107] FIG. **64** is a block diagram of a localization component of an XR system, according to some embodiments; and

[0108] FIG. **65** is alternative embodiment of a localization component of an XR system, according to some alternative embodiments.

DETAILED DESCRIPTION

[0109] Described herein are methods and apparatus for providing XR scenes. To provide realistic XR experiences to multiple users, an XR system must know the users' location within the physical world in order to correctly correlate locations of virtual objects to real objects. The inventors have recognized and appreciated methods and apparatus that are computationally efficient and quick in localizing XR devices, even in large and very large scale environments (e.g., a neighborhood, a city, a country, the globe).

[0110] An XR system may build a map of an environment in which user devices may operate. The environment map may be created from image information collected with sensors that are part of XR devices worn by users of the XR system. Each XR device may develop a local map of its

physical environment by integrating information from one or more images collected as the device operates. In some embodiments, the coordinate system of the local map is tied to the position and/or orientation of the device when the device first initiates scanning the physical world (e.g. starts a new session). That position and/or orientation of the device may change from session to session as a user interacts with the XR system, whether different sessions are associated with different users, each with their own wearable device with sensors that scan the environment, or the same user who uses the same device at different times.

[0111] The XR system may implement one or more techniques to enable persistent operation across sessions based on persistent spatial information. The techniques, for example, may provide XR scenes for a more computationally efficient and immersive experiences for a single or multiple users by enabling persistent spatial information to be created, stored, and retrieved by any of multiple users of an XR system. When shared by multiple users, persistent spatial information provides a more immersive experience as it enables multiple users to experience virtual content in the same location with respect to the physical world. Even when used by a single user, persistent spatial information may enable quickly recovering and resetting headposes on an XR device in a computationally efficient way.

[0112] The persistent spatial information may be represented by a persistent map. The persistent map may be stored in a remote storage medium (e.g., a cloud). A wearable device worn by a user, after being turned on, may retrieve from persistent storage an appropriate map that was previously created and stored. That previously stored map may have been based on data about the environment collected with sensors on the user's wearable device during prior sessions. Retrieving a stored map may enable use of the wearable device without completing a scan of the physical world with the sensors on the wearable device. Alternatively or additionally, the device, upon entering a new region of the physical world, may similarly retrieve an appropriate stored map.

[0113] The stored map may be represented in a canonical form to which a local frame of reference on each XR device may be related. In a multi-device XR system, the stored map accessed by one device may have been created and stored by another device and/or may have been constructed by aggregating data about the physical world collected by sensors on multiple wearable devices that were previously present in at least a portion of the physical world represented by the stored map.

[0114] In some embodiments, persistent spatial information may be represented in a way that may be readily shared among users and among the distributed components, including applications. Canonical maps may provide information about the physical world, which may be formatted, for example, as persistent coordinate frames (PCFs). A PCF may be defined based on a set of features recognized in the physical world. The features may be selected such that they are likely to be the same from user session to user session of the XR system. PCFs may be sparse, providing less than all of the available information about the physical world, such that they may be efficiently processed and transferred.

[0115] Techniques for processing persistent spatial information also may include creating dynamic maps based on the local coordinate systems of one or more devices. These maps may be sparse maps, representing the physical world with feature points detected in images used in forming the maps. Canonical maps may be formed by merging multiple such maps created by one or more XR devices.

[0116] The relationship between a canonical map and a local map for each device may be determined through a localization process. That localization process may be performed on each XR device based on a set of canonical maps selected and sent to the device. Alternatively or additionally, a localization service may be provided on remote processors, such as might be implemented in the cloud.

[0117] Two XR devices that have access to the same stored map, for example, may both localize with respect to the stored map. Once localized, a user device may render virtual content that has a location specified by reference to the stored map by translating that location to a frame of reference

maintained by the user device. The user device may use this local frame of reference to control the display of the user device to render the virtual content in the specified location.

[0118] The XR system may be configured to create, share, and use persistent spatial information with low usage of computational resources and/or low latency to provide a more immersive user experience. To support these operations, the system may use techniques for efficient comparison of spatial information. Such comparisons may arise, for example, as part of localization in which a collection of features from a local device is matched to a collection of features in a canonical map. Similarly, in map merge, attempts may be made to match one or more collections of features in a tracking map from a device to corresponding features in a canonical map.

[0119] Map selection may aid in efficient comparison of spatial information. By selecting a set of candidate maps from a universe of stored maps for comparison to spatial information provided from a user device, the number of computationally intensive comparison operations that are performed to identify a best match may be reduced. In some embodiments, location metadata may be associated with maps, and/or portions of maps, that might be compared. Candidate maps may be quickly selected based on matching location metadata. Computationally intensive feature matching may then be limited to only the candidate maps, which requires less computation than were comparisons required to an entire universe of stored maps.

[0120] In some embodiments, comparisons of collections of features also may be streamlined, such as through the use of feature descriptors. The descriptors may have numerical values, assigned by a trained neural network, that enable comparison of the features. Features that likely represent the same portion of an object in the physical world are assigned feature descriptors of similar values, such that matching features may be quickly identified based on descriptors with similar values.

[0121] In scenarios in which collections of features are derived from image frames, further processing simplifications can be made by providing a simple mechanism to find similar frames. Comparisons of sets of features may be restricted to just collections of features derived from similar frames. Finding similar image frames may also be simplified by representing image frames in toto by numeric descriptors. The descriptors may be computed via a transformation that maps a collection of features identified in the image to the descriptor.

[0122] Even with these, and possibly other, techniques to reduce the number of comparisons to determine whether two collections of features match, at least some computationally intensive comparisons of collections of features may be performed. For example, two frames with matching frame descriptors may be determined to match only after finding correspondence, with a sufficiently low error, between sets of features in those frames. Finding correspondence between sets of features with low error may serve as confirmation that a collection of features received from a local device matches a stored map. Alternatively or additionally, correspondence may be based on alignment of sets of feature points so that a transformation between the coordinate frames for those sets of feature points can be determined.

[0123] To reduce processing time spent searching for alignment between collections of features, a rough localization between two collections may be performed. The rough localization may be used to bound the search space for a refined localization. In this way, both the rough localization and refined localization may be computationally simple relative to a full search for a transformation that aligns a full collection of features with a matching collection of features.

[0124] The inventors have recognized and appreciated techniques for further reduction in processing to match spatial information, such as occurs during a localization process. To ensure sufficient accuracy, localization may be performed based on a batch of image frames, each represented by a collection of features. The localization result may be determined based on consensus among the localization results for the batch.

[0125] A reduction in processing time, with suitable accuracy in localization result, may be achieved by selectively performing a refined localization for some collections in the batch based on rough localization for other collections in the batch. Propagating the rough localization result of a

first collection to compute a refined localization for a second collection of features can avoid the processing required to compute the rough localization for the second collection.

[0126] The inventors have further recognized and appreciated criteria that are predictive of suitable accuracy of a localization result when processing is simplified by propagating rough localization results computed for some collections of features to compute refined localization results for other sets of features. Where those criteria are not met, localization may be performed without propagating a rough localization for one set in refined localization of another set.

[0127] Techniques as described herein may be used together or separately with many types of devices and for many types of scenes, including wearable or portable devices with limited computational resources that provide an augmented or mixed reality scene. In some embodiments, the techniques may be implemented by one or more services that form a portion of an XR system.

AR System Overview

[0128] FIGS. 1 and 2 illustrate scenes with virtual content displayed in conjunction with a portion of the physical world. For purposes of illustration, an AR system is used as an example of an XR system. FIGS. 3-6B illustrate an exemplary AR system, including one or more processors, memory, sensors and user interfaces that may operate according to the techniques described herein.

[0129] Referring to FIG. 1, an outdoor AR scene 354 is depicted in which a user of an AR technology sees a physical world park-like setting 356, featuring people, trees, buildings in the background, and a concrete platform 358. In addition to these items, the user of the AR technology also perceives that they “see” a robot statue 357 standing upon the physical world concrete platform 358, and a cartoon-like avatar character 352 flying by which seems to be a personification of a bumble bee, even though these elements (e.g., the avatar character 352, and the robot statue 357) do not exist in the physical world. Due to the extreme complexity of the human visual perception and nervous system, it is challenging to produce an AR technology that facilitates a comfortable, natural-feeling, rich presentation of virtual image elements amongst other virtual or physical world imagery elements.

[0130] Such an AR scene may be achieved with a system that builds maps of the physical world based on tracking information, enable users to place AR content in the physical world, determine locations in the maps of the physical world where AR content are placed, preserve the AR scenes such that the placed AR content can be reloaded to display in the physical world during, for example, a different AR experience session, and enable multiple users to share an AR experience. The system may build and update a digital representation of the physical world surfaces around the user. This representation may be used to render virtual content so as to appear fully or partially occluded by physical objects between the user and the rendered location of the virtual content, to place virtual objects, in physics based interactions, and for virtual character path planning and navigation, or for other operations in which information about the physical world is used.

[0131] FIG. 2 depicts another example of an indoor AR scene 400, showing exemplary use cases of an XR system, according to some embodiments. The exemplary scene 400 is a living room having walls, a bookshelf on one side of a wall, a floor lamp at a corner of the room, a floor, a sofa, and coffee table on the floor. In addition to these physical items, the user of the AR technology also perceives virtual objects such as images on the wall behind the sofa, birds flying through the door, a deer peeking out from the book shelf, and a decoration in the form of a windmill placed on the coffee table.

[0132] For the images on the wall, the AR technology requires information about not only surfaces of the wall but also objects and surfaces in the room such as lamp shape, which are occluding the images to render the virtual objects correctly. For the flying birds, the AR technology requires information about all the objects and surfaces around the room for rendering the birds with realistic physics to avoid the objects and surfaces or bounce off them if the birds collide. For the deer, the AR technology requires information about the surfaces such as the floor or coffee table to compute where to place the deer. For the windmill, the system may identify that is an object separate from

the table and may determine that it is movable, whereas corners of shelves or corners of the wall may be determined to be stationary. Such a distinction may be used in determinations as to which portions of the scene are used or updated in each of various operations.

[0133] The virtual objects may be placed in a previous AR experience session. When new AR experience sessions start in the living room, the AR technology requires the virtual objects being accurately displayed at the locations previously placed and realistically visible from different viewpoints. For example, the windmill should be displayed as standing on the books rather than drifting above the table at a different location without the books. Such drifting may happen if the locations of the users of the new AR experience sessions are not accurately localized in the living room. As another example, if a user is viewing the windmill from a viewpoint different from the viewpoint when the windmill was placed, the AR technology requires corresponding sides of the windmill being displayed.

[0134] A scene may be presented to the user via a system that includes multiple components, including a user interface that can stimulate one or more user senses, such as sight, sound, and/or touch. In addition, the system may include one or more sensors that may measure parameters of the physical portions of the scene, including position and/or motion of the user within the physical portions of the scene. Further, the system may include one or more computing devices, with associated computer hardware, such as memory. These components may be integrated into a single device or may be distributed across multiple interconnected devices. In some embodiments, some or all of these components may be integrated into a wearable device.

[0135] FIG. 3 depicts an AR system **502** configured to provide an experience of AR contents interacting with a physical world **506**, according to some embodiments. The AR system **502** may include a display **508**. In the illustrated embodiment, the display **508** may be worn by the user as part of a headset such that a user may wear the display over their eyes like a pair of goggles or glasses. At least a portion of the display may be transparent such that a user may observe a see-through reality **510**. The see-through reality **510** may correspond to portions of the physical world **506** that are within a present viewpoint of the AR system **502**, which may correspond to the viewpoint of the user in the case that the user is wearing a headset incorporating both the display and sensors of the AR system to acquire information about the physical world.

[0136] AR contents may also be presented on the display **508**, overlaid on the see-through reality **510**. To provide accurate interactions between AR contents and the see-through reality **510** on the display **508**, the AR system **502** may include sensors **522** configured to capture information about the physical world **506**.

[0137] The sensors **522** may include one or more depth sensors that output depth maps **512**. Each depth map **512** may have multiple pixels, each of which may represent a distance to a surface in the physical world **506** in a particular direction relative to the depth sensor. Raw depth data may come from a depth sensor to create a depth map. Such depth maps may be updated as fast as the depth sensor can form a new image, which may be hundreds or thousands of times per second. However, that data may be noisy and incomplete, and have holes shown as black pixels on the illustrated depth map.

[0138] The system may include other sensors, such as image sensors. The image sensors may acquire monocular or stereoscopic information that may be processed to represent the physical world in other ways. For example, the images may be processed in world reconstruction component **516** to create a mesh, representing connected portions of objects in the physical world. Metadata about such objects, including for example, color and surface texture, may similarly be acquired with the sensors and stored as part of the world reconstruction.

[0139] The system may also acquire information about the headpose (or “pose”) of the user with respect to the physical world. In some embodiments, a headpose tracking component of the system may be used to compute headposes in real time. The headpose tracking component may represent a headpose of a user in a coordinate frame with six degrees of freedom including, for example,

translation in three perpendicular axes (e.g., forward/backward, up/down, left/right) and rotation about the three perpendicular axes (e.g., pitch, yaw, and roll). In some embodiments, sensors **522** may include inertial measurement units that may be used to compute and/or determine a headpose **514**. A headpose **514** for a depth map may indicate a present viewpoint of a sensor capturing the depth map with six degrees of freedom, for example, but the headpose **514** may be used for other purposes, such as to relate image information to a particular portion of the physical world or to relate the position of the display worn on the user's head to the physical world.

[0140] In some embodiments, the headpose information may be derived in other ways than from an IMU, such as from analyzing objects in an image. For example, the headpose tracking component may compute relative position and orientation of an AR device to physical objects based on visual information captured by cameras and inertial information captured by IMUs. The headpose tracking component may then compute a headpose of the AR device by, for example, comparing the computed relative position and orientation of the AR device to the physical objects with features of the physical objects. In some embodiments, that comparison may be made by identifying features in images captured with one or more of the sensors **522** that are stable over time such that changes of the position of these features in images captured over time can be associated with a change in headpose of the user.

[0141] The inventors have realized and appreciated techniques for operating XR systems to provide XR scenes for a more immersive user experience such as estimating headpose at a frequency of 1 kHz, with low usage of computational resources in connection with an XR device, that may be configured with, for example, four video graphic array (VGA) cameras operating at 30 Hz, one inertial measurement unit (IMU) operating at 1 kHz, compute power of a single advanced RISC machine (ARM) core, memory less than 1 GB, and network bandwidth less than 100 Mbp. These techniques relate to reducing processing required to generate and maintain maps and estimate headpose as well as to providing and consuming data with low computational overhead. The XR system may calculate its pose based on the matched visual features. U.S. patent application Ser. No. 16/221,065 describes hybrid tracking and is hereby incorporated herein by reference in its entirety.

[0142] In some embodiments, the AR device may construct a map from the feature points recognized in successive images in a series of image frames captured as a user moves throughout the physical world with the AR device. Though each image frame may be taken from a different pose as the user moves, the system may adjust the orientation of the features of each successive image frame to match the orientation of the initial image frame by matching features of the successive image frames to previously captured image frames. Translations of the successive image frames so that points representing the same features will match corresponding feature points from previously collected image frames, can be used to align each successive image frame to match the orientation of previously processed image frames. The frames in the resulting map may have a common orientation established when the first image frame was added to the map. This map, with sets of feature points in a common frame of reference, may be used to determine the user's pose within the physical world by matching features from current image frames to the map. In some embodiments, this map may be called a tracking map.

[0143] In addition to enabling tracking of the user's pose within the environment, this map may enable other components of the system, such as world reconstruction component **516**, to determine the location of physical objects with respect to the user. The world reconstruction component **516** may receive the depth maps **512** and headposes **514**, and any other data from the sensors, and integrate that data into a reconstruction **518**. The reconstruction **518** may be more complete and less noisy than the sensor data. The world reconstruction component **516** may update the reconstruction **518** using spatial and temporal averaging of the sensor data from multiple viewpoints over time.

[0144] The reconstruction **518** may include representations of the physical world in one or more data formats including, for example, voxels, meshes, planes, etc. The different formats may

represent alternative representations of the same portions of the physical world or may represent different portions of the physical world. In the illustrated example, on the left side of the reconstruction **518**, portions of the physical world are presented as a global surface; on the right side of the reconstruction **518**, portions of the physical world are presented as meshes.

[0145] In some embodiments, the map maintained by headpose component **514** may be sparse relative to other maps that might be maintained of the physical world. Rather than providing information about locations, and possibly other characteristics, of surfaces, the sparse map may indicate locations of interest points and/or structures, such as corners or edges. In some embodiments, the map may include image frames as captured by the sensors **522**. These frames may be reduced to features, which may represent the interest points and/or structures. In conjunction with each frame, information about a pose of a user from which the frame was acquired may also be stored as part of the map. In some embodiments, every image acquired by the sensor may or may not be stored. In some embodiments, the system may process images as they are collected by sensors and select subsets of the image frames for further computation. The selection may be based on one or more criteria that limits the addition of information yet ensures that the map contains useful information. The system may add a new image frame to the map, for example, based on overlap with a prior image frame already added to the map or based on the image frame containing a sufficient number of features determined as likely to represent stationary objects. In some embodiments, the selected image frames, or groups of features from selected image frames may serve as key frames for the map, which are used to provide spatial information.

[0146] In some embodiments, the amount of data that is processed when constructing maps may be reduced, such as by constructing sparse maps with a collection of mapped points and keyframes and/or dividing the maps into blocks to enable updates by blocks. A mapped point may be associated with a point of interest in the environment. A keyframe may include selected information from camera-captured data. U.S. patent application Ser. No. 16/520,582 describes determining and/or evaluating localization maps and is hereby incorporated herein by reference in its entirety.

[0147] The AR system **502** may integrate sensor data over time from multiple viewpoints of a physical world. The poses of the sensors (e.g., position and orientation) may be tracked as a device including the sensors is moved. As the sensor's frame pose is known and how it relates to the other poses, each of these multiple viewpoints of the physical world may be fused together into a single, combined reconstruction of the physical world, which may serve as an abstract layer for the map and provide spatial information. The reconstruction may be more complete and less noisy than the original sensor data by using spatial and temporal averaging (i.e. averaging data from multiple viewpoints over time), or any other suitable method.

[0148] In the illustrated embodiment in FIG. **3**, a map represents the portion of the physical world in which a user of a single, wearable device is present. In that scenario, headpose associated with frames in the map may be represented as a local headpose, indicating orientation relative to an initial orientation for a single device at the start of a session. For example, the headpose may be tracked relative to an initial headpose when the device was turned on or otherwise operated to scan an environment to build a representation of that environment.

[0149] In combination with content characterizing that portion of the physical world, the map may include metadata. The metadata, for example, may indicate time of capture of the sensor information used to form the map. Metadata alternatively or additionally may indicate location of the sensors at the time of capture of information used to form the map. Location may be expressed directly, such as with information from a GPS chip, or indirectly, such as with a wireless (e.g. Wi-Fi) signature indicating strength of signals received from one or more wireless access points while the sensor data was being collected and/or with identifiers, such as BSSID's, of wireless access points to which the user device connected while the sensor data was collected.

[0150] The reconstruction **518** may be used for AR functions, such as producing a surface

representation of the physical world for occlusion processing or physics-based processing. This surface representation may change as the user moves or objects in the physical world change. Aspects of the reconstruction **518** may be used, for example, by a component **520** that produces a changing global surface representation in world coordinates, which may be used by other components.

[0151] The AR content may be generated based on this information, such as by AR applications **504**. An AR application **504** may be a game program, for example, that performs one or more functions based on information about the physical world, such as visual occlusion, physics-based interactions, and environment reasoning. It may perform these functions by querying data in different formats from the reconstruction **518** produced by the world reconstruction component **516**. In some embodiments, component **520** may be configured to output updates when a representation in a region of interest of the physical world changes. That region of interest, for example, may be set to approximate a portion of the physical world in the vicinity of the user of the system, such as the portion within the view field of the user, or is projected (predicted/determined) to come within the view field of the user.

[0152] The AR applications **504** may use this information to generate and update the AR contents. The virtual portion of the AR contents may be presented on the display **508** in combination with the see-through reality **510**, creating a realistic user experience.

[0153] In some embodiments, an AR experience may be provided to a user through an XR device, which may be a wearable display device, which may be part of a system that may include remote processing and or remote data storage and/or, in some embodiments, other wearable display devices worn by other users. FIG. 4 illustrates an example of system **580** (hereinafter referred to as “system **580**”) including a single wearable device for simplicity of illustration. The system **580** includes a head mounted display device **562** (hereinafter referred to as “display device **562**”), and various mechanical and electronic modules and systems to support the functioning of the display device **562**. The display device **562** may be coupled to a frame **564**, which is wearable by a display system user or viewer **560** (hereinafter referred to as “user **560**”) and configured to position the display device **562** in front of the eyes of the user **560**. According to various embodiments, the display device **562** may be a sequential display. The display device **562** may be monocular or binocular. In some embodiments, the display device **562** may be an example of the display **508** in FIG. 3.

[0154] In some embodiments, a speaker **566** is coupled to the frame **564** and positioned proximate an ear canal of the user **560**. In some embodiments, another speaker, not shown, is positioned adjacent another ear canal of the user **560** to provide for stereo/shapeable sound control. The display device **562** is operatively coupled, such as by a wired lead or wireless connectivity **568**, to a local data processing module **570** which may be mounted in a variety of configurations, such as fixedly attached to the frame **564**, fixedly attached to a helmet or hat worn by the user **560**, embedded in headphones, or otherwise removably attached to the user **560** (e.g., in a backpack-style configuration, in a belt-coupling style configuration).

[0155] The local data processing module **570** may include a processor, as well as digital memory, such as non-volatile memory (e.g., flash memory), both of which may be utilized to assist in the processing, caching, and storage of data. The data include data a) captured from sensors (which may be, e.g., operatively coupled to the frame **564**) or otherwise attached to the user **560**, such as image capture devices (such as cameras), microphones, inertial measurement units, accelerometers, compasses, GPS units, radio devices, and/or gyros; and/or b) acquired and/or processed using remote processing module **572** and/or remote data repository **574**, possibly for passage to the display device **562** after such processing or retrieval.

[0156] In some embodiments, the wearable device may communicate with remote components. The local data processing module **570** may be operatively coupled by communication links **576**, **578**, such as via a wired or wireless communication links, to the remote processing module **572** and

remote data repository **574**, respectively, such that these remote modules **572**, **574** are operatively coupled to each other and available as resources to the local data processing module **570**. In further embodiments, in addition or as alternative to remote data repository **574**, the wearable device can access cloud based remote data repositories, and/or services. In some embodiments, the headpose tracking component described above may be at least partially implemented in the local data processing module **570**. In some embodiments, the world reconstruction component **516** in FIG. 3 may be at least partially implemented in the local data processing module **570**. For example, the local data processing module **570** may be configured to execute computer executable instructions to generate the map and/or the physical world representations based at least in part on at least a portion of the data.

[0157] In some embodiments, processing may be distributed across local and remote processors. For example, local processing may be used to construct a map on a user device (e.g. tracking map) based on sensor data collected with sensors on that user's device. Such a map may be used by applications on that user's device. Additionally, previously created maps (e.g., canonical maps) may be stored in remote data repository **574**. Where a suitable stored or persistent map is available, it may be used instead of or in addition to the tracking map created locally on the device. In some embodiments, a tracking map may be localized to the stored map, such that a correspondence is established between a tracking map, which might be oriented relative to a position of the wearable device at the time a user turned the system on, and the canonical map, which may be oriented relative to one or more persistent features. In some embodiments, the persistent map might be loaded on the user device to allow the user device to render virtual content without a delay associated with scanning a location to build a tracking map of the user's full environment from sensor data acquired during the scan. In some embodiments, the user device may access a remote persistent map (e.g., stored on a cloud) without the need to download the persistent map on the user device.

[0158] In some embodiments, spatial information may be communicated from the wearable device to remote services, such as a cloud service that is configured to localize a device to stored maps maintained on the cloud service. According to one embodiment, the localization processing can take place in the cloud matching the device location to existing maps, such as canonical maps, and return transforms that link virtual content to the wearable device location. In such embodiments, the system can avoid communicating maps from remote resources to the wearable device. Other embodiments can be configured for both device-based and cloud-based localization, for example, to enable functionality where network connectivity is not available or a user opts not to enable cloud-based localization.

[0159] Alternatively or additionally, the tracking map may be merged with previously stored maps to extend or improve the quality of those maps. The processing to determine whether a suitable previously created environment map is available and/or to merge a tracking map with one or more stored environment maps may be done in local data processing module **570** or remote processing module **572**.

[0160] In some embodiments, the local data processing module **570** may include one or more processors (e.g., a graphics processing unit (GPU)) configured to analyze and process data and/or image information. In some embodiments, the local data processing module **570** may include a single processor (e.g., a single-core or multi-core ARM processor), which would limit the local data processing module **570**'s compute budget but enable a more miniature device. In some embodiments, the world reconstruction component **516** may use a compute budget less than a single Advanced RISC Machine (ARM) core to generate physical world representations in real-time on a non-predefined space such that the remaining compute budget of the single ARM core can be accessed for other uses such as, for example, extracting meshes.

[0161] In some embodiments, the remote data repository **574** may include a digital data storage facility, which may be available through the Internet or other networking configuration in a "cloud"

resource configuration. In some embodiments, all data is stored and all computations are performed in the local data processing module **570**, allowing fully autonomous use from a remote module. In some embodiments, all data is stored and all or most computations are performed in the remote data repository **574**, allowing for a smaller device. A world reconstruction, for example, may be stored in whole or in part in this repository **574**.

[0162] In embodiments in which data is stored remotely, and accessible over a network, data may be shared by multiple users of an augmented reality system. For example, user devices may upload their tracking maps to augment a database of environment maps. In some embodiments, the tracking map upload occurs at the end of a user session with a wearable device. In some embodiments, the tracking map uploads may occur continuously, semi-continuously, intermittently, at a pre-defined time, after a pre-defined period from the previous upload, or when triggered by an event. A tracking map uploaded by any user device may be used to expand or improve a previously stored map, whether based on data from that user device or any other user device. Likewise, a persistent map downloaded to a user device may be based on data from that user device or any other user device. In this way, high quality environment maps may be readily available to users to improve their experiences with the AR system.

[0163] In further embodiments, persistent map downloads can be limited and/or avoided based on localization executed on remote resources (e.g., in the cloud). In such configurations, a wearable device or other XR device communicates to the cloud service feature information coupled with pose information (e.g., positioning information for the device at the time the features represented in the feature information were sensed). One or more components of the cloud service may match the feature information to respective stored maps (e.g., canonical maps) and generates transforms between a tracking map maintained by the XR device and the coordinate system of the canonical map. Each XR device that has its tracking map localized with respect to the canonical map may accurately render virtual content in locations specified with respect to the canonical map based on its own tracking.

[0164] In some embodiments, the local data processing module **570** is operatively coupled to a battery **582**. In some embodiments, the battery **582** is a removable power source, such as over the counter batteries. In other embodiments, the battery **582** is a lithium-ion battery. In some embodiments, the battery **582** includes both an internal lithium-ion battery chargeable by the user **560** during non-operation times of the system **580** and removable batteries such that the user **560** may operate the system **580** for longer periods of time without having to be tethered to a power source to charge the lithium-ion battery or having to shut the system **580** off to replace batteries.

[0165] FIG. 5A illustrates a user **530** wearing an AR display system rendering AR content as the user **530** moves through a physical world environment **532** (hereinafter referred to as “environment **532**”). The information captured by the AR system along the movement path of the user may be processed into one or more tracking maps. The user **530** positions the AR display system at positions **534**, and the AR display system records ambient information of a passable world (e.g., a digital representation of the real objects in the physical world that can be stored and updated with changes to the real objects in the physical world) relative to the positions **534**. That information may be stored as poses in combination with images, features, directional audio inputs, or other desired data. The positions **534** are aggregated to data inputs **536**, for example, as part of a tracking map, and processed at least by a passable world module **538**, which may be implemented, for example, by processing on a remote processing module **572** of FIG. 4. In some embodiments, the passable world module **538** may include the headpose component **514** and the world reconstruction component **516**, such that the processed information may indicate the location of objects in the physical world in combination with other information about physical objects used in rendering virtual content.

