



(51) International Patent Classification:

G06V 20/52 (2022.01) H04N 7/15 (2006.01)
G06V 40/16 (2022.01) H04N 23/90 (2023.01)

(21) International Application Number:

PCT/US2023/075906

(22) International Filing Date:

04 October 2023 (04.10.2023)

(25) Filing Language:

English

(26) Publication Language:

English

(71) Applicant: **HEWLETT-PACKARD DEVELOPMENT COMPANY, L.P.** [US/US]; 10300 Energy Drive, Spring, Texas 77389 (US).

(72) Inventors: **BHATT, Rajen B.**; 6414 Interlaken Drive, McDonald, Pennsylvania 15057 (US). **ZHANG, Kui**;

7700 West Parmer Lane, Austin, Texas 78729-8101 (US). **KRISHNAMURTHY, Raghavendra Balavalikar**; 7700 W. Parmer Lane, Austin, Texas 78729-8101 (US). **MING, Chen**; 10300 Energy Drive, Spring, Texas 77389 (US).

(74) Agent: **COOK, Victoria C.**; Quarles & Brady LLP, 411 East Wisconsin Avenue, Suite 2400, Milwaukee, Wisconsin 53202 (US).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH,

(54) Title: SYSTEMS AND METHODS FOR PARTICIPANT REIDENTIFICATION IN MULTI-CAMERA VIDEOCONFERENCING

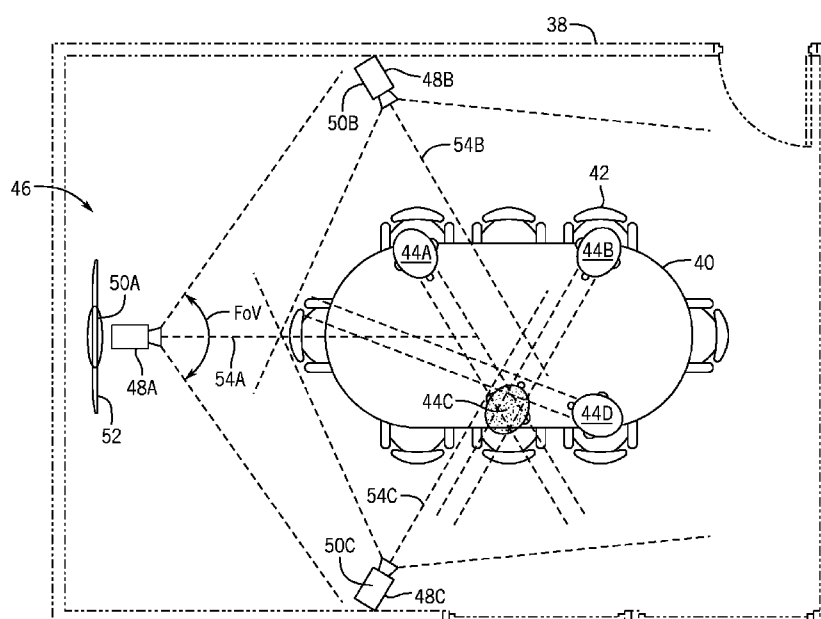


FIG. 1

(57) Abstract: Examples of the disclosure provide a device, system, and method for identifying participants in a location using a plurality of cameras. The method includes capturing images of the location using a primary camera and a secondary camera, applying a machine learning human head detector model to the images to detect human heads in the images and identify coordinates for each human head detected in the images, and determining a centroid of the coordinates for each human head detected in the images and a centroid angle for each human head detected in the images relative to the centroid. The method further includes ranking identification labels for each human head based on the centroid angle for each human head detected in the images, identifying a reference human head in a primary image captured by the primary camera, and aligning the identification labels across each of the images to reidentify each human head.



TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS,
ZA, ZM, ZW.

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

Published:

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

SYSTEMS AND METHODS FOR PARTICIPANT REIDENTIFICATION IN MULTI-CAMERA VIDEOCONFERENCING

BACKGROUND

[0001] Videoconferencing systems typically connect people at a videoconferencing endpoint, such as a videoconference room, with people at other videoconferencing endpoints. In some videoconferencing modes, all participants detected in a videoconference room are separated and put into a gallery view to create equality with the remote participants.

BRIEF DESCRIPTION OF THE DRAWINGS

[0002] FIG. 1 is a top view of an example conference room including multiple cameras, according to some aspects of the present disclosure.

[0003] FIG. 2 is a schematic perspective view of another example conference room with three individuals located at different coordinate positions in relation to a videoconference camera.

[0004] FIG. 3 is a top view of the example conference room of FIG. 2.

[0005] FIG. 4 is a front view of yet another example conference room with three individuals located at different coordinate positions, according to some examples of the present disclosure.

[0006] FIG. 5 is a side view of an example vertical dimension determination for a human head of a subject, according to an example of the present disclosure.

[0007] FIG. 6 is a top view of an example horizontal dimension determination for a human head of a subject, according to an example of the present disclosure.

[0008] FIG. 7 is a perspective view of subjects in a conference room with different vertical dimensions based on the example determinations of FIGS. 5 and 6.

[0009] FIG. 8 is a schematic diagram of a camera and a two-dimensional image plane with an example determination of room coordinates for a head bounding box, according to an example of the present disclosure.

[0010] FIG. 9 is a top view of an example videoconference room plotted on an example two-dimensional world plane with three cameras arranged in an outside-in configuration.

[0011] FIG. 10 is another top view of the example videoconference room of FIG. 9 with two cameras arranged in an inside-out configuration.

[0012] FIG. 11 is a flowchart of a method of implementing a multi-camera calibration system in a conference room, according to an example of the present disclosure.

[0013] FIG. 12A is a rear view of a still another example conference room with a first camera configuration, according to an example of the present disclosure.

[0014] FIG. 12B is a rear view of the example conference room of FIG. 12A with a second camera configuration.

[0015] FIG. 13 is a pixel coordinate plot of a front view of the example conference room of FIG. 12A.

[0016] FIG. 14 is a polar coordinate plot of the front view of the example conference room of FIG. 13.

[0017] FIG. 15 is a pixel coordinate plot of a left view of the example conference room of FIG. 12A.

[0018] FIG. 16 is a polar coordinate plot of the left view of the example conference room of FIG. 15.

[0019] FIG. 17 is a pixel coordinate plot of a right view of the example conference room of FIG. 12A.

[0020] FIG. 18 is a polar coordinate plot of the right view of the example conference room of FIG. 17.

[0021] FIG. 19 is the pixel plot of FIG. 13 with a reference person having been identified, according to an example of the present disclosure.

[0022] FIG. 20 is the pixel plot of FIG. 15 with a reference person having been identified, according to an example of the present disclosure.

[0023] FIG. 21 is the pixel plot of FIG. 17 with a reference person having been identified, according to an example of the present disclosure.

[0024] FIG. 22 is the pixel plot of FIG. 13 where identification (ID) labels are aligned with ID labels in other views of the example conference room.

[0025] FIG. 23 is the pixel plot of FIG. 15 where the ID labels are aligned with the ID labels in the other views of the example conference room.

[0026] FIG. 24 is the pixel plot of FIG. 17 where the ID labels are aligned with the ID labels in the other views of the example conference room.

[0027] FIG. 25 is a flowchart of a method of implementing a centroid-based identification system in a conference room, according to an example of the present disclosure.

[0028] FIG. 26 is a front view of a camera according to an example of the present disclosure.

[0029] FIG. 27 is a flowchart of a method of determining image plane coordinates for a detected subject, according to an example of the present disclosure.

[0030] FIG. 28 is a schematic of an example codec, according to an example of the present disclosure.

[0031] FIG. 29 is a flowchart of a method of reidentifying participants across multiple views of a videoconference and selecting an optimal view of each participant, according to an example of the present disclosure.

DETAILED DESCRIPTION

[0032] In a video conference room equipped with multiple cameras, the same participant may appear in the field of view of more than one camera. Thus, an issue in videoconferencing using multiple cameras is the duplication of participants, which includes transmitting more than one image of a particular participant to a far end of the videoconference. Correspondingly, the framing of individuals in a videoconference room can be improved by determining the location of individual participants in the room relative to one another or a particular reference point. For example, if Person A is sitting at 2.5 meters from the camera and Person B is sitting at 4 meters from the camera, the ability to detect this location information can enable various advanced framing and tracking experiences. For example, participant location information can be used to compare multiple camera views and reidentify the same participant in each camera view to determine an optimal camera view of the participant.

[0033] More specifically, when multiple cameras of a videoconference system are used in a conference room with two or more participants, multiple views of the participants are captured, *e.g.*, multiple views at different angles. This is particularly true when the videoconference system includes a front camera and at least one side camera which are used to record visual data for the videoconference system. For example, a primary camera may capture a first or frontal view of the participants, and the side camera(s) may capture a side views and/or view(s) of the meeting participant. As a result, multiple images of the same participant and/or participants may be transmitted to a far end of the videoconference, which in turn may be confusing to the other participants in the videoconference. Correspondingly, it is undesirable to transmit a particular view to a far end of the videoconference if a participant's face is not fully visible in that particular view. Further, no adequate industry standard or specification has been developed to track individuals across multiple camera views in a videoconferencing system based on their relative distance from one another or their relative position in a conference room.

[0034] For many applications, it is useful to know the horizontal and vertical location of the participants in the room to provide for a more comprehensive and complete understanding of the videoconference room environment. The ability to determine two-dimensional room

distance parameters for each meeting participant and use such parameters to identify each participant can be enabled by using an external feature anchor or computationally intensive machine learning-based monocular depth estimation models, but such approaches impose significant hardware and/or processing costs without providing the accuracy for identifying participants across multiple camera views. Further, such approaches are limited in their adaptability to a variety of different room geometries and locations due to their reliance on dedicated hardware.

[0035] For example, various conventional techniques attempt to perform what is called person reidentification (ReID) to match a participant captured by one camera to the same participant captured on another camera. In one variation, pre-trained deep learning ReID models are utilized to identify embeddings of a participant in a particular camera view and match those embeddings with embeddings found in other camera views. However, this technique relies upon resource-intensive computations that require considerable memory in order to be performed, and such computations still prove ineffective to differentiate between participants who have similar clothing to one another, *e.g.*, participants who are wearing uniforms. In addition, deep learning ReID models are primarily trained by targeting upright pedestrians, so their performance is further compromised when such models are applied to seated participants.

[0036] In another variation, an ReID-based approach includes mapping three-dimensional (3D) world coordinate systems to two-dimensional (2D) camera planes by using an external object or feature vector matching to determine a transformation matrix. The reliance upon external reference points in a room limits the applicability of such models to the physical room in which they are in. As a result, ReID techniques which rely on external reference points to calculate a transformation matrix lack adaptability and are unable to be used in a wide variety of videoconference room settings.

[0037] Accordingly, in some examples, the present disclosure provides methods of and apparatus for reidentifying participants across multiple camera views in a videoconference. In particular, the present disclosure provides a method of identifying, from multiple camera views, each participant in a videoconferencing room, preventing duplicate participant images from being transmitted to a far end of the videoconference, and/or identifying an optimal view of a participant to be transmitted to the far end instead of another, less desirable duplicate view. Thus, while a conference room is covered by multiple cameras, the far end can be shown a single manipulated stream that does not have duplicated people and that has the best view of everyone in the videoconferencing room. By utilizing the disclosed ReID methods,

communication between participants in the videoconference may be clearer, and the overall videoconferencing experience may be more enjoyable for the participants. Further, the methods discussed herein are applicable to a wide variety of different locations and room designs, meaning that the disclosed methods may be easily assembled and applied to any particular conference room.

[0038] By way of example, FIG. 1 illustrates a conference room 38 for use in videoconferencing. The conference room 38 includes a conference table 40 and a series of chairs 42. Videoconference participants 44 are seated in the chairs 42 around the conference table 40. In the non-limiting example illustrated in FIG. 1, a first participant 44A, a second participant 44B, a third participant 44C, and a fourth participant 44D are seated around the conference table 40.