[0166] The passable world module **538** determines, at least in part, where and how AR content **540** can be placed in the physical world as determined from the data inputs **536**. The AR content is

“placed” in the physical world by presenting via the user interface both a representation of the physical world and the AR content, with the AR content rendered as if it were interacting with objects in the physical world and the objects in the physical world presented as if the AR content were, when appropriate, obscuring the user's view of those objects. In some embodiments, the AR content may be placed by appropriately selecting portions of a fixed element **542** (e.g., a table) from a reconstruction (e.g., the reconstruction **518**) to determine the shape and position of the AR content **540**. As an example, the fixed element may be a table and the virtual content may be positioned such that it appears to be on that table. In some embodiments, the AR content may be placed within structures in a field of view **544**, which may be a present field of view or an estimated future field of view. In some embodiments, the AR content may be persisted relative to a model **546** of the physical world (e.g. a mesh).

[0167] As depicted, the fixed element **542** serves as a proxy (e.g. digital copy) for any fixed element within the physical world which may be stored in the passable world module **538** so that the user **530** can perceive content on the fixed element **542** without the system having to map to the fixed element **542** each time the user **530** sees it. The fixed element **542** may, therefore, be a mesh model from a previous modeling session or determined from a separate user but nonetheless stored by the passable world module **538** for future reference by a plurality of users. Therefore, the passable world module **538** may recognize the environment **532** from a previously mapped environment and display AR content without a device of the user **530** mapping all or part of the environment **532** first, saving computation process and cycles and avoiding latency of any rendered AR content.

[0168] The mesh model **546** of the physical world may be created by the AR display system and appropriate surfaces and metrics for interacting and displaying the AR content **540** can be stored by the passable world module **538** for future retrieval by the user **530** or other users without the need to completely or partially recreate the model. In some embodiments, the data inputs **536** are inputs such as geolocation, user identification, and current activity to indicate to the passable world module **538** which fixed element **542** of one or more fixed elements are available, which AR content **540** has last been placed on the fixed element **542**, and whether to display that same content (such AR content being “persistent” content regardless of user viewing a particular passable world model).

[0169] Even in embodiments in which objects are considered to be fixed (e.g. a kitchen table), the passable world module **538** may update those objects in a model of the physical world from time to time to account for the possibility of changes in the physical world. The model of fixed objects may be updated with a very low frequency. Other objects in the physical world may be moving or otherwise not regarded as fixed (e.g. kitchen chairs). To render an AR scene with a realistic feel, the AR system may update the position of these non-fixed objects with a much higher frequency than is used to update fixed objects. To enable accurate tracking of all of the objects in the physical world, an AR system may draw information from multiple sensors, including one or more image sensors.

[0170] FIG. **5B** is a schematic illustration of a viewing optics assembly **548** and attendant components. In some embodiments, two eye tracking cameras **550**, directed toward user eyes **549**, detect metrics of the user eyes **549**, such as eye shape, eyelid occlusion, pupil direction and glint on the user eyes **549**.

[0171] In some embodiments, one of the sensors may be a depth sensor **551**, such as a time of flight sensor, emitting signals to the world and detecting reflections of those signals from nearby objects to determine distance to given objects. A depth sensor, for example, may quickly determine whether objects have entered the field of view of the user, either as a result of motion of those objects or a change of pose of the user. However, information about the position of objects in the field of view of the user may alternatively or additionally be collected with other sensors. Depth information, for example, may be obtained from stereoscopic visual image sensors or plenoptic

sensors.

[0172] In some embodiments, world cameras **552** record a greater-than-peripheral view to map and/or otherwise create a model of the environment **532** and detect inputs that may affect AR content. In some embodiments, the world camera **552** and/or camera **553** may be grayscale and/or color image sensors, which may output grayscale and/or color image frames at fixed time intervals. Camera **553** may further capture physical world images within a field of view of the user at a specific time. Pixels of a frame-based image sensor may be sampled repetitively even if their values are unchanged. Each of the world cameras **552**, the camera **553** and the depth sensor **551** have respective fields of view of **554**, **555**, and **556** to collect data from and record a physical world scene, such as the physical world environment **532** depicted in FIG. **34A**.

[0173] Inertial measurement units **557** may determine movement and orientation of the viewing optics assembly **548**. In some embodiments, inertial measurement units **557** may provide an output indicating a direction of gravity. In some embodiments, each component is operatively coupled to at least one other component. For example, the depth sensor **551** is operatively coupled to the eye tracking cameras **550** as a confirmation of measured accommodation against actual distance the user eyes **549** are looking at.

[0174] It should be appreciated that a viewing optics assembly **548** may include some of the components illustrated in FIG. **34B** and may include components instead of or in addition to the components illustrated. In some embodiments, for example, a viewing optics assembly **548** may include two world camera **552** instead of four. Alternatively or additionally, cameras **552** and **553** need not capture a visible light image of their full field of view. A viewing optics assembly **548** may include other types of components. In some embodiments, a viewing optics assembly **548** may include one or more dynamic vision sensor (DVS), whose pixels may respond asynchronously to relative changes in light intensity exceeding a threshold.

[0175] In some embodiments, a viewing optics assembly **548** may not include the depth sensor **551** based on time of flight information. In some embodiments, for example, a viewing optics assembly **548** may include one or more plenoptic cameras, whose pixels may capture light intensity and an angle of the incoming light, from which depth information can be determined. For example, a plenoptic camera may include an image sensor overlaid with a transmissive diffraction mask (TDM). Alternatively or additionally, a plenoptic camera may include an image sensor containing angle-sensitive pixels and/or phase-detection auto-focus pixels (PDAF) and/or micro-lens array (MLA). Such a sensor may serve as a source of depth information instead of or in addition to depth sensor **551**.

[0176] It also should be appreciated that the configuration of the components in FIG. **5B** is provided as an example. A viewing optics assembly **548** may include components with any suitable configuration, which may be set to provide the user with the largest field of view practical for a particular set of components. For example, if a viewing optics assembly **548** has one world camera **552**, the world camera may be placed in a center region of the viewing optics assembly instead of at a side.

[0177] Information from the sensors in viewing optics assembly **548** may be coupled to one or more of processors in the system. The processors may generate data that may be rendered so as to cause the user to perceive virtual content interacting with objects in the physical world. That rendering may be implemented in any suitable way, including generating image data that depicts both physical and virtual objects. In other embodiments, physical and virtual content may be depicted in one scene by modulating the opacity of a display device that a user looks through at the physical world. The opacity may be controlled so as to create the appearance of the virtual object and also to block the user from seeing objects in the physical world that are occluded by the virtual objects. In some embodiments, the image data may only include virtual content that may be modified such that the virtual content is perceived by a user as realistically interacting with the physical world (e.g. clip content to account for occlusions), when viewed through the user

interface.

[0178] The location on the viewing optics assembly **548** at which content is displayed to create the impression of an object at a particular location may depend on the physics of the viewing optics assembly. Additionally, the pose of the user's head with respect to the physical world and the direction in which the user's eyes are looking may impact where in the physical world content displayed at a particular location on the viewing optics assembly content will appear. Sensors as described above may collect this information, and or supply information from which this information may be calculated, such that a processor receiving sensor inputs may compute where objects should be rendered on the viewing optics assembly **548** to create a desired appearance for the user.

[0179] Regardless of how content is presented to a user, a model of the physical world may be used so that characteristics of the virtual objects, which can be impacted by physical objects, including the shape, position, motion, and visibility of the virtual object, can be correctly computed. In some embodiments, the model may include the reconstruction of a physical world, for example, the reconstruction **518**.

[0180] That model may be created from data collected from sensors on a wearable device of the user. Though, in some embodiments, the model may be created from data collected by multiple users, which may be aggregated in a computing device remote from all of the users (and which may be “in the cloud”).

[0181] The model may be created, at least in part, by a world reconstruction system such as, for example, the world reconstruction component **516** of FIG. 3 depicted in more detail in FIG. 6A. The world reconstruction component **516** may include a perception module **660** that may generate, update, and store representations for a portion of the physical world. In some embodiments, the perception module **660** may represent the portion of the physical world within a reconstruction range of the sensors as multiple voxels. Each voxel may correspond to a 3D cube of a predetermined volume in the physical world, and include surface information, indicating whether there is a surface in the volume represented by the voxel. Voxels may be assigned values indicating whether their corresponding volumes have been determined to include surfaces of physical objects, determined to be empty or have not yet been measured with a sensor and so their value is unknown. It should be appreciated that values indicating that voxels that are determined to be empty or unknown need not be explicitly stored, as the values of voxels may be stored in computer memory in any suitable way, including storing no information for voxels that are determined to be empty or unknown.

[0182] In addition to generating information for a persisted world representation, the perception module **660** may identify and output indications of changes in a region around a user of an AR system. Indications of such changes may trigger updates to volumetric data stored as part of the persisted world, or trigger other functions, such as triggering components **604** that generate AR content to update the AR content.

[0183] In some embodiments, the perception module **660** may identify changes based on a signed distance function (SDF) model. The perception module **660** may be configured to receive sensor data such as, for example, depth maps **660a** and headposes **660b**, and then fuse the sensor data into a SDF model **660c**. Depth maps **660a** may provide SDF information directly, and images may be processed to arrive at SDF information. The SDF information represents distance from the sensors used to capture that information. As those sensors may be part of a wearable unit, the SDF information may represent the physical world from the perspective of the wearable unit and therefore the perspective of the user. The headposes **660b** may enable the SDF information to be related to a voxel in the physical world.

[0184] In some embodiments, the perception module **660** may generate, update, and store representations for the portion of the physical world that is within a perception range. The perception range may be determined based, at least in part, on a sensor's reconstruction range,

which may be determined based, at least in part, on the limits of a sensor's observation range. As a specific example, an active depth sensor that operates using active IR pulses may operate reliably over a range of distances, creating the observation range of the sensor, which may be from a few centimeters or tens of centimeters to a few meters.

[0185] The world reconstruction component **516** may include additional modules that may interact with the perception module **660**. In some embodiments, a persisted world module **662** may receive representations for the physical world based on data acquired by the perception module **660**. The persisted world module **662** also may include various formats of representations of the physical world. For example, volumetric metadata **662b** such as voxels may be stored as well as meshes **662c** and planes **662d**. In some embodiments, other information, such as depth maps could be saved.

[0186] In some embodiments, representations of the physical world, such as those illustrated in FIG. **6A** may provide relatively dense information about the physical world in comparison to sparse maps, such as a tracking map based on feature points as described above.

[0187] In some embodiments, the perception module **660** may include modules that generate representations for the physical world in various formats including, for example, meshes **660d**, planes and semantics **660e**. The representations for the physical world may be stored across local and remote storage mediums. The representations for the physical world may be described in different coordinate frames depending on, for example, the location of the storage medium. For example, a representation for the physical world stored in the device may be described in a coordinate frame local to the device. The representation for the physical world may have a counterpart stored in a cloud. The counterpart in the cloud may be described in a coordinate frame shared by all devices in an XR system.

[0188] In some embodiments, these modules may generate representations based on data within the perception range of one or more sensors at the time the representation is generated as well as data captured at prior times and information in the persisted world module **662**. In some embodiments, these components may operate on depth information captured with a depth sensor. However, the AR system may include vision sensors and may generate such representations by analyzing monocular or binocular vision information.

[0189] In some embodiments, these modules may operate on regions of the physical world. Those modules may be triggered to update a subregion of the physical world, when the perception module **660** detects a change in the physical world in that subregion. Such a change, for example, may be detected by detecting a new surface in the SDF model **660c** or other criteria, such as changing the value of a sufficient number of voxels representing the subregion.

[0190] The world reconstruction component **516** may include components **664** that may receive representations of the physical world from the perception module **660**. Information about the physical world may be pulled by these components according to, for example, a use request from an application. In some embodiments, information may be pushed to the use components, such as via an indication of a change in a pre-identified region or a change of the physical world representation within the perception range. The components **664**, may include, for example, game programs and other components that perform processing for visual occlusion, physics-based interactions, and environment reasoning.

[0191] Responding to the queries from the components **664**, the perception module **660** may send representations for the physical world in one or more formats. For example, when the component **664** indicates that the use is for visual occlusion or physics-based interactions, the perception module **660** may send a representation of surfaces. When the component **664** indicates that the use is for environmental reasoning, the perception module **660** may send meshes, planes and semantics of the physical world.

[0192] In some embodiments, the perception module **660** may include components that format information to provide the component **664**. An example of such a component may be raycasting

component **660f**. A use component (e.g., component **664**), for example, may query for information about the physical world from a particular point of view. Raycasting component **660f** may select from one or more representations of the physical world data within a field of view from that point of view.

[0193] As should be appreciated from the foregoing description, the perception module **660**, or another component of an AR system, may process data to create 3D representations of portions of the physical world. Data to be processed may be reduced by culling parts of a 3D reconstruction volume based at least in part on a camera frustum and/or depth image, extracting and persisting plane data, capturing, persisting, and updating 3D reconstruction data in blocks that allow local update while maintaining neighbor consistency, providing occlusion data to applications generating such scenes, where the occlusion data is derived from a combination of one or more depth data sources, and/or performing a multi-stage mesh simplification. The reconstruction may contain data of different levels of sophistication including, for example, raw data such as live depth data, fused volumetric data such as voxels, and computed data such as meshes.

[0194] In some embodiments, components of a passable world model may be distributed, with some portions executing locally on an XR device and some portions executing remotely, such as on a network connected server, or otherwise in the cloud. The allocation of the processing and storage of information between the local XR device and the cloud may impact functionality and user experience of an XR system. For example, reducing processing on a local device by allocating processing to the cloud may enable longer battery life and reduce heat generated on the local device. But, allocating too much processing to the cloud may create undesirable latency that causes an unacceptable user experience.

[0195] FIG. **6B** depicts a distributed component architecture **600** configured for spatial computing, according to some embodiments. The distributed component architecture **600** may include a passable world component **602** (e.g., PW **538** in FIG. **5A**), a Lumin OS **604**, API's **606**, SDK **608**, and Application **610**. The Lumin OS **604** may include a Linux-based kernel with custom drivers compatible with an XR device. The API's **606** may include application programming interfaces that grant XR applications (e.g., Applications **610**) access to the spatial computing features of an XR device. The SDK **608** may include a software development kit that allows the creation of XR applications.

[0196] One or more components in the architecture **600** may create and maintain a model of a passable world. In this example sensor data is collected on a local device. Processing of that sensor data may be performed in part locally on the XR device and partially in the cloud. PW **538** may include environment maps created based, at least in part, on data captured by AR devices worn by multiple users. During sessions of an AR experience, individual AR devices (such as wearable devices described above in connection with FIG. **4** may create tracking maps, which is one type of map.

[0197] In some embodiments, the device may include components that construct both sparse maps and dense maps. A tracking map may serve as a sparse map and may include headposes of the AR device scanning an environment as well as information about objects detected within that environment at each headpose. Those headposes may be maintained locally for each device. For example, the headpose on each device may be relative to an initial headpose when the device was turned on for its session. As a result, each tracking map may be local to the device creating it and may have its own frame of reference defined by its own local coordinate system. In some embodiments, however, the tracking map on each device may be formed such that one coordinate of its local coordinate system is aligned with the direction of gravity as measured by its sensors, such as inertial measurement unit **557**.

[0198] The dense map may include surface information, which may be represented by a mesh or depth information. Alternatively or additionally, a dense map may include higher level information derived from surface or depth information, such as the location and/or characteristics of planes

and/or other objects.

[0199] Creation of the dense maps may be independent of the creation of sparse maps, in some embodiments. The creation of dense maps and sparse maps, for example, may be performed in separate processing pipelines within an AR system. Separating processing, for example, may enable generation or processing of different types of maps to be performed at different rates. Sparse maps, for example, may be refreshed at a faster rate than dense maps. In some embodiments, however, the processing of dense and sparse maps may be related, even if performed in different pipelines. Changes in the physical world revealed in a sparse map, for example, may trigger updates of a dense map, or vice versa. Further, even if independently created, the maps might be used together. For example, a coordinate system derived from a sparse map may be used to define position and/or orientation of objects in a dense map.

[0200] The sparse map and/or dense map may be persisted for re-use by the same device and/or sharing with other devices. Such persistence may be achieved by storing information in the cloud. The AR device may send the tracking map to a cloud to, for example, merge with environment maps selected from persisted maps previously stored in the cloud. In some embodiments, the selected persisted maps may be sent from the cloud to the AR device for merging. In some embodiments, the persisted maps may be oriented with respect to one or more persistent coordinate frames. Such maps may serve as canonical maps, as they can be used by any of multiple devices. In some embodiments, a model of a passable world may comprise or be created from one or more canonical maps. Devices, even though they perform some operations based on a coordinate frame local to the device, may nonetheless use the canonical map by determining a transformation between their coordinate frame local to the device and the canonical map.

[0201] A canonical map may originate as a tracking map (TM) (e.g., TM **1102** in FIG. **31A**), which may be promoted to a canonical map. The canonical map may be persisted such that devices that access the canonical map may, once determining a transformation between their local coordinate system and a coordinate system of the canonical map, use the information in the canonical map to determine locations of objects represented in the canonical map in the physical world around the device. In some embodiments, a TM may be a headpose sparse map created by an XR device. In some embodiments, the canonical map may be created when an XR device sends one or more TMs to a cloud server for merging with additional TMs captured by the XR device at a different time or by other XR devices.

[0202] In embodiments in which tracking maps are formed on local devices with one coordinate of a local coordinate frame aligned with gravity, this orientation with respect to gravity may be preserved upon creation of a canonical map. For example, when a tracking map that is submitted for merging does not overlap with any previously stored map, that tracking map may be promoted to a canonical map. Other tracking maps, which may also have an orientation relative to gravity, may be subsequently merged with that canonical map. The merging may be done so as to ensure that the resulting canonical map retains its orientation relative to gravity. Two maps, for example, may not be merged, regardless of correspondence of feature points in those maps, if coordinates of each map aligned with gravity do not align with each other with a sufficiently close tolerance.

[0203] The canonical maps, or other maps, may provide information about the portions of the physical world represented by the data processed to create respective maps. FIG. **7** depicts an exemplary tracking map **700**, according to some embodiments. The tracking map **700** may provide a floor plan **706** of physical objects in a corresponding physical world, represented by points **702**. In some embodiments, a map point **702** may represent a feature of a physical object that may include multiple features. For example, each corner of a table may be a feature that is represented by a point on a map. The features may be derived from processing images, such as may be acquired with the sensors of a wearable device in an augmented reality system. The features, for example, may be derived by processing an image frame output by a sensor to identify features based on large gradients in the image or other suitable criteria. Further processing may limit the number of

features in each frame. For example, processing may select features that likely represent persistent objects. One or more heuristics may be applied for this selection.

[0204] The tracking map **700** may include data on points **702** collected by a device. For each image frame with data points included in a tracking map, a pose may be stored. The pose may represent the orientation from which the image frame was captured, such that the feature points within each image frame may be spatially correlated. The pose may be determined by positioning information, such as may be derived from the sensors, such as an IMU sensor, on the wearable device.

Alternatively or additionally, the pose may be determined from matching image frames to other image frames that depict overlapping portions of the physical world. By finding such positional correlation, which may be accomplished by matching subsets of features points in two frames, the relative pose between the two frames may be computed. A relative pose may be adequate for a tracking map, as the map may be relative to a coordinate system local to a device established based on the initial pose of the device when construction of the tracking map was initiated.

[0205] Not all of the feature points and image frames collected by a device may be retained as part of the tracking map, as much of the information collected with the sensors is likely to be redundant. Rather, only certain frames may be added to the map. Those frames may be selected based on one or more criteria, such as degree of overlap with image frames already in the map, the number of new features they contain or a quality metric for the features in the frame. Image frames not added to the tracking map may be discarded or may be used to revise the location of features. As a further alternative, all or most of the image frames, represented as a set of features may be retained, but a subset of those frames may be designated as key frames, which are used for further processing.

[0206] The key frames may be processed to produce keyrigs **704**. The key frames may be processed to produce three dimensional sets of feature points and saved as keyrigs **704**. Such processing may entail, for example, comparing image frames derived simultaneously from two cameras to stereoscopically determine the 3D position of feature points. Metadata may be associated with these keyframes and/or keyrigs, such as poses.

[0207] The environment maps may have any of multiple formats depending on, for example, the storage locations of an environment map including, for example, local storage of AR devices and remote storage. For example, a map in remote storage may have higher resolution than a map in local storage on a wearable device where memory is limited. To send a higher resolution map from remote storage to local storage, the map may be down sampled or otherwise converted to an appropriate format, such as by reducing the number of poses per area of the physical world stored in the map and/or the number of feature points stored for each pose. In some embodiments, a slice or portion of a high resolution map from remote storage may be sent to local storage, where the slice or portion is not down sampled.

[0208] A database of environment maps may be updated as new tracking maps are created. To determine which of a potentially very large number of environment maps in a database is to be updated, updating may include efficiently selecting one or more environment maps stored in the database relevant to the new tracking map. The selected one or more environment maps may be ranked by relevance and one or more of the highest ranking maps may be selected for processing to merge higher ranked selected environment maps with the new tracking map to create one or more updated environment maps. When a new tracking map represents a portion of the physical world for which there is no preexisting environment map to update, that tracking map may be stored in the database as a new environment map.

View Independent Display

[0209] Described herein are methods and apparatus for providing virtual contents using an XR system, independent of locations of eyes viewing the virtual content. Conventionally, a virtual content is re-rendered upon any motion of the displaying system. For example, if a user wearing a display system views a virtual representation of a three-dimensional (3D) object on the display and walks around the area where the 3D object appears, the 3D object should be re-rendered for each

viewpoint such that the user has the perception that he or she is walking around an object that occupies real space. However, the re-rendering consumes significant computational resources of a system and causes artifacts due to latency.

[0210] The inventors have recognized and appreciated that headpose (e.g., the location and orientation of a user wearing an XR system) may be used to render a virtual content independent of eye rotations within a head of the user. In some embodiments, dynamic maps of a scene may be generated based on multiple coordinate frames in real space across one or more sessions such that virtual contents interacting with the dynamic maps may be rendered robustly, independent of eye rotations within the head of the user and/or independent of sensor deformations caused by, for example, heat generated during high-speed, computation-intensive operation. In some embodiments, the configuration of multiple coordinate frames may enable a first XR device worn by a first user and a second XR device worn by a second user to recognize a common location in a scene. In some embodiments, the configuration of multiple coordinate frames may enable users wearing XR devices to view a virtual content in a same location of a scene.

[0211] In some embodiments, a tracking map may be built in a world coordinate frame, which may have a world origin. The world origin may be the first pose of an XR device when the XR device is powered on. The world origin may be aligned to gravity such that a developer of an XR application can get gravity alignment without extra work. Different tracking maps may be built in different world coordinate frames because the tracking maps may be captured by a same XR device at different sessions and/or different XR devices worn by different users. In some embodiments, a session of an XR device may span from powering on to powering off the device. In some embodiments, an XR device may have a head coordinate frame, which may have a head origin. The head origin may be the current pose of an XR device when an image is taken. The difference between headpose of a world coordinate frame and of a head coordinate frame may be used to estimate a tracking route.

[0212] In some embodiments, an XR device may have a camera coordinate frame, which may have a camera origin. The camera origin may be the current pose of one or more sensors of an XR device. The inventors have recognized and appreciated that the configuration of a camera coordinate frame enables robust displaying virtual contents independent of eye rotation within a head of a user. This configuration also enables robust displaying of virtual contents independent of sensor deformation due to, for example, heat generated during operation.

[0213] In some embodiments, an XR device may have a head unit with a head-mountable frame that a user can secure to their head and may include two waveguides, one in front of each eye of the user. The waveguides may be transparent so that ambient light from real-world objects can transmit through the waveguides and the user can see the real-world objects. Each waveguide may transmit projected light from a projector to a respective eye of the user. The projected light may form an image on the retina of the eye. The retina of the eye thus receives the ambient light and the projected light. The user may simultaneously see real-world objects and one or more virtual objects that are created by the projected light. In some embodiments, XR devices may have sensors that detect real-world objects around a user. These sensors may, for example, be cameras that capture images that may be processed to identify the locations of real-world objects.

[0214] In some embodiments, an XR system may assign a coordinate frame to a virtual content, as opposed to attaching the virtual content in a world coordinate frame. Such configuration enables a virtual content to be described without regard to where it is rendered for a user, but it may be attached to a more persistent frame position such as a persistent coordinate frame (PCF) described in relation to, for example, FIGS. **14-20C**, to be rendered in a specified location. When the locations of the objects change, the XR device may detect the changes in the environment map and determine movement of the head unit worn by the user relative to real-world objects.

[0215] FIG. **8** illustrates a user experiencing virtual content, as rendered by an XR system **10**, in a physical environment, according to some embodiments. The XR system may include a first XR

device **12.1** that is worn by a first user **14.1**, a network **18** and a server **20**. The user **14.1** is in a physical environment with a real object in the form of a table **16**.

[0216] In the illustrated example, the first XR device **12.1** includes a head unit **22**, a belt pack **24** and a cable connection **26**. The first user **14.1** secures the head unit **22** to their head and the belt pack **24** remotely from the head unit **22** on their waist. The cable connection **26** connects the head unit **22** to the belt pack **24**. The head unit **22** includes technologies that are used to display a virtual object or objects to the first user **14.1** while the first user **14.1** is permitted to see real objects such as the table **16**. The belt pack **24** includes primarily processing and communications capabilities of the first XR device **12.1**. In some embodiments, the processing and communication capabilities may reside entirely or partially in the head unit **22** such that the belt pack **24** may be removed or may be located in another device such as a backpack.

[0217] In the illustrated example, the belt pack **24** is connected via a wireless connection to the network **18**. The server **20** is connected to the network **18** and holds data representative of local content. The belt pack **24** downloads the data representing the local content from the server **20** via the network **18**. The belt pack **24** provides the data via the cable connection **26** to the head unit **22**. The head unit **22** may include a display that has a light source, for example, a laser light source or a light emitting diode (LED), and a waveguide that guides the light.

[0218] In some embodiments, the first user **14.1** may mount the head unit **22** to their head and the belt pack **24** to their waist. The belt pack **24** may download image data representing virtual content over the network **18** from the server **20**. The first user **14.1** may see the table **16** through a display of the head unit **22**. A projector forming part of the head unit **22** may receive the image data from the belt pack **24** and generate light based on the image data. The light may travel through one or more of the waveguides forming part of the display of the head unit **22**. The light may then leave the waveguide and propagates onto a retina of an eye of the first user **14.1**. The projector may generate the light in a pattern that is replicated on a retina of the eye of the first user **14.1**. The light that falls on the retina of the eye of the first user **14.1** may have a selected field of depth so that the first user **14.1** perceives an image at a preselected depth behind the waveguide. In addition, both eyes of the first user **14.1** may receive slightly different images so that a brain of the first user **14.1** perceives a three-dimensional image or images at selected distances from the head unit **22**. In the illustrated example, the first user **14.1** perceives a virtual content **28** above the table **16**. The proportions of the virtual content **28** and its location and distance from the first user **14.1** are determined by the data representing the virtual content **28** and various coordinate frames that are used to display the virtual content **28** to the first user **14.1**.

[0219] In the illustrated example, the virtual content **28** is not visible from the perspective of the drawing and is visible to the first user **14.1** through using the first XR device **12.1**. The virtual content **28** may initially reside as data structures within vision data and algorithms in the belt pack **24**. The data structures may then manifest themselves as light when the projectors of the head unit **22** generate light based on the data structures. It should be appreciated that although the virtual content **28** has no existence in three-dimensional space in front of the first user **14.1**, the virtual content **28** is still represented in FIG. 1 in three-dimensional space for illustration of what a wearer of head unit **22** perceives. The visualization of computer data in three-dimensional space may be used in this description to illustrate how the data structures that facilitate the renderings are perceived by one or more users relate to one another within the data structures in the belt pack **24**.