[0039] Referring still to FIG. 1, in some aspects, a videoconferencing system 46 includes a camera 48, a microphone array 50, and a monitor 52. More specifically, as shown in the example of FIG. 1, the videoconferencing system 46 can include a primary or front camera 48A and a secondary camera (or cameras), such as a secondary or left camera 48B and a tertiary or right camera 48C. However, it is contemplated that the videoconferencing system 46 may include more or fewer cameras 48 than those illustrated in FIG. 1. Each of the cameras 48 has a field-of-view (FOV), horizontal and vertical, and an axis or centerline (CL) extending in a direction that corresponds to the direction in which the respective camera 48 is pointing (*i.e.*, the respective camera's 48 line of sight that is straight at 90-degrees from its focal point). In some aspects, each of the cameras 48 includes a corresponding microphone array 50 (that is, a primary microphone array 50A corresponding to the primary camera 48A, a secondary microphone array 50B corresponding to the secondary camera 48B, and a tertiary microphone array 50C corresponding to the tertiary camera 48C). The microphone arrays 50 may be used to record and transmit audio data in the videoconference using sound source localization (SSL). In some examples, SSL is used in a way that is similar to the uses described in Int'l. App. No. PCT/US2023/016764 and U.S. Patent App. Pub. No. 2023/0053202, which are incorporated herein by reference in their entirety. In some examples, each of the microphone arrays 50 are housed on or within housings of each of the cameras 48. In addition, the videoconferencing system 46 can include a monitor 52 or television that is provided to display a far end conference site or sites and generally to provide loudspeaker output. The monitor 52 can be coupled to the front camera 48A and the front microphone array 50A. However, it is contemplated that each of the left and right cameras 48B, 48C may also be coupled to a separate monitor (not shown),

and that the videoconferencing system 46 may include any number of other monitors (not shown) in addition to the monitor 52.

[0040] As illustrated in FIG. 1, the cameras 48 are positioned such that the front camera 48A defines a front CL 54A centered on the length of the conference table 40, and the left and right cameras 48B, 48C are angled on either side of the conference table 40 with respect to the front CL 54A. In some aspects, a central microphone (not shown) is also provided on the conference table 40 to capture the speaker (*i.e.*, the participant speaking) for transmission to a far end of the videoconference. Angling the left and right cameras 48B, 48C on either side of the conference table 40 provides the videoconferencing system 46 with additional views of the conference room 38. This in turn may provide a better opportunity to see the faces of videoconference participants 44 seated on the sides of the conference table 40 when the participants 44 are facing one another in the conference room 38.

[0041] In some aspects, a participant 44 is located within more than one FOV of the cameras 48, meaning that the participant 44 may be duplicated when the views of the cameras 48 are transmitted to a far end of the videoconference. This in turn may cause confusion in the videoconference and/or result in non-optimal views of the participant 44 being transmitted to a far end of the videoconference. Thus, it is advantageous to reidentify the participants 44 across each of the views of the cameras 48, thereby reducing confusion in the videoconference and ensuring that only one, optimal view of each participant 44 is shown in the videoconference. Additionally, while FIG. 1 illustrates an example of a videoconference room with four participants 44, more or fewer participants 44 may be seated around the conference table 40 at any given time, and further examples of videoconference rooms, locations of participants in videoconference rooms, and camera arrangements will be discussed below in greater detail.

[0042] Correspondingly, the front camera 48A may provide a better view of the faces of certain participants 44 when they are looking forward, *i.e.*, toward the front of the conference room 38. In some aspects, the cameras 48 and microphone arrays 50 are used in combination to provide multiple views of the conference room 38, and one or more of the processes described herein are applied to each of the views to reidentify participants thereacross and prevent duplicate images of the participants from being transmitted to a far end of the videoconference. For example, the processes described herein allow an optimal view of each meeting participant to be identified and chosen for transmission, while non-ideal views are not transmitted to the far end of the videoconference. This is accomplished using an artificial intelligence (AI) or machine learning human head detector model, as discussed below. As used

herein, an “optimal” view may be a view that provides a best frontal view of a participant, *i.e.*, a best view of a face of a participant. For example, an optimal view may be determined by applying facial recognition techniques to an image, such as those discussed in U.S. Patent App. Pub. No. 2023/0216988, which is incorporated herein by reference in its entirety. In particular, an optimal view may be a view in which the quality of facial features recognized by a facial recognition technique is greater than any other view.

[0043] In some aspects, the AI human head detector model is substantially similar to that described in Int’l. App. No. PCT/US2023/016764, which is incorporated herein by reference in its entirety. For example, referring now to FIGS. 2 and 3, a conference room 64 is illustrated with three videoconference participants 66, 68, 70 located at different coordinate positions. In the conference room 64, the front camera 48A has horizontal and vertical FOV, and the camera location with respect to the room 64 is denoted by the three-dimensional (3D) coordinates $\{0, 0, 0\}$. Further, the front camera 48A captures a view of all three participants 66, 68, 70 having locations that can be characterized in terms of a pan angle Φ_{PAN} relative to a centerline 56 of the front camera 48A and a distance measure between the front camera 48A and each participant 66, 68, 70. In particular, a first participant 66 has a location defined by a first pan angle 76 and a first distance 78. In addition, a second participant 68 has a location defined by pan angle 80 and a second distance 82, and a third participant 70 has a location defined by pan angle 84 and a third distance measure 86.

[0044] Referring now specifically to FIG. 3, a top view is illustrated of the conference room 64 of FIG. 2. In some examples, the location of each participant 66, 68, 70 may be characterized in terms of the pan angles 76, 80, 84 and distances 78, 82, 86 that are derived from an x_{ROOM} dimension or axis 88 and a y_{ROOM} dimension or axis 90, where the front camera 48A is located at $\{x_{\text{ROOM}}, y_{\text{ROOM}}\}$ coordinate positions of $\{0, 0\}$. In particular, the first participant 66 has a location defined by the first pan angle 76 and a first distance measure 78 which is characterized by two-dimensional room distance parameters $\{-0.5, 1\}$ to indicate that the participant is located at a vertical distance of 1 meter, measured from the front camera 48A along the y_{ROOM} axis 90, and at a horizontal distance of -0.5 meters, measured along the x_{ROOM} axis 88 that is perpendicular to the y_{ROOM} axis 90. In addition, the second participant 68 has a location defined by the second pan angle 80 and a second distance measure 82 which is characterized by two-dimensional room distance parameters $\{0, 3\}$ to indicate that the participant is located at a vertical distance of 3 meters (measured along the y_{ROOM} axis 90) and at a horizontal distance of 0 meters (measured along the x_{ROOM} axis 88) to indicate that the second person is located along the centerline 56 of the front camera 48. Finally, the third participant 70 has a location

defined by the third pan angle 84 and a third distance measure 86 which is characterized by two-dimensional room distance parameters {1, 2.5} to indicate that the participant is located at a vertical distance of 2.5 meters (measured along the y_{ROOM} axis 90) and at a horizontal distance of 1 meter (measured along the x_{ROOM} axis 88).

[0045] The relationship between the pan angle values (Φ_{PAN}) and the two-dimensional room distance parameters $\{x_{ROOM}, y_{ROOM}\}$ may be determined by using a reference coordinate table (not shown) in which pan angle Φ_{PAN} values for the videoconference front camera 48A are computed for meeting participants located at different coordinate positions $\{x_{ROOM}, y_{ROOM}\}$ in the example conference room 64 of FIGS. 2 and 3. An identical table (not shown) of negative pan angle Φ_{PAN} values (e.g., $-\Phi_{PAN}$) can be computed for coordinate positions of $\{-x_{ROOM}, y_{ROOM}\}$ in the example conference room 64. Thus, it will be understood that the same pan angle Φ_{PAN} value (e.g., $\Phi_{PAN} = 0$) will be generated for a meeting participant located along the centerline 56 of the front camera 48A (e.g., $x_{ROOM} = 0$) at any depth measure (e.g., $y_{ROOM} = 0.5-8$). Similarly, the same pan angle Φ_{PAN} value (e.g., $\Phi_{PAN} = 45$) will be generated for a meeting participant located at any coordinate position where $x_{ROOM} = y_{ROOM}$. As illustrated, the pan angle Φ_{PAN} alone may not be sufficient information for determining the two-dimensional room distance parameters $\{x_{ROOM}, y_{ROOM}\}$ for the location of a participant. For example, the first participant 66 may appear larger to the front camera 48A than the second participant 68 due to vanishing points perspective. Thus, as a meeting participant moves further away from the front camera 48A, the apparent height and width of the participant become smaller to the videoconferencing system, and when projected to a camera image sensor 92, meeting participants are represented with a smaller number of pixels compared to participants that are nearer to the front camera 48. Further, if two heads are seen by the front camera 48A as having the same size, they are not necessarily located at the same distance, and their locations in a two-dimensional x_{ROOM} - y_{ROOM} plane 94, as illustrated in FIG. 3, may be different due to the pan angle Φ_{PAN} and distortion in the height and width.

[0046] To illustrate an example of the challenges posed by perspective projection effects when determining the locations of meeting participants, reference is now made to FIG. 4 which illustrates a picture image 96 of another example conference room 98. Three subjects or meeting participants 100, 102, 104 are located in the conference room 98 at different coordinate positions and with corresponding head frames or bounding boxes 106, 108, 110 identified in terms of the coordinate positions for each of the participants 100, 102, 104. As depicted, the coordinate positions may be measured with reference to a room width dimension x_{ROOM} and a room depth dimension y_{ROOM} . The room width dimension x_{ROOM} extends across a width of the

conference room 98 from the centerline 56 (see FIG. 1) of the front camera 48A (see FIG. 1) so that negative values of x_{ROOM} are located to the left of the centerline 56 (see FIG. 1) and positive values of x_{ROOM} are located to the right of the centerline 56 (see FIG. 1). In addition, the room depth dimension y_{ROOM} extends down a length of the room 98 parallel with the centerline 56 of the front camera 48A (see FIG. 2). By applying computer vision processing to the picture image 96, a first meeting participant 100 is detected in the back left corner of the room 98, and an interest region around the head of the first meeting participant 100 is framed with a first head bounding box 106, where the first meeting participant 100 is located at the two-dimensional room distance parameters $\{x_{ROOM} = -3, y_{ROOM} = 21\}$. In similar fashion, a second meeting participant 102 seated at a table 116 is detected with the head of the second meeting participant 102 framed with a second head bounding box 108, where the second meeting participant 102 is located at the two-dimensional room distance parameters $\{x_{ROOM} = -1, y_{ROOM} = 13\}$. Finally, a third meeting participant 104 standing to the right is detected with the head of the third meeting participant 104 framed with a third head bounding box 110, where the third meeting participant 104 is located at the two-dimensional room distance parameters $\{x_{ROOM} = 5, y_{ROOM} = 14\}$.

[0047] In particular, the statistical distribution of human head height and width measurements may be used to determine a min-median-max measure for the head size in centimeters. Additionally, by knowing the FOV resolution of the front camera 48A in both horizontal and vertical directions with the respective horizontal and vertical pixel counts, the measured angular extent of each head can be used to compute the percentage of the overall frame occupied by the head and the number of pixels for the head height and width measures. Using this information to compute a look-up table for min-median-max head sizes (height and width) at various distances, an artificial (AI) human head detector model can be applied to detect the location of each head in a two-dimensional viewing plane with specified image plane coordinates and associated width and height measures for a head frame or bounding box (e.g., $\{x_{box}, y_{box}, width, height\}$). By using the reverse look-up table operation, the distance can be determined between the front camera 48A and each head that is located on the centerline 56 the front camera 48.

[0048] In some examples, the subject detection process is similar to the AI head detection process as disclosed in U.S. Patent Application No. 17/971,564, filed on October 22, 2022, which is incorporated by reference herein in its entirety. Referring specifically to FIG. 5, a side view is illustrated of an example vertical dimension determination 200 for a human head 202. In some examples, a front camera 48A is positioned to capture an image of the human head

202 so that a vertical head height measure V that can be calculated based on an angular extent angle $\theta_{\text{FRAME_V}/2}$ of the upper half of a vertical head height $V/2$ and a distance d between the front camera 48A and the human head 202. As illustrated, the human head 202 has a head height V which corresponds to the vertical dimension of a head bounding box (not shown). From the vantage of the front camera 48A, the vertical head height V makes an angle $\theta_{\text{FRAME_V}}$ extending from the bottom to the top of the head 202. Upon bisecting the angle $\theta_{\text{FRAME_V}}$, the upper half of the vertical head height $V/2$ makes an angle $\theta_{\text{FRAME_V}/2}$ with the camera's focal point (*i.e.*, the centerline 56 of the camera). As a result, the vertical head height measure V can be calculated based on an angular extent $\theta_{\text{FRAME_V}/2}$ and the distance d between the front camera 48A and head 202 using the equation $\tan(\theta_{\text{FRAME_V}/2}) = (V/2)/d$. Solving for V , the vertical head height measure V may be computed as $V = 2d \times \tan(\theta_{\text{FRAME_V}/2})$.

[0049] FIG. 6 illustrates of an example horizontal dimension determination 206 for a human head 208. In some examples, a front camera 48A is positioned to capture an image of the human head 208 so that a horizontal head width measure H that can be calculated based on an angular extent angle $\theta_{\text{FRAME_H}/2}$ of one half of the horizontal head width $H/2$ and the distance d between the front camera 48A and human head 208. As depicted, the human head 208 has a head width H which corresponds to the horizontal dimension of a head bounding box. From the vantage of the front camera 48A, the horizontal head width H makes an angle $\theta_{\text{FRAME_H}}$ extending from the sides of the head 208. Upon bisecting the angle $\theta_{\text{FRAME_H}}$, the upper half of the horizontal head width $H/2$ makes an angle $\theta_{\text{FRAME_H}/2}$ with the camera's focal point (or the line of sight that is straight at 90-degrees from the camera focal point). As a result, the horizontal head width measure H can be calculated based on an angular extent $\theta_{\text{FRAME_H}/2}$ and the distance d between the front camera 48A and the head 208 using the equation $\tan(\theta_{\text{FRAME_H}/2}) = (H/2)/d$. Solving for H , the horizontal head width measure H may be computed as $H = 2d \times \tan(\text{FRAME_H}/2)$. As the human head is moved laterally or sideways from the centerline 56 of the camera focal point at a pan angle Φ_{PAN} , the perspective projection causes the head to appear to be smaller than its original vertical and horizontal head measures V , H .

[0050] FIG. 7 illustrates a front camera 48A and a two-dimensional image plane 210. In some examples, a front camera 48A is used to provide an image of a meeting participant located in a first, centered position 212 and a second, panned position 214 that is shifted laterally in the X_{ROOM} direction. In the first, centered position, the meeting participant is located along the

centerline 56 of the camera (e.g., $\Phi_{\text{PAN}} = 0$) at $d_0 = Y$ meters, so the two-dimensional room distance parameters for the first, centered position 212 are $\{x_{\text{ROOM}} = 0, y_{\text{ROOM}} = Y\}$. In the second, panned position, the meeting participant is shifted laterally in the x_{ROOM} direction by a panned angle Φ_{PAN} and is located at $d_1 > d_0$ meters, so the two-dimensional room distance parameters for the second, panned position 214 are $\{x_{\text{ROOM}} = P, y_{\text{ROOM}} = Y\}$. Further, the same vertical head height measure $V/2$ for the meeting participant positions 212, 214 will result in an angular extent $\theta_{\text{FRAME_V1/2}}$ for the first meeting participant position 212 that is larger than the angular extent $\theta_{\text{FRAME_V2/2}}$ for the second meeting participant position 214. In effect, the fact that the second, panned position 214 is located further away from the front camera 48A than the first, centered position 212 ($d_1 > d_0$) results in the angular extent for the second, panned position 214 appearing to be smaller than the angular extent for the first, centered position 212 so that $\theta_{\text{FRAME_V1/2}} > \theta_{\text{FRAME_V2/2}}$.

[0051] From the foregoing, the issue is to find an angular extent for the entire head height θ_{HH} and then represent it as a percentage of the full frame vertical field of view (VFrame_Percentage) which is then translated into the number of pixels the head will occupy (VHead_Pixel_Count) at a particular distance and at a pan angle Φ_{PAN} . To this end, the angular extent for the entire head height θ_{HH1} for the first meeting participant location 212 may be calculated by starting with the equation, $\tan(\theta_{\text{HH1}}/2) = (V/2)/d_0$. Solving for the angular extent θ_1 , the angular extent for the entire head height θ_{HH1} may be calculated as $\theta_{\text{HH1}} = 2 \arctan((V/2)/d_0)$. In similar fashion, the angular extent for the entire head height θ_{HH2} for the second meeting participant location 214 located at the pan angle Φ_{PAN} may be calculated by starting with the equation, $\tan(\theta_{\text{HH2}}/2) = (V/2)/d_1$, where $d_1 = \sqrt{d_0^2 + P^2}$. Solving for the angular extent θ_{HH2} , the angular extent for the entire head height θ_{HH2} may be calculated as $\theta_{\text{HH2}} = 2 \times \arctan((V/2)/d_1) = 2 \times \arctan((V/2)/\sqrt{d_0^2 + P^2})$. Based on this computation, the percentage of the frame occupied by the head height for the second meeting participant location 214 can be computed as $\text{VFrame_Percentage} = \theta_{\text{HH2}}/\text{Vertical FOV}$. In addition, the corresponding number of pixels for the head height for the second meeting participant location 214 can be computed as $\text{VHead_Pixel_Count} = \text{VFrame_Percentage} \times \text{Vertical FOV}$ in pixels. Based on the foregoing calculations, the angular extent for the entire head height $\theta_{\text{HH}} = \theta_{\text{FRAME_V}}$ may be calculated at discrete distances of 0.5 meters in each of the x_{ROOM} and y_{ROOM} directions that are equivalent to various angular pan angles Φ_{PAN} which may be listed in a lookup table (not shown).

[0052] FIG. 8 illustrates a front camera 48A and a videoconference room 300 including a two-dimensional image plane 310 to illustrate how to calculate a vertical or depth room distance Y_{ROOM} (meters) to the meeting participant location from the distance measure X_{ROOM} (meters) by calculating a direct distance measure HYP between the front camera 48A and the meeting participant location. The two-dimensional image plane 310 includes a plurality of two-dimensional coordinate points 312, 314, 316 that are defined with image plane 310 coordinates $\{x_i, y_i\}$ as described above. In addition, a head bounding box 318 is defined with reference to the starting coordinate point $\{x_1, y_1\}$ for the head bounding box 318, a Width dimension (measured along the x_i axis), and a Height dimension (measured along the y_i axis). To locate the vertical or depth room distance Y_{ROOM} (meters) from the front camera 48A, a vertical angular extent (Θ) for the head bounding box 318 is computed as $\Theta = \text{Height} * V_FOV / V_PIXELS$, where Height is the height of the head bounding box in pixels, where V_FOV is the Vertical FOV in degrees, and where V_PIXELS is the Vertical FOV in Pixels. Next, a vertical angular extent for the upper half of the head bounding box is computed ($\Theta/2$) and used to derive the direct distance measure HYP between the front camera 48A and the meeting participant location, $HYP = V_HEAD / (2 * \tan(\Theta/2))$, where HYP is the direct distance measure to the meeting participant location at the pan angle Φ_{PAN} . Finally, the vertical or depth room distance Y_{ROOM} (meters) is derived from the direct distance measure HYP and the distance measure X_{ROOM} (meters) using Pythagorean's Theorem, $Y_{ROOM} = \sqrt{HYP^2 - X_{ROOM}^2}$.

[0053] With this understanding of the AI human head detector model, the present disclosure provides methods, devices, systems, and computer readable media to accurately detect and reidentify participants in a videoconferencing system using multiple cameras. The location of each meeting participant is determined by the AI human head detector model using room distance parameters, as discussed above. In particular, coordinates, *e.g.*, image and/or world coordinates, are determined for each participant in each camera view. In some aspects, the world coordinates identified by the AI human detector model are referred to as world coordinate points. Further, identification (ID) labels are assigned to each meeting participant detected by the AI human head detector model, and each identification label in a first view captured by a first camera is grouped or paired with each identification label in at least one second view captured by at least one second camera. In some aspects, the identification labels in the second image are paired with the identification labels in the first image based on a distance between the identification labels in the first and second images. However, in another example, the participant coordinates locations are transformed into 2D world coordinates and

projected back onto a coordinate system of one of the cameras, *e.g.*, the first or primary camera. By mapping the locations of the meeting participants from each camera view onto a single camera coordinate system, it becomes possible to identify the same participants in each camera view, thus preventing duplicate views of participants from being transmitted to a far end of the videoconference. In yet another example, a centroid for each camera view may be calculated using the image coordinates of the detected human heads, and the image coordinates are then transformed into polar coordinates. In some aspects, ID labels are assigned to each human head in each camera view based on their respective polar coordinates, *e.g.*, angles from the respective centroids. Subsequently, a reference human head may be determined for each camera view which is known to be the same participant in each camera view. The ID labels may be aligned across all camera views by re-ranking the ID labels in a counterclockwise order about each centroid, starting with the reference human head. Accordingly, it will be understood that multiple methods may be used to identify human heads across different camera views without departing from the scope of the present disclosure.

[0054] According to some aspects of the present disclosure, a multi-camera calibration system is provided that can identify and map participant coordinates across different camera views and coordinate systems. In some examples, a multi-camera calibration system includes a plurality of cameras, such as, *e.g.*, a primary camera and a secondary camera, which are used to capture images of a location. In some aspects, a location is a conference room, an enclosed room, an open concept workspace, and/or a portion of an open concept workspace. The primary camera and the secondary camera may each define a respective coordinate system. According to one aspect of the present disclosure, a point in one coordinate system, *e.g.*, the secondary camera coordinate system, can be mapped onto another coordinate system, *e.g.*, the primary camera coordinate system, by considering or evaluating scale, translation, and rotation factors between the two coordinate systems. For example, the below transformation equations may govern coordinate mapping between a primary coordinate system A of the primary camera and a secondary coordinate system B of the secondary camera:

$$X_a = O_x + S * \cos(\alpha) * X_b - S * \sin(\alpha) * Y_b$$

$$Y_a = O_y + S * \sin(\alpha) * X_b + S * \cos(\alpha) * Y_b$$

In the above equations, (X_a, Y_a) is a point represented in the primary coordinate system A, (X_b, Y_b) is the same point represented in the secondary coordinate system B, (O_x, O_y) is an origin of the secondary coordinate system B, α is an angle between axes of the two coordinate systems A and B, and S is a scaling factor between the two coordinate systems A and B.

[0055] Further, the multi-camera calibration system determines the coordinates (O_x , O_y) of each camera in the room relative to one another, as well as the corresponding offset angles α between each of the camera coordinate systems. Accordingly, the above equations can be simplified to arrive at the below simplified transformation equations:

$$X_a = a * X_b - b * Y_b + c$$

$$Y_a = b * X_b + a * Y_b + d$$

To determine the parameters a, b, c, d, in the above equations, the cameras may include a calibration setting to capture calibration points. In some aspects, calibration points are defined as world coordinates of a single participant, *i.e.*, a single human head, who is visible by each camera in the videoconference room. Further, the world coordinates of the participant may be determined using the AI head detector model as described above. During a calibration phase, each of the cameras can capture images of the single participant over a period of time to record a pre-determined number of calibration points. In some aspects, the calibration phase may require that at least 1500 frames of the participant be taken by each camera, or about 50 seconds of a video being taken at 30 frames-per-second (FPS). Thus, it will be understood that each calibration point corresponds to a frame for a camera, meaning that the total number of calibration points captured by a camera corresponds to the total number of frames captured during the calibration phase.