[0220] FIG. 9 illustrates components of the first XR device **12.1**, according to some embodiments. The first XR device **12.1** may include the head unit **22**, and various components forming part of the vision data and algorithms including, for example, a rendering engine **30**, various coordinate systems **32**, various origin and destination coordinate frames **34**, and various origin to destination coordinate frame transformers **36**. The various coordinate systems may be based on intrinsic of the XR device or may be determined by reference to other information, such as a persistent pose or a persistent coordinate system, as described herein.

[0221] The head unit **22** may include a head-mountable frame **40**, a display system **42**, a real object detection camera **44**, a movement tracking camera **46**, and an inertial measurement unit **48**.

[0222] The head-mountable frame **40** may have a shape that is securable to the head of the first user **14.1** in FIG. **8**. The display system **42**, real object detection camera **44**, movement tracking camera **46**, and inertial measurement unit **48** may be mounted to the head-mountable frame **40** and therefore move together with the head-mountable frame **40**.

[0223] The coordinate systems **32** may include a local data system **52**, a world frame system **54**, a head frame system **56**, and a camera frame system **58**.

[0224] The local data system **52** may include a data channel **62**, a local frame determining routine **64** and a local frame storing instruction **66**. The data channel **62** may be an internal software routine, a hardware component such as an external cable or a radio frequency receiver, or a hybrid component such as a port that is opened up. The data channel **62** may be configured to receive image data **68** representing a virtual content.

[0225] The local frame determining routine **64** may be connected to the data channel **62**. The local frame determining routine **64** may be configured to determine a local coordinate frame **70**. In some embodiments, the local frame determining routine may determine the local coordinate frame based on real world objects or real world locations. In some embodiments, the local coordinate frame may be based on a top edge relative to a bottom edge of a browser window, head or feet of a character, a node on an outer surface of a prism or bounding box that encloses the virtual content, or any other suitable location to place a coordinate frame that defines a facing direction of a virtual content and a location (e.g. a node, such as a placement node or PCF node) with which to place the virtual content, etc.

[0226] The local frame storing instruction **66** may be connected to the local frame determining routine **64**. One skilled in the art will understand that software modules and routines are “connected” to one another through subroutines, calls, etc. The local frame storing instruction **66** may store the local coordinate frame **70** as a local coordinate frame **72** within the origin and destination coordinate frames **34**. In some embodiments, the origin and destination coordinate frames **34** may be one or more coordinate frames that may be manipulated or transformed in order for a virtual content to persist between sessions. In some embodiments, a session may be the period of time between a boot-up and shut-down of an XR device. Two sessions may be two start-up and shut-down periods for a single XR device, or may be a start-up and shut-down for two different XR devices.

[0227] In some embodiments, the origin and destination coordinate frames **34** may be the coordinate frames involved in one or more transformations required in order for a first user's XR device and a second user's XR device to recognize a common location. In some embodiments, the destination coordinate frame may be the output of a series of computations and transformations applied to the target coordinate frame in order for a first and second user to view a virtual content in the same location.

[0228] The rendering engine **30** may be connected to the data channel **62**. The rendering engine **30** may receive the image data **68** from the data channel **62** such that the rendering engine **30** may render virtual content based, at least in part, on the image data **68**.

[0229] The display system **42** may be connected to the rendering engine **30**. The display system **42** may include components that transform the image data **68** into visible light. The visible light may form two patterns, one for each eye. The visible light may enter eyes of the first user **14.1** in FIG. **8** and may be detected on retinas of the eyes of the first user **14.1**.

[0230] The real object detection camera **44** may include one or more cameras that may capture images from different sides of the head-mountable frame **40**. The movement tracking camera **46** may include one or more cameras that capture images on sides of the head-mountable frame **40**. One set of one or more cameras may be used instead of the two sets of one or more cameras representing the real object detection camera(s) **44** and the movement tracking camera(s) **46**. In

some embodiments, the cameras **44**, **46** may capture images. As described above these cameras may collect data that is used to construct a tacking map.

[0231] The inertial measurement unit **48** may include a number of devices that are used to detect movement of the head unit **22**. The inertial measurement unit **48** may include a gravitation sensor, one or more accelerometers and one or more gyroscopes. The sensors of the inertial measurement unit **48**, in combination, track movement of the head unit **22** in at least three orthogonal directions and about at least three orthogonal axes.

[0232] In the illustrated example, the world frame system **54** includes a world surface determining routine **78**, a world frame determining routine **80**, and a world frame storing instruction **82**. The world surface determining routine **78** is connected to the real object detection camera **44**. The world surface determining routine **78** receives images and/or key frames based on the images that are captured by the real object detection camera **44** and processes the images to identify surfaces in the images. A depth sensor (not shown) may determine distances to the surfaces. The surfaces are thus represented by data in three dimensions including their sizes, shapes, and distances from the real object detection camera.

[0233] In some embodiments, a world coordinate frame **84** may be based on the origin at the initialization of the headpose session. In some embodiments, the world coordinate frame may be located where the device was booted up, or could be somewhere new if headpose was lost during the boot session. In some embodiments, the world coordinate frame may be the origin at the start of a headpose session.

[0234] In the illustrated example, the world frame determining routine **80** is connected to the world surface determining routine **78** and determines a world coordinate frame **84** based on the locations of the surfaces as determined by the world surface determining routine **78**. The world frame storing instruction **82** is connected to the world frame determining routine **80** to receive the world coordinate frame **84** from the world frame determining routine **80**. The world frame storing instruction **82** stores the world coordinate frame **84** as a world coordinate frame **86** within the origin and destination coordinate frames **34**.

[0235] The head frame system **56** may include a head frame determining routine **90** and a head frame storing instruction **92**. The head frame determining routine **90** may be connected to the movement tracking camera **46** and the inertial measurement unit **48**. The head frame determining routine **90** may use data from the movement tracking camera **46** and the inertial measurement unit **48** to calculate a head coordinate frame **94**. For example, the inertial measurement unit **48** may have a gravitation sensor that determines the direction of gravitational force relative to the head unit **22**. The movement tracking camera **46** may continually capture images that are used by the head frame determining routine **90** to refine the head coordinate frame **94**. The head unit **22** moves when the first user **14.1** in FIG. **8** moves their head. The movement tracking camera **46** and the inertial measurement unit **48** may continuously provide data to the head frame determining routine **90** so that the head frame determining routine **90** can update the head coordinate frame **94**.

[0236] The head frame storing instruction **92** may be connected to the head frame determining routine **90** to receive the head coordinate frame **94** from the head frame determining routine **90**. The head frame storing instruction **92** may store the head coordinate frame **94** as a head coordinate frame **96** among the origin and destination coordinate frames **34**. The head frame storing instruction **92** may repeatedly store the updated head coordinate frame **94** as the head coordinate frame **96** when the head frame determining routine **90** recalculates the head coordinate frame **94**. In some embodiments, the head coordinate frame may be the location of the wearable XR device **12.1** relative to the local coordinate frame **72**.

[0237] The camera frame system **58** may include camera intrinsics **98**. The camera intrinsics **98** may include dimensions of the head unit **22** that are features of its design and manufacture. The camera intrinsics **98** may be used to calculate a camera coordinate frame **100** that is stored within the origin and destination coordinate frames **34**.

[0238] In some embodiments, the camera coordinate frame **100** may include all pupil positions of a left eye of the first user **14.1** in FIG. **8**. When the left eye moves from left to right or up and down, the pupil positions of the left eye are located within the camera coordinate frame **100**. In addition, the pupil positions of a right eye are located within a camera coordinate frame **100** for the right eye. In some embodiments, the camera coordinate frame **100** may include the location of the camera relative to the local coordinate frame when an image is taken.

[0239] The origin to destination coordinate frame transformers **36** may include a local-to-world coordinate transformer **104**, a world-to-head coordinate transformer **106**, and a head-to-camera coordinate transformer **108**. The local-to-world coordinate transformer **104** may receive the local coordinate frame **72** and transform the local coordinate frame **72** to the world coordinate frame **86**. The transformation of the local coordinate frame **72** to the world coordinate frame **86** may be represented as a local coordinate frame transformed to world coordinate frame **110** within the world coordinate frame **86**.

[0240] The world-to-head coordinate transformer **106** may transform from the world coordinate frame **86** to the head coordinate frame **96**. The world-to-head coordinate transformer **106** may transform the local coordinate frame transformed to world coordinate frame **110** to the head coordinate frame **96**. The transformation may be represented as a local coordinate frame transformed to head coordinate frame **112** within the head coordinate frame **96**.

[0241] The head-to-camera coordinate transformer **108** may transform from the head coordinate frame **96** to the camera coordinate frame **100**. The head-to-camera coordinate transformer **108** may transform the local coordinate frame transformed to head coordinate frame **112** to a local coordinate frame transformed to camera coordinate frame **114** within the camera coordinate frame **100**. The local coordinate frame transformed to camera coordinate frame **114** may be entered into the rendering engine **30**. The rendering engine **30** may render the image data **68** representing the local content **28** based on the local coordinate frame transformed to camera coordinate frame **114**.

[0242] FIG. **10** is a spatial representation of the various origin and destination coordinate frames **34**. The local coordinate frame **72**, world coordinate frame **86**, head coordinate frame **96**, and camera coordinate frame **100** are represented in the figure. In some embodiments, the local coordinate frame associated with the XR content **28** may have a position and rotation (e.g. may provide a node and facing direction) relative to a local and/or world coordinate frame and/or PCF when the virtual content is placed in the real world so the virtual content may be viewed by the user. Each camera may have its own camera coordinate frame **100** encompassing all pupil positions of one eye. Reference numerals **104A** and **106A** represent the transformations that are made by the local-to-world coordinate transformer **104**, world-to-head coordinate transformer **106**, and head-to-camera coordinate transformer **108** in FIG. **9**, respectively.

[0243] FIG. **11** depicts a camera render protocol for transforming from a head coordinate frame to a camera coordinate frame, according to some embodiments. In the illustrated example, a pupil for a single eye moves from position A to B. A virtual object that is meant to appear stationary will project onto a depth plane at one of the two positions A or B depending on the position of the pupil (assuming that the camera is configured to use a pupil-based coordinate frame). As a result, using a pupil coordinate frame transformed to a head coordinate frame will cause jitter in a stationary virtual object as the eye moves from position A to position B. This situation is referred to as view dependent display or projection.

[0244] As depicted in FIG. **12**, a camera coordinate frame (e.g., CR) is positioned and encompasses all pupil positions and object projection will now be consistent regardless of pupil positions A and B. The head coordinate frame transforms to the CR frame, which is referred to as view independent display or projection. An image reprojection may be applied to the virtual content to account for a change in eye position, however, as the rendering is still in the same position, jitter is minimized.

[0245] FIG. **13** illustrates the display system **42** in more detail. The display system **42** includes a stereoscopic analyzer **144** that is connected to the rendering engine **30** and forms part of the vision

data and algorithms.

[0246] The display system **42** further includes left and right projectors **166A** and **166B** and left and right waveguides **170A** and **170B**. The left and right projectors **166A** and **166B** are connected to power supplies. Each projector **166A** and **166B** has a respective input for image data to be provided to the respective projector **166A** or **166B**. The respective projector **166A** or **166B**, when powered, generates light in two-dimensional patterns and emanates the light therefrom. The left and right waveguides **170A** and **170B** are positioned to receive light from the left and right projectors **166A** and **166B**, respectively. The left and right waveguides **170A** and **170B** are transparent waveguides.

[0247] In use, a user mounts the head mountable frame **40** to their head. Components of the head mountable frame **40** may, for example, include a strap (not shown) that wraps around the back of the head of the user. The left and right waveguides **170A** and **170B** are then located in front of left and right eyes **220A** and **220B** of the user.

[0248] The rendering engine **30** enters the image data that it receives into the stereoscopic analyzer **144**. The image data is three-dimensional image data of the local content **28** in FIG. **8**. The image data is projected onto a plurality of virtual planes. The stereoscopic analyzer **144** analyzes the image data to determine left and right image data sets based on the image data for projection onto each depth plane. The left and right image data sets are data sets that represent two-dimensional images that are projected in three-dimensions to give the user a perception of a depth.

[0249] The stereoscopic analyzer **144** enters the left and right image data sets into the left and right projectors **166A** and **166B**. The left and right projectors **166A** and **166B** then create left and right light patterns. The components of the display system **42** are shown in plan view, although it should be understood that the left and right patterns are two-dimensional patterns when shown in front elevation view. Each light pattern includes a plurality of pixels. For purposes of illustration, light rays **224A** and **226A** from two of the pixels are shown leaving the left projector **166A** and entering the left waveguide **170A**. The light rays **224A** and **226A** reflect from sides of the left waveguide **170A**. It is shown that the light rays **224A** and **226A** propagate through internal reflection from left to right within the left waveguide **170A**, although it should be understood that the light rays **224A** and **226A** also propagate in a direction into the paper using refractory and reflective systems.

[0250] The light rays **224A** and **226A** exit the left light waveguide **170A** through a pupil **228A** and then enter a left eye **220A** through a pupil **230A** of the left eye **220A**. The light rays **224A** and **226A** then fall on a retina **232A** of the left eye **220A**. In this manner, the left light pattern falls on the retina **232A** of the left eye **220A**. The user is given the perception that the pixels that are formed on the retina **232A** are pixels **234A** and **236A** that the user perceives to be at some distance on a side of the left waveguide **170A** opposing the left eye **220A**. Depth perception is created by manipulating the focal length of the light.

[0251] In a similar manner, the stereoscopic analyzer **144** enters the right image data set into the right projector **166B**. The right projector **166B** transmits the right light pattern, which is represented by pixels in the form of light rays **224B** and **226B**. The light rays **224B** and **226B** reflect within the right waveguide **170B** and exit through a pupil **228B**. The light rays **224B** and **226B** then enter through a pupil **230B** of the right eye **220B** and fall on a retina **232B** of a right eye **220B**. The pixels of the light rays **224B** and **226B** are perceived as pixels **134B** and **236B** behind the right waveguide **170B**.

[0252] The patterns that are created on the retinas **232A** and **232B** are individually perceived as left and right images. The left and right images differ slightly from one another due to the functioning of the stereoscopic analyzer **144**. The left and right images are perceived in a mind of the user as a three-dimensional rendering.

[0253] As mentioned, the left and right waveguides **170A** and **170B** are transparent. Light from a real-life object such as the table **16** on a side of the left and right waveguides **170A** and **170B** opposing the eyes **220A** and **220B** can project through the left and right waveguides **170A** and **170B** and fall on the retinas **232A** and **232B**.

Persistent Coordinate Frame (PCF)

[0254] Described herein are methods and apparatus for providing spatial persistence across user instances within a shared space. Without spatial persistence, virtual content placed in the physical world by a user in a session may not exist or may be misplaced in the user's view in a different session. Without spatial persistence, virtual content placed in the physical world by one user may not exist or may be out of place in a second user's view, even if the second user is intended to be sharing an experience of the same physical space with the first user.

[0255] The inventors have recognized and appreciated that spatial persistence may be provided through persistent coordinate frames (PCFs). A PCF may be defined based on one or more points, representing features recognized in the physical world (e.g., corners, edges). The features may be selected such that they are likely to be the same from a user instance to another user instance of an XR system.

[0256] Further, drift during tracking, which causes the computed tracking path (e.g., camera trajectory) to deviate from the actual tracking path, can cause the location of virtual content, when rendered with respect to a local map that is based solely on a tracking map to appear out of place. A tracking map for the space may be refined to correct the drifts as an XR device collects more information of the scene overtime. However, if virtual content is placed on a real object before a map refinement and saved with respect to the world coordinate frame of the device derived from the tracking map, the virtual content may appear displaced, as if the real object has been moved during the map refinement. PCFs may be updated according to map refinement because the PCFs are defined based on the features and are updated as the features move during map refinements.

[0257] A PCF may comprise six degrees of freedom with translations and rotations relative to a map coordinate system. A PCF may be stored in a local and/or remote storage medium. The translations and rotations of a PCF may be computed relative to a map coordinate system depending on, for example, the storage location. For example, a PCF used locally by a device may have translations and rotations relative to a world coordinate frame of the device. A PCF in the cloud may have translations and rotations relative to a canonical coordinate frame of a canonical map.

[0258] PCFs may provide a sparse representation of the physical world, providing less than all of the available information about the physical world, such that they may be efficiently processed and transferred. Techniques for processing persistent spatial information may include creating dynamic maps based on one or more coordinate systems in real space across one or more sessions, generating persistent coordinate frames (PCF) over the sparse maps, which may be exposed to XR applications via, for example, an application programming interface (API).

[0259] FIG. 14 is a block diagram illustrating the creation of a persistent coordinate frame (PCF) and the attachment of XR content to the PCF, according to some embodiments. Each block may represent digital information stored in a computer memory. In the case of applications 1180, the data may represent computer-executable instructions. In the case of virtual content 1170, the digital information may define a virtual object, as specified by the application 1180, for example. In the case of the other boxes, the digital information may characterize some aspect of the physical world.

[0260] In the illustrated embodiment, one or more PCFs are created from images captured with sensors on a wearable device. In the embodiment of FIG. 14, the sensors are visual image cameras. These cameras may be the same cameras used for forming a tracking map. Accordingly, some of the processing suggested by FIG. 14 may be performed as part of updating a tracking map. However, FIG. 14 illustrates that information that provides persistence is generated in addition to the tracking map.

[0261] In order to derive a 3D PCF, two images 1110 from two cameras mounted to a wearable device in a configuration that enables stereoscopic image analysis are processed together. FIG. 14 illustrates an Image 1 and an Image 2, each derived from one of the cameras. A single image from each camera is illustrated for simplicity. However, each camera may output a stream of image

frames and the processing illustrated in FIG. **14** may be performed for multiple image frames in the stream.

[0262] Accordingly, Image 1 and Image 2 may each be one frame in a sequence of image frames. Processing as depicted in FIG. **14** may be repeated on successive image frames in the sequence until image frames containing feature points providing a suitable image from which to form persistent spatial information is processed. Alternatively or additionally, the processing of FIG. **14** might be repeated as a user moves such that the user is no longer close enough to a previously identified PCF to reliably use that PCF for determining positions with respect to the physical world. For example, an XR system may maintain a current PCF for a user. When that distance exceeds a threshold, the system may switch to a new current PCF, closer to the user, which may be generated according to the process of FIG. **14**, using image frames acquired in the user's current location.

[0263] Even when generating a single PCF, a stream of image frames may be processed to identify image frames depicting content in the physical world that is likely stable and can be readily identified by a device in the vicinity of the region of the physical world depicted in the image frame. In the embodiment of FIG. **14**, this processing begins with the identification of features **1120** in the image. Features may be identified, for example, by finding locations of gradients in the image above a threshold or other characteristics, which may correspond to a corner of an object, for example. In the embodiment illustrated, the features are points, but other recognizable features, such as edges, may alternatively or additionally be used.

[0264] In the embodiment illustrated, a fixed number, N, of features **1120** are selected for further processing. Those feature points may be selected based on one or more criteria, such as magnitude of the gradient, or proximity to other feature points. Alternatively or additionally, the feature points may be selected heuristically, such as based on characteristics that suggest the feature points are persistent. For example, heuristics may be defined based on the characteristics of feature points that likely correspond to a corner of a window or a door or a large piece of furniture. Such heuristics may take into account the feature point itself and what surrounds it. As a specific example, the number of feature points per image may be between 100 and 500 or between 150 and 250, such as 200.

[0265] Regardless of the number of feature points selected, descriptors **1130** may be computed for the feature points. In this example, a descriptor is computed for each selected feature point, but a descriptor may be computed for groups of feature points or for a subset of the feature points or for all features within an image. The descriptor characterizes a feature point such that feature points representing the same object in the physical world are assigned similar descriptors. The descriptors may facilitate alignment of two frames, such as may occur when one map is localized with respect to another. Rather than searching for a relative orientation of the frames that minimizes the distance between feature points of the two images, an initial alignment of the two frames may be made by identifying feature points with similar descriptors. Alignment of the image frames may be based on aligning points with similar descriptors, which may entail less processing than computing an alignment of all the feature points in the images.

[0266] The descriptors may be computed as a mapping of the feature points or, in some embodiments a mapping of a patch of an image around a feature point, to a descriptor. The descriptor may be a numeric quantity. U.S. patent application Ser. No. 16/190,948 describes computing descriptors for feature points and is hereby incorporated herein by reference in its entirety.

[0267] In the example of FIG. **14**, a descriptor **1130** is computed for each feature point in each image frame. Based on the descriptors and/or the feature points and/or the image itself, the image frame may be identified as a key frame **1140**. In the embodiment illustrated, a key frame is an image frame meeting certain criteria that is then selected for further processing. In making a tracking map, for example, image frames that add meaningful information to the map may be

selected as key frames that are integrated into the map. On the other hand, image frames that substantially overlap a region for which an image frame has already been integrated into the map may be discarded such that they do not become key frames. Alternatively or additionally, key frames may be selected based on the number and/or type of feature points in the image frame. In the embodiment of FIG. **14**, key frames **1150** selected for inclusion in a tracking map may also be treated as key frames for determining a PCF, but different or additional criteria for selecting key frames for generation of a PCF may be used.

[0268] Though FIG. **14** shows that a key frame is used for further processing, information acquired from an image may be processed in other forms. For example, the feature points, such as in a key rig, may alternatively or additionally be processed. Moreover, though a key frame is described as being derived from a single image frame, it is not necessary that there be a one to one relationship between a key frame and an acquired image frame. A key frame, for example, may be acquired from multiple image frames, such as by stitching together or aggregating the image frames such that only features appearing in multiple images are retained in the key frame.

[0269] A key frame may include image information and/or metadata associated with the image information. In some embodiments, images captured by the cameras **44**, **46** (FIG. **9**) may be computed into one or more key frames (e.g., key frames 1, 2). In some embodiments, a key frame may include a camera pose. In some embodiments, a key frame may include one or more camera images captured at the camera pose. In some embodiments, an XR system may determine a portion of the camera images captured at the camera pose as not useful and thus not include the portion in a key frame. Therefore, using key frames to align new images with earlier knowledge of a scene reduces the use of computational resource of the XR system. In some embodiments, a key frame may include an image, and/or image data, at a location with a direction/angle. In some embodiments, a key frame may include a location and a direction from which one or more map points may be observed. In some embodiments, a key frame may include a coordinate frame with an ID. U.S. patent application Ser. No. 15/877,359 describes key frames and is hereby incorporated herein by reference in its entirety.

[0270] Some or all of the key frames **1140** may be selected for further processing, such as the generation of a persistent pose **1150** for the key frame. The selection may be based on the characteristics of all, or a subset of, the feature points in the image frame. Those characteristics may be determined from processing the descriptors, features and/or image frame, itself. As a specific example, the selection may be based on a cluster of feature points identified as likely to relate to a persistent object.

[0271] Each key frame is associated with a pose of the camera at which that key frame was acquired. For key frames selected for processing into a persistent pose, that pose information may be saved along with other metadata about the key frame, such as a WiFi fingerprint and/or GPS coordinates at the time of acquisition and/or at the location of acquisition.

[0272] The persistent poses are a source of information that a device may use to orient itself relative to previously acquired information about the physical world. For example, if the key frame from which a persistent pose was created is incorporated into a map of the physical world, a device may orient itself relative to that persistent pose using a sufficient number of feature points in the key frame that are associated with the persistent pose. The device may align a current image that it takes of its surroundings to the persistent pose. This alignment may be based on matching the current image to the image **1110**, the features **1120**, and/or the descriptors **1130** that gave rise to the persistent pose, or any subset of that image or those features or descriptors. In some embodiments, the current image frame that is matched to the persistent pose may be another key frame that has been incorporated into the device's tracking map.

[0273] Information about a persistent pose may be stored in a format that facilitates sharing among multiple applications, which may be executing on the same or different devices. In the example of FIG. **14**, some or all of the persistent poses may be reflected as a persistent coordinate frames

(PCF) **1160**. Like a persistent pose, a PCF may be associated with a map and may comprise a set of features, or other information, that a device can use to determine its orientation with respect to that PCF. The PCF may include a transformation that defines its transformation with respect to the origin of its map, such that, by correlating its position to a PCF, the device can determine its position with respect to any objects in the physical world reflected in the map.

[0274] As the PCF provides a mechanism for determining locations with respect to the physical objects, an application, such as applications **1180**, may define positions of virtual objects with respect to one or more PCFs, which serve as anchors for the virtual content **1170**. FIG. **14** illustrates, for example, that App 1 has associated its virtual content 2 with PCF 1.2. Likewise, App 2 has associated its virtual content 3 with PCF 1.2. App 1 is also shown associating its virtual content 1 to PCF 4.5, and App 2 is shown associating its virtual content 4 with PCF 3. In some embodiments, PCF 3 may be based on Image 3 (not shown), and PCF 4.5 may be based on Image 4 and Image 5 (not shown) analogously to how PCF 1.2 is based on Image 1 and Image 2. When rendering this virtual content, a device may apply one or more transformations to compute information, such as the location of the virtual content with respect to the display of the device and/or the location of physical objects with respect to the desired location of the virtual content. Using the PCFs as reference may simplify such computations.

[0275] In some embodiments, a persistent pose may be a coordinate location and/or direction that has one or more associated key frames. In some embodiments, a persistent pose may be automatically created after the user has traveled a certain distance, e.g., three meters. In some embodiments, the persistent poses may act as reference points during localization. In some embodiments, the persistent poses may be stored in a passable world (e.g., the passable world module **538**).

[0276] In some embodiments, a new PCF may be determined based on a pre-defined distance allowed between adjacent PCFs. In some embodiments, one or more persistent poses may be computed into a PCF when a user travels a pre-determined distance, e.g. five meters. In some embodiments, PCFs may be associated with one or more world coordinate frames and/or canonical coordinate frames, e.g., in the passable world. In some embodiments, PCFs may be stored in a local and/or remote database depending on, for example, security settings.

[0277] FIG. **15** illustrates a method **4700** of establishing and using a persistence coordinate frame, according to some embodiments. The method **4700** may start from capturing (Act **4702**) images (e.g., Image 1 and Image 2 in FIG. **14**) about a scene using one or more sensors of an XR device. Multiple cameras may be used and one camera may generate multiple images, for example, in a stream.

[0278] The method **4700** may include extracting (**4704**) interest points (e.g., map points **702** in FIG. **7**, features **1120** in FIG. **14**) from the captured images, generating (Act **4706**) descriptors (e.g., descriptors **1130** in FIG. **14**) for the extracted interest points, and generating (Act **4708**) key frames (e.g., key frames **1140**) based on the descriptors. In some embodiments, the method may compare interest points in the key frames, and form pairs of key frames that share a predetermined amount of interest points. The method may reconstruct parts of the physical world using individual pairs of key frames. Mapped parts of the physical world may be saved as 3D features (e.g., keyrig **704** in FIG. **7**). In some embodiments, a selected portion of the pairs of key frames may be used to build 3D features. In some embodiments, results of the mapping may be selectively saved. Key frames not used for building 3D features may be associated with the 3D features through poses, for example, representing distances between key frames with a covariance matrix between poses of keyframes. In some embodiments, pairs of key frames may be selected to build the 3D features such that distances between each two of the build 3D features are within a predetermined distance, which may be determined to balance the amount of computation needed and the level of accuracy of a resulting model. Such approaches enable providing a model of the physical world with the amount of data that is suitable for efficient and accurate computation with an XR system. In some

embodiments, a covariance matrix of two images may include covariances between poses of the two images (e.g., six degrees of freedom).

[0279] The method **4700** may include generating (Act **4710**) persistent poses based on the key frames. In some embodiments, the method may include generating the persistent poses based on the 3D features reconstructed from pairs of key frames. In some embodiments, a persistent pose may be attached to a 3D feature. In some embodiments, the persistent pose may include a pose of a key frame used to construct the 3D feature. In some embodiments, the persistent pose may include an average pose of key frames used to construct the 3D feature. In some embodiments, persistent poses may be generated such that distances between neighboring persistent poses are within a predetermined value, for example, in the range of one meter to five meters, any value in between, or any other suitable value. In some embodiments, the distances between neighboring persistent poses may be represented by a covariance matrix of the neighboring persistent poses.