[0056] After the calibration points have been captured, a least square solver $AX=L$ as depicted below is used to solve the parameters a, b, c, d in the above simplified transformation equations:

$$\begin{bmatrix} X_{b1} & Y_{b1} & 1 & 0 \\ Y_{b1} & X_{b1} & 0 & 1 \\ X_{b2} & Y_{b2} & 1 & 0 \\ Y_{b2} & X_{b2} & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix} * \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = \begin{bmatrix} X_{a1} \\ Y_{a1} \\ X_{a2} \\ Y_{a2} \\ \vdots \end{bmatrix}$$

Thus, with reference to the above matrix equation, L in the least square solver $AX=L$ is defined as a vector of the calibration points captured by the primary camera, X is defined as the vector [a, b, c, d], and A is defined as a matrix of the calibration points captured by the secondary camera. Once the parameters a, b, c, d have been determined, α , S, and (O_x , O_y) can be calculated using the below relationships:

$$\alpha = \text{atan}\left(\frac{b}{a}\right)$$

$$S = \frac{a}{\cos(\alpha)}$$

$$O_x = c \text{ and } O_y = d.$$

[0057] After the calibration phase has been completed, the locations of the primary and secondary cameras are known to one another, meaning that the first set of equations above may be used to map any point in the secondary coordinate system onto the primary coordinate system, and vice versa. Thus, the world coordinates of each human head detected in the room using the AI head detector method as discussed above are all geometrically transformed, *i.e.*, projected, onto the primary coordinate system. Relatedly, each frame that is captured by the cameras is projected onto the primary coordinate system, and pairs of adjacent frames are further analyzed using a reidentification process. In the reidentification phase, Euclidean distances are measured between all points captured by pairs of adjacent frames, and points which have minimum Euclidean distances therebetween are subsequently grouped together in pairs or tuples. Put another way, points in adjacent frames are compared in the reidentification phase to determine which points are closest to each other. Any two points which are closer to one another than any other point in the frames, *i.e.*, points with a minimum Euclidean distance therebetween, are determined to correspond to the same participant. In this way, pairs of points with minimum Euclidean distances are assigned a common ID, thereby identifying each participant in each camera view.

[0058] Referring now to FIG. 9, a top view of a videoconference room 400 is plotted on a two-dimensional world plane 402. Specifically, the world plane 402 corresponds to the actual spatial location of objects and/or participants in the videoconference room 400. The world plane 402 defines a coordinate system 404 with world plane coordinates $\{x_i, y_i\}$ as described above, with x_i denoting the x-axis of the world plane 402 and y_i denoting the y-axis of the world plane 402. As such, it will be understood that a right side 406 of the room 400 is indicated by negative y_i coordinates, and a left side 408 of the room 400 is indicated by positive y_i coordinates. Put another way, the room 400 is bisected by the x_i axis. In a first example, the videoconference room 400 includes a first camera Cam_0 , a second camera Cam_1 , and a third camera Cam_2 . In some aspects, the first camera Cam_0 is located at the origin O of the coordinate system 404, the second camera Cam_1 is located on the right side 406 of the room 400, and the third camera Cam_3 is located on the left side 408 of the room 400. Each of the cameras Cam_0 , Cam_1 , Cam_2 is configured to capture a view of the room 400. Specifically, the first camera Cam_0 defines a first FOV 410 which is directed along the x_i axis, the second camera Cam_1 defines a second FOV 412 which is angled toward the x_i axis from the right side 406, and the

third camera Cam₂ defines a third FOV 414 which is angled toward the x₁ axis from the left side 408. In the illustrated example, the first and third cameras Cam₀, Cam₂ are located within the second FOV 412, and the first and second cameras Cam₀, Cam₁ are located within the third FOV 414. In some aspects, the FOVs 410, 412, 414 overlap with one another, and the above arrangement of the cameras Cam₀, Cam₁, and Cam₂ is referred to as an outside-in arrangement. In some aspects, the area enclosed by each of the FOVs 410, 412, 414 is defined as an intersection area 416. As illustrated, participants A, B, C, D, E, F, G are located within each of the FOVs 410, 412, 414, meaning that each of the participants are visible in each view captured by the cameras Cam₀, Cam₁, and Cam₂. In some aspects, the first camera Cam₀ includes a codec with a processing unit (see FIG. 28) which can maintain videoconferencing calls or events (e.g., streaming video to far endpoints) and further apply the AI head detection model as discussed above to the views captured by the cameras Cam₀, Cam₁, and Cam₂. Correspondingly, the AI head detection model determines 2D world coordinates for each of the participants relative to the coordinate systems of each of the cameras Cam₀, Cam₁, and Cam₂. For example, participant A has world coordinates of (3, 2), participant B has world coordinates of (5, 2), participant C has world coordinates of (6.6, 4), *etc.*

[0059] In some aspects, according to a calibration phase, the outside-in calibration system first determines the locations, *i.e.*, world coordinates, of the second and third cameras Cam₁, Cam₂ relative to the first camera Cam₀ in the first camera Cam₀ coordinate system. Specifically, the transformation equations discussed above are applied to the images captured by each camera Cam₀, Cam₁, Cam₂ and using a single participant, *e.g.*, participant A, to create calibration points as discussed above. In this way, the coordinates of the second and third cameras Cam₁ and Cam₂ can be mapped to the first camera Cam₀ coordinate system using the below equations which are derived from the previously discussed transformation equations:

$$\begin{aligned} X_{01} &= O_{x1} - Y_1 \\ Y_{01} &= O_{y1} + X_1 \\ X_{02} &= O_{x2} + Y_2 \\ Y_{02} &= O_{y2} - X_2 \end{aligned}$$

In the above equations, X_{01}, Y_{01} denote coordinates that have been mapped from the second camera Cam₁ coordinate system to the first camera Cam₀ coordinate system, and X_{02}, Y_{02} denote coordinates that have been mapped from the third camera Cam₂ coordinate system to the first camera Cam₀ coordinate system. Correspondingly, $O_{x1}, O_{y1}, O_{x2}, O_{y2}$ are the coordinates of the second and third cameras Cam₁, Cam₂ when viewed from the first camera

Cam₀. After the calibration points have been mapped onto the first camera Cam₀ coordinate system, the least square solver discussed above is used to determine each of the parameters in the simplified transformation equations, which in turn are used to determine the angle α , scaling factor S, and origin coordinates O_{x1} , O_{y1} , O_{x2} , O_{y2} of the second and third cameras Cam₁, Cam₂. Once these parameters are known, the calibration phase is completed.

[0060] After the calibration phase, the cameras enter a reidentification phase which remains active during the videoconference. During the re-identification phase, the AI head detector model as discussed above is applied to the images captured by the cameras Cam₀, Cam₁, and Cam₂ to determine world coordinates for each participant detected in each image. The world coordinates of the participants in each view are mapped onto the primary coordinate system of the first camera Cam₀ using the transformation equations discussed above. This in turn results in clusters of points forming around the coordinates identified by the first camera Cam₀. Euclidean distances are measured between each of the points in the clusters, *i.e.*, points corresponding to participant location in the room 400, and pairs or tuples of points with minimum Euclidean distances are assigned a common ID, thus identifying the pairs or tuples of points as a single meeting participant. In this way, individual participants are reidentified across each captured camera view.

[0061] Referring now to FIG. 10, another top view is illustrated of the videoconference room 400. In this example, the videoconference room 400 includes a primary or front camera Cam_f and a secondary or center camera Cam_c. In some aspects, the front camera Cam_f is located at the origin O of the coordinate system 404, and the center camera Cam_c is located at the center of the room 400 along the x_1 axis. For example, the center camera Cam_c has world coordinates of (6,0). In some aspects, the arrangement of the front camera Cam_f and the center camera Cam_c is referred to as an inside-out arrangement. The front camera Cam_f defines the first FOV 410 which is directed along the x_1 axis, and the center camera Cam_c can be a 360-degree camera, meaning that the center camera Cam_c defines a fourth, circular FOV 418 (as opposed to the fan-shaped FOVs 410-414) which extends circumferentially therearound, *e.g.*, around a perimeter of the videoconference room 400 and/or a portion of the videoconference room 400. Accordingly, the intersection area 416 in the inside-out arrangement illustrated in FIG. 10 is defined entirely by the first FOV 410 of the front camera Cam_f. In some aspects, the multi-camera calibration system first requires that the coordinates of the center camera Cam_c be known to the front camera Cam_f, *i.e.*, the primary camera, before the participants can be reidentified in the images captured by the cameras Cam_f, Cam_c. Because the center camera

Cam_c defines a circular FOV 418, the transformation equations discussed above can be used to convert the world cartesian coordinates into world polar coordinates with respect to the center camera Cam_c. For example, the above transformation equations may be converted into polar form as shown below:

[0062] $X_{fc} = O_{xc} + r * \cos\theta_{fc} = O_{yc} + r * \sin\theta_{fc}$ polar transformation equations, X_{fc}, Y_{fc} denote coordinates that have been mapped from the center camera Cam_c to the front camera Cam_f, and O_{xc}, O_{yc} denote the coordinates of the center camera Cam_c when viewed from the front camera Cam_f. Similar to the outside-in arrangement discussed above, a least square solver and calibration points are used to solve for the parameters during the calibration phase, and the AI head detector model as discussed above is applied to the images captured by the cameras Cam_f, Cam_c to determine world coordinates for each image. The cartesian world coordinates are converted to polar world coordinates and input to the above polar transformation equations to map all participant location coordinates onto the front camera Cam_f coordinate system. In this way, point clusters are generated. During the reidentification phase, Euclidean distances are measured between each of the points in the clusters, *i.e.*, points corresponding to participant location in the room 400, and pairs or tuples of points with minimum Euclidean distances are assigned a common ID, thus identifying the pairs or tuples of points as a single meeting participant. Accordingly, individual participants are reidentified across each captured camera view. Thus, as shown above, the multi-camera calibration system of some aspects may be applied to a variety of different camera setups, meaning that the multi-camera calibration system is scalable and may include fewer or additional cameras than those illustrated in FIGS. 9 and 10.

[0063] In light of the above, FIG. 11 illustrates a method 500 of using a multi-camera calibration system, as discussed above, in a conference room. At step 502, calibration frames are captured using a primary camera and a secondary camera (or cameras). As discussed above, a calibration frame contains a calibration point corresponding to a single participant, and the single participant may move around the conference room as the cameras capture the calibration frames. At step 504, the world coordinates of the secondary camera(s) are determined relative to the primary camera, *i.e.*, relative to the view captured by the primary camera. In some aspects, the transformation equations discussed above are used to determine the world coordinates of the secondary camera(s). Once the world coordinates of the secondary camera(s) are known, the multi-camera calibration system exits the calibration phase (which may contain steps 502 and 504) and enters a normal operation phase in which human heads are

detected using an AI head detection model, as shown at step 506. For example, the AI head detection model is applied to images captured by each of the cameras in order to design, for each detected human head, a head bounding box with specified room coordinates and dimension information, which are then used to calculate horizontal pan distance and depth dimension distance measured from the camera from a top-down perspective of the room. In this way, a two-dimensional world coordinate location for each detected human head is determined.

[0064] At step 508, the multi-camera calibration system enters a reidentification phase, and the world coordinates of the detected human heads in each captured view are projected, *i.e.*, mapped, onto the coordinate system of the primary camera. This in turn generates clusters around the points corresponding to heads detected in images captured by the primary camera. At step 510, Euclidean distances between all points are measured. At step 512, points with minimum Euclidean distances therebetween are clustered, *i.e.*, ordered in pairs and/or tuples, and assigned a shared ID. As such, points with the same ID are determined to correspond to the same participant, and this information is relayed to the multi-camera calibration system. In some aspects, the method 500 then returns to step 506 and proceeds as discussed above. Put another way, the multi-calibration system returns to a normal operation phase and repeats steps 506, 508, 510, 512 for the remainder of the conference. In this way, conference participants are constantly reidentified by the multi-calibration system, which in turn eliminates duplicate views from being transmitted to a far end of the conference. That is, when all participants are displayed at the far end of the conference, for example, in a gallery view, using the above method 500, each participant is only transmitted once despite appearing in multiple FOVs. Generally, the method 500 can be performed in real-time or near real-time. For example, in some aspects, the multi-calibration system enters the reidentification mode after a period of time has elapsed, such as, *e.g.*, at least every 30 seconds, or at least every 15 seconds, or at least every 10 seconds, or at least every 5 seconds, or at least every 3 seconds, or at least every second, or at least every 0.5 seconds.

[0065] It should be noted that any of the cameras used in systems and methods described here may have machine-learning computing capability such that they can run machine learning models, or such a computation can be offloaded to a primary camera or a separate, centralized codec. Accordingly, steps of the method 500 described above (or any of the other methods described herein), may be machine readable instructions carried out by a processor or coupled to the camera(s) of the system and/or a separate codec.

[0066] According to another aspect of the present disclosure, a centroid-based identification system is provided to identify and track participants across multiple camera views based on their relative position around a conference table. In some examples, a conference room includes a plurality of cameras, such as, *e.g.*, a primary camera and a secondary camera. The primary camera and the secondary camera may capture images of a conference room and, more specifically, participants seated around a conference table in conference room. Images that are captured by the cameras may define image planes that are represented by pixel coordinate systems. In some aspects, an AI head detection model is applied to the centroid-based identification system to create bounding boxes around heads of the participants, and a centroid of all of the bounding boxes in a single image plane is determined and displayed in the corresponding pixel coordinate system. In each image, the pixel coordinates of each bounding box are transformed to polar coordinates and ranked in counterclockwise order starting with the bounding box with the lowest polar angle to the centroid relative to the other polar angles in the image. Put another way, the bounding boxes are ranked in counterclockwise order starting with the bounding box with a minimum centroid angle. A reference person is chosen in each image, and the rankings for each image are rearranged starting with the reference person in each image. The resulting rankings are stored in the system and/or appended to the original rankings. In this way, the rankings for each view are aligned with one another, and ID labels are assigned to each participant. Correspondingly, the ID labels are the same across all camera views, meaning that each participant is identified across all camera views.

[0067] FIGS. 12A-24 illustrate an example of the centroid-based identification system as discussed above. In particular, FIGS. 12A and 12B illustrate a conference room 600 which is defined by a front wall 602, a left wall 604, and a right wall 606. A table 608 is located in the center of the conference room 600, and six participants 610 are located within the conference room 600. For example, a first participant 610A, a second participant 610B, a third participant 610C, a fourth participant 610D, a fifth participant 610E, and a sixth participant 610F are located in the conference room 600. Specifically, the first and second participants 610A, 610B are seated along a left side 612 of the table 608, the third and fourth participants 610C, 610D are seated along a rear side 614 of the table 608, and the fifth and sixth participants 610E, 610F are seated along a right side 616 of the table 608. In addition, the centroid-based identification system includes a first camera 618, a second camera 620, and a third camera 622. In some aspects, the first camera 618 is coupled to a monitor 624, *e.g.*, fastened on top of the monitor 624, and the monitor 624 is located at the front of the conference room 600 adjacent a front

side 626 of the table 608. In some aspects, the second and third cameras 620, 622 are also coupled to monitors (not shown).

[0068] Referring specifically now to FIG. 12A, a first example arrangement of the cameras 618, 620, 622 can include mounting the first camera 618 towards the front, *e.g.*, directly in front of the front wall 602 and/or on the front wall 602, the second camera 620 to the left wall 604, and the third camera 622 to the right wall 606. In the illustrated example, each of the cameras 618, 620, 622 is angled towards the table 608 in order to view the participants 610. Put another way, each of the participants 610 are within a FOV of the first camera 618, a FOV of the second camera 620, and/or a FOV of the third camera 622. However, it is contemplated that the cameras 618, 620, 622 can be arranged in various other configurations without departing from the scope of the present disclosure. To that end, FIG. 12B illustrates a second example arrangement of the cameras 618, 620, 622 in which the first camera 618 is mounted to a top of the monitor 624 and faces a rear wall (not shown) of the conference room 600. In addition, the second camera 620 can be mounted to a left side of the monitor 624 and angled diagonally toward the rear wall (not shown) and the right wall 604 (see FIG. 12A) so as to focus on the right side 616 of the table 608. In a similar way, the third camera 622 can be mounted to a right side of the monitor 624 and angled diagonally toward the rear wall (not shown) and left wall 606 (see FIG. 12A) so as to focus on the left side 614 of the table 608. Thus, it will be understood that the cameras 618, 620, 622 may be arranged in a variety of different positions including those illustrated in FIGS. 12A and 12B. In some aspects, the cameras 618, 620, 622 are configured to rotate to focus on an active speaker, or the cameras are fixed in their respective positions and do not rotate.

[0069] Referring now to FIG. 13, a front image 628, *i.e.*, an image of the conference room 600 captured by the first camera 618, is overlaid on a first pixel coordinate system 630, which may be a cartesian coordinate system. As illustrated, the table 608, the participants 610, the second camera 620, and the third camera 622, are all visible in the front image 628, *i.e.*, within the FOV of the first camera 618. In some aspects, the arrangement of the cameras 618, 620, 622 is substantially similar to the arrangement illustrated in FIG. 12A. After the front image 628 is captured, the AI head detector model, as discussed above, can be applied to the image, which generates head bounding boxes 640A, 640B, 640C, 640D, 640E, 640F for the participants 610A, 610B, 610C, 610D, 610E, 610F, respectively, and plots the bounding boxes 640 on the first pixel coordinate system 630. Once the pixel coordinates are known for each bounding box 640, the center of each bounding box is calculated and stored in the centroid-based identification system. In some aspects, the center of each bounding box 640 is used to

describe the location of each participant 610 using pixel coordinates, as will be discussed below in greater detail. A first centroid 642 is calculated using the centers of the bounding boxes 640 using the below formulas:

$$x_c = \frac{\sum_{k=0}^n x_k}{n} \text{ and } y_c = \frac{\sum_{k=0}^n y_k}{n}$$

In the above formulas, $(x_k, y_k), k = 0, 1, \dots, n$ corresponds to the pixel coordinates of the center of each bounding box 640, and (x_c, y_c) denotes the pixel coordinates of the first centroid 642.

[0070] Once the first centroid 642 has been determined, ID labels ID_N, *e.g.*, ID_0, ID_1, ID_2, ID_3, ID_4, ID_5, are assigned to each of the participants 610, respectively. For example, the ID labels may be assigned arbitrarily to each participant 610 by the AI head detector model, or the ID labels may be assigned to each participant 610 based on a predetermined order. In one example, the ID labels may be assigned to each participant 610 based on the order in which the bounding boxes 640 are calculated and generated, such as a clockwise order starting with the participant with the largest sum of (x_k, y_k) pixel coordinate values, *e.g.*, the sixth participant 610F, or another method. Thus, it will be understood that a variety of methods may be used to assign the ID labels to the participant 610. Once the ID labels are assigned to the participants 610, the centroid-based identification system measures the distance of each bounding box 640 center to the first centroid 642 and subsequently transforms the first pixel coordinate system 630, including the cartesian bounding box 640 center coordinates and the first centroid 642 coordinates, into polar coordinates. Specifically, the pixel coordinates are transformed into polar coordinates such that the first centroid 642 is the polar origin O by using the below equations:

$$x'_k = x_k - x_c, y'_k = y_k - y_c$$

$$\rho = \sqrt{x'^2_k + y'^2_k}, \varphi = \frac{180^\circ}{\pi} \tan^{-1} \left(\frac{y'_k}{x'_k} \right)$$

[0071] In the above equations, $[(x'_k, y'_k)]$ correspond to updated pixel coordinates of the center of each bounding box 640 when using the pixel coordinates (x_c, y_c) as the origin O. Further, ρ corresponds to the distance between each bounding box 610 center and the origin O, *i.e.*, the first centroid 642, and φ corresponds to the angle between each bounding box 640 center and the origin O. Thus, it will be understood that (ρ, φ) denotes the new polar coordinates of each bounding box 610 center relative to the first centroid 642. In some aspects, the angle φ between each bounding box 610 center and the origin O may be referred to as a

centroid angle. In some examples, all of the centroid angles φ are made to be positive, which may be accomplished using the following conditional equation: $\varphi = \varphi + 360^\circ$ if $\varphi < 0^\circ$. Put another way, 360 degrees is added to a centroid angle φ if the centroid angle φ is negative.

[0072] FIG. 14 illustrates a 2-D representation of the locations of the participants 610 relative the first centroid 642 using a polar coordinate system 644. As discussed above, polar coordinates (ρ, φ) are determined for each of the participants 610 relative to the first centroid 642. For example, the fourth participant 610E may have polar coordinates (ρ_E, φ_E) , as illustrated in FIG. 14. Once the polar coordinates (ρ, φ) for each bounding box 640 center have been determined, the centroid-based identification system determines which bounding box 640 center has the minimum centroid angle φ . In the non-limiting example illustrated in FIGS. 13 and 14, the fifth participant 610E has the minimum centroid angle φ and is subsequently identified as such. The centroid-based identification then ranks each of the participants 610 in a counterclockwise order about the first centroid 642, beginning with the participant 610 with the minimum centroid angle φ , *e.g.*, the fifth participant 610E. For example, the participants 610 are ranked in the following order: fifth participant 610E, fourth participant 610D, third participant 610C, second participant 610B, first participant 610A, sixth participant 610F. This ranking is stored in the centroid-based identification system and corresponds to the front image 628. However, it is contemplated that the participants 610 may be ranked using another regiment, such as, *e.g.*, clockwise about the first centroid 642. In addition, it is contemplated that a different participant 610 may be used instead of the participant with the minimum centroid angle φ , such as, *e.g.*, the participant with the largest centroid angle φ . While a variety of different techniques may be used to rank the participants, it will be understood that the technique used to rank participants is consistent across all camera views, *e.g.*, the views captured by the first camera 618 as well as the second and third cameras 620, 622, to ensure the accurate identification and reidentification of participants, as will be discussed below in greater detail.

[0073] Referring now to FIG. 15, a left image 648, *i.e.*, an image of the conference room 600 captured by the second camera 620, is overlaid on a second pixel coordinate system 650. As performed for the front image 628 (see FIG. 13), the centroid-based identification system applies the AI head detector model to the left image 648, which in turn generates the head bounding boxes 640 for the participants 610 and calculates a second centroid 652 based on the centers of the bounding boxes 640. In addition, ID labels ID_N are applied to each of the participants 610 arbitrarily or using a specific regimen, as discussed above. In some aspects,

the ID labels are applied to the participants 610 in a different order than as illustrated in FIG. 13. The second pixel coordinate system 650 is then converted to a polar coordinate (ρ, φ) system 654 using the second centroid 652 as the origin O, as illustrated in FIG. 16. The centroid-based identification system determines which bounding box 640 center has the minimum centroid angle φ . In the non-limiting example illustrated in FIGS. 15 and 16, the sixth participant 610F has the minimum centroid angle φ and is subsequently identified as such. The centroid-based identification then ranks each of the participants 610 in a counterclockwise order about the second centroid 652, beginning with the participant 610 with the minimum centroid angle φ , *e.g.*, the sixth participant 610F. For example, the participants 610 are ranked in the following order: sixth participant 610F, fifth participant 610E, fourth participant 610D, third participant 610C, second participant 610B, first participant 610A. Thus, it will be understood that the participants 610 are ranked using the same regimen in both the front image 628 and the left image 648, but the ranked order of the participants 610 is different between the front image 628 and the left image 648 due to the different camera perspectives of the conference room 600.

[0074] Referring now to FIG. 17, a right image 658, *i.e.*, an image of the conference room 600 captured by the third camera 622, is overlaid on a third pixel coordinate system 660. As performed for the front and left images 628, 648 (see FIGS. 13 and 15), the centroid-based identification system applies the AI head detector model to the right image 658, which in turn generates the head bounding boxes 640 for the participants 610 and calculates a third centroid 662 based on the centers of the bounding boxes 640. In addition, ID labels ID_N are applied to each of the participants 610 arbitrarily or using a specific regimen as discussed above. In some aspects, the ID labels are applied to the participants 610 in a different order than as illustrated in FIG. 13. The third pixel coordinate system 660 is then converted to a polar coordinate (ρ, φ) system 664 using the third centroid 662 as the origin O, as illustrated in FIG. 18. The centroid-based identification system determines which bounding box 640 center has the minimum centroid angle φ . In the non-limiting example illustrated in FIGS. 17 and 18, the third participant 610C has the minimum centroid angle φ and is subsequently identified as such. The centroid-based identification then ranks each of the participants 610 in a counterclockwise order about the second centroid 652, beginning with the participant 610 with the minimum centroid angle φ , *e.g.*, the third participant 610C. For example, the participants 610 are ranked in the following order: third participant 610C, second participant 610B, first participant 610A, sixth participant 610F, fifth participant 610E, fourth participant 610D. In some aspects, the

participants 610 are ranked using the same regimen in the front, left, and right images 628, 648, 658, but the ranked order of the participants 610 is different between the images 628, 648, 658 due to the different camera perspectives of the conference room 600.