[0280] The method **4700** may include generating (Act **4712**) PCFs based on the persistent poses. In some embodiments, a PCF may be attached to a 3D feature. In some embodiments, a PCF may be associated with one or more persistent poses. In some embodiments, a PCF may include a pose of one of the associated persistent poses. In some embodiments, a PCF may include an average pose of the poses of the associated persistent poses. In some embodiments, PCFs may be generated such that distances between neighboring PCFs are within a predetermined value, for example, in the range of three meters to ten meters, any value in between, or any other suitable value. In some embodiments, the distances between neighboring PCFs may be represented by a covariance matrix of the neighboring PCFs. In some embodiments, PCFs may be exposed to XR applications via, for example, an application programming interface (API) such that the XR applications can access a model of the physical world through the PCFs without accessing the model itself.

[0281] The method **4700** may include associating (Act **4714**) image data of a virtual object to be displayed by the XR device to at least one of the PCFs. In some embodiments, the method may include computing translations and orientations of the virtual object with respect to the associated PCF. It should be appreciated that it is not necessary to associate a virtual object to a PCF generated by the device placing the virtual object. For example, a device may retrieve saved PCFs in a canonical map in a cloud and associate a virtual object to a retrieved PCF. It should be appreciated that the virtual object may move with the associated PCF as the PCF is adjusted overtime.

[0282] FIG. **16** illustrates the first XR device **12.1** and vision data and algorithms of a second XR device **12.2** and the server **20**, according to some embodiments. The components illustrated in FIG. **16** may operate to perform some or all of the operations associated with generating, updating, and/or using spatial information, such as persistent poses, persistent coordinate frames, tracking maps, or canonical maps, as described herein. Although not illustrated, the first XR device **12.1** may be configured the same as the second XR device **12.2**. The server **20** may have a map storing routine **118**, a canonical map **120**, a map transmitter **122**, and a map merge algorithm **124**.

[0283] The second XR device **12.2**, which may be in the same scene as the first XR device **12.1**, may include a persistent coordinate frame (PCF) integration unit **1300**, an application **1302** that generates the image data **68** that may be used to render a virtual object, and a frame embedding generator **308** (See FIG. **21**). In some embodiments, a map download system **126**, PCF identification system **128**, Map 2, localization module **130**, canonical map incorporator **132**, canonical map **133**, and map publisher **136** may be grouped into a passable world unit **1304**. The PCF integration unit **1300** may be connected to the passable world unit **1304** and other components of the second XR device **12.2** to allow for the retrieval, generation, use, upload, and download of PCFs.

[0284] A map, comprising PCFs, may enable more persistence in a changing world. In some embodiments, localizing a tracking map including, for example, matching features for images, may include selecting features that represent persistent content from the map constituted by PCFs,

which enables fast matching and/or localizing. For example, a world where people move into and out of the scene and objects such as doors move relative to the scene, requires less storage space and transmission rates, and enables the use of individual PCFs and their relationships relative to one another (e.g., integrated constellation of PCFs) to map a scene.

[0285] In some embodiments, the PCF integration unit **1300** may include PCFs **1306** that were previously stored in a data store on a storage unit of the second XR device **12.2**, a PCF tracker **1308**, a persistent pose acquirer **1310**, a PCF checker **1312**, a PCF generation system **1314**, a coordinate frame calculator **1316**, a persistent pose calculator **1318**, and three transformers, including a tracking map and persistent pose transformer **1320**, a persistent pose and PCF transformer **1322**, and a PCF and image data transformer **1324**.

[0286] In some embodiments, the PCF tracker **1308** may have an on-prompt and an off-prompt that are selectable by the application **1302**. The application **1302** may be executable by a processor of the second XR device **12.2** to, for example, display a virtual content. The application **1302** may have a call that switches the PCF tracker **1308** on via the on-prompt. The PCF tracker **1308** may generate PCFs when the PCF tracker **1308** is switched on. The application **1302** may have a subsequent call that can switch the PCF tracker **1308** off via the off-prompt. The PCF tracker **1308** terminates PCF generation when the PCF tracker **1308** is switched off.

[0287] In some embodiments, the server **20** may include a plurality of persistent poses **1332** and a plurality of PCFs **1330** that have previously been saved in association with a canonical map **120**. The map transmitter **122** may transmit the canonical map **120** together with the persistent poses **1332** and/or the PCFs **1330** to the second XR device **12.2**. The persistent poses **1332** and PCFs **1330** may be stored in association with the canonical map **133** on the second XR device **12.2**. When Map 2 localizes to the canonical map **133**, the persistent poses **1332** and the PCFs **1330** may be stored in association with Map 2.

[0288] In some embodiments, the persistent pose acquirer **1310** may acquire the persistent poses for Map 2. The PCF checker **1312** may be connected to the persistent pose acquirer **1310**. The PCF checker **1312** may retrieve PCFs from the PCFs **1306** based on the persistent poses retrieved by the persistent pose acquirer **1310**. The PCFs retrieved by the PCF checker **1312** may form an initial group of PCFs that are used for image display based on PCFs.

[0289] In some embodiments, the application **1302** may require additional PCFs to be generated. For example, if a user moves to an area that has not previously been mapped, the application **1302** may switch the PCF tracker **1308** on. The PCF generation system **1314** may be connected to the PCF tracker **1308** and begin to generate PCFs based on Map 2 as Map 2 begins to expand. The PCFs generated by the PCF generation system **1314** may form a second group of PCFs that may be used for PCF-based image display.

[0290] The coordinate frame calculator **1316** may be connected to the PCF checker **1312**. After the PCF checker **1312** retrieved PCFs, the coordinate frame calculator **1316** may invoke the head coordinate frame **96** to determine a headpose of the second XR device **12.2**. The coordinate frame calculator **1316** may also invoke the persistent pose calculator **1318**. The persistent pose calculator **1318** may be directly or indirectly connected to the frame embedding generator **308**. In some embodiments, an image/frame may be designated a key frame after a threshold distance from the previous key frame, e.g. 3 meters, is traveled. The persistent pose calculator **1318** may generate a persistent pose based on a plurality, for example three, key frames. In some embodiments, the persistent pose may be essentially an average of the coordinate frames of the plurality of key frames.

[0291] The tracking map and persistent pose transformer **1320** may be connected to Map 2 and the persistent pose calculator **1318**. The tracking map and persistent pose transformer **1320** may transform Map 2 to the persistent pose to determine the persistent pose at an origin relative to Map 2.

[0292] The persistent pose and PCF transformer **1322** may be connected to the tracking map and

persistent pose transformer **1320** and further to the PCF checker **1312** and the PCF generation system **1314**. The persistent pose and PCF transformer **1322** may transform the persistent pose (to which the tracking map has been transformed) to the PCFs from the PCF checker **1312** and the PCF generation system **1314** to determine the PCFs relative to the persistent pose.

[0293] The PCF and image data transformer **1324** may be connected to the persistent pose and PCF transformer **1322** and to the data channel **62**. The PCF and image data transformer **1324** transforms the PCFs to the image data **68**. The rendering engine **30** may be connected to the PCF and image data transformer **1324** to display the image data **68** to the user relative to the PCFs.

[0294] The PCF integration unit **1300** may store the additional PCFs that are generated with the PCF generation system **1314** within the PCFs **1306**. The PCFs **1306** may be stored relative to persistent poses. The map publisher **136** may retrieve the PCFs **1306** and the persistent poses associated with the PCFs **1306** when the map publisher **136** transmits Map 2 to the server **20**, the map publisher **136** also transmits the PCFs and persistent poses associated with Map 2 to the server **20**. When the map storing routine **118** of the server **20** stores Map 2, the map storing routine **118** may also store the persistent poses and PCFs generated by the second viewing device **12.2**. The map merge algorithm **124** may create the canonical map **120** with the persistent poses and PCFs of Map 2 associated with the canonical map **120** and stored within the persistent poses **1332** and PCFs **1330**, respectively.

[0295] The first XR device **12.1** may include a PCF integration unit similar to the PCF integration unit **1300** of the second XR device **12.2**. When the map transmitter **122** transmits the canonical map **120** to the first XR device **12.1**, the map transmitter **122** may transmit the persistent poses **1332** and PCFs **1330** associated with the canonical map **120** and originating from the second XR device **12.2**. The first XR device **12.1** may store the PCFs and the persistent poses within a data store on a storage device of the first XR device **12.1**. The first XR device **12.1** may then make use of the persistent poses and the PCFs originating from the second XR device **12.2** for image display relative to the PCFs. Additionally or alternatively, the first XR device **12.1** may retrieve, generate, make use, upload, and download PCFs and persistent poses in a manner similar to the second XR device **12.2** as described above.

[0296] In the illustrated example, the first XR device **12.1** generates a local tracking map (referred to hereinafter as “Map 1”) and the map storing routine **118** receives Map 1 from the first XR device **12.1**. The map storing routine **118** then stores Map 1 on a storage device of the server **20** as the canonical map **120**.

[0297] The second XR device **12.2** includes a map download system **126**, an anchor identification system **128**, a localization module **130**, a canonical map incorporator **132**, a local content position system **134**, and a map publisher **136**.

[0298] In use, the map transmitter **122** sends the canonical map **120** to the second XR device **12.2** and the map download system **126** downloads and stores the canonical map **120** as a canonical map **133** from the server **20**.

[0299] The anchor identification system **128** is connected to the world surface determining routine **78**. The anchor identification system **128** identifies anchors based on objects detected by the world surface determining routine **78**. The anchor identification system **128** generates a second map (Map 2) using the anchors. As indicated by the cycle **138**, the anchor identification system **128** continues to identify anchors and continues to update Map 2. The locations of the anchors are recorded as three-dimensional data based on data provided by the world surface determining routing **78**. The world surface determining routine **78** receives images from the real object detection camera **44** and depth data from depth sensors **135** to determine the locations of surfaces and their relative distance from the depth sensors **135**.

[0300] The localization module **130** is connected to the canonical map **133** and Map 2. The localization module **130** repeatedly attempts to localize Map 2 to the canonical map **133**. The canonical map incorporator **132** is connected to the canonical map **133** and Map 2. When the

localization module **130** localizes Map 2 to the canonical map **133**, the canonical map incorporator **132** incorporates the canonical map **133** into anchors of Map 2. Map 2 is then updated with missing data that is included in the canonical map.

[0301] The local content position system **134** is connected to Map 2. The local content position system **134** may, for example, be a system wherein a user can locate local content in a particular location within a world coordinate frame. The local content then attaches itself to one anchor of Map 2. The local-to-world coordinate transformer **104** transforms the local coordinate frame to the world coordinate frame based on the settings of the local content position system **134**. The functioning of the rendering engine **30**, display system **42**, and data channel **62** have been described with reference to FIG. 2.

[0302] The map publisher **136** uploads Map 2 to the server **20**. The map storing routine **118** of the server **20** then stores Map 2 within a storage medium of the server **20**.

[0303] The map merge algorithm **124** merges Map 2 with the canonical map **120**. When more than two maps, for example, three or four maps relating to the same or adjacent regions of the physical world, have been stored, the map merge algorithm **124** merges all the maps into the canonical map **120** to render a new canonical map **120**. The map transmitter **122** then transmits the new canonical map **120** to any and all devices **12.1** and **12.2** that are in an area represented by the new canonical map **120**. When the devices **12.1** and **12.2** localize their respective maps to the canonical map **120**, the canonical map **120** becomes the promoted map.

[0304] FIG. 17 illustrates an example of generating key frames for a map of a scene, according to some embodiments. In the illustrated example, a first key frame, KF1, is generated for a door on a left wall of the room. A second key frame, KF2, is generated for an area in a corner where a floor, the left wall, and a right wall of the room meet. A third key frame, KF3, is generated for an area of a window on the right wall of the room. A fourth key frame, KF4, is generated for an area at a far end of a rug on a floor of the wall. A fifth key frame, KF5, is generated for an area of the rug closest to the user.

[0305] FIG. 18 illustrates an example of generating persistent poses for the map of FIG. 17, according to some embodiments. In some embodiments, a new persistent pose is created when the device measures a threshold distance traveled, and/or when an application requests a new persistent pose (PP). In some embodiments, the threshold distance may be 3 meters, 5 meters, 20 meters, or any other suitable distance. Selecting a smaller threshold distance (e.g., 1 m) may result in an increase in compute load since a larger number of PPs may be created and managed compared to larger threshold distances. Selecting a larger threshold distance (e.g. 40 m) may result in increased virtual content placement error since a smaller number of PPs would be created, which would result in fewer PCFs being created, which means the virtual content attached to the PCF could be a relatively large distance (e.g. 30 m) away from the PCF, and error increases with increasing distance from a PCF to the virtual content.

[0306] In some embodiments, a PP may be created at the start of a new session. This initial PP may be thought of as zero, and can be visualized as the center of a circle that has a radius equal to the threshold distance. When the device reaches the perimeter of the circle, and, in some embodiments, an application requests a new PP, a new PP may be placed at the current location of the device (at the threshold distance). In some embodiments, a new PP will not be created at the threshold distance if the device is able to find an existing PP within the threshold distance from the device's new position. In some embodiments, when a new PP (e.g., PP1150 in FIG. 14) is created, the device attaches one or more of the closest key frames to the PP. In some embodiments, the location of the PP relative to the key frames may be based on the location of the device at the time a PP is created. In some embodiments, a PP will not be created when the device travels a threshold distance unless an application requests a PP.

[0307] In some embodiments, an application may request a PCF from the device when the application has virtual content to display to the user. The PCF request from the application may

trigger a PP request, and a new PP would be created after the device travels the threshold distance. FIG. 18 illustrates a first persistent pose PP1 which may have the closest key frames, (e.g. KF1, KF2, and KF3) attached by, for example, computing relative poses between the key frames to the persistent pose. FIG. 18 also illustrates a second persistent pose PP2 which may have the closest key frames (e.g. KF4 and KF5) attached.

[0308] FIG. 19 illustrates an example of generating a PCF for the map of FIG. 17, according to some embodiments. In the illustrated example, a PCF 1 may include PP1 and PP2. As described above, the PCF may be used for displaying image data relative to the PCF. In some embodiments, each PCF may have coordinates in another coordinate frame (e.g., a world coordinate frame) and a PCF descriptor, for example, uniquely identifying the PCF. In some embodiments, the PCF descriptor may be computed based on feature descriptors of features in frames associated with the PCF. In some embodiments, various constellations of PCFs may be combined to represent the real world in a persistent manner and that requires less data and less transmission of data.

[0309] FIGS. 20A to 20C are schematic diagrams illustrating an example of establishing and using a persistent coordinate frame. FIG. 20A shows two users 4802A, 4802B with respective local tracking maps 4804A, 4804B that have not localized to a canonical map. The origins 4806A, 4806B for individual users are depicted by the coordinate system (e.g., a world coordinate system) in their respective areas. These origins of each tracking map may be local to each user, as the origins are dependent on the orientation of their respective devices when tracking was initiated.

[0310] As the sensors of the user device scan the environment, the device may capture images that, as described above in connection with FIG. 14, may contain features representing persistent objects such that those images may be classified as key frames, from which a persistent pose may be created. In this example, the tracking map 4802A includes a persistent pose (PP) 4808A; the tracking 4802B includes a PP 4808B.

[0311] Also as described above in connection with FIG. 14, some of the PP's may be classified as PCFs which are used to determine the orientation of virtual content for rendering it to the user. FIG. 20B shows that XR devices worn by respective users 4802A, 4802B may create local PCFs 4810A, 4810B based on the PP 4808A, 4808B. FIG. 20C shows that persistent content 4812A, 4812B (e.g., a virtual content) may be attached to the PCFs 4810A, 4810B by respective XR devices.

[0312] In this example, virtual content may have a virtual content coordinate frame, that may be used by an application generating virtual content, regardless of how the virtual content should be displayed. The virtual content, for example, may be specified as surfaces, such as triangles of a mesh, at particular locations and angles with respect to the virtual content coordinate frame. To render that virtual content to a user, the locations of those surfaces may be determined with respect to the user that is to perceive the virtual content.

[0313] Attaching virtual content to the PCFs may simplify the computation involved in determining locations of the virtual content with respect to the user. The location of the virtual content with respect to a user may be determined by applying a series of transformations. Some of those transformations may change, and may be updated frequently. Others of those transformations may be stable and may be updated frequently or not at all. Regardless, the transformations may be applied with relatively low computational burden such that the location of the virtual content can be updated with respect to the user frequently, providing a realistic appearance to the rendered virtual content.

[0314] In the example of FIGS. 20A-20C, user 1's device has a coordinate system that can be related to the coordinate system that defines the origin of the map by the transformation $rig1_T_w1$. User 2's device has a similar transformation $rig2_T_w2$. These transformations may be expressed as six degrees of transformation, specifying translation and rotation to align the device coordinate systems with the map coordinate systems. In some embodiments, the transformation may be expressed as two separate transformations, one specifying translation and the other specifying rotation. Accordingly, it should be appreciated that the transformations may be

expressed in a form that simplifies computation or otherwise provides an advantage.

[0315] Transformations between the origins of the tracking maps and the PCFs identified by the respective user devices are expressed as $pcf1_T_w1$ and $pcf2_T_w2$. In this example the PCF and the PP are identical, such that the same transformation also characterizes the PP's.

[0316] The location of the user device with respect to the PCF can therefore be computed by the serial application of these transformations, such as $rig1_T_pcf1 = (rig1_T_w1) * (pcf1_T_w1)$.

[0317] As shown in FIG. 20C, the virtual content is located with respect to the PCFs, with a transformation of $obj1_T_pcf1$. This transformation may be set by an application generating the virtual content that may receive information from a world reconstruction system describing physical objects with respect to the PCF. To render the virtual content to the user, a transformation to the coordinate system of the user's device is computed, which may be computed by relating the virtual content coordinate frame to the origin of the tracking map through the transformation $obj1_t_w1 = (obj1_T_pcf1) * (pcf1_T_w1)$. That transformation may then be related to the user's device through further transformation $rig1_T_w1$.

[0318] The location of the virtual content may change, based on output from an application generating the virtual content. When that changes, the end-to-end transformation, from a source coordinate system to a destination coordinate system, may be recomputed. Additionally, the location and/or headpose of the user may change as the user moves. As a result, the transformation $rig1_T_w1$ may change, as would any end-to-end transformation that depends on the location or headpose of the user.

[0319] The transformation $rig1_T_w1$ may be updated with motion of the user based on tracking the position of the user with respect to stationary objects in the physical world. Such tracking may be performed by a headphone tracking component processing a sequence of images, as described above, or other component of the system. Such updates may be made by determining pose of the user with respect to a stationary frame of reference, such as a PP.

[0320] In some embodiments, the location and orientation of a user device may be determined relative to the nearest persistent pose, or, in this example, a PCF, as the PP is used as a PCF. Such a determination may be made by identifying in current images captured with sensors on the device feature points that characterize the PP. Using image processing techniques, such as stereoscopic image analysis, the location of the device with respect to those feature points may be determined. From this data, the system could calculate the change in transformation associated with the user's motions based on the relationship $rig1_T_pcf1 = (rig1_T_w1) * (pcf1_T_w1)$.

[0321] A system may determine and apply transformations in an order that is computationally efficient. For example, the need to compute $rig1_T_w1$ from a measurement yielding $rig1_T_pcf1$ might be avoided by tracking both user pose and defining the location of virtual content relative to the PP or a PCF built on a persistent pose. In this way the transformation from a source coordinate system of the virtual content to the destination coordinate system of the user's device may be based on the measured transformation according to the expression $(rig1_T_pcf1) * (obj1_t_pcf1)$, with the first transformation being measured by the system and the latter transformation being supplied by an application specifying virtual content for rendering. In embodiments in which the virtual content is positioned with respect to the origin of the map, the end-to-end transformation may relate the virtual object coordinate system to the PCF coordinate system based on a further transformation between the map coordinates and the PCF coordinates. In embodiments in which the virtual content is positioned with respect to a different PP or PCF than the one against which user position is being tracked, a transformation between the two may be applied. Such a transformation may be fixed and may be determined, for example, from a map in which both appear.

[0322] A transform-based approach may be implemented, for example, in a device with components that process sensor data to build a tracking map. As part of that process, those components may identify feature points that may be used as persistent poses, which in turn may be turned into PCFs. Those components may limit the number of persistent poses generated for the

map, to provide a suitable spacing between persistent poses, while allowing the user, regardless of location in the physical environment, to be close enough to a persistent pose location to accurately compute the user's pose, as described above in connection with FIGS. 17-19. As the closest persistent pose to a user is updated, as a result of user movement, refinements to the tracking map or other causes, any of the transformations that are used to compute the location of virtual content relative to the user that depend on the location of the PP (or PCF if being used) may be updated and stored for use, at least until the user moves away from that persistent pose. Nonetheless, by computing and storing transformations, the computational burden each time the location of virtual content is update may be relatively low, such that it may be performed with relatively low latency. [0323] FIGS. 20A-20C illustrate positioning with respect to a tracking map, and each device had its own tracking map. However, transformations may be generated with respect to any map coordinate system. Persistence of content across user sessions of an XR system may be achieved by using a persistent map. Shared experiences of users may also be facilitated by using a map to which multiple user devices may be oriented.

[0324] In some embodiments, described in greater detail below, the location of virtual content may be specified in relation to coordinates in a canonical map, formatted such that any of multiple devices may use the map. Each device might maintain a tracking map and may determine the change of pose of the user with respect to the tracking map. In this example, a transformation between the tracking map and the canonical map may be determined through a process of “localization”-which may be performed by matching structures in the tracking map (such as one or more persistent poses) to one or more structures of the canonical map (such as one or more PCFs). [0325] Described in greater below are techniques for creating and using canonical maps in this way.

Deep Key Frame

[0326] Techniques as described herein rely on comparison of image frames. For example, to establish the position of a device with respect to a tracking map, a new image may be captured with sensors worn by the user and an XR system may search, in a set of images that were used to create the tracking map, images that share at least a predetermined amount of interest points with the new image. As an example of another scenario involving comparisons of image frames, a tracking map might be localized to a canonical map by first finding image frames associated with a persistent pose in the tracking map that is similar to an image frame associated with a PCF in the canonical map. Alternatively, a transformation between two canonical maps may be computed by first finding similar image frames in the two maps.

[0327] Deep key frames provide a way to reduce the amount of processing required to identify similar image frames. For example, in some embodiments, the comparison may be between image features in a new 2D image (e.g., “2D features”) and 3D features in the map. Such a comparison may be made in any suitable way, such as by projecting the 3D images into a 2D plane. A conventional method such as Bag of Words (BoW) searches the 2D features of a new image in a database including all 2D features in a map, which may require significant computing resources especially when a map represents a large area. The conventional method then locates the images that share at least one of the 2D features with the new image, which may include images that are not useful for locating meaningful 3D features in the map. The conventional method then locates 3D features that are not meaningful with respect to the 2D features in the new image.

[0328] The inventors have recognized and appreciated techniques to retrieve images in the map using less memory resource (e.g., a quarter of the memory resource used by BoW), higher efficiency (e.g., 2.5 ms processing time for each key frame, 100 us for comparing against 500 key frames), and higher accuracy (e.g., 20% better retrieval recall than BoW for 1024 dimensional model, 5% better retrieval recall than BoW for 256 dimensional model).

[0329] To reduce computation, a descriptor may be computed for an image frame that may be used to compare an image frame to other image frames. The descriptors may be stored instead of or in

addition to the image frames and feature points. In a map in which persistent poses and/or PCFs may be generated from image frames, the descriptor of the image frame or frames from which each persistent pose or PCF was generated may be stored as part of the persistent pose and/or PCF. [0330] In some embodiments, the descriptor may be computed as a function of feature points in the image frame. In some embodiments, a neural network is configured to compute a unique frame descriptor to represent an image. The image may have a resolution higher than 1 Megabyte such that enough details of a 3D environment within a field-of-view of a device worn by a user is captured in the image. The frame descriptor may be much shorter, such as a string of numbers, for example, in the range of 128 Bytes to 512 Bytes or any number in between.

[0331] In some embodiments, the neural network is trained such that the computed frame descriptors indicate similarity between images. Images in a map may be located by identifying, in a database comprising images used to generate the map, the nearest images that may have frame descriptors within a predetermined distance to a frame descriptor for a new image. In some embodiments, the distances between images may be represented by a difference between the frame descriptors of the two images.

[0332] FIG. 21 is a block diagram illustrating a system for generating a descriptor for an individual image, according to some embodiments. In the illustrated example, a frame embedding generator 308 is shown. The frame embedding generator 308, in some embodiments, may be used within the server 20, but may alternatively or additionally execute in whole or in part within one of the XR devices 12.1 and 12.2, or any other device processing images for comparison to other images.

[0333] In some embodiments, the frame embedding generator may be configured to generate a reduced data representation of an image from an initial size (e.g., 76,800 bytes) to a final size (e.g., 256 bytes) that is nonetheless indicative of the content in the image despite a reduced size. In some embodiments, the frame embedding generator may be used to generate a data representation for an image which may be a key frame or a frame used in other ways. In some embodiments, the frame embedding generator 308 may be configured to convert an image at a particular location and orientation into a unique string of numbers (e.g., 256 bytes). In the illustrated example, an image 320 taken by an XR device may be processed by feature extractor 324 to detect interest points 322 in the image 320. Interest points may be or may not be derived from feature points identified as described above for features 1120 (FIG. 14) or as otherwise described herein. In some embodiments, interest points may be represented by descriptors as described above for descriptors 1130 (FIG. 14), which may be generated using a deep sparse feature method. In some embodiments, each interest point 322 may be represented by a string of numbers (e.g., 32 bytes). There may, for example, be n features (e.g., 100) and each feature is represented by a string of 32 bytes.

[0334] In some embodiments, the frame embedding generator 308 may include a neural network 326. The neural network 326 may include a multi-layer perceptron unit 312 and a maximum (max) pool unit 314. In some embodiments, the multi-layer perceptron (MLP) unit 312 may comprise a multi-layer perceptron, which may be trained. In some embodiments, the interest points 322 (e.g., descriptors for the interest points) may be reduced by the multi-layer perceptron 312, and may output as weighted combinations 310 of the descriptors. For example, the MLP may reduce n features to m feature that is less than n features.

[0335] In some embodiments, the MLP unit 312 may be configured to perform matrix multiplication. The multi-layer perceptron unit 312 receives the plurality of interest points 322 of an image 320 and converts each interest point to a respective string of numbers (e.g., 256). For example, there may be 100 features and each feature may be represented by a string of 256 numbers. A matrix, in this example, may be created having 100 horizontal rows and 256 vertical columns. Each row may have a series of 256 numbers that vary in magnitude with some being smaller and others being larger. In some embodiments, the output of the MLP may be an $n \times 256$ matrix, where n represents the number of interest points extracted from the image. In some

embodiments, the output of the MLP may be an $m \times 256$ matrix, where m is the number of interest points reduced from n .

[0336] In some embodiments, the MLP **312** may have a training phase, during which model parameters for the MLP are determined, and a use phase. In some embodiments, the MLP may be trained as illustrated in FIG. **25**. The input training data may comprise data in sets of three, the set of three comprising 1) a query image, 2) a positive sample, and 3) a negative sample. The query image may be considered the reference image.

[0337] In some embodiments, the positive sample may comprise an image that is similar to the query image. For example, in some embodiments, similar may be having the same object in both the query and positive sample image but viewed from a different angle. In some embodiments, similar may be having the same object in both the query and positive sample images but having the object shifted (e.g., left, right, up, down) relative to the other image.

[0338] In some embodiments, the negative sample may comprise an image that is dissimilar to the query image. For example, in some embodiments, a dissimilar image may not contain any objects that are prominent in the query image or may contain only a small portion of a prominent object in the query image (e.g., $<10\%$, 1%). A similar image, in contrast, may have most of an object (e.g. $>50\%$, or $>75\%$) in the query image, for example.

[0339] In some embodiments, interest points may be extracted from the images in the input training data and may be converted to feature descriptors. These descriptors may be computed both for the training images as shown in FIG. **25** and for extracted features in operation of frame embedding generator **308** of FIG. **21**. In some embodiments, a deep sparse feature (DSF) process may be used to generate the descriptors (e.g., DSF descriptors) as described in U.S. patent application Ser. No. 16/190,948. In some embodiments, DSF descriptors are $n \times 32$ dimension. The descriptors may then be passed through the model/MLP to create a 256 byte output. In some embodiments, the model/MLP may have the same structure as MLP **312** such that once the model parameters are set through training, the resulting trained MLP may be used as MLP **312**.

[0340] In some embodiments, the feature descriptors (e.g., the 256 byte output from the MLP model) may then be sent to a triplet margin loss module (which may only be used during the training phase, not during use phase of the MLP neural network). In some embodiments, the triplet margin loss module may be configured to select parameters for the model so as to reduce the difference between the 256 byte output from the query image and the 256 byte output from the positive sample, and to increase the difference between the 256 byte output from the query image and the 256 byte output from the negative sample. In some embodiments, the training phase may comprise feeding a plurality of triplet input images into the learning process to determine model parameters. This training process may continue, for example, until the differences for positive images is minimized and the difference for negative images is maximized or until other suitable exit criteria are reached.

[0341] Referring back to FIG. **21**, the frame embedding generator **308** may include a pooling layer, here illustrated as maximum (max) pool unit **314**. The max pool unit **314** may analyze each column to determine a maximum number in the respective column. The max pool unit **314** may combine the maximum value of each column of numbers of the output matrix of the MLP **312** into a global feature string **316** of, for example, 256 numbers. It should be appreciated that images processed in XR systems might, desirably, have high-resolution frames, with potentially millions of pixels. The global feature string **316** is a relatively small number that takes up relatively little memory and is easily searchable compared to an image (e.g., with a resolution higher than 1 Megabyte). It is thus possible to search for images without analyzing each original frame from the camera and it is also cheaper to store 256 bytes instead of complete frames.

[0342] FIG. **22** is a flow chart illustrating a method **2200** of computing an image descriptor, according to some embodiments. The method **2200** may start from receiving (Act **2202**) a plurality of images captured by an XR device worn by a user. In some embodiments, the method **2200** may

include determining (Act **2204**) one or more key frames from the plurality of images. In some embodiments, Act **2204** may be skipped and/or may occur after step **2210** instead.

[0343] The method **2200** may include identifying (Act **2206**) one or more interest points in the plurality of images with an artificial neural network, and computing (Act **2208**) feature descriptors for individual interest points with the artificial neural network. The method may include computing (Act **2210**), for each image, a frame descriptor to represent the image based, at least in part, on the computed feature descriptors for the identified interest points in the image with the artificial neural network.

[0344] FIG. **23** is a flow chart illustrating a method **2300** of localization using image descriptors, according to some embodiments. In this example, a new image frame, depicting the current location of the XR device may be compared to image frames stored in connection with points in a map (such as a persistent pose or a PCF as described above). The method **2300** may start from receiving (Act **2302**) a new image captured by an XR device worn by a user. The method **2300** may include identifying (Act **2304**) one or more nearest key frames in a database comprising key frames used to generate one or more maps. In some embodiments, a nearest key frame may be identified based on coarse spatial information and/or previously determined spatial information. For example, coarse spatial information may indicate that the XR device is in a geographic region represented by a 50 m×50 m area of a map. Image matching may be performed only for points within that area. As another example, based on tracking, the XR system may know that an XR device was previously proximate a first persistent pose in the map and was moving in a direction of a second persistent pose in the map. That second persistent pose may be considered the nearest persistent pose and the key frame stored with it may be regarded as the nearest key frame. Alternatively or additionally, other metadata, such as GPS data or WiFi fingerprints, may be used to select a nearest key frame or set of nearest key frames.

[0345] Regardless of how the nearest key frames are selected, frame descriptors may be used to determine whether the new image matches any of the frames selected as being associated with a nearby persistent pose. The determination may be made by comparing a frame descriptor of the new image with frame descriptors of the closest key frames, or a subset of key frames in the database selected in any other suitable way, and selecting key frames with frame descriptors that are within a predetermined distance of the frame descriptor of the new image. In some embodiments, a distance between two frame descriptors may be computed by obtaining the difference between two strings of numbers that may represent the two frame descriptors. In embodiments in which the strings are processed as strings of multiple quantities, the difference may be computed as a vector difference.

[0346] Once a matching image frame is identified, the orientation of the XR device relative to that image frame may be determined. The method **2300** may include performing (Act **2306**) feature matching against 3D features in the maps that correspond to the identified nearest key frames, and computing (Act **2308**) pose of the device worn by the user based on the feature matching results. In this way, the computationally intensive matching of features points in two images may be performed for as few as one image that has already been determined to be a likely match for the new image.

[0347] FIG. **24** is a flow chart illustrating a method **2400** of training a neural network, according to some embodiments. The method **2400** may start from generating (Act **2402**) a dataset comprising a plurality of image sets. Each of the plurality of image sets may include a query image, a positive sample image, and a negative sample image. In some embodiments, the plurality of image sets may include synthetic recording pairs configured to, for example, teach the neural network basic information such as shapes. In some embodiments, the plurality of image sets may include real recording pairs, which may be recorded from a physical world.

[0348] In some embodiments, inliers may be computed by fitting a fundamental matrix between two images. In some embodiments, sparse overlap may be computed as the intersection over union

(IoU) of interest points seen in both images. In some embodiments, a positive sample may include at least twenty interest points, serving as inliers, that are the same as in the query image. A negative sample may include less than ten inlier points. A negative sample may have less than half of the sparse points overlapping with the sparse points of the query image.

[0349] The method **2400** may include computing (Act **2404**), for each image set, a loss by comparing the query image with the positive sample image and the negative sample image. The method **2400** may include modifying (Act **2406**) the artificial neural network based on the computed loss such that a distance between a frame descriptor generated by the artificial neural network for the query image and a frame descriptor for the positive sample image is less than a distance between the frame descriptor for the query image and a frame descriptor for the negative sample image.

[0350] It should be appreciated that although methods and apparatus configured to generate global descriptors for individual images are described above, methods and apparatus may be configured to generate descriptors for individual maps. For example, a map may include a plurality of key frames, each of which may have a frame descriptor as described above. A max pool unit may analyze the frame descriptors of the map's key frames and combines the frame descriptors into a unique map descriptor for the map.

[0351] Further, it should be appreciated that other architectures may be used for processing as described above. For example, separate neural networks are described for generating DSF descriptors and frame descriptors. Such an approach is computationally efficient. However, in some embodiments, the frame descriptors may be generated from selected feature points, without first generating DSF descriptors.

Ranking and Merging Maps

[0352] Described herein are methods and apparatus for ranking and merging a plurality of environment maps in an cross reality (XR) system. Map merging may enable maps representing overlapping portions of the physical world to be combined to represent a larger area. Ranking maps may enable efficiently performing techniques as described herein, including map merging, that involve selecting a map from a set of maps based on similarity. In some embodiments, for example, a set of canonical maps formatted in a way that they may be accessed by any of a number of XR devices, may be maintained by the system. These canonical maps may be formed by merging selected tracking maps from those devices with other tracking maps or previously stored canonical maps. The canonical maps may be ranked, for example, for use in selecting one or more canonical maps to merge with a new tracking map and/or to select one or more canonical maps from the set to use within a device.

[0353] To provide realistic XR experiences to users, the XR system must know the user's physical surroundings in order to correctly correlate locations of virtual objects in relation to real objects. Information about a user's physical surroundings may be obtained from an environment map for the user's location.

[0354] The inventors have recognized and appreciated that an XR system could provide an enhanced XR experience to multiple users sharing a same world, comprising real and/or virtual content, by enabling efficient sharing of environment maps of the real/physical world collected by multiple users, whether those users are present in the world at the same or different times. However, there are significant challenges in providing such a system. Such a system may store multiple maps generated by multiple users and/or the system may store multiple maps generated at different times. For operations that might be performed with a previously generated map, such as localization, for example as described above, substantial processing may be required to identify a relevant environment map of a same world (e.g. same real world location) from all the environment maps collected in an XR system. In some embodiments, there may only be a small number of environment maps a device could access, for example for localization. In some embodiments, there may be a large number of environment maps a device could access. The inventors have recognized

and appreciated techniques to quickly and accurately rank the relevance of environment maps out of all possible environment maps, such as the universe of all canonical maps **120** in FIG. **28**, for example. A high ranking map may then be selected for further processing, such as to render virtual objects on a user display realistically interacting with the physical world around the user or merging map data collected by that user with stored maps to create larger or more accurate maps. [0355] In some embodiments, a stored map that is relevant to a task for a user at a location in the physical world may be identified by filtering stored maps based on multiple criteria. Those criteria may indicate comparisons of a tracking map, generated by the wearable device of the user in the location, to candidate environment maps stored in a database. The comparisons may be performed based on metadata associated with the maps, such as a Wi-Fi fingerprint detected by the device generating the map and/or set of BSSID's to which the device was connected while forming the map. The comparisons may also be performed based on compressed or uncompressed content of the map. Comparisons based on a compressed representation may be performed, for example, by comparison of vectors computed from map content. Comparisons based on un-compressed maps may be performed, for example, by localizing the tracking map within the stored map, or vice versa. Multiple comparisons may be performed in an order based on computation time needed to reduce the number of candidate maps for consideration, with comparisons involving less computation being performed earlier in the order than other comparisons requiring more computation.

[0356] FIG. **26** depicts an AR system **800** configured to rank and merge one or more environment maps, according to some embodiments. The AR system may include a passable world model **802** of an AR device. Information to populate the passable world model **802** may come from sensors on the AR device, which may include computer executable instructions stored in a processor **804** (e.g., a local data processing module **570** in FIG. **4**), which may perform some or all of the processing to convert sensor data into a map. Such a map may be a tracking map, as it can be built as sensor data is collected as the AR device operates in a region. Along with that tracking map, area attributes may be supplied so as to indicate the area that the tracking map represents. These area attributes may be a geographic location identifier, such as coordinates presented as latitude and longitude or an ID used by the AR system to represent a location. Alternatively or additionally, the area attributes may be measured characteristics that have a high likelihood of being unique for that area. The area attributes, for example, may be derived from parameters of wireless networks detected in the area. In some embodiments, the area attribute may be associated with a unique address of an access-point the AR system is nearby and/or connected to. For example, the area attribute may be associated with a MAC address or basic service set identifiers (BSSIDs) of a 5G base station/router, a Wi-Fi router, and the like.

[0357] In the example of FIG. **26**, the tracking maps may be merged with other maps of the environment. A map rank portion **806** receives tracking maps from the device PW **802** and communicates with a map database **808** to select and rank environment maps from the map database **808**. Higher ranked, selected maps are sent to a map merge portion **810**.

[0358] The map merge portion **810** may perform merge processing on the maps sent from the map rank portion **806**. Merge processing may entail merging the tracking map with some or all of the ranked maps and transmitting the new, merged maps to a passable world model **812**. The map merge portion may merge maps by identifying maps that depict overlapping portions of the physical world. Those overlapping portions may be aligned such that information in both maps may be aggregated into a final map. Canonical maps may be merged with other canonical maps and/or tracking maps.

[0359] The aggregation may entail extending one map with information from another map. Alternatively or additionally, aggregation may entail adjusting the representation of the physical world in one map, based on information in another map. A later map, for example, may reveal that objects giving rise to feature points have moved, such that the map may be updated based on later

information. Alternatively, two maps may characterize the same region with different feature points and aggregating may entail selecting a set of feature points from the two maps to better represent that region. Regardless of the specific processing that occurs in the merging process, in some embodiments, PCFs from all maps that are merged may be retained, such that applications positioning content with respect to them may continue to do so. In some embodiments, merging of maps may result in redundant persistent poses, and some of the persistent poses may be deleted. When a PCF is associated with a persistent pose that is to be deleted, merging maps may entail modifying the PCF to be associated with a persistent pose remaining in the map after merging.

[0360] In some embodiments, as maps are extended and or updated, they may be refined.

Refinement may entail computation to reduce internal inconsistency between feature points that likely represent the same object in the physical world. Inconsistency may result from inaccuracies in the poses associated with key frames supplying feature points that represent the same objects in the physical world. Such inconsistency may result, for example, from an XR device computing poses relative to a tracking map, which in turn is built based on estimating poses, such that errors in pose estimation accumulate, creating a “drift” in pose accuracy over time. By performing a bundle adjustment or other operation to reduce inconsistencies of the feature points from multiple key frames, the map may be refined.

[0361] Upon a refinement, the location of a persistent point relative to the origin of a map may change. Accordingly, the transformation associated with that persistent point, such as a persistent pose or a PCF, may change. In some embodiments, the XR system, in connection with map refinement (whether as part of a merge operation or performed for other reasons) may re-compute transformations associated with any persistent points that have changed. These transformations might be pushed from a component computing the transformations to a component using the transformation such that any uses of the transformations may be based on the updated location of the persistent points.

[0362] Passable world model **812** may be a cloud model, which may be shared by multiple AR devices. Passable world model **812** may store or otherwise have access to the environment maps in map database **808**. In some embodiments, when a previously computed environment map is updated, the prior version of that map may be deleted so as to remove out of date maps from the database. In some embodiments, when a previously computed environment map is updated, the prior version of that map may be archived enabling retrieving/viewing prior versions of an environment. In some embodiments, permissions may be set such that only AR systems having certain read/write access may trigger prior versions of maps being deleted/archived.

[0363] These environment maps created from tracking maps supplied by one or more AR devices/systems may be accessed by AR devices in the AR system. The map rank portion **806** also may be used in supplying environment maps to an AR device. The AR device may send a message requesting an environment map for its current location, and map rank portion **806** may be used to select and rank environment maps relevant to the requesting device.

[0364] In some embodiments, the AR system **800** may include a downsample portion **814** configured to receive the merged maps from the cloud PW **812**. The received merged maps from the cloud PW **812** may be in a storage format for the cloud, which may include high resolution information, such as a large number of PCFs per square meter or multiple image frames or a large set of feature points associated with a PCF. The downsample portion **814** may be configured to downsample the cloud format maps to a format suitable for storage on AR devices. The device format maps may have less data, such as fewer PCFs or less data stored for each PCF to accommodate the limited local computing power and storage space of AR devices.

[0365] FIG. **27** is a simplified block diagram illustrating a plurality of canonical maps **120** that may be stored in a remote storage medium, for example, a cloud. Each canonical map **120** may include a plurality of canonical map identifiers indicating the canonical map's location within a physical space, such as somewhere on the planet earth. These canonical map identifiers may include one or

more of the following identifiers: area identifiers represented by a range of longitudes and latitudes, frame descriptors (e.g., global feature string **316** in FIG. **21**), Wi-Fi fingerprints, feature descriptors (e.g., feature descriptors **310** in FIG. **21**), and device identities indicating one or more devices that contributed to the map.

[0366] In the illustrated example, the canonical maps **120** are disposed geographically in a two-dimensional pattern as they may exist on a surface of the earth. The canonical maps **120** may be uniquely identifiable by corresponding longitudes and latitudes because any canonical maps that have overlapping longitudes and latitudes may be merged into a new canonical map.

[0367] FIG. **28** is a schematic diagram illustrating a method of selecting canonical maps, which may be used for localizing a new tracking map to one or more canonical maps, according to some embodiment. The method may start from accessing (Act **120**) a universe of canonical maps **120**, which may be stored, as an example, in a database in a passable world (e.g., the passable world module **538**). The universe of canonical maps may include canonical maps from all previously visited locations. An XR system may filter the universe of all canonical maps to a small subset or just a single map. It should be appreciated that, in some embodiments, it may not be possible to send all the canonical maps to a viewing device due to bandwidth restrictions. Selecting a subset selected as being likely candidates for matching the tracking map to send to the device may reduce bandwidth and latency associated with accessing a remote database of maps.

[0368] The method may include filtering (Act **300**) the universe of canonical maps based on areas with predetermined size and shapes. In the illustrated example in FIG. **27**, each square may represent an area. Each square may cover 50 m×50 m. Each square may have six neighboring areas. In some embodiments, Act **300** may select at least one matching canonical map **120** covering longitude and latitude that include that longitude and latitude of the position identifier received from an XR device, as long as at least one map exists at that longitude and latitude. In some embodiments, the Act **300** may select at least one neighboring canonical map covering longitude and latitude that are adjacent the matching canonical map. In some embodiments, the Act **300** may select a plurality of matching canonical maps and a plurality of neighboring canonical maps. The Act **300** may, for example, reduce the number of canonical maps approximately ten times, for example, from thousands to hundreds to form a first filtered selection. Alternatively or additionally, criteria other than latitude and longitude may be used to identify neighboring maps. An XR device, for example, may have previously localized with a canonical map in the set as part of the same session. A cloud service may retain information about the XR device, including maps previously localized to. In this example, the maps selected at Act **300** may include those that cover an area adjacent to the map to which the XR device localized to.

[0369] The method may include filtering (Act **302**) the first filtered selection of canonical maps based on Wi-Fi fingerprints. The Act **302** may determine a latitude and longitude based on a Wi-Fi fingerprint received as part of the position identifier from an XR device. The Act **302** may compare the latitude and longitude from the Wi-Fi fingerprint with latitude and longitude of the canonical maps **120** to determine one or more canonical maps that form a second filtered selection. The Act **302** may reduce the number of canonical maps approximately ten times, for example, from hundreds to tens of canonical maps (e.g., 50) that form a second selection. For example, a first filtered selection may include 130 canonical maps and the second filtered selection may include 50 of the **130** canonical maps and may not include the other 80 of the **130** canonical maps.

[0370] The method may include filtering (Act **304**) the second filtered selection of canonical maps based on key frames. The Act **304** may compare data representing an image captured by an XR device with data representing the canonical maps **120**. In some embodiments, the data representing the image and/or maps may include feature descriptors (e.g., DSF descriptors in FIG. **25**) and/or global feature strings (e.g., **316** in FIG. **21**). The Act **304** may provide a third filtered selection of canonical maps. In some embodiments, the output of Act **304** may only be five of the 50 canonical maps identified following the second filtered selection, for example. The map transmitter **122** then

transmits the one or more canonical maps based on the third filtered selection to the viewing device. The Act **304** may reduce the number of canonical maps for approximately ten times, for example, from tens to single digits of canonical maps (e.g., 5) that form a third selection. In some embodiments, an XR device may receive canonical maps in the third filtered selection, and attempt to localize into the received canonical maps.

[0371] For example, the Act **304** may filter the canonical maps **120** based on the global feature strings **316** of the canonical maps **120** and the global feature string **316** that is based on an image that is captured by the viewing device (e.g. an image that may be part of the local tracking map for a user). Each one of the canonical maps **120** in FIG. **27** thus has one or more global feature strings **316** associated therewith. In some embodiments, the global feature strings **316** may be acquired when an XR device submits images or feature details to the cloud and the cloud processes the image or feature details to generate global feature strings **316** for the canonical maps **120**.

[0372] In some embodiments, the cloud may receive feature details of a live/new/current image captured by a viewing device, and the cloud may generate a global feature string **316** for the live image. The cloud may then filter the canonical maps **120** based on the live global feature string **316**. In some embodiments, the global feature string may be generated on the local viewing device. In some embodiments, the global feature string may be generated remotely, for example on the cloud. In some embodiments, a cloud may transmit the filtered canonical maps to an XR device together with the global feature strings **316** associated with the filtered canonical maps. In some embodiments, when the viewing device localizes its tracking map to the canonical map, it may do so by matching the global feature strings **316** of the local tracking map with the global feature strings of the canonical map.

[0373] It should be appreciated that an operation of an XR device may not perform all of the Acts (**300**, **302**, **304**). For example, if a universe of canonical maps are relatively small (e.g., **500** maps), an XR device attempting to localize may filter the universe of canonical maps based on Wi-Fi fingerprints (e.g., Act **302**) and Key Frame (e.g., Act **304**), but omit filtering based on areas (e.g., Act **300**). Moreover, it is not necessary that maps in their entirety be compared. In some embodiments, for example, a comparison of two maps may result in identifying common persistent points, such as persistent poses or PCFs that appear in both the new map the selected map from the universe of maps. In that case, descriptors may be associated with persistent points, and those descriptors may be compared.

[0374] FIG. **29** is flow chart illustrating a method **900** of selecting one or more ranked environment maps, according to some embodiments. In the illustrated embodiment, the ranking is performed for a user's AR device that is creating a tracking map. Accordingly, the tracking map is available for use in ranking environment maps. In embodiments in which the tracking map is not available, some or all of portions of the selection and ranking of environment maps that do not expressly rely on the tracking map may be used.

[0375] The method **900** may start at Act **902**, where a set of maps from a database of environment maps (which may be formatted as canonical maps) that are in the neighborhood of the location where the tracking map was formed may be accessed and then filtered for ranking. Additionally, at Act **902**, at least one area attribute for the area in which the user's AR device is operating is determined. In scenarios in which the user's AR device is constructing a tracking map, the area attributes may correspond to the area over which the tracking map was created. As a specific example, the area attributes may be computed based on received signals from access points to computer networks while the AR device was computing the tracking map.

[0376] FIG. **30** depicts an exemplary map rank portion **806** of the AR system **800**, according to some embodiments. The map rank portion **806** may be executing in a cloud computing environment, as it may include portions executing on AR devices and portions executing on a remote computing system such as a cloud. The map rank portion **806** may be configured to perform at least a portion of the method **900**.

[0377] FIG. 31A depicts an example of area attributes AA1-AA8 of a tracking map (TM) **1102** and environment maps CM1-CM4 in a database, according to some embodiments. As illustrated, an environment map may be associated to multiple area attributes. The area attributes AA1-AA8 may include parameters of wireless networks detected by the AR device computing the tracking map **1102**, for example, basic service set identifiers (BSSIDs) of networks to which the AR device are connected and/or the strength of the received signals of the access points to the wireless networks through, for example, a network tower **1104**. The parameters of the wireless networks may comply with protocols including Wi-Fi and 5G NR. In the example illustrated in FIG. 32, the area attributes are a fingerprint of the area in which the user AR device collected sensor data to form the tracking map.

[0378] FIG. 31B depicts an example of the determined geographic location **1106** of the tracking map **1102**, according to some embodiments. In the illustrated example, the determined geographic location **1106** includes a centroid point **1110** and an area **1108** circling around the centroid point. It should be appreciated that the determination of a geographic location of the present application is not limited to the illustrated format. A determined geographic location may have any suitable formats including, for example, different area shapes. In this example, the geographic location is determined from area attributes using a database relating area attributes to geographic locations. Databases are commercially available, for example, databases that relate Wi-Fi fingerprints to locations expressed as latitude and longitude and may be used for this operation.

[0379] In the embodiment of FIG. 29, a map database, containing environment maps may also include location data for those maps, including latitude and longitude covered by the maps. Processing at Act **902** may entail selecting from that database a set of environment maps that covers the same latitude and longitude determined for the area attributes of the tracking map.

[0380] Act **904** is a first filtering of the set of environment maps accessed in Act **902**. In Act **902**, environment maps are retained in the set based on proximity to the geolocation of the tracking map. This filtering step may be performed by comparing the latitude and longitude associated with the tracking map and the environment maps in the set.

[0381] FIG. 32 depicts an example of Act **904**, according to some embodiments. Each area attribute may have a corresponding geographic location **1202**. The set of environment maps may include the environment maps with at least one area attribute that has a geographic location overlapping with the determined geographic location of the tracking map. In the illustrated example, the set of identified environment maps includes environment maps CM1, CM2, and CM4, each of which has at least one area attribute that has a geographic location overlapping with the determined geographic location of the tracking map **1102**. The environment map CM3 associated with the area attribute AA6 is not included in the set because it is outside the determined geographic location of the tracking map.

[0382] Other filtering steps may also be performed on the set of environment maps to reduce/rank the number of environment maps in the set that is ultimately processed (such as for map merge or to provide passable world information to a user device). The method **900** may include filtering (Act **906**) the set of environment maps based on similarity of one or more identifiers of network access points associated with the tracking map and the environment maps of the set of environment maps. During the formation of a map, a device collecting sensor data to generate the map may be connected to a network through a network access point, such as through Wi-Fi or similar wireless communication protocol. The access points may be identified by BSSID. The user device may connect to multiple different access points as it moves through an area collecting data to form a map. Likewise, when multiple devices supply information to form a map, the devices may have connected through different access points, so there may be multiple access points used in forming the map for that reason too. Accordingly, there may be multiple access points associated with a map, and the set of access points may be an indication of location of the map. Strength of signals from an access point, which may be reflected as an RSSI value, may provide further geographic

information. In some embodiments, a list of BSSID and RSSI values may form the area attribute for a map.

[0383] In some embodiments, filtering the set of environment maps based on similarity of the one or more identifiers of the network access points may include retaining in the set of environment maps environment maps with the highest Jaccard similarity to the at least one area attribute of the tracking map based on the one or more identifiers of network access points. FIG. 33 depicts an example of Act **906**, according to some embodiments. In the illustrated example, a network identifier associated with the area attribute AA7 may be determined as the identifier for the tracking map **1102**. The set of environment maps after Act **906** includes environment map CM2, which may have area attributes within higher Jaccard similarity to AA7, and environment map CM4, which also include the area attributes AA7. The environment map CM1 is not included in the set because it has the lowest Jaccard similarity to AA7.

[0384] Processing at Acts **902-906** may be performed based on metadata associated with maps and without actually accessing the content of the maps stored in a map database. Other processing may involve accessing the content of the maps. Act **908** indicates accessing the environment maps remaining in the subset after filtering based on metadata. It should be appreciated that this act may be performed either earlier or later in the process, if subsequent operations can be performed with accessed content.

[0385] The method **900** may include filtering (Act **910**) the set of environment maps based on similarity of metrics representing content of the tracking map and the environment maps of the set of environment maps. The metrics representing content of the tracking map and the environment maps may include vectors of values computed from the contents of the maps. For example, the Deep Key Frame descriptor, as described above, computed for one or more key frames used in forming a map may provide a metric for comparison of maps, or portions of maps. The metrics may be computed from the maps retrieved at Act **908** or may be pre-computed and stored as metadata associated with those maps. In some embodiments, filtering the set of environment maps based on similarity of metrics representing content of the tracking map and the environment maps of the set of environment maps, may include retaining in the set of environment maps environment maps with the smallest vector distance between a vector of characteristics of the tracking map and vectors representing environment maps in the set of environment maps.

[0386] The method **900** may include further filtering (Act **912**) the set of environment maps based on degree of match between a portion of the tracking map and portions of the environment maps of the set of environment maps. The degree of match may be determined as a part of a localization process. As a non-limiting example, localization may be performed by identifying critical points in the tracking map and the environment map that are sufficiently similar as they could represent the same portion of the physical world. In some embodiments, the critical points may be features, feature descriptors, key frames/key rigs, persistent poses, and/or PCFs. The set of critical points in the tracking map might then be aligned to produce a best fit with the set of critical points in the environment map. A mean square distance between the corresponding critical points might be computed and, if below a threshold for a particular region of the tracking map, used as an indication that the tracking map and the environment map represent the same region of the physical world.

[0387] In some embodiments, filtering the set of environment maps based on degree of match between a portion of the tracking map and portions of the environment maps of the set of environment maps may include computing a volume of a physical world represented by the tracking map that is also represented in an environment map of the set of environment maps, and retaining in the set of environment maps environment maps with a larger computed volume than environment maps filtered out of the set. FIG. 34 depicts an example of Act **912**, according to some embodiments. In the illustrated example, the set of environment maps after Act **912** includes environment map CM4, which has an area **1402** matched with an area of the tracking map **1102**.

The environment map CM1 is not included in the set because it has no area matched with an area of the tracking map **1102**.

[0388] In some embodiments, the set of environment maps may be filtered in the order of Act **906**, Act **910**, and Act **912**. In some embodiments, the set of environment maps may be filtered based on Act **906**, Act **910**, and Act **912**, which may be performed in an order based on processing required to perform the filtering, from lowest to highest. The method **900** may include loading (Act **914**) the set of environment maps and data.