[0075] Thus, it will be understood that the centroid-based identification system identifies and stores rankings of the participants 610 for each of the images 628, 648, 658. In some aspects, the centroid-based identification system identifies a reference person in the images 628, 648, 658 and rearranges the rankings based thereon. To that end, FIG. 19 illustrates the front image 628 of the conference room 600 captured by the first camera 618, FIG. 20 illustrates the left image 648 of the conference room 600 captured by the second camera 620, and FIG. 21 illustrates the right image 658 of the conference room 600 captured by the third camera 622. Referring now to FIGS. 19-21, the centroid-based identification system identifies a reference person using the known pixel coordinates of the participants 610. It is contemplated that the specific method of determining a reference person may depend upon the geometry of the room, meaning that the method of determining a reference person may be modified to best suit a particular conference room. In the non-limiting example illustrated in FIGS. 19-21, the pixel coordinate $\{x_i, y_i\}$ locations of the cameras 618, 620, 622 are known to one another. By manipulating the pixel coordinate $\{x_i, y_i\}$ values of the participants 610, it becomes possible to identify the same participant across each of the images 628, 648, 658. In some aspects, the centroid-based identification system includes applying the equations in Table 1 to identify a reference person in the images 628, 648, 658:

Table 1

Front view – bottom right	$\operatorname{argmax}(x_i + y_i)$
Left view – top right	$\operatorname{argmin}(x_i - y_i)$
Right view – bottom left	$\operatorname{argmax}(x_i - y_i)$

In the above table, $\operatorname{argmax}(x_i + y_i)$ is used to identify the participant who is farthest toward a bottom-right corner 664 of the front image 628, $\operatorname{argmin}(x_i - y_i)$ is used to identify the participant who is farthest toward a top-right corner 668 of the left image 648, and $\operatorname{argmax}(x_i - y_i)$ is used to identify the participant who is farthest toward a bottom-left corner 670 of the right image 658. As a result, the centroid-based identification system identifies the same participant, *e.g.*, the sixth participant 610F, as the reference person. However, it will be readily understood that a different participant may be identified by adjusting the equations in the above table. For example, the participant in a bottom-left corner 672 of the front image 628,

a bottom-right corner 674 of the left image 648, and a top-left corner 676 of the right image 658, *e.g.*, the first participant 610A, may be identified as a reference person 680.

[0076] Once the reference person 680 has been identified, the centroid-based identification system re-arranges the ID labels ID_N in each image, *i.e.*, re-ranks the participants 610. Specifically, the identification system re-ranks the participants 610 by starting with the reference person 680, *i.e.*, the sixth participant 610F, and proceeds counterclockwise about the centroids 642, 652, 662 for each of the images 628, 648, 658. For example, the identification system appends the new rankings to the original ID label rankings, or the identification system stores the new rankings as separate rankings, as depicted below in Table 2:

Table 2

	Front Image 628	Left Image 648	Right Image 658
Original Participant 610 Ranking	E→D→C→B→A→F	F→E→D→C→B→A	C→B→A→F→E→D
Original ID Label Ranking	ID_4→ID_3→ID_2→ID_1→ID_0→ID_5	ID_4→ID_5→ID_1→ID_0→ID_3→ID_2	ID_0→ID_3→ID_5→ID_1→ID_4→ID_2
New ID Label Ranking	ID_5→ID_4→ID_3→ID_2→ID_1→ID_0	ID_4→ID_5→ID_1→ID_0→ID_3→ID_2	ID_1→ID_4→ID_2→ID_0→ID_3→ID_5

[0077] Thus, it will be understood that re-ranking the ID labels ID_N in a clockwise order starting with the common reference person 680 aligns the ID labels ID_N across the images 628, 648, 658. Put another way, the step of re-ranking the ID labels ID_N based on the reference person 680 allows the centroid-based identification system to reidentify the participants 610 across each of the images 628, 648, 658. Thus, once the rankings are aligned, the centroid-based identification system correlates the rankings with one another. For example: the sixth participant 610F is known to correspond to ID_5 in the front image 628, ID_4 in the left image 648, and ID_1 in the right image 658; the fifth participant 610E is known to correspond to ID_4 in the front image 628, ID_5 in the left image 648, and ID_4 in the right image 658; the fourth participant 610D is known to correspond to ID_3 in the front image 628, ID_1 in the left image 648, and ID_2 in the right image 658; *etc.*

[0078] In some aspects, the centroid-based identification system reassigns the ID labels ID_N in the images 648, 658 taken by the left and right cameras 620, 622, to match the ID labels ID_N in the front image 628 taken by the front camera 618, *e.g.*, the primary camera. In other examples, the centroid-based identification system creates reidentification labels ReID_N for the left and right images 648, 658, *e.g.*, the secondary images, which are identical to the original ID labels ID_N in the front image 628, *e.g.*, the primary image. FIGS. 22-24 illustrate the images 628, 648, 658, respectively, with reidentification labels ReID_N aligned with the original ID labels ID_N in the front image 628. In the left and right images 648, 658 of FIGS. 23 and 24, respectively, the first participant 610A is assigned reidentification ID label ReID_0, the second participant 610B is assigned reidentification ID label ReID_1, the third participant 610C is assigned reidentification ID label ReID_2, *etc.* In this way, the centroid-based identification system identifies and re-identifies each of the participants 610 in each of the images 628, 648, 658.

[0079] Therefore, the centroid-based identification system disclosed herein is capable of reidentifying videoconference participants across different camera views. Correspondingly, the centroid-based identification system prevents multiple views of the same participant from being transmitted to a far end of a videoconference, which in turn may reduce confusion in the videoconference. In some aspects, the centroid-based identification system disclosed herein is particularly advantageous in crowded conference rooms and/or when there is little distance between participants in a conference room. Further, it is contemplated that FIGS. 12A-24 illustrate non-limiting examples of the centroid-based identification system, and that the centroid-based identification system may be applied to a variety of different conference rooms and is compatible with a variety of different camera arrangements.

[0080] FIG. 25 illustrates a method 700 of implementing the centroid-based identification system discussed above. At step 702 images of a location are captured using a primary camera and a secondary camera (or cameras). As discussed above, the primary camera can be arranged at a front of a conference room, and the secondary camera(s) may be arranged on left and right sides of a conference room, respectively. In some aspects, a primary camera is a camera that is in communication with and/or connected to a monitor and/or a codec that includes a memory and a processor, as will be discussed below in greater detail. At step 704, human heads in the images are detected using an AI head detection model, as described above. For example, the AI head detection model is applied to images captured by each of the cameras in order to identify, for each detected human head, a head bounding box with specified room and/or pixel coordinates. The AI head detection model also determines pixel coordinates for the center of

each bounding box, as will be discussed below in greater detail. At step 706, a centroid is determined based on pixel coordinates of the bounding boxes, or, more specifically, based on the pixel coordinates of the center of each bounding box. At step 708, the pixel coordinates are transformed to polar coordinates and ID labels are assigned to each bounding box. In particular, the pixel coordinates of each bounding box are transformed to polar coordinates using the centroid as the origin. At step 710, the ID labels are ranked in counterclockwise order starting with the ID label associated with the bounding box that has a minimum polar angle with respect to the centroid. In some aspects, this ranking is stored in the centroid-based identification system. At step 712, a reference human head, *i.e.*, a reference bounding box, is identified based on the particular pixel coordinates thereof. In some aspects, the reference human head corresponds to the same participant in each image captured by the primary and secondary cameras. At step 714, the ID labels are rearranged in counterclockwise order starting with the reference human head, thereby aligning the ID labels across the images captured by the primary and secondary cameras. In this way, the centroid-based identification system correlates the ID labels across the images captured by the primary and secondary cameras, which in turn allows the system to track participants across the images. In some aspects, the centroid-based identification system repeats each step in the method 700 during normal operation, meaning that the centroid-based identification reidentifies participants continuously as the primary and secondary cameras capture images of the location. Generally, the method 500 can be performed in real-time or near real-time. For example, in some aspects, the steps 702, 704, 706, 708, 710, 712, 714 of the method 700 are repeated after a period of time has elapsed, such as, *e.g.*, at least every 30 seconds, or at least every 15 seconds, or at least every 10 seconds, or at least every 5 seconds, or at least every 3 seconds, or at least every second, or at least every 0.5 seconds.

[0081] FIG. 26 illustrates an example camera 848, which may be similar to front cameras 48A, Cam₀, Cam_f, 618 (see FIGS. 1, 10, 11, 12), and an example microphone array 850, similar to microphone array 50 (see FIG. 1). The camera 848 has a housing 852 with a lens 854 provided in the center to operate with an imager 856. A series of openings 858, such as five openings 858, are provided as ports to microphones in the microphone array 850. In some examples, the microphone openings 858 form a horizontal line 860 to provide a desired angular determination for the SSL process, as discussed above. FIG. 26 is an example illustration of a camera 848, though numerous other configurations are possible, with varying lens and microphone configurations. Additionally, in some examples, aspects of the technology, including computerized implementations of methods according to the technology, can be

implemented as a system, method, apparatus, or article of manufacture using standard programming or engineering techniques to produce software, firmware, hardware, machine readable instructions, or any combination thereof to control a processor device (*e.g.*, a serial or parallel general purpose or specialized processor chip, a single- or multi-core chip, a microprocessor, a field programmable gate array, any variety of combinations of a control unit, arithmetic logic unit, and processor register, and so on), a computer (*e.g.*, a processor device operatively coupled to a memory), or another electronically operated controller to implement aspects detailed herein. Accordingly, for example, the technology can be implemented as a set of instructions, tangibly embodied on a non-transitory computer-readable media, such that a processor device can implement the instructions based upon reading the instructions from the computer-readable media. Some examples of the technology can include (or utilize) a control device such as, *e.g.*, an automation device, a special purpose or general-purpose computer including various computer hardware, software, firmware, and so on, consistent with the discussion below. As specific examples, a control device can include a processor, a microcontroller, a field-programmable gate array, a programmable logic controller, logic gates *etc.*, and other suitable components for implementation of appropriate functionality (*e.g.*, memory, communication systems, power sources, user interfaces and other inputs, *etc.*).

[0082] The above description assumes that the axes of front camera 48A and the microphone array 50 (see FIG. 1) are collocated. If the axes are displaced, the displacement is used in translating the determined sound angle from the microphone array 850 to the camera frames of reference.

[0083] As described above, the methods of some aspects include detecting a location of individual meeting participants using an AI human head detector model. Referring now to FIG. 27, an example process 900 is illustrated for determining coordinates for a detected human head using such an AI human head detector process. The AI human head detector process analyzes incoming room-view video frame images 902 of a meeting room scene with a machine-learning, AI human head detector model 904 to detect and display human heads with corresponding head bounding boxes 906, 908, 910. In some aspects, the AI human head detector process further identifies and displays centers 912 of the head bounding boxes 906, 908, 910. As depicted, each incoming room-view video frame image 902 may be captured by a front camera 48A in the video conferencing system. For example, a first view of the meeting participants is captured by a first camera (not shown) in a first profile image or video frame 902a, a second camera (not shown) captures a second profile image or video frame 902b, and a third camera (not shown) captures a third profile image or video frame 902c. Each incoming

room-view video frame image 902 may be processed with an on-device AI human head detector model 904 that may be located at the respective camera which captures the video frame images. However, in other examples, the AI human head detector model 904 may be located at a remote or centralized location, or at only a single camera. Wherever located, the AI human head detector model 904 may include a plurality of processing modules 914, 916, 918, 920 which implement a machine learning model which is trained to detect or classify human heads from the incoming video frame images, and to identify, for each detected human head, a head bounding box with specified image plane coordinate and dimension information.