[0389] In the illustrated example, a user database stores area identities indicating areas that AR devices were used in. The area identities may be area attributes, which may include parameters of wireless networks detected by the AR devices when in use. A map database may store multiple environment maps constructed from data supplied by the AR devices and associated metadata. The associated metadata may include area identities derived from the area identities of the AR devices that supplied data from which the environment maps were constructed. An AR device may send a message to a PW module indicating a new tracking map is created or being created. The PW module may compute area identifiers for the AR device and updates the user database based on the received parameters and/or the computed area identifiers. The PW module may also determine area identifiers associated with the AR device requesting the environment maps, identify sets of environment maps from the map database based on the area identifiers, filter the sets of environment maps, and transmit the filtered sets of environment maps to the AR devices. In some embodiments, the PW module may filter the sets of environment maps based on one or more criteria including, for example, a geographic location of the tracking map, similarity of one or more identifiers of network access points associated with the tracking map and the environment maps of the set of environment maps, similarity of metrics representing contents of the tracking map and the environment maps of the set of environment maps, and degree of match between a portion of the tracking map and portions of the environment maps of the set of environment maps.

[0390] Having thus described several aspects of some embodiments, it is to be appreciated that various alterations, modifications, and improvements will readily occur to those skilled in the art. As one example, embodiments are described in connection with an augmented (AR) environment. It should be appreciated that some or all of the techniques described herein may be applied in an MR environment or more generally in other XR environments, and in VR environments.

[0391] As another example, embodiments are described in connection with devices, such as wearable devices. It should be appreciated that some or all of the techniques described herein may be implemented via networks (such as cloud), discrete applications, and/or any suitable combinations of devices, networks, and discrete applications.

[0392] Further, FIG. **29** provides examples of criteria that may be used to filter candidate maps to yield a set of high ranking maps. Other criteria may be used instead of or in addition to the described criteria. For example, if multiple candidate maps have similar values of a metric used for filtering out less desirable maps, characteristics of the candidate maps may be used to determine which maps are retained as candidate maps or filtered out. For example, larger or more dense candidate maps may be prioritized over smaller candidate maps. In some embodiments, FIGS. **27-28** may describe all or part of the systems and methods described in FIGS. **29-34**.

[0393] FIGS. **35** and **36** are schematic diagrams illustrating an XR system configured to rank and merge a plurality of environment maps, according to some embodiments. In some embodiments, a passable world (PW) may determine when to trigger ranking and/or merging the maps. In some embodiments, determining a map to be used may be based at least partly on deep key frames described above in relation to FIGS. **21-25**, according to some embodiments.

[0394] FIG. **37** is a block diagram illustrating a method **3700** of creating environment maps of a physical world, according to some embodiments. The method **3700** may start from localizing (Act **3702**) a tracking map captured by an XR device worn by a user to a group of canonical maps (e.g., canonical maps selected by the method of FIG. **28** and/or the method **900** of FIG. **29**). The Act

3702 may include localizing keyrigs of the tracking map into the group of canonical maps. The localization result of each keyrig may include the keyrig's localized pose and a set of 2D-to-3D feature correspondences.

[0395] In some embodiments, the method **3700** may include splitting (Act **3704**) a tracking map into connected components, which may enable merging maps robustly by merging connected pieces. Each connected component may include keyrigs that are within a predetermined distance. The method **3700** may include merging (Act **3706**) the connected components that are larger than a predetermined threshold into one or more canonical maps, and removing the merged connected components from the tracking map.

[0396] In some embodiments, the method **3700** may include merging (Act **3708**) canonical maps of the group that are merged with the same connected components of the tracking map. In some embodiments, the method **3700** may include promoting (Act **3710**) the remaining connected components of the tracking map that has not been merged with any canonical maps to be a canonical map. In some embodiments, the method **3700** may include merging (Act **3712**) persistent poses and/or PCFs of the tracking maps and the canonical maps that are merged with at least one connected component of the tracking map. In some embodiments, the method **3700** may include finalizing (Act **3714**) the canonical maps by, for example, fusing map points and pruning redundant keyrigs.

[0397] FIGS. **38A** and **38B** illustrate an environment map **3800** created by updating a canonical map **700**, which may be promoted from the tracking map **700** (FIG. **7**) with a new tracking map, according to some embodiments. As illustrated and described with respect to FIG. **7**, the canonical map **700** may provide a floor plan **706** of reconstructed physical objects in a corresponding physical world, represented by points **702**. In some embodiments, a map point **702** may represent a feature of a physical object that may include multiple features. A new tracking map may be captured about the physical world and uploaded to a cloud to merge with the map **700**. The new tracking map may include map points **3802**, and keyrigs **3804**, **3806**. In the illustrated example, keyrigs **3804** represent keyrigs that are successfully localized to the canonical map by, for example, establishing a correspondence with a keyrig **704** of the map **700** (as illustrated in FIG. **38B**). On the other hand, keyrigs **3806** represent keyrigs that have not been localized to the map **700**. Keyrigs **3806** may be promoted to a separate canonical map in some embodiments.

[0398] FIGS. **39A** to **39F** are schematic diagrams illustrating an example of a cloud-based persistent coordinate system providing a shared experience for users in the same physical space. FIG. **39A** shows that a canonical map **4814**, for example, from a cloud, is received by the XR devices worn by the users **4802A** and **4802B** of FIGS. **20A-20C**. The canonical map **4814** may have a canonical coordinate frame **4806C**. The canonical map **4814** may have a PCF **4810C** with a plurality of associated PPs (e.g., **4818A**, **4818B** in FIG. **39C**).

[0399] FIG. **39B** shows that the XR devices established relationships between their respective world coordinate system **4806A**, **4806B** with the canonical coordinate frame **4806C**. This may be done, for example, by localizing to the canonical map **4814** on the respective devices. Localizing the tracking map to the canonical map may result, for each device, a transformation between its local world coordinate system and the coordinate system of the canonical map.

[0400] FIG. **39C** shows that, as a result of localization, a transformation can be computed (e.g., transformation **4816A**, **4816B**) between a local PCF (e.g., PCFs **4810A**, **4810B**) on the respective device to a respective persistent pose (e.g., PPs **4818A**, **4818B**) on the canonical map. With these transformations, each device may use its local PCFs, which can be detected locally on the device by processing images detected with sensors on the device, to determine where with respect to the local device to display virtual content attached to the PPs **4818A**, **4818B** or other persistent points of the canonical map. Such an approach may accurately position virtual content with respect to each user and may enable each user to have the same experience of the virtual content in the physical space.

[0401] FIG. 39D shows a persistent pose snapshot from the canonical map to the local tracking maps. As can be seen, the local tracking maps are connected to one another via the persistent poses. FIG. 39E shows that the PCF 4810A on the device worn by the user 4802A is accessible in the device worn by the user 4802B through PPs 4818A. FIG. 39F shows that the tracking maps 4804A, 4804B and the canonical 4814 may merge. In some embodiments, some PCFs may be removed as a result of merging. In the illustrated example, the merged map includes the PCF 4810C of the canonical map 4814 but not the PCFs 4810A, 4810B of the tracking maps 4804A, 4804B. The PPs previously associated with the PCFs 4810A, 4810B may be associated with the PCF 4810C after the maps merge.

EXAMPLES

[0402] FIGS. 40 and 41 illustrate an example of using a tracking map by the first XR device 12.1 of FIG. 9. FIG. 40 is a two-dimensional representation of a three-dimensional first local tracking map (Map 1), which may be generated by the first XR device of FIG. 9, according to some embodiments. FIG. 41 is a block diagram illustrating uploading Map 1 from the first XR device to the server of FIG. 9, according to some embodiments.

[0403] FIG. 40 illustrates Map 1 and virtual content (Content123 and Content456) on the first XR device 12.1. Map 1 has an origin (Origin 1). Map 1 includes a number of PCFs (PCF a to PCF d). From the perspective of the first XR device 12.1, PCF a, by way of example, is located at the origin of Map 1 and has X, Y, and Z coordinates of (0,0,0) and PCF b has X, Y, and Z coordinates (-1,0,0). Content123 is associated with PCF a. In the present example, Content123 has an X, Y, and Z relationship relative to PCF a of (1,0,0). Content456 has a relationship relative to PCF b. In the present example, Content456 has an X, Y, and Z relationship of (1,0,0) relative to PCF b.

[0404] In FIG. 41, the first XR device 12.1 uploads Map 1 to the server 20. In this example, as the server stores no canonical map for the same region of the physical world represented by the tracking map, and the tracking map is stored as an initial canonical map. The server 20 now has a canonical map based on Map 1. The first XR device 12.1 has a canonical map that is empty at this stage. The server 20, for purposes of discussion, and in some embodiments, includes no other maps other than Map 1. No maps are stored on the second XR device 12.2.

[0405] The first XR device 12.1 also transmits its Wi-Fi signature data to the server 20. The server 20 may use the Wi-Fi signature data to determine a rough location of the first XR device 12.1 based on intelligence gathered from other devices that have, in the past, connected to the server 20 or other servers together with the GPS locations of such other devices that have been recorded. The first XR device 12.1 may now end the first session (See FIG. 8) and may disconnect from the server 20.

[0406] FIG. 42 is a schematic diagram illustrating the XR system of FIG. 16, showing the second user 14.2 has initiated a second session using a second XR device of the XR system after the first user 14.1 has terminated a first session, according to some embodiments.

[0407] FIG. 43A is a block diagram showing the initiation of a second session by a second user 14.2. The first user 14.1 is shown in phantom lines because the first session by the first user 14.1 has ended. The second XR device 12.2 begins to record objects. Various systems with varying degrees of granulation may be used by the server 20 to determine that the second session by the second XR device 12.2 is in the same vicinity of the first session by the first XR device 12.1. For example, Wi-Fi signature data, global positioning system (GPS) positioning data, GPS data based on Wi-Fi signature data, or any other data that indicates a location may be included in the first and second XR devices 12.1 and 12.2 to record their locations. Alternatively, the PCFs that are identified by the second XR device 12.2 may show a similarity to the PCFs of Map 1.

[0408] As shown in FIG. 43B, the second XR device boots up and begins to collect data, such as images 1110 from one or more cameras 44, 46. As shown in FIG. 14, in some embodiments, an XR device (e.g. the second XR device 12.2) may collect one or more images 1110 and perform image processing to extract one or more features/interest points 1120. Each feature may be converted to a

descriptor **1130**. In some embodiments, the descriptors **1130** may be used to describe a key frame **1140**, which may have the position and direction of the associated image attached. One or more key frames **1140** may correspond to a single persistent pose **1150**, which may be automatically generated after a threshold distance from the previous persistent pose **1150**, e.g., 3 meters. One or more persistent poses **1150** may correspond to a single PCF **1160**, which may be automatically generated after a pre-determined distance, e.g. every 5 meters. Over time as the user continues to move around the user's environment, and the XR device continues to collect more data, such as images **1110**, additional PCFs (e.g., PCF 3 and PCF 4, 5) may be created. One or more applications **1180** may run on the XR device and provide virtual content **1170** to the XR device for presentation to the user. The virtual content may have an associated content coordinate frame which may be placed relative to one or more PCFs. As shown in FIG. **43B**, the second XR device **12.2** creates three PCFs. In some embodiments, the second XR device **12.2** may try to localize into one or more canonical maps stored on the server **20**.

[0409] In some embodiments, as shown in FIG. **43C**, the second XR device **12.2** may download the canonical map **120** from the server **20**. Map 1 on the second XR device **12.2** includes PCFs a to d and Origin 1. In some embodiments, the server **20** may have multiple canonical maps for various locations and may determine that the second XR device **12.2** is in the same vicinity as the vicinity of the first XR device **12.1** during the first session and sends the second XR device **12.2** the canonical map for that vicinity.

[0410] FIG. **44** shows the second XR device **12.2** beginning to identify PCFs for purposes of generating Map 2. The second XR device **12.2** has only identified a single PCF, namely PCF 1,2. The X, Y, and Z coordinates of PCF 1,2 for the second XR device **12.2** may be (1,1,1). Map 2 has its own origin (Origin 2), which may be based on the headpose of device 2 at device start-up for the current headpose session. In some embodiments, the second XR device **12.2** may immediately attempt to localize Map 2 to the canonical map. In some embodiments, Map 2 may not be able to localize into Canonical Map (Map 1) (i.e. localization may fail) because the system does not recognize any or enough overlap between the two maps. Localization may be performed by identifying a portion of the physical world represented in a first map that is also represented in a second map, and computing a transformation between the first map and the second map required to align those portions. In some embodiments, the system may localize based on PCF comparison between the local and canonical maps. In some embodiments, the system may localize based on persistent pose comparison between the local and canonical maps. In some embodiments, the system may localize based on key frame comparison between the local and canonical maps.

[0411] FIG. **45** shows Map 2 after the second XR device **12.2** has identified further PCFs (PCF 1,2, PCF 3, PCF 4,5) of Map 2. The second XR device **12.2** again attempts to localize Map 2 to the canonical map. Because Map 2 has expanded to overlap with at least a portion of the Canonical Map, the localization attempt will succeed. In some embodiments, the overlap between the local tracking map, Map 2, and the Canonical Map may be represented by PCFs, persistent poses, key frames, or any other suitable intermediate or derivative construct.

[0412] Furthermore, the second XR device **12.2** has associated Content123 and Content456 to PCFs 1,2 and PCF 3 of Map 2. Content123 has X, Y, and Z coordinates relative to PCF 1,2 of (1,0,0). Similarly, the X, Y, and Z coordinates of Content456 relative to PCF 3 in Map 2 are (1,0,0).

[0413] FIGS. **46A** and **46B** illustrate a successful localization of Map 2 to the canonical map. Localization may be based on matching features in one map to the other. With an appropriate transformation, here involving both translation and rotation of one map with respect to the other, the overlapping area/volume/section of the maps **1410** represent the common parts to Map 1 and the canonical map. Since Map 2 created PCFs 3 and 4, 5 before localizing, and the Canonical map created PCFs a and c before Map 2 was created, different PCFs were created to represent the same volume in real space (e.g., in different maps).

[0414] As shown in FIG. **47**, the second XR device **12.2** expands Map 2 to include PCFs a-d from

the Canonical Map. The inclusion of PCFs a-d represents the localization of Map 2 to the Canonical Map. In some embodiments, the XR system may perform an optimization step to remove duplicate PCFs from overlapping areas, such as the PCFs in **1410**, PCF 3 and PCF 4,5. After Map 2 localizes, the placement of virtual content, such as Content456 and Content123 will be relative to the closest updated PCFs in the updated Map 2. The virtual content appears in the same real-world location relative to the user, despite the changed PCF attachment for the content, and despite the updated PCFs for Map 2.

[0415] As shown in FIG. **48**, the second XR device **12.2** continues to expand Map 2 as further PCFs (e.g., PCFs e, f, g, and h) are identified by the second XR device **12.2**, for example as the user walks around the real world. It can also be noted that Map 1 has not expanded in FIGS. **47** and **48**.

[0416] Referring to FIG. **49**, the second XR device **12.2** uploads Map 2 to the server **20**. The server **20** stores Map 2 together with the canonical map. In some embodiments, Map 2 may upload to the server **20** when the session ends for the second XR device **12.2**.

[0417] The canonical map within the server **20** now includes PCF i which is not included in Map 1 on the first XR device **12.1**. The canonical map on the server **20** may have expanded to include PCF i when a third XR device (not shown) uploaded a map to the server **20** and such a map included PCF i.

[0418] In FIG. **50**, the server **20** merges Map 2 with the canonical map to form a new canonical map. The server **20** determines that PCFs a to d are common to the canonical map and Map 2. The server expands the canonical map to include PCFs e to h and PCF 1,2 from Map 2 to form a new canonical map. The canonical maps on the first and second XR devices **12.1** and **12.2** are based on Map 1 and are outdated.

[0419] In FIG. **51**, the server **20** transmits the new canonical map to the first and second XR devices **12.1** and **12.2**. In some embodiments, this may occur when the first XR device **12.1** and second device **12.2** try to localize during a different or new or subsequent session. The first and second XR devices **12.1** and **12.2** proceed as described above to localize their respective local maps (Map1 and Map 2 respectively) to the new canonical map.

[0420] As shown in FIG. **52**, the head coordinate frame **96** or “headpose” is related to the PCFs in Map 2. In some embodiments, the origin of the map, Origin 2, is based on the headpose of second XR device **12.2** at the start of the session. As PCFs are created during the session, the PCFs are placed relative to the world coordinate frame, Origin 2. The PCFs of Map 2 serve as a persistent coordinate frames relative to a canonical coordinate frame, where the world coordinate frame may be a previous session's world coordinate frame (e.g. Map 1's Origin 1 in FIG. **40**). These coordinate frames are related by the same transformation used to localize Map 2 to the canonical map, as discussed above in connection with FIG. **46B**.

[0421] The transformation from the world coordinate frame to the head coordinate frame **96** has been previously discussed with reference to FIG. **9**. The head coordinate frame **96** shown in FIG. **52** only has two orthogonal axes that are in a particular coordinate position relative to the PCFs of Map 2, and at particular angles relative to Map 2. It should however be understood that the head coordinate frame **96** is in a three-dimensional location relative to the PCFs of Map 2 and has three orthogonal axes within three-dimensional space.

[0422] In FIG. **53**, the head coordinate frame **96** has moved relative to the PCFs of Map 2. The head coordinate frame **96** has moved because the second user **14.2** has moved their head. The user can move their head in six degrees of freedom (6dof). The head coordinate frame **96** can thus move in 6dof, namely in three-dimensions from its previous location in FIG. **52** and about three orthogonal axes relative to the PCFs of Map 2. The head coordinate frame **96** is adjusted when the real object detection camera **44** and inertial measurement unit **48** in FIG. **9** respectively detect real objects and motion of the head unit **22**. More information regarding headpose tracking is disclosed in U.S. patent application Ser. No. 16/221,065 entitled “Enhanced Pose Determination for Display

Device” and is hereby incorporated by reference in its entirety.

[0423] FIG. 54 shows that sound may be associated with one or more PCFs. A user may, for example, wear headphones or earphones with stereoscopic sound. The location of sound through headphones can be simulated using conventional techniques. The location of sound may be located in a stationary position so that, when the user rotates their head to the left, the location of sound rotates to the right so that the user perceives the sound coming from the same location in the real world. In the present example, location of sound is represented by Sound123 and Sound456. For purposes of discussion, FIG. 54 is similar to FIG. 48 in its analysis. When the first and second users 14.1 and 14.2 are located in the same room at the same or different times, they perceive Sound123 and Sound456 coming from the same locations within the real world.

[0424] FIGS. 55 and 56 illustrate a further implementation of the technology described above. The first user 14.1 has initiated a first session as described with reference to FIG. 8. As shown in FIG. 55, the first user 14.1 has terminated the first session as indicated by the phantom lines. At the end of the first session, the first XR device 12.1 uploaded Map 1 to the server 20. The first user 14.1 has now initiated a second session at a later time than the first session. The first XR device 12.1 does not download Map 1 from the server 20 because Map 1 is already stored on the first XR device 12.1. If Map 1 is lost, then the first XR device 12.1 downloads Map 1 from the server 20. The first XR device 12.1 then proceeds to build PCFs for Map 2, localizes to Map 1, and further develops a canonical map as described above. Map 2 of the first XR device 12.1 is then used for relating local content, a head coordinate frame, local sound, etc. as described above.

[0425] Referring to FIGS. 57 and 58, it may also be possible that more than one user interacts with the server in the same session. In the present example, the first user 14.1 and the second user 14.2 are joined by a third user 14.3 with a third XR device 12.3. Each XR device 12.1, 12.2, and 12.3 begins to generate its own map, namely Map 1, Map 2, and Map 3, respectively. As the XR devices 12.1, 12.2, and 12.3 continue to develop Maps 1, 2, and 3, the maps are incrementally uploaded to the server 20. The server 20 merges Maps 1, 2, and 3 to form a canonical map. The canonical map is then transmitted from the server 20 to each one of the XR devices 12.1, 12.2 and 12.3.

[0426] FIG. 59 illustrates aspects of a viewing method to recover and/or reset headpose, according to some embodiments. In the illustrated example, at Act 1400, the viewing device is powered on. At Act 1410, in response to being powered on, a new session is initiated. In some embodiments, a new session may include establishing headpose. One or more capture devices on a head-mounted frame secured to a head of a user capture surfaces of an environment by first capturing images of the environment and then determining the surfaces from the images. In some embodiments, surface data may be combined with a data from a gravitational sensor to establish headpose. Other suitable methods of establishing headpose may be used.

[0427] At Act 1420, a processor of the viewing device enters a routine for tracking of headpose. The capture devices continue to capture surfaces of the environment as the user moves their head to determine an orientation of the head-mounted frame relative to the surfaces.

[0428] At Act 1430, the processor determines whether headpose has been lost. Headpose may become lost due to “edge” cases, such as too many reflective surfaces, low light, blank walls, being outdoor, etc. that may result in low feature acquisition, or because of dynamic cases such as a crowd that moves and forms part of the map. The routine at 1430 allows for a certain amount of time, for example 10 seconds, to pass to allow enough time to determine whether headpose has been lost. If headpose has not been lost, then the processor returns to 1420 and again enters tracking of headpose.

[0429] If headpose has been lost at Act 1430, the processor enters a routine at 1440 to recover headpose. If headpose is lost due to low light, then a message such as the following message is displayed to the user through a display of the viewing device:

[0430] THE SYSTEM IS DETECTING A LOW LIGHT CONDITION. PLEASE MOVE TO AN AREA WHERE THERE IS MORE LIGHT.

[0431] The system will continue to monitor whether there is sufficient light available and whether headpose can be recovered. The system may alternatively determine that low texture of surfaces is causing headpose to be lost, in which case the user is given the following prompt in the display as a suggestion to improve capturing of surfaces:

[0432] THE SYSTEM CANNOT DETECT ENOUGH SURFACES WITH FINE TEXTURES. PLEASE MOVE TO AN AREA WHERE THE SURFACES ARE LESS ROUGH IN TEXTURE AND MORE REFINED IN TEXTURE.

[0433] At Act **1450**, the processor enters a routine to determine whether headpose recovery has failed. If headpose recovery has not failed (i.e. headpose recovery has succeeded), then the processor returns to Act **1420** by again entering tracking of headpose. If headpose recovery has failed, the processor returns to Act **1410** to establish a new session. As part of the new session, all cached data is invalidated, whereafter headpose is established anew. Any suitable method of head tracking may be used in combination with the process described in FIG. **59**. U.S. patent application Ser. No. 16/221,065 describes head tracking and is hereby incorporated by reference in its entirety.

Remote Localization

[0434] Various embodiments may utilize remote resources to facilitate persistent and consistent cross reality experiences between individual and/or groups of users. The inventors have recognized and appreciated that the benefits of operation of an XR device with canonical maps as described herein can be achieved without downloading a set of canonical maps, such as is illustrated in FIG. **30**. The benefit, for example, may be achieved by sending feature and pose information to a remote service that maintains a set of canonical maps. A device seeking to use a canonical map to position virtual content in locations specified relative to the canonical map may receive from the remote service one or more transformations between the features and the canonical maps. Those transformations may be used on the device, which maintains information about the positions of those features in the physical world, to position virtual content in locations specified with respect to one or more canonical maps or to otherwise identify locations in the physical world that are specified with respect to the canonical map.

[0435] In some embodiments, spatial information is captured by an XR device and communicated to a remote service, such as a cloud based service, which uses the spatial information to localize the XR device to a canonical map used by applications or other components of an XR system to specify the location of virtual content with respect to the physical world. Once localized, transforms that link a tracking map maintained by the device to the canonical map can be communicated to the device. The transforms may be used, in conjunction with the tracking map, to determine a position in which to render virtual content specified with respect to the canonical map, or otherwise identify locations in the physical world that are specified with respect to the canonical map.

[0436] The inventors have realized that the data needed to be exchanged between a device and a remote localization service can be quite small relative to communicating map data, as might occur when a device communicates a tracking map to a remote service and receives from that service a set of canonical maps for device based localization). In some embodiments, performing localization functions on cloud resources requires only small amounts of information to be transmitted from the device to the remote service. It is not a requirement, for example, that a full tracking map be communicated to the remote service to perform localization. In some embodiments, features and pose information, such as might be stored in connection with a persistent pose, as described above, might be transmitted to the remote server. In embodiments in which features are represented by descriptors, as described above, the information uploaded may be even smaller.

[0437] The results returned to the device from the localization service may be one or more transformations that relate the uploaded features to portions of a matching canonical map. Those transformations may be used within the XR system, in conjunction with its tracking map, for identifying locations of virtual content or otherwise identifying locations in the physical world. In

embodiments in which persistent spatial information, such as PCFs as described above, are used to specify locations with respect to a canonical map, the localization service may download to the device transformations between the features and one or more PCFs after a successful localization. [0438] As a result, network bandwidth consumed by communications between an XR device and a remote service for performing localization may be low. The system may therefore support frequent localization, enabling each device interacting with the system to quickly obtain information for positioning virtual content or performing other location-based functions. As a device moves within the physical environment, it may repeat requests for updated localization information. Additionally, a device may frequently obtain updates to the localization information, such as when the canonical maps change, such as through merging of additional tracking maps to expand the map or increase their accuracy.

[0439] Further, uploading features and downloading transformations can enhance privacy in an XR system that shares map information among multiple users by increasing the difficulty of obtaining maps by spoofing. An unauthorized user, for example, may be thwarted from obtaining a map from the system by sending a fake request for a canonical map representing a portion of the physical world in which that unauthorized user is not located. An unauthorized user would be unlikely to have access to the features in the region of the physical world for which it is requesting map information if not physically present in that region. In embodiments in which feature information is formatted as feature descriptions, the difficulty in spoofing feature information in a request for map information would be compounded. Further, when the system returns a transformation intended to be applied to a tracking map of a device operating in the region about which location information is requested, the information returned by the system is likely to be of little or no use to an imposter.

[0440] According to one embodiment, a localization service is implemented as a cloud based micro-service. In some examples, implementing a cloud-based localization service can help save device compute resources and may enable computations required for localization to be performed with very low latency. Those operations can be supported by nearly infinite compute power or other computing resources available by provisioning additional cloud resources, ensuring scalability of the XR system to support numerous devices. In one example, many canonical maps can be maintained in memory for nearly instant access, or alternatively stored in high availability devices reducing system latency.

[0441] Further, performing localization for multiple devices in a cloud service may enable refinements to the process. Localization telemetry and statistics can provide information on which canonical maps to have in active memory and/or high availability storage. Statistics for multiple devices may be used, for example, to identify most frequently accessed canonical maps.

[0442] Additional accuracy may also be achieved as a result of processing in a cloud environment or other remote environment with substantial processing resources relative to a remote device. For example, localization can be made on higher density canonical maps in the cloud relative to processing performed on local devices. Maps may be stored in the cloud, for example, with more PCFs or a greater density of feature descriptors per PCF, increasing the accuracy of a match between a set of features from a device and a canonical map.

[0443] FIG. 61 is a schematic diagram of an XR system 6100. The user devices that display cross reality content during user sessions can come in a variety of forms. For example, a user device can be a wearable XR device (e.g., 6102) or a handheld mobile device (e.g., 6104). As discussed above, these devices can be configured with software, such as applications or other components, and/or hardwired to generate local position information (e.g., a tracking map) that can be used to render virtual content on their respective displays.

[0444] Virtual content positioning information may be specified with respect to global location information, which may be formatted as a canonical map containing one or more PCFs, for example. According to some embodiments, for example the embodiment shown in FIG. 61, the system 6100 is configured with cloud based services that support the functioning and display of the

virtual content on the user device.

[0445] In one example, localization functions are provided as a cloud-based service **6106**, which may be a micro-service. Cloud-based service **6106** may be implemented on any of multiple computing devices, from which computing resources may be allocated to one or more services executing in the cloud. Those computing devices may be interconnected with each other and accessibly to devices, such as a wearable XR device **6102** and hand held device **6104**. Such connections may be provided over one or more networks.

[0446] In some embodiments, the cloud-based service **6106** is configured to accept descriptor information from respective user devices and “localize” the device to a matching canonical map or maps. For example, the cloud-based localization service matches descriptor information received to descriptor information for respective canonical map(s). The canonical maps may be created using techniques as described above that create canonical maps by merging maps provided by one or more devices that have image sensors or other sensors that acquire information about a physical world. However, it is not a requirement that the canonical maps be created by the devices that access them, as such maps may be created by a map developer, for example, who may publish the maps by making them available to localization service **6106**.

[0447] According to some embodiments, the cloud service handles canonical map identification, and may include operations to filter a repository of canonical maps to a set of potential matches. Filtering may be performed as illustrated in FIG. **29**, or by using any subset of the filter criteria and other filter criteria instead of or in addition to the filter criteria shown in FIG. **29**. In one embodiment, geographic data can be used to limit a search for matching canonical map to maps representing areas proximate to the device requesting localization. For example, area attributes such as Wi-Fi signal data, Wi-Fi fingerprint information, GPS data, and/or other device location information can be used as a coarse filter on stored canonical maps, and thereby limit analysis of descriptors to canonical maps known or likely to be in proximity to the user device. Similarly, location history of each device may be maintained by the cloud service such that canonical maps in the vicinity of the device's last location are preferentially searched. In some examples, filtering can include the functions discussed above with respect to FIGS. **31B**, **32**, **33**, and **34**.

[0448] FIG. **62** is an example process flow that can be executed by a device to use a cloud-based service to localize the device's position with canonical map(s) and receive transform information specifying one or more transformations between the device local coordinate system and the coordinate system of a canonical map. Various embodiments and examples may describe the one or more transforms as specifying transforms from a first coordinate frame to a second coordinate frame. Other embodiments include transforms from the second coordinate frame to the first coordinate frame. In yet other embodiments, the transforms enable transition from one coordinate frame to another, the resulting coordinate frames depend only on the desired coordinate frame output (including, for example, the coordinate frame in which to display content). In yet further embodiments, the coordinate system transforms may enable determination of a first coordinate frame from the second coordinate frame and the second coordinate frame from the first coordinate frame.

[0449] According to some embodiments, information reflecting a transformation for each persistent pose defined with respect to the canonical map can be communicated to device.

[0450] According to one embodiment, process **6200** can begin at **6202** with a new session. Starting new session on the device may initiate capture of image information to build a tracking map for the device. Additionally, the device may send a message, registering with a server of a localization service, prompting the server to create a session for that device.

[0451] In some embodiments, starting a new session on a device optionally may include sending adjustment data from the device to the localization service. The localization service returns to the device one or more transforms computed based on the set of features and associated poses. If the poses of the features are adjusted based on device-specific information before computation of the

transformation and/or the transformations are adjusted based on device-specific information after computation of the transformation, rather than perform those computations on the device, the device specific information might be sent to the localization service, such that the localization service may apply the adjustments. As a specific example, sending device-specific adjustment information may include capturing calibration data for sensors and/or displays. The calibration data may be used, for example, to adjust the locations of feature points relative to a measured location. Alternatively or additionally, the calibration data may be used to adjust the locations at which the display is commanded to render virtual content so as to appear accurately positioned for that particular device. This calibration data may be derived, for example, from multiple images of the same scene taken with sensors on the device. The locations of features detected in those images may be expressed as a function of sensor location, such that multiple images yield a set of equations that may be solved for the sensor location. The computed sensor location may be compared to a nominal position, and the calibration data may be derived from any differences. In some embodiments, intrinsic information about the construction of the device may also enable calibration data to be computed for the display, in some embodiments.

[0452] In embodiments in which calibration data is generated for the sensors and/or display, the calibration data may be applied at any point in the measurement or display process. In some embodiments, the calibration data may be sent to the localization server, which may store the calibration data in a data structure established for each device that has registered with the localization server and is therefore in a session with the server. The localization server may apply the calibration data to any transformations computed as part of a localization process for the device supplying that calibration data. The computational burden of using the calibration data for greater accuracy of sensed and/or displayed information is thus borne by the calibration service, providing a further mechanism to reduce processing burden on the devices.

[0453] Once the new session is established, process **6200** may continue at **6204** with capture of new frames of the device's environment. Each frame can be processed to generate descriptors (including for example, DSF values discussed above) for the captured frame at **6206**. These values may be computed using some or all of the techniques described above, including techniques as discussed above with respect to FIGS. **14**, **22** and **23**. As discussed, the descriptors may be computed as a mapping of the feature points or, in some embodiments a mapping of a patch of an image around a feature point, to a descriptor. The descriptor may have a value that enables efficient matching between newly acquired frames/images and stored maps. Moreover, the number of features extracted from an image may be limited to a maximum number of features points per image, such as 200 feature points per image. The feature points may be selected to represent interest points, as described above. Accordingly, acts **6204** and **6206** may be performed as part of a device process of forming a tracking map or otherwise periodically collecting images of the physical world around the device, or may be, but need not be, separately performed for localization.

[0454] Feature extraction at **6206** may include appending pose information to the extracted features at **6206**. The pose information may be a pose in the device's local coordinate system. In some embodiments, the pose may be relative to a reference point in the tracking map, such as a persistent pose, as discussed above. Alternatively or additionally, the pose may be relative to the origin of a tracking map of the device. Such an embodiment may enable the localization service as described herein to provide localization services for a wide range of devices, even if they do not utilize persistent poses. Regardless, pose information may be appended to each feature or each set of features, such that the localization service may use the pose information for computing a transformation that can be returned to the device upon matching the features to features in a stored map.

[0455] The process **6200** may continue to decision block **6207** where a decision is made whether to request localization. One or more criteria may be applied to determine whether to request

localization. The criteria may include passage of time, such that a device may request localization after some threshold amount of time. For example, if localization has not been attempted within a threshold amount of time, the process may continue from decision block **6207** to act **6208** where localization is requested from the cloud. That threshold amount of time may be between ten and thirty seconds, such as twenty-five seconds, for example. Alternatively or additionally, localization may be triggered by motion of a device. A device executing the process **6200** may track its motion using an IMU and/or its tracking map, and initiate localization upon detection motion exceeding a threshold distance from the location where the device last requested localization. The threshold distance may be between one and ten meters, such as between three and five meters, for example. As yet a further alternative, localization may be triggered in response to an event, such as when a device creates a new persistent pose or the current persistent pose for the device changes, as described above.

[0456] In some embodiments, decision block **6207** may be implemented such that the thresholds for triggering localization may be established dynamically. For example, in environments in which features are largely uniform such that there may be a low confidence in matching a set of extracted features to features of a stored map, localization may be requested more frequently to increase the chances that at least one attempt at localization will succeed. In such a scenario, the thresholds applied at decision block **6207** may be decreased. Similarly, in an environment in which there are relatively few features, the thresholds applied at decision block **6207** may be decreased so as to increase the frequency of localization attempts.

[0457] Regardless of how the localization is triggered, when triggered, the process **6200** may proceed to act **6208** where the device sends a request to the localization service, including data used by the localization service to perform localization. In some embodiments, data from multiple image frames may be provided for a localization attempt. The localization service, for example, may not deem localization successful unless features in multiple image frames yield consistent localization results. In some embodiments, process **6200** may include saving feature descriptors and appended pose information into a buffer. The buffer may, for example, be a circular buffer, storing sets of features extracted from the most recently captured frames. Accordingly, the localization request may be sent with a number of sets of features accumulated in the buffer. In some settings, a buffer size is implemented to accumulate a number of sets of data that will be more likely to yield successful localization. In some embodiments, a buffer size may be set to accumulate features from two, three, four, five, six, seven, eight, nine, or ten frames, for example). Optionally, the buffer size can have a baseline setting which can be increased responsive to localization failures. In some examples, increasing the buffer size and corresponding number of sets of features transmitted reduces the likelihood that subsequent localization functions fail to return a result.

[0458] Regardless of how the buffer size is set, the device may transfer the contents of the buffer to the localization service as part of a localization request. Other information may be transmitted in conjunction with the feature points and appended pose information. For example, in some embodiments, geographic information may be transmitted. The geographic information may include, for example, GPS coordinates or a wireless signature associated with the devices tracking map or current persistent pose.

[0459] In response to the request sent at **6208**, a cloud localization service may analyze the feature descriptors to localize the device into a canonical map or other persistent map maintained by the service. For example, the descriptors are matched to a set of features in a map to which the device is localized. The cloud based localization service may perform localization as described above with respect to device based localization (e.g., can rely on any of the functions discussed above for localization including, map ranking, map filtering, location estimation, filtered map selection, examples in FIGS. **44-46**, and/or discussed with respect to a localization module, PCF and/or PP identification and matching etc.). However, instead of communicating identified canonical maps to

a device (e.g., in device localization), the cloud-based localization service may proceed to generate transforms based on the relative orientation of feature sets sent from the device and the matching features of the canonical maps. The localization service may return these transforms to the device, which may be received at block **6210**.

[0460] In some embodiments, the canonical maps maintained by the localization service may employ PCFs, as described above. In such embodiments, the feature points of the canonical maps that match the feature points sent from the device may have positions specified with respect to one or more PCFs. Accordingly, the localization service may identify one or more canonical maps and may compute a transformation between the coordinate frame represented in the poses sent with the request for localization and the one or more PCFs. In some embodiments, identification of the one or more canonical maps is assisted by filtering potential maps based on geographic data for a respective device. For example, once filtered to a candidate set (e.g., by GPS coordinate, among other options) the candidate set of canonical maps can be analyzed in detail to determine matching feature points or PCFs as described above.

[0461] The data returned to the requesting device at act **6210** may be formatted as a table of persistent pose transforms. The table can be accompanied by one or more canonical map identifiers, indicating the canonical maps to which the device was localized by the localization service. However, it should be appreciated that the localization information may be formatted in other ways, including as a list of transforms, with associated PCF and/or canonical map identifiers.

[0462] Regardless of how the transforms are formatted, at act **6212** the device may use these transforms to compute the location at which to render virtual content for which a location has been specified by an application or other component of the XR system relative to any of the PCFs. This information may alternatively or additionally be used on the device to perform any location based operation in which a location is specified based on the PCFs.

[0463] In some scenarios, the localization service may be unable to match features sent from a device to any stored canonical map or may not be able to match a sufficient number of the sets of features communicated with the request for the localization service to deem a successful localization occurred. In such a scenario, rather than returning transformations to the device as described above in connection with act **6210**, the localization service may indicate to the device that localization failed. In such a scenario, the process **6200** may branch at decision block **6209** to act **6230**, where the device may take one or more actions for failure processing. These actions may include increasing the size of the buffer holding feature sets sent for localization. For example, if the localization service does not deem a successful localization unless three sets of features match, the buffer size may be increased from five to six, increasing the chances that three of the transmitted sets of features can be matched to a canonical map maintained by the localization service.

[0464] Alternatively or additionally, failure processing may include adjusting an operating parameter of the device to trigger more frequent localization attempts. The threshold time between localization attempts and/or the threshold distance may be decreased, for example. As another example, the number of feature points in each set of features may be increased. A match between a set of features and features stored within a canonical map may be deemed to occur when a sufficient number of features in the set sent from the device match features of the map. Increasing the number of features sent may increase the chances of a match. As a specific example, the initial feature set size may be 50, which may be increased to 100, 150, and then 200, on each successive localization failure. Upon successful match, the set size may then be returned to its initial value.

[0465] Failure processing may also include obtaining localization information other than from the localization service. According to some embodiments, the user device can be configured to cache canonical maps. Cached maps permit devices to access and display content where the cloud is unavailable. For example, cached canonical maps permit device based localization in the event of communication failure or other unavailability.

[0466] According to various embodiments, FIG. 62 describes a high level flow for a device initiating cloud based localization. In other embodiments, various ones or more of the illustrated steps can be combined, omitted, or invoke other processes to accomplish localization and ultimately visualization of virtual content in a view of a respective device.

[0467] Further, it should be appreciated that, though the process **6200** shows the device determining whether to initiate localization at decision block **6207**, the trigger for initiating localization may come from outside the device, including from the localization service. The localization service, for example, may maintain information about each of the devices that is in a session with it. That information, for example, may include an identifier of a canonical map to which each device most recently localized. The localization service, or other components of the XR system, may update canonical maps, including using techniques as described above in connection with FIG. 26. When a canonical map is updated, the localization service may send a notification to each device that most recently localized to that map. That notification may serve as a trigger for the device to request localization and/or may include updated transformations, recomputed using the most recently sent sets of features from the device.

[0468] FIGS. 63A, B, and C are an example process flow showing operations and communication between a device and cloud services. Shown at blocks **6350**, **6352** **6354**, and **6456** are example architecture and separation between components participating in the cloud based localization process. For example, the modules, components, and/or software that are configured to handle perception on the user device are shown at **6350** (e.g., **660**, FIG. 6A). Device functionality for persisted world operations are shown at **6352** (including, for example, as described above and with respect to persisted world module (e.g., **662**, FIG. 6A)). In other embodiments, the separation between **6350** and **6352** is not needed and the communication shown can be between processes executing on the device.

[0469] Similarly, shown at block **6354** is a cloud process configured to handle functionality associated with passable world/passable world modeling (e.g., **802**, **812**, FIG. 26). Shown at block **6356** is a cloud process configured to handle functionality associated with localizing a device, based on information sent from a device, to one or more maps of a repository of stored canonical maps.

[0470] In the illustrated embodiment, process **6300** begins at **6302** when a new session starts. At **6304** sensor calibration data is obtained. The calibration data obtained can be dependent on the device represented at **6350** (e.g., number of cameras, sensors, positioning devices, etc.). Once the sensor calibration is obtained for the device, the calibrations can be cached at **6306**. If device operation resulted in a change in frequency parameters (e.g., collection frequency, sampling frequency, matching frequency, among other options) the frequency parameters are reset to baseline at **6308**.

[0471] Once the new session functions are complete (e.g., calibration, steps **6302-6306**) process **6300** can continue with capture of a new frame **6312**. Features and their corresponding descriptors are extracted from the frame at **6314**. In some examples, descriptors can comprise DSF's, as discussed above. According to some embodiments, the descriptors can have spatial information attached to them to facilitate subsequent processing (e.g., transformation generation). Pose information (e.g., information, specified relative to the device's tracking map for locating the features in the physical world as discussed above) generated on the device can be appended to the extracted descriptors at **6316**.

[0472] At **6318**, the descriptor and pose information is added to a buffer. New frame capture and addition to the buffer shown in steps **6312-6318** is executed in a loop until a buffer size threshold is exceeded at **6319**. Responsive to a determination that the buffer size has been met, a localization request is communicated from the device to the cloud at **6320**. According to some embodiments, the request can be handled by a passable world service instantiated in the cloud (e.g. **6354**). In further embodiments, functional operations for identifying candidate canonical maps can be

segregated from operations for actual matching (e.g., shown as blocks **6354** and **6356**). In one embodiment, a cloud service for map filtering and/or map ranking can be executed at **6354** and process the received localization request from **6320**. According to one embodiment, the map ranking operations are configured to determine a set of candidate maps at **6322** that are likely to include a device's location.

[0473] In one example, the map ranking function includes operations to identify candidate canonical maps based on geographic attributes or other location data (e.g., observed or inferred location information). For example, other location data can include Wi-Fi signatures or GPS information.

[0474] According to other embodiments, location data can be captured during a cross reality session with the device and user. Process **6300** can include additional operations to populate a location for a given device and/or session (not shown). For example, the location data may be stored as device area attribute values and the attribute values used to select candidate canonical maps proximate to the device's location.

[0475] Any one or more of the location options can be used to filter sets of canonical maps to those likely to represent an area including the location of a user device. In some embodiments, the canonical maps may cover relatively large regions of the physical world. The canonical maps may be segmented into areas such that selection of a map may entail selection of a map area. A map area, for example may be on the order of tens of meters squared. Thus, the filtered set of canonical maps may be a set of areas of the maps.

[0476] According to some embodiments, a localization snapshot can be built from the candidate canonical maps, posed features, and sensor calibration data. For example, an array of candidate canonical maps, posed features, and sensor calibration information can be sent with a request to determine specific matching canonical maps. Matching to a canonical map can be executed based on descriptors received from a device and stored PCF data associated with the canonical maps.

[0477] In some embodiments, a set of features from the device is compared to sets of features stored as part of the canonical map. The comparison may be based on the feature descriptors and/or pose. For example, a candidate set of features of a canonical map may be selected based on the number of features in the candidate set that have descriptors similar enough to the descriptors of the feature set from the device that they could be the same feature. The candidate set, for example, may be features derived from an image frame used in forming the canonical map.

[0478] In some embodiments, if the number of similar features exceeds a threshold, further processing may be performed on the candidate set of features. Further processing may determine the degree to which the set of posed features from the device can be aligned with the features of the candidate set. The set of features from the canonical map, like the features from the device, may be posed.

[0479] In some embodiments, features are formatted as a highly dimensional embedding (e.g., DSF, etc.) and may be compared using a nearest neighbor search. In one example, the system is configured (e.g., by executing process **6200** and/or **6300**) to find the top two nearest neighbors using Euclidian distance, and may execute a ratio test. If the closest neighbor is much closer than the second closest neighbor, the system considers the closest neighbor to be a match. "Much closer" in this context may be determined, for example, by the ratio of Euclidean distance relative to the second nearest neighbor is more than a threshold times the Euclidean distance relative to the nearest neighbor. Once a feature from the device is considered to be a "match" to a feature in canonical map, the system may be configured to use the pose of the matching features to compute a relative transformation. The transformation developed from the pose information may be used to indicate the transformation required to localize the device to the canonical map.

[0480] The number of inliers may serve as an indication of the quality of the match. For example, in the case of DSF matching, the number of inliers reflects the number of features that were matched between received descriptor information and stored/canonical maps. In further

embodiments, inliers may be determined in this embodiment by counting the number of features in each set that “match.”

[0481] An indication of the quality of a match may alternatively or additionally be determined in other ways. In some embodiments, for example, when a transformation is computed to localize a map from a device, which may contain multiple features, to a canonical map, based on relative pose of matching features, statistics of the transformation computed for each of multiple matching features may serve as quality indication. A large variance, for example, may indicate a poor quality of match. Alternatively or additionally, the system may compute, for a determined transformation, a mean error between features with matching descriptors. The mean error may be computed for the transformation, reflecting the degree of positional mismatch. A mean squared error is a specific example of an error metric. Regardless of the specific error metric, if the error is below a threshold, the transformation may be determined to be usable for the features received from the device, and the computed transformation is used for localizing the device. Alternatively or additionally, the number of inliers may also be used in determining whether there is a map that matches a device's positional information and/or descriptors received from a device.

[0482] As noted above, in some embodiments, a device may send multiple sets of features for localization. Localization may be deemed successful when at least a threshold number of sets of features match, with an error below a threshold, and/or a number of inliers above a threshold, a set of features from the canonical map. That threshold number, for example, may be three sets of features. However, it should be appreciated that the threshold used for determining whether a sufficient number of sets of feature have suitable values may be determined empirically or in other suitable ways. Likewise, other thresholds or parameters of the matching process, such as degree of similarity between feature descriptors to be deemed matching, the number of inliers for selection of a candidate set of features, and/or the magnitude of the mismatch error, may similarly be determined empirically or in other suitable ways.

[0483] Once a match is determined, a set of persistent map features associated with the matched canonical map or maps is identified. In embodiments in which the matching is based on areas of maps, the persistent map features may be the map features in the matching areas. The persistent map features may be persistent poses or PCFs as described above. In the example of FIG. 63, the persistent map features are persistent poses.

[0484] Regardless of the format of the persistent map features, each persistent map feature may have a predetermined orientation relative to the canonical map in which it is a part. This relative orientation may be applied to the transformation computed to align the set of features from the device with the set of features from the canonical map to determine a transformation between the set of features from the device and the persistent map feature. Any adjustments, such as might be derived from calibration data, may then be applied to this computed transformation. The resulting transformation may be the transformation between the local coordinate frame of the device and the persistent map feature. This computation may be performed for each persistent map feature of a matching map area, and the results may be stored in a table, denoted as the `persistent_pose_table` in **6326**.

[0485] In one example, block **6326** returns a table of persistent pose transforms, canonical map identifiers, and number of inliers. According to some embodiments, the canonical map ID is an identifier for uniquely identifying a canonical map and a version of the canonical map (or area of a map, in embodiments in which localization is based on map areas).

[0486] In various embodiments, the computed localization data can be used to populate localization statistics and telemetry maintained by the localization service at **6328**. This information may be stored for each device, and may be updated for each localization attempt, and may be cleared when the device's session ends. For example, which maps were matched by a device can be used to refine map ranking operations. For example, maps covering the same area to which the device previously matched may be prioritized in the ranking. Likewise, maps covering adjacent areas may be give

higher priority over more remote areas. Further, the adjacent maps might be prioritized based on a detected trajectory of the device over time, with map areas in the direction of motion being given higher priority over other map areas. The localization service may use this information, for example, upon a subsequent localization request from the device to limit the maps or map areas searched for candidate sets of features in the stored canonical maps. If a match, with low error metrics and/or a large number or percentage of inliers, is identified in this limited area, processing of maps outside the area may be avoided.

[0487] Process **6300** can continue with communication of information from the cloud (e.g., **6354**) to the user device (e.g., **6352**). According to one embodiment, a persistent pose table and canonical map identifiers are communicated to the user device at **6330**. In one example, the persistent pose table can be constructed of elements including at least a string identifying a persistent pose ID and a transformation linking the device's tracking map and the persistent pose. In embodiments in which the persistent map features are PCFs the table may, instead, indicate transformations to the PCFs of the matching maps.

[0488] If localization fails at **6336**, process **6300** continues by adjusting parameters that may increase the amount of data sent from a device to the localization service to increase the chances that localization will succeed. Failure, for example, may be indicated when no sets of features in the canonical map can be found with more than a threshold number of similar descriptors or when the error metric associated with all transformed sets of candidate features is above a threshold. As an example of a parameter that may be adjusted, the size constraint for the descriptor buffer may be increased (of **6319**). For example, where the descriptor buffer size is five, localization failure can trigger an increase to at least six sets of features, extracted from at least six image frames. In some embodiments, process **6300** can include a descriptor buffer increment value. In one example, the increment value can be used to control the rate of increase in the buffer size, for example, responsive to localization failures. Other parameters, such as parameters controlling the rate of localization requests, may be changed upon a failure to find matching canonical maps.

[0489] In some embodiments, execution of **6300** can generate an error condition at **6340**, which includes execution where the localization request fails to work, rather than return a no match result. An error, for example, may occur as a result of a network error making the storage holding a database of canonical maps unavailable to a server executing the localization service or a received request for localization services containing incorrectly formatted information. In the event of an error condition, in this example, the process **6300** schedules a retry of the request at **6342**.

[0490] When a localization request is successful, any parameters adjusted in response to a failure may be reset. At **6332**, process **6300** can continue with an operation to reset frequency parameters to any default or baseline. In some embodiments **6332** is executed regardless of any changes thus ensuring baseline frequency is always established.

[0491] The received information can be used by the device at **6334** to update a cache localization snapshot. According to various embodiments, the respective transforms, canonical maps identifiers, and other localization data can be stored by the device and used to relate locations specified with respect to the canonical maps, or persistent map features of them such as persistent poses or PCFs to locations determined by the device with respect to its local coordinate frame, such as might be determined from its tracking map.

[0492] Various embodiments of processes for localization in the cloud can implement any one or more of the preceding steps and be based on the preceding architecture. Other embodiments may combine various ones or more of the preceding steps, execute steps simultaneously, in parallel, or in another order.

[0493] According to some embodiments, localization services in the cloud in the context of cross reality experiences can include additional functionality. For example, canonical map caching may be executed to resolve issues with connectivity. In some embodiments, the device may periodically download and cache canonical maps to which it has localized. If the localization services in the

cloud are unavailable, the device may run localizations itself (e.g., as discussed above-including with respect to FIG. 26). In other embodiments, the transformations returned from localization requests can be chained together and applied in subsequent sessions. For example, a device may cache a train of transformations and use the sequence of transformations to establish localization. [0494] Various embodiments of the system can use the results of localization operations to update transformation information. For example, the localization service and/or a device can be configured to maintain state information on a tracking map to canonical map transformations. The received transformations can be averaged over time. According to one embodiment, the averaging operations can be limited to occur after a threshold number of localizations are successful (e.g., three, four, five, or more times). In further embodiments, other state information can be tracked in the cloud, for example, by a passable world module. In one example, state information can include a device identifier, tracking map ID, canonical map reference (e.g., version and ID), and the canonical map to tracking map transformation. In some examples, the state information can be used by the system to continuously update and get more accurate canonical map to tracking map transforms with every execution of the cloud-based localization functions.

[0495] Additional enhancements to cloud-based localization can include communicating to devices outliers in the sets of features that did not match features in the canonical maps. The device may use this information, for example, to improve its tracking map, such as by removing the outliers from the sets of features used to build its tracking map. Alternatively or additionally, the information from the localization service may enable the device to limit bundle adjustments for its tracking map to computing adjustments based on inlier features or to otherwise impose constraints on the bundle adjustment process.

[0496] According to another embodiment, various sub-processes or additional operations can be used in conjunction and/or as alternatives to the processes and/or steps discussed for cloud based localization. For example, candidate map identification may include accessing canonical maps based on area identifiers and/or area attributes stored with respective maps.

Localization with Candidate Pose Propagation

[0497] In the operation of an XR system, which may contain multiple components, a coordinate frame of one component may be related to a coordinate frame of another component. Such a relationship, defining the relative pose of the two components, may be developed through a localization process. In a localization process as described herein, information expressed in a coordinate frame of one component (e.g. a first map) is transformed so as to align with corresponding information expressed in coordinate frame of another component (e.g. a second map). The transformation may be used to relate locations specified in the coordinate frame of one component in the coordinate frame of the other and vice versa.

[0498] Such a localization operation may occur, for example, when an XR device uploads information to a cloud containing a localization service and receives its pose with respect to one or more persisted maps, for example, as described above with respect to FIGS. 61-63C. For simplicity of explanation of simplified processing that can result from candidate pose propagation, an example of that technique as applied to a localization service is used as an example. The technique, however, may be applied in other scenarios in which collections of features, associated with one component of an XR system, are localized with respect to another set of feature points that may be associated with another component. Such an operation may occur, for example, when relating a tracking map of a portable device to a stored map, such as may occur in map merge. FIG. 64 is a block diagram of this example.

[0499] The information used for localization may be sparse information, reflected as a collection of features. These features may be extracted from images taken of a 3D environment. In the example system with a localization service, a portable device 6410 may accumulate a batch 6430 of multiple snapshots 6432, collected over time with sensors, such as image sensors, on the device.

[0500] A snapshot 6432, for example, may contain a collection of features extracted from select

images. Those images for example, may be key frames as described above and the features may be selected and represented by descriptors, such as is described above as DSF's or another suitable descriptor. Those features may be posed with respect to a coordinate frame of the user device **6410**, such as in a posed feature rig in which features are posed relative to a coordinate frame of a tracking map on the user device **6410**. A snapshot may include information in addition to the features, including location metadata. Location metadata may indicate the device's location in a 3D environment at the time of the snapshot and may be, for example, a wireless fingerprint or geolocation data or other suitable metadata.

[0501] As a device moves through a 3D environment, it may collect snapshots **6432** and accumulate them in a batch **6430**. The snapshots may be collected periodically, such as every 2 seconds, or in response to events, such as motion of the device by more than 3 meters. The snapshots may be added to a buffer on the device until a localization request is sent from the device, at which time the snapshots accumulated in the buffer may be sent to a localization service. The localization request may be triggered by passage of time or another event.

[0502] FIG. **64** illustrates a process **6400** that may be performed in a localization service. The localization service may be implemented on cloud resources, such as one or more cloud servers that each has one or more processors that collectively perform the processing depicted in FIG. **64**. Additionally or alternatively, at least a portion of the localization service may be implemented on the device itself or elsewhere in the XR system.

[0503] The localization service may compare the batch **6430** of snapshots **6432** to persistent spatial information, which may also be sparse information likewise represented as collections of features. For a localization service, the persistent spatial information may be one or more shared maps **6420** that may be used for localization of multiple user devices. For simplicity, only one such user device is illustrated in FIG. **64**.