[0084] In this example, the AI human head detector model 904 may include a first pre-processing module 914 that applies image pre-processing (such as color conversion, image scaling, image enhancement, image resizing, *etc.*) so that the input video frame image is prepared for subsequent AI processing. In addition, a second module 916 may include training data parameters and/or model architecture definitions which may be pre-defined and used to train and define the human head detection model 904 to accurately detect or classify human heads from the incoming video frame images. In selected examples, a human head detection model module 918 may be implemented as a model inference software or machine learning model, such as a Convolutional Neural Network (CNN) model that is specially trained for video codec operations to detect heads in an input image by generating pixel-wise locations for each detected head and by generating, for each detected head, a corresponding head bounding box which frames the detected head. Finally, the AI human head detector model 904 may include a post-processing module 920 which is applies image post-processing to the output from the AI human head detector model module 918 to make the processed images suitable for human viewing and understanding. In addition, the post-processing module 920 may also reduce the size of the data outputs generated by the human head detection model module 918, such as by consolidating or grouping a plurality of head bounding boxes or frames which are generated from a single meeting participant so that a single head bounding box or frame is specified.

[0085] Based on the results of the processing modules 914, 916, 918, 920, the AI human head detector model 904 may generate output video frame images 902 in which the detected human heads are framed with corresponding head bounding boxes 906, 908, 910. As depicted, the first output video frame image 902a includes head bounding boxes 906a-c which are superimposed around each detected human head. In addition, the second output video frame image 902b includes head bounding boxes 908a-c which are superimposed around each detected human head, and the third output video frame image 902c includes head bounding boxes 910a, 910b which are superimposed around each detected human head. The AI human

head detector model 904 may specify each head bounding box using any suitable pixel-based parameters, such as defining the x and y pixel coordinates of a head bounding box or frame in combination with the height and width dimensions of the head bounding box or frame. In addition, the AI human head detector model 904 may specify a distance measure between the camera location and the location of the detected human head using any suitable measurement technique. The AI human head detector model 904 may also compute, for each head bounding box, a corresponding confidence measure or score which quantifies the model's confidence that a human head is detected.

[0086] In some examples of the present disclosure, the AI human head detector model 904 may specify all head detections in a data structure that holds the coordinates of each detected human head along with their detection confidence. More specifically, the human head data structure for a number, n, of human heads may be generated as follows:

$$\left\{ \begin{array}{l} x_1 \ y_1 \ Width_1 \ Height_1 \ Score_1 \\ x_2 \ y_2 \ Width_2 \ Height_2 \ Score_2 \\ \vdots \\ x_n \ y_n \ Width_n \ Height_n \ Score_n \end{array} \right\}$$

In this example, x_i and y_i refer to the image plane coordinates of the i^{th} detected head, and where $Width_i$ and $Height_i$ refer to the width and height information for the head bounding box of the i^{th} detected head. In addition, $Score_i$ is in the range [0, 100] and reflects confidence as a percentage for the i^{th} detected head. This data structure may be used as an input to various applications, such as framing, tracking, composing, recording, switching, reporting, encoding, *etc.* In this example data structure, the first detected head is in the image frame in a head bounding box located at pixel location parameters x_1, y_1 and extending laterally by $Width_1$ and vertically down by $Height_1$. In addition, the second detected head is in the image frame in a head bounding box located at pixel location parameters x_2, y_2 and extending laterally by $Width_2$ and vertically down by $Height_2$, and the n^{th} detected head is in the image frame in a head bounding box located at pixel location parameters x_n, y_n and extending laterally by $Width_n$ and vertically down by $Height_n$. In some aspects, the center of each head bounding box is determined using the following equation:

$$\left(\left\lfloor (x) \right\rfloor_i + \frac{w_i}{2}, y_i + \frac{h_i}{2} \right).$$

[0087] This human head data structure may then be used as an input to the distance estimation process that takes the {Width, Height} parameters of each head bounding box to

pick the best matching distance in terms of meeting room coordinates $\{x_{\text{ROOM}}, y_{\text{ROOM}}\}$ from the look-up table by first using one of the Width or Height parameters with a first lookup table, and then using the other parameter as a tie breaking if multiple meeting room coordinates $\{x_{\text{ROOM}}, y_{\text{ROOM}}\}$ are determined by the one. The human head data structure itself may then be modified to also embed the distance information with each Head, resulting in a modified human head data structure that looks like the following:

$$\left\{ \begin{array}{l} x_1 \ y_1 \ \text{Width}_1 \ \text{Height}_1 \ \text{Score}_1 \ x_{1\text{physicalroom}} \ y_{1\text{physicalroom}} \\ x_2 \ y_2 \ \text{Width}_2 \ \text{Height}_2 \ \text{Score}_2 \ x_{2\text{physicalroom}} \ y_{2\text{physicalroom}} \\ \vdots \\ x_n \ y_n \ \text{Width}_n \ \text{Height}_n \ \text{Score}_n \ x_{n\text{physicalroom}} \ y_{n\text{physicalroom}} \end{array} \right\}$$

where $\{x_{\text{ROOM}1}, y_{\text{ROOM}1}\}$, $\{x_{\text{ROOM}2}, y_{\text{ROOM}2}\}$, ..., $\{x_{\text{ROOM}n}, y_{\text{ROOM}n}\}$ specify the distance of Head₁, Head₂, ..., Head_n, from the camera, respective, in two-dimensional coordinates.

[0088] FIG. 28 illustrates aspects of a codec 1000 according to some examples of the present disclosure. As discussed above, a codec 1000 may be a separate device of a videoconferencing system or may be incorporated into the camera(s) within the videoconferencing system, such as a primary camera. Generally, the codec 100 includes machine readable instructions to maintain a video call with a videoconferencing end point, receive streams from secondary cameras (and a primary camera if not integrated with the primary camera), and encode and composite the streams, according to the methods described herein, to send to the end point.

[0089] As shown in FIG. 28, the codec 1000 may include loudspeaker(s) 1002, though in many cases the loudspeaker 1002 is provided in the monitor 1004. The codec 1000 may include microphone(s) 1006 interfaced via a bus 1008. The microphones 1006 are connected through an analog to digital (A/D) converter 1010, and the loudspeaker 1002 is connected through a digital to analog (D/A) converter 1012. The codec 1000 also includes a processing unit 1014, a network interface 1016, a flash or other non-transitory memory 1018, RAM 1020, and an input/output (I/O) general interface 1022, all coupled by a bus 1008. A camera 1024 is connected to the I/O general interface 1022. Microphone(s) 1006 are connected to the network interface 1016. An HDMI interface 1026 is connected to the bus 1008 and to the external display or monitor 1004. Bus 1008 is illustrative and any interconnect between the elements can be used, such as Peripheral Component Interconnect Express (PCIe) links and switches, Universal Serial Bus (USB) links and hubs, and combinations thereof. The camera 1024 and

microphones 1006, 1006 can be contained in housings containing the other components or can be external and removable, connected by wired or wireless connections.

[0090] The processing unit 1014 can include digital signal processors (DSPs), central processing units (CPUs), graphics processing units (GPUs), dedicated hardware elements, such as neural network accelerators and hardware codecs.

[0091] The flash memory 1018 stores modules of varying functionality in the form of software and firmware, generically programs or machine readable instructions, for controlling the codec 1000. Illustrated modules include a video codec 1028, camera control 1030, framing 1032, other video processing 1034, audio codec 1036, audio processing 1038, network operations 1040, user interface 1042 and operating system, and various other modules 1044. In some examples, an AI head detector module is included with the modules included in the flash memory 1018. Furthermore, in some examples, machine readable instructions can be stored in the flash memory 1018 that cause the processing unit 1014 to carry out any of the methods described above. The RAM 1020 is used for storing any of the modules in the flash memory 1018 when the module is executing, storing video images of video streams and audio samples of audio streams and can be used for scratchpad operation of the processing unit 1014.

[0092] The network interface 1016 enables communications between the codec 1000 and other devices and can be wired, wireless or a combination. In one example, the network interface 1016 is connected or coupled to the Internet 1046 to communicate with remote endpoints 1048 in a videoconference. In one example, the general interface 1022 provides data transmission with local devices (not shown) such as a keyboard, mouse, printer, projector, display, external loudspeakers, additional cameras, and microphone pods, *etc.*

[0093] In one example, the camera 1024 and the microphones 1006 capture video and audio, respectively, in the videoconference environment and produce video and audio streams or signals transmitted through the bus 1008 to the processing unit 1014. As discussed herein, capturing “views” or “images” of a location may include capturing individual frames and/or frames within a video stream. For example, the camera 1024 may be instructed to continuously capture a particular view, *e.g.*, images within a video stream, of a location for the duration of a videoconference. In one example of this disclosure, the processing unit 1014 processes the video and audio using processes in the modules stored in the flash memory 1018. Processed audio and video streams can be sent to and received from remote devices coupled to network interface 1016 and devices coupled to general interface 1022.

[0094] Microphones in the microphone array used for SSL can be used as the microphones providing speech to the far site, or separate microphones, such as microphone 1006, can be used.

[0095] In light of the above, FIG. 29 illustrates a method 1100 of reidentifying participants in a videoconference and transmitting a composite stream of optimal views of the participants to a far end of the videoconference, according to an example of the disclosure. At step 1102, first and second images of a location are captured using a primary camera and a secondary camera (or cameras). In some aspects, the primary and secondary cameras are in communication with and/or are connected to a monitor and/or a codec that includes a memory and a processing unit, as discussed above. In some aspects, the method 1100 is executable via machine readable instructions stored on the codec and/or executed on the processing unit, meaning that the processor instructs the primary and secondary cameras to capture the first and second images of the location.

[0096] At step 1104, human heads in the images are detected using an AI head detection model, as described above. For example, the AI head detection model is applied to images captured by each of the cameras in order to identify, for each detected human head, a head bounding box with specified room and/or pixel coordinates. In some aspects, the AI head detection model includes a facial feature recognition model which is configured to evaluate the detected head based on the facial features visible in each image. At step 1106, ID labels are assigned to each bounding box using the AI head detection model. At step 1108, the same detected human head is found in the first and second images using the ID labels, for example using one of the methods described above. In some aspects, finding the same detected human head at step 1108 includes grouping together an ID label in the second image with an associated ID label in the first image, thus re-identifying a participant across the first and second images. In some examples, the ID labels are grouped together based on a distance of the identification label in the second image with the associated ID label in the first image. At step 1110, images with optimal views of each participant are selected. For example, the image with the best frontal view of a participant, *i.e.*, the image in which the facial features of the participant are most visible, is selected, and step 1110 is repeated so that an optimal view of each participant is selected. At step 1112, a composite stream of the optimal views of the participants is transmitted to a far end of the videoconference. In some aspects, only the optimal view of each participant is visible in the composite stream, meaning that duplicate views of the participants are not transmitted to a far end of the videoconference. As discussed above, it is contemplated that the entirety of the method 1100 (including any of the other methods described above) may

be performed within the primary camera and/or secondary camera, and/or the method 1100 is executable via machine readable instructions stored on the codec and/or executed on the processing unit. Thus, it will be understood that the methods described herein may be computationally light-weight and may be performed entirely in the primary camera, thus reducing the need for a resource-heavy GPU and/or other specialized computational machinery.

[0097] Certain operations of methods according to the technology, or of systems executing those methods, can be represented schematically in the figures or otherwise discussed herein. Unless otherwise specified or limited, representation in the figures of particular operations in particular spatial order can not necessarily require those operations to be executed in a particular sequence corresponding to the particular spatial order. Correspondingly, certain operations represented in the figures, or otherwise disclosed herein, can be executed in different orders than are expressly illustrated or described, as appropriate for particular examples of the technology. Further, in some examples, certain operations can be executed in parallel, including by dedicated parallel processing devices, or separate computing devices that interoperate as part of a large system.