[0504] In some embodiments, shared maps **6420** may be a database of stored maps. Those maps may be canonical maps, as described above. Further, also as described above, one or more criteria may be applied to select a map from the shared maps **6420**. Location metadata, for example, associated with the snapshots **6432** may be used to select a shared map that likely represents the same region of the 3D environment as is depicted in snapshots **6432** of batch **6430**.

[0505] Those criteria and/or other criteria may be used to select a segment of the selected map for localization processing. For example, a tile of the map may be selected based on a comparison of location metadata. As another example, a PCF in the stored maps may be selected based on similarity of a key frame descriptor for a keyframe from which the PCF was generated and a keyframe from which features in one or more of the snapshots **6432** were extracted. As yet a further example, if selection criteria result in identification of multiple segments of one or more maps, those segments may be processed separately and a matching segment may be selected from the localization results computed separately for the segments.

[0506] Regardless of the one or more criteria used to select a segment of a map, the collections of features within batch **6430** of snapshots **6432** may be compared to a collection of features in the selected segment of the selected map. In the example of FIG. **64**, a candidate localization **6448** may be computed for each collection of features in the batch **6430** with respect to the collection of features in the selected segment of the selected map. A localization result **6452** may be determined via a location consensus function **6450** among the candidate localizations **6448** computed for each collection of features in batch **6430**.

[0507] Consensus function **6450** may be based on one or more criteria that provide sufficiently high confidence that the localization result **6452** accurately reflects the relative pose of user device **6410** with respect to the selected map. As an example, the consensus function **6450** may be based on the mode of the distribution of candidate localizations **6448**. As a specific example, batch **6430** may contain ten snapshots **6432** and, if at least three of the candidate localizations **6448** have the value identified as the mode of that distribution, a candidate localization **6448** at the mode may be

selected as final localization **6452**.

[0508] In the illustrated embodiment, a localization result may be expressed as a transformation that aligns the coordinate frame of snapshots **6432** with respect to the selected map. This transformation may be returned to user device **6410** where it may be applied, as either a forward or inverse transformation, to relate locations specified with respect to the shared map to the coordinate frame used by the user device **6410**, or vice versa, as described above.

[0509] In determining whether there is consensus among localization results, inexact matching techniques may be used such that localization results that are sufficiently similar (e.g. differences can be considered within the inherent variation in the system). For example, a localization result may be reflected as a multi-dimensional transformation between the coordinate frame of the user device and the coordinate frame of the selected canonical map. Two transformations may be compared by computing a vector difference, and transformations that differ by an amount or percentage less than a threshold may be regarded as the same.

[0510] In scenarios in which there are insufficient candidate localizations **6448** at the mode to meet the selection criteria, there may be no localization result with respect to the selected segment of the selected map. The XR system may respond by attempting localization as illustrated in FIG. **64** with respect to other segments or other maps or taking other corrective action. If no localization result is identified, an indication of localization failure may be returned to the user device **6410**, such as is described above.

[0511] In the example of FIG. **64** candidate localizations **6448** may be computed for each of the collection of features in the batch **6430** in a two-step process, with a rough localization **6442** and a refined localization **6446**. In the example of FIG. **64**, localization processing **6440** may be performed in accordance with a random sample consensus (RANSAC) approach.

[0512] That processing is performed based on corresponding features within the collections of featured being compared. In this example, the comparison is based on a set of features in the collection of features in a snapshot **6432** that match features in the collection of features in the selected shared map. In the exemplary embodiment illustrated in FIG. **64**, sets of matching features may be formed by selecting features in the two collections that have matching descriptors. As with other parameters discussed herein, matching descriptors may be defined based on identical values or values within an acceptable range of variability.

[0513] Rough localization **6442** may be performed in such a way that it requires little processing relative to processing to find a low error alignment between the full sets of features. As a specific example, rough localization may require 10% or less the processing to align the full sets of features. Rough localization **6442**, for example, may be performed on only subsets of the features in each set of corresponding features. The number of features in a subset may be small such that a relatively large number of subsets may be compared without significant processing. For example, a subset may contain less than 10 features, such as 6 features.

[0514] The comparison may entail searching for a transformation that aligns the features of the subsets such that there is a low measure of error between features deemed to correspond based on similar descriptors. Searching may entail iteratively testing transformations applied to one subset of features and computing the resulting error of alignment in the corresponding features of the subset. The transformations tested may be selected through successive refinement to arrive at a transformation that provides the lowest result error for the subsets being processed.

[0515] Multiple subsets may be selected and processed such that a transformation and a resulting error in correspondence of features of the subset are computed for each subset. Based on these results, the subset with the lowest measure of error may be identified. The transformation associated with that subset may be the rough localization result **6444**. Rough localization result **6444** may serve as a rough transformation that is an input to a refined localization process **6446**.

[0516] Refined localization process **6446** may be performed similarly based on the same sets of corresponding features that were processed in rough localization. However, in the refined

localization process **6446**, the selected subsets of features may be larger. For example, subsets with four to ten times more features may be processed in refined localization process **6446**. As a specific example, rough localization may be based on subsets with six features and refined localization may be computed with subsets with 25 features.

[0517] Processing for refined localization may be limited because the search space for transformations that aligns the subsets of features may be limited based on the rough transformation. As a specific example, the rough localization result **6444** may indicate that a snapshot must be translated by a specific distance, X , to correspond with features of a shared map. The search for transformations may be limited to transformations that impose a translation $X \pm \Delta$. The parameter Δ may be, for example, 5 meters. Similar constraints may be imposed on other dimensions of the transformation tested, such as rotation.

[0518] From among the subsets for refined localization process **6446**, a refined localization result may be provided as a candidate localization **6448**. As illustrated in FIG. **64**, such processing may be performed multiple times, once for each snapshot **6432** in batch **6430**. The refined localization computed for the collection of features in each snapshot **6432** leads to one candidate localization **6448**, such that there are multiple candidate locations from which consensus function **6450** may select a localization result **6452**.

[0519] The inventors have recognized and appreciated techniques for further reduction in processing to match spatial information, such as occurs during a localization process. The inventors have recognized that, in many instances, a rough localization result for a collection of features in a batch will be applicable to other collections of features in the batch. Overall processing of the batch may be reduced by computing a rough localization result for a collection of features and then, selectively, propagating that rough localization result to use it for refined localization processing of other collections of features in the batch.

[0520] FIG. **65** illustrates a process **6500** that is a variation of the process **6400**. In this example, the inputs and the outputs of the localization process **6540** are the same as for localization process **6440** (FIG. **64**). Localization process **6540** differs from localization process **6440** in that rough localization is performed for only a first portion of the snapshots **6432** in batch **6430**. A rough localization result computed for one of the snapshots in the first portion is propagated to a second portion of the snapshots for use in computing refined localizations for the second portion of the snapshots. Such an approach avoids the rough localization processing for the second portion of the snapshots.

[0521] In the example of FIG. **65**, rough localization is shown to be performed on one of the snapshots, and the rough localization result **6444** is propagated to the collections of features associated with all of the other snapshots **6432**. However, the inventors have recognized and appreciated that reduced compute load required for a given level of accuracy may be achieved by selectively propagating the rough localization to other collections of features.

[0522] Whether to use the rough localization results from a first collection of features in computing a refined localization result for a second collection of features may be determined based on one or more criteria that may be evaluated based on the collections of features. Accordingly, for any batch of snapshots, processing may be as illustrated in FIG. **65** where rough localization is performed for one collection of features and propagated to all other collections of features for refined localization processing based on that rough localization result. At the other extreme, based on a runtime evaluation of criteria, no rough localization result may be propagated and processing for all of the collections of features may be based on the rough localization results separately computed for that collection of features (e.g. as illustrated in FIG. **64**). Any alternative in between is also possible, with refined localization for any number of collections of features in a first portion of the collections of features being performed based on rough localization for that collection of features and refined localization for any number of collections of features in a second portion of the collections of features being performed based on rough localization for a collection of features in

the first portion.

[0523] The criteria used to determine whether to propagate a rough pose from a first collection of features to a second collection of features for refined localization may relate to different parameters and may be applied at different times in the processing. The criteria, for example, may relate to the candidate localization computed for the collection of features from which the rough localization may be propagated. Alternatively or additionally, the criteria may relate to the joint characteristics of the collection of features from which the rough localization is to be propagated and to which the rough localization is to be propagated. As a further alternative, the criteria may relate to characteristics of the batch of candidate localizations **6448** that results from propagating a rough localization to some or all of the other collections of features.

[0524] One such criteria that may be applied to determine whether to apply the rough localization results for a first snapshot to a second snapshot, for example, may be confidence in the localization result computed for the first snapshot. Confidence in the refined localization result for the first snapshot may be computed based on positional correspondence between a first set of features associated with the first snapshot and corresponding set of features in the map to which a localization attempt is being made.

[0525] A confidence metric, for example, may be computed by applying the transformation computed as part of the refined localization to the first set of features. A measure of positional deviation may then be computed between features of the first set and the features in the map selected as matching those features based on similar feature descriptors. As a specific example, the metric may be computed by counting a number of features identified as matching that are close enough together to be considered to have coincident locations (“inliers”). In other embodiments, the metric might be the ratio of inliers to outliers.

[0526] The same confidence metric may be applied as is used to determine whether a computed refined localization has sufficiently low error to be provided as a candidate localization **6448**. However, a higher threshold may be used. For example, to use a refined localization as a candidate localization, the threshold number of inliers may be 25 or at least 13% of the features in the set of features with matching features in the map. To propagate a rough localization, the threshold number of inliers may be 35 or at least 18% of the features in the set of features with matching features in the map.

[0527] A further criteria may be whether the parallax between the snapshots is above a threshold. In the described embodiment, the snapshots are each derived from an image of a 3D environment of a portable device. If the device has moved between acquisition of the image from which the first collection of features was derived and acquisition of the image from which the second collection of features was derived, features in the two images should be offset with respect to each other which creates parallax. Parallax may be determined by a comparison of the position of corresponding features in the first and second collection of features. Alternatively, a measure of parallax may be determined based on motion of the portable device detected between the time of capture of the images corresponding to the first and second collection of features. As described above, portable devices used in an XR system may include an IMU that outputs movement information. Parallax may be expressed in pixels of the image or amount of movement of the device between image captures, which may be expressed in feet or degrees. Regardless of how parallax is measured, in some embodiments, the rough localization for a first collection of features may be used for a refined localization of a second collection of features only when the parallax between the associated images exceeds a threshold.

[0528] A further criteria may relate to the degree of consensus among the candidate localizations **6448** when rough localization results are propagated from some of the snapshots to others of the snapshots. In some embodiments, the degree of consensus used as a threshold may be changed depending on whether propagation was used. For example, in the example of FIG. **64**, consensus function **6450** may output a localization result when at least three of the candidate localizations, or

at least 30%, match. With propagation of a rough localization, the threshold number may be increased to be at least five, or at least 50%, of the candidate localizations that match.

[0529] In some embodiments, the threshold may be adaptive, depending on the number of candidate localizations that match had a rough localization propagated to them. If none of the candidate localizations at the mode had a rough localization propagated to them, then the lower thresholds as in FIG. 64 might be used. Conversely, if more than half of the candidate localizations at the mode had a rough localization propagated to them, then the higher thresholds may be applied. If the number of candidate localizations at the mode that had a rough localization propagated to them is between none and half, then the threshold may be proportionately between the lower and the higher threshold.

[0530] In scenarios in which there is insufficient consensus among candidate localizations 6448 when rough localizations are propagated from some collections of features to others, the full localization process, such as is illustrated in FIG. 64, may be performed. After finding inconsistent consensus, candidate localizations may be computed for all of the snapshots, using both rough localization and refined localization, as shown in FIG. 64. Alternatively, only the refined localizations for the second portion of the collections of features which were computed based on a rough localization propagated from a collection of features in the first portion may be recomputed. Similarly, if other criteria for propagating a rough localization from one collection of features to another are not satisfied for any collections of features, processing may default to the full process as shown in FIG. 64.

[0531] In scenarios in which the processing as described in connection with FIGS. 64 and 65 does not result in computations of a localization result 6452, other processing as elsewhere described herein may be performed. For example, the result returned to user device 6410 may be an indication of a localization failure. Alternatively, further localization attempts may be made with respect to other maps or other segments of the same map. In some embodiments, processing with propagation of rough localization results as illustrated in FIG. 65 may be used in place of other localization processing, and may require less computation in many circumstances, which leads, on average, to less processing.

[0532] The process of FIG. 65 is illustrated in connection with localizing a device with respect to a stored map. The processing, alternatively or additionally, may be applied in other contexts. In FIG. 65, the localization result is a transformation that may be used, for example, in positioning virtual content with respect to the portable device such that each of multiple portable devices may render the virtual content with the same positional relationship with respect to the physical world so as to provide a shared experience of the virtual content for multiple users. However, the process may be used whenever a batch of snapshots is available for computing a transformation between a coordinate frame in which those snapshots were acquired and a coordinate frame of a map or other collection of features.

[0533] For example in merging a tracking map to a canonical map, a set of 2D features from the tracking map, such as those associated with a persistent pose, may be compared to 3D features in the canonical map. The processing of FIG. 65 may be applied to determine whether sets of features match and, if so, a transformation between the sets of features. Computation of that transformation may be used to align the tracking map to the canonical map so that the maps may be combined.

[0534] As another example, the output of the process of FIG. 65, though described as a transformation, more generally describes a relative pose between two components of a XR system. A relative pose may be computed between a first component of an XR system, establishing the coordinate frame in which the snapshots 6432 were acquired, and a second component, establishing the coordinate frame of the collection of features to which the snapshots were compared.

Processing as illustrated in FIG. 65 does not depend on the nature of those components and may be used in other circumstances to determine a relative pose between two components of an XR system.

[0535] Regardless of how the source of the inputs and the use of the output of the processing of FIG. **65**, that processing may, on average, require less computations resources than other techniques for computing a relative pose between two components. That computation may nonetheless have comparable accuracy, providing benefits such as a better user experience in an XR system.

FURTHER CONSIDERATIONS

[0536] FIG. **60** shows a diagrammatic representation of a machine in the exemplary form of a computer system **1900** within which a set of instructions, for causing the machine to perform any one or more of the methodologies discussed herein, may be executed, according to some embodiments. In alternative embodiments, the machine operates as a standalone device or may be connected (e.g., networked) to other machines. Further, while only a single machine is illustrated, the term “machine” shall also be taken to include any collection of machines that individually or jointly execute a set (or multiple sets) of instructions to perform any one or more of the methodologies discussed herein.

[0537] The exemplary computer system **1900** includes a processor **1902** (e.g., a central processing unit (CPU), a graphics processing unit (GPU) or both), a main memory **1904** (e.g., read only memory (ROM), flash memory, dynamic random access memory (DRAM) such as synchronous DRAM (SDRAM) or Rambus DRAM (RDRAM), etc.), and a static memory **1906** (e.g., flash memory, static random access memory (SRAM), etc.), which communicate with each other via a bus **1908**.

[0538] The computer system **1900** may further include a disk drive unit **1916**, and a network interface device **1920**.

[0539] The disk drive unit **1916** includes a machine-readable medium **1922** on which is stored one or more sets of instructions **1924** (e.g., software) embodying any one or more of the methodologies or functions described herein. The software may also reside, completely or at least partially, within the main memory **1904** and/or within the processor **1902** during execution thereof by the computer system **1900**, the main memory **1904** and the processor **1902** also constituting machine-readable media.

[0540] The software may further be transmitted or received over a network **18** via the network interface device **1920**.

[0541] The computer system **1900** includes a driver chip **1950** that is used to drive projectors to generate light. The driver chip **1950** includes its own data store **1960** and its own processor **1962**.

[0542] While the machine-readable medium **1922** is shown in an exemplary embodiment to be a single medium, the term “machine-readable medium” should be taken to include a single medium or multiple media (e.g., a centralized or distributed database, and/or associated caches and servers) that store the one or more sets of instructions. The term “machine-readable medium” shall also be taken to include any medium that is capable of storing, encoding, or carrying a set of instructions for execution by the machine and that cause the machine to perform any one or more of the methodologies of the present invention. The term “machine-readable medium” shall accordingly be taken to include, but not be limited to, solid-state memories, optical and magnetic media, and carrier wave signals.

[0543] Having thus described several aspects of some embodiments, it is to be appreciated that various alterations, modifications, and improvements will readily occur to those skilled in the art.

[0544] As one example, embodiments are described in connection with an augmented (AR) environment. It should be appreciated that some or all of the techniques described herein may be applied in an MR environment or more generally in other XR environments, and in VR environments.

[0545] As another example, embodiments are described in connection with devices, such as wearable devices. It should be appreciated that some or all of the techniques described herein may be implemented via networks (such as cloud), discrete applications, and/or any suitable

combinations of devices, networks, and discrete applications.

[0546] Further, FIG. 29 provides examples of criteria that may be used to filter candidate maps to yield a set of high ranking maps. Other criteria may be used instead of or in addition to the described criteria. For example, if multiple candidate maps have similar values of a metric used for filtering out less desirable maps, characteristics of the candidate maps may be used to determine which maps are retained as candidate maps or filtered out. For example, larger or more dense candidate maps may be prioritized over smaller candidate maps.

[0547] Such alterations, modifications, and improvements are intended to be part of this disclosure, and are intended to be within the spirit and scope of the disclosure. Further, though advantages of the present disclosure are indicated, it should be appreciated that not every embodiment of the disclosure will include every described advantage. Some embodiments may not implement any features described as advantageous herein and in some instances. Accordingly, the foregoing description and drawings are by way of example only.

[0548] The above-described embodiments of the present disclosure can be implemented in any of numerous ways. For example, the embodiments may be implemented using hardware, software or a combination thereof. When implemented in software, the software code can be executed on any suitable processor or collection of processors, whether provided in a single computer or distributed among multiple computers. Such processors may be implemented as integrated circuits, with one or more processors in an integrated circuit component, including commercially available integrated circuit components known in the art by names such as CPU chips, GPU chips, microprocessor, microcontroller, or co-processor. In some embodiments, a processor may be implemented in custom circuitry, such as an ASIC, or semicustom circuitry resulting from configuring a programmable logic device. As yet a further alternative, a processor may be a portion of a larger circuit or semiconductor device, whether commercially available, semi-custom or custom. As a specific example, some commercially available microprocessors have multiple cores such that one or a subset of those cores may constitute a processor. Though, a processor may be implemented using circuitry in any suitable format.

[0549] Further, it should be appreciated that a computer may be embodied in any of a number of forms, such as a rack-mounted computer, a desktop computer, a laptop computer, or a tablet computer. Additionally, a computer may be embedded in a device not generally regarded as a computer but with suitable processing capabilities, including a Personal Digital Assistant (PDA), a smart phone or any other suitable portable or fixed electronic device.

[0550] Also, a computer may have one or more input and output devices. These devices can be used, among other things, to present a user interface. Examples of output devices that can be used to provide a user interface include printers or display screens for visual presentation of output and speakers or other sound generating devices for audible presentation of output. Examples of input devices that can be used for a user interface include keyboards, and pointing devices, such as mice, touch pads, and digitizing tablets. As another example, a computer may receive input information through speech recognition or in other audible format. In the embodiment illustrated, the input/output devices are illustrated as physically separate from the computing device. In some embodiments, however, the input and/or output devices may be physically integrated into the same unit as the processor or other elements of the computing device. For example, a keyboard might be implemented as a soft keyboard on a touch screen. In some embodiments, the input/output devices may be entirely disconnected from the computing device, and functionally integrated through a wireless connection.

[0551] Such computers may be interconnected by one or more networks in any suitable form, including as a local area network or a wide area network, such as an enterprise network or the Internet. Such networks may be based on any suitable technology and may operate according to any suitable protocol and may include wireless networks, wired networks or fiber optic networks.

[0552] Also, the various methods or processes outlined herein may be coded as software that is

executable on one or more processors that employ any one of a variety of operating systems or platforms. Additionally, such software may be written using any of a number of suitable programming languages and/or programming or scripting tools, and also may be compiled as executable machine language code or intermediate code that is executed on a framework or virtual machine.

[0553] In this respect, the disclosure may be embodied as a computer readable storage medium (or multiple computer readable media) (e.g., a computer memory, one or more floppy discs, compact discs (CD), optical discs, digital video disks (DVD), magnetic tapes, flash memories, circuit configurations in Field Programmable Gate Arrays or other semiconductor devices, or other tangible computer storage medium) encoded with one or more programs that, when executed on one or more computers or other processors, perform methods that implement the various embodiments of the disclosure discussed above. As is apparent from the foregoing examples, a computer readable storage medium may retain information for a sufficient time to provide computer-executable instructions in a non-transitory form. Such a computer readable storage medium or media can be transportable, such that the program or programs stored thereon can be loaded onto one or more different computers or other processors to implement various aspects of the present disclosure as discussed above. As used herein, the term “computer-readable storage medium” encompasses only a computer-readable medium that can be considered to be a manufacture (i.e., article of manufacture) or a machine. In some embodiments, the disclosure may be embodied as a computer readable medium other than a computer-readable storage medium, such as a propagating signal.

[0554] The terms “program” or “software” are used herein in a generic sense to refer to any type of computer code or set of computer-executable instructions that can be employed to program a computer or other processor to implement various aspects of the present disclosure as discussed above. Additionally, it should be appreciated that according to one aspect of this embodiment, one or more computer programs that when executed perform methods of the present disclosure need not reside on a single computer or processor, but may be distributed in a modular fashion amongst a number of different computers or processors to implement various aspects of the present disclosure.

[0555] Computer-executable instructions may be in many forms, such as program modules, executed by one or more computers or other devices. Generally, program modules include routines, programs, objects, components, data structures, etc. that perform particular tasks or implement particular abstract data types. Typically the functionality of the program modules may be combined or distributed as desired in various embodiments.

[0556] Also, data structures may be stored in computer-readable media in any suitable form. For simplicity of illustration, data structures may be shown to have fields that are related through location in the data structure. Such relationships may likewise be achieved by assigning storage for the fields with locations in a computer-readable medium that conveys relationship between the fields. However, any suitable mechanism may be used to establish a relationship between information in fields of a data structure, including through the use of pointers, tags or other mechanisms that establish relationship between data elements.

[0557] Various aspects of the present disclosure may be used alone, in combination, or in a variety of arrangements not specifically discussed in the embodiments described in the foregoing and is therefore not limited in its application to the details and arrangement of components set forth in the foregoing description or illustrated in the drawings. For example, aspects described in one embodiment may be combined in any manner with aspects described in other embodiments.

[0558] Also, the disclosure may be embodied as a method, of which an example has been provided. The acts performed as part of the method may be ordered in any suitable way. Accordingly, embodiments may be constructed in which acts are performed in an order different than illustrated, which may include performing some acts simultaneously, even though shown as sequential acts in illustrative embodiments.

[0559] Use of ordinal terms such as “first,” “second,” “third,” etc., in the claims to modify a claim element does not by itself connote any priority, precedence, or order of one claim element over another or the temporal order in which acts of a method are performed, but are used merely as labels to distinguish one claim element having a certain name from another element having a same name (but for use of the ordinal term) to distinguish the claim elements.

[0560] Also, the phraseology and terminology used herein is for the purpose of description and should not be regarded as limiting. The use of “including,” “comprising,” or “having,” “containing,” “involving,” and variations thereof herein, is meant to encompass the items listed thereafter and equivalents thereof as well as additional items.

Claims

1. A method for aligning a collection of a plurality of collections of features to spatial information, the method comprising: using at least one processor to perform: computing a first rough alignment relative to the spatial information for a first collection of the plurality of collections of features; computing, based on the first rough alignment, a first refined alignment for the first collection of features; and determining, based on one or more criteria, whether to compute a second rough alignment for a second collection of the plurality of collections of features; when it is determined to not compute the second rough alignment for the second collection of features: computing, based on the first rough alignment, a second refined alignment for the second collection of features based on the first rough alignment; and when it is determined to perform the second rough alignment for the second collection of features: computing the second rough alignment relative to the spatial information for the second collection of features; and computing, based on the second rough alignment, the second refined alignment for the second collection of features.
2. The method of claim 1, wherein computing the second rough alignment relative to the spatial information for the second collection of features comprises computing the second rough alignment with a subset of the second collection of features.
3. The method of claim 1, further comprising evaluating the one or more criteria by computing a confidence metric for the first collection of features.
4. The method of claim 3, wherein the confidence metric is computed based on the first collection of features and a set of features in the spatial information corresponding to the first collection of features.
5. The method of claim 4, wherein the confidence metric is computed at least in part by determining a number of features in the first collection of features that are coincident with features of the set of features.
6. The method of claim 1, further comprising evaluating the one or more criteria by computing a measure of parallax determined between the first collection of features and the second collection of features.
7. The method of claim 1, further comprising evaluating the one or more criteria based on motion of a portable device.
8. The method of claim 1, further comprising evaluating the one or more criteria by computing a measure of consensus among sets of features in the spatial information.
9. The method of claim 8, wherein determining, based on one or more criteria, whether to compute the second rough alignment for the second collection of features comprises determining whether the measure of consensus among the sets of features is below a threshold.
10. The method of claim 1, wherein computing the second refined alignment comprises computing a transformation between a local coordinate frame of a device and a coordinate frame of a persisted map in the spatial information.
11. The method of claim 1, wherein the plurality of collections of features each comprise a posed feature rig, wherein the posed feature rig comprises a plurality of feature descriptors and a pose

associated with each of the plurality of features descriptors.

12. The method of claim 1, wherein computing the first rough alignment for the first collection of features comprises selecting a first subset of spatial information from the spatial information based on location metadata associated with the first collection of features.

13. At least one non-transitory computer readable medium comprising instructions that, when executed by at least one processor, perform a method for aligning a collection of a plurality of collections of features to spatial information, the method comprising: computing a first rough alignment relative to the spatial information for a first collection of the plurality of collections of features; computing, based on the first rough alignment, a first refined alignment for the first collection of features; determining, based on one or more criteria, whether to compute a second rough alignment for a second collection of the plurality of collections of features; and when it is determined to not compute the second rough alignment for the second collection of features: computing, based on the first rough alignment, a second refined alignment for the second collection of features based on the first rough alignment; and when it is determined to perform the second rough alignment for the second collection of features: computing the second rough alignment relative to the spatial information for the second collection of features; and computing, based on the second rough alignment, the second refined alignment for the second collection of features.

14. The at least one non-transitory computer readable medium of claim 13, wherein computing the second rough alignment relative to the spatial information for the second collection of features comprises computing the second rough alignment with a subset of the second collection of features.

15. The at least one non-transitory computer readable medium of claim 13, wherein the method further comprises: evaluating the one or more criteria by computing a confidence metric for the first collection of features.

16. The at least one non-transitory computer readable medium of claim 13, wherein the method further comprises evaluating the one or more criteria by computing a measure of parallax determined between the first collection of features and the second collection of features.

17. The at least one non-transitory computer readable medium of claim 13, wherein computing the second refined alignment comprises computing a transformation between a local coordinate frame of a device and a coordinate frame of a persisted map in the spatial information.

18. The at least one non-transitory computer readable medium of claim 13, wherein computing the first rough alignment for the first collection of features comprises selecting a first subset of spatial information from the spatial information based on location metadata associated with the first collection of features.

19. A system for aligning a collection of a plurality of collections of features to spatial information, the system comprising: at least one processor; and at least one non-transitory computer-readable storage medium storing instructions that, when executed by the at least one processor, cause the at least one processor to: compute a first rough alignment relative to the spatial information for a first collection of the plurality of collections of features; compute, based on the first rough alignment, a first refined alignment for the first collection of features; and determine, based on one or more criteria, whether to compute a second rough alignment for a second collection of the plurality of collections of features; when it is determined to not compute the second rough alignment for the second collection of features: compute, based on the first rough alignment, a second refined alignment for the second collection of features based on the first rough alignment; and when it is determined to perform the second rough alignment for the second collection of features: compute the second rough alignment relative to the spatial information for the second collection of features; and compute, based on the second rough alignment, the second refined alignment for the second collection of features.

20. The system of claim 19, wherein computing the second rough alignment relative to the spatial information for the second collection of features comprises computing the second rough alignment with a subset of the second collection of features.