[0098] The disclosed technology is not limited in its application to the details of construction and the arrangement of components set forth in the following description or illustrated in the following drawings. Other examples of the disclosed technology are possible and examples described and/or illustrated here are capable of being practiced or of being carried out in various ways.

[0099] A plurality of hardware and software-based devices, as well as a plurality of different structural components can be used to implement the disclosed technology. In addition, examples of the disclosed technology can include hardware, software, and electronic components or modules that, for purposes of discussion, can be illustrated and described as if the majority of the components were implemented solely in hardware. However, in one example, the electronic based aspects of the disclosed technology can be implemented in software (for example, stored on non-transitory computer-readable medium) executable by a processor. Although certain drawings illustrate hardware and software located within particular devices, these depictions are for illustrative purposes. In some examples, the illustrated components can be combined or divided into separate software, firmware, hardware, or combinations thereof. As one example, instead of being located within and performed by a single electronic processor, logic and processing can be distributed among multiple electronic processors. Regardless of how they are combined or divided, hardware and software

components can be located on the same computing device or can be distributed among different computing devices connected by a network or other suitable communication links.

[00100] Any suitable non-transitory computer usable or computer readable medium may be utilized. The computer-usable or computer-readable medium may be, for example, but not limited to, an electronic, magnetic, optical, electromagnetic, infrared, or semiconductor system, apparatus, or device. More specific examples (a non-exhaustive list) of the computer-readable medium would include the following: a portable computer diskette, a hard disk, a random-access memory (RAM), a read-only memory (ROM), an erasable programmable read-only memory (EPROM or Flash memory), a portable compact disc read-only memory (CD-ROM), an optical storage device, or a magnetic storage device. In the context of this disclosure, a computer-usable or computer-readable medium may be any medium that can contain, store, communicate, or transport the program for use by or in connection with the instruction execution system, apparatus, or device.

[00101] As used herein in the context of computer implementation, unless otherwise specified or limited, the terms “component,” “system,” “module,” “block,” and the like are intended to encompass part or all of computer-related systems that include hardware, software, a combination of hardware and software, or software in execution. For example, a component can be, but is not limited to being, a processor device, a process being executed (or executable) by a processor device, an object, an executable, a thread of execution, a computer program, or a computer. By way of illustration, both an application running on a computer and the computer can be a component. Components (or system, module, and so on) can reside within a process or thread of execution, can be localized on one computer, can be distributed between two or more computers or other processor devices, or can be included within another component (or system, module, and so on).

CLAIMS

1. A method of identifying participants in a location and using a plurality of cameras, the method comprising:
 - capturing images of the location using a primary camera and a secondary camera;
 - applying a machine learning human head detector model to the images to identify coordinates for each human head detected in the images;
 - determining a centroid of the coordinates for each human head detected in the images and a centroid angle for each human head detected in the images relative to the centroid;
 - ranking, for each image, identification labels for each human head based on the centroid angle for each human head detected in the images;
 - identifying a reference human head in a primary image captured by the primary camera;
 - and
 - aligning the identification labels across each of the images to reidentify each human head.
2. The method of claim 1, wherein ranking the identification labels based on the centroid angle for each human head detected in the images includes ranking the identification labels for each human head in counterclockwise order starting with a human head having a lowest centroid angle.
3. The method of claim 1, wherein applying the machine learning human head detector model to identify the coordinates for each human head detected includes determining cartesian coordinates for each human head and subsequently transforming the cartesian coordinates into polar coordinates.
4. The method of claim 3, wherein the polar coordinates include a distance from the centroid and the centroid angle.
5. The method of claim 1, wherein the reference human head is located farthest toward a bottom-right corner of the primary image, and wherein the reference human head is subsequently identified in a secondary image captured by the secondary camera.
6. The method of claim 5, wherein aligning the identification labels further includes re-ranking the identification labels in the secondary image in counterclockwise order starting with the reference human head.

7. The method of claim 6, further comprising:

creating reidentification labels for each human head detected in the secondary image, wherein the reidentification labels match the identification labels in the primary image.

8. A method of identifying participants in a location, the method comprising:

capturing images of the location using a primary camera and a secondary camera, the primary camera defining a primary coordinate system and the secondary camera defining a secondary coordinate system;

applying a machine learning human head detector model to the images to identify world coordinate points associated with each human head detected in the images, including world coordinate points on a secondary coordinate system associated with the images captured by the secondary camera and world coordinate points on a primary coordinate system associated with the images captured by the primary camera;

projecting the world coordinate points from the secondary coordinate system onto the primary coordinate system;

measuring distances between each of the world coordinate points; and assigning an identification label to a grouping of the world coordinate points having a minimum distance therebetween, the identification label identifying a first participant.

9. The method of claim 8, wherein projecting the world coordinate points includes evaluating scale, translation, and rotation factors between the primary coordinate system and the secondary coordinate system.

10. The method of claim 8, wherein projecting the world coordinate points further includes capturing, using the primary and the secondary camera, calibration points in the images, and wherein the calibration points are world coordinates of a single human head detected by the machine learning human head detector model.

11. The method of claim 10, wherein capturing the calibration points includes capturing at least 1500 frames using the primary camera and the secondary camera.

12. The method of claim 8, wherein the secondary camera includes a plurality of secondary cameras.

13. The method of claim 8, wherein projecting the world coordinate points further includes determining world coordinates of the secondary camera relative to the primary camera in the primary coordinate system.

14. A non-transitory computer-readable medium containing instructions that when executed cause a processor to:

instruct a plurality of cameras to capture a first image and a second image of a location;

apply a machine learning human head detector model to the first image and the second image to detect human heads in the first image and the second image and identify coordinates for each human head detected;

assign identification labels to each human head detected in the first image and the second image; and

group together each identification label in the second image with an associated identification label in the first image.

15. The non-transitory computer-readable medium of claim 14, wherein the identification labels in the second image are grouped with the identification labels in the first image based on a distance between each of the identification labels to one another.

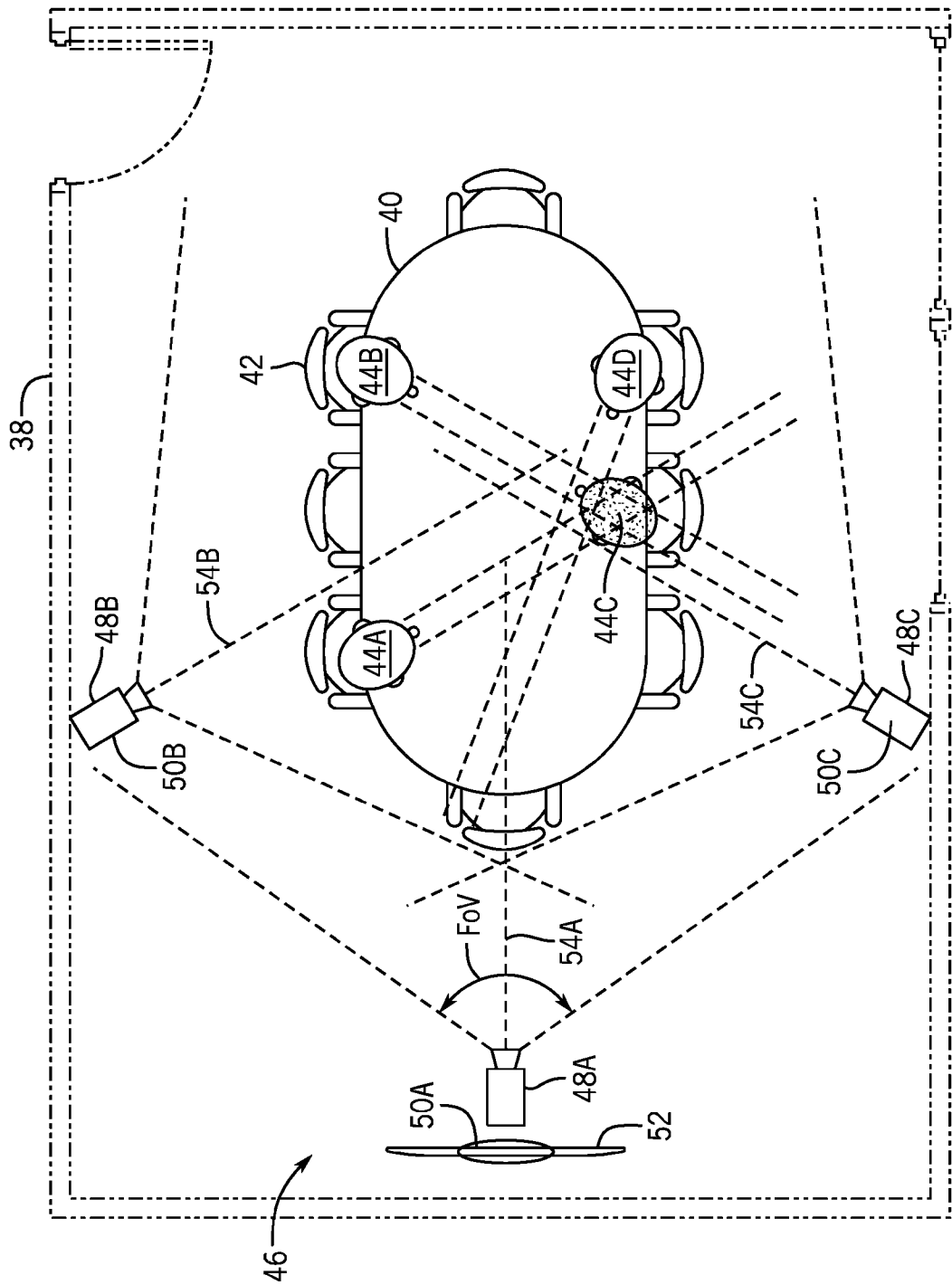


FIG. 1

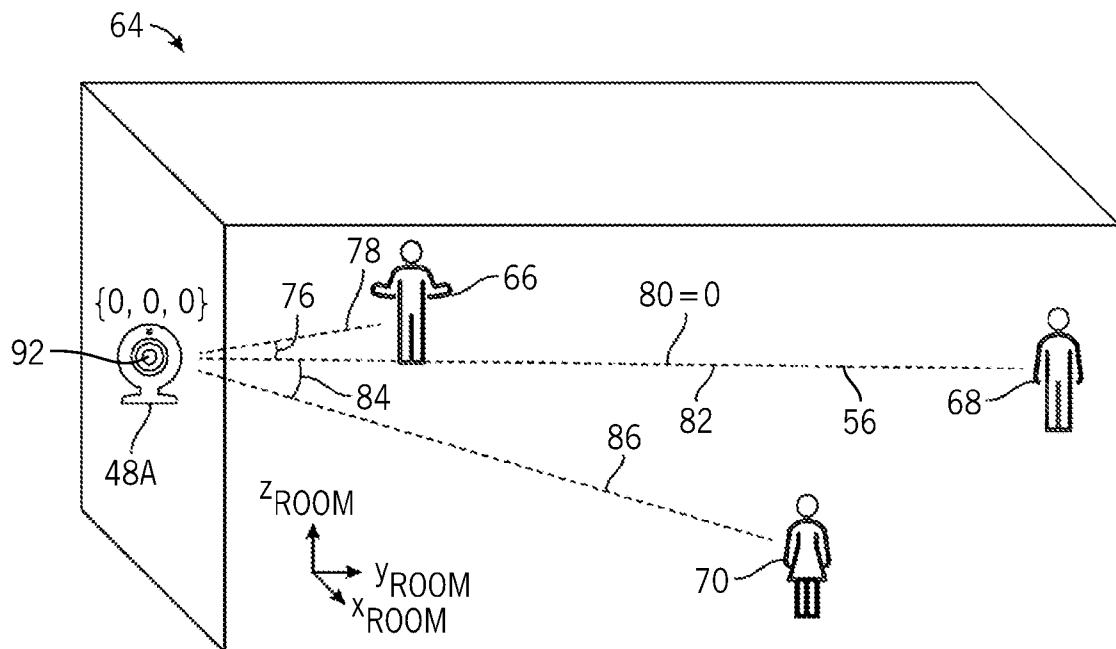


FIG. 2

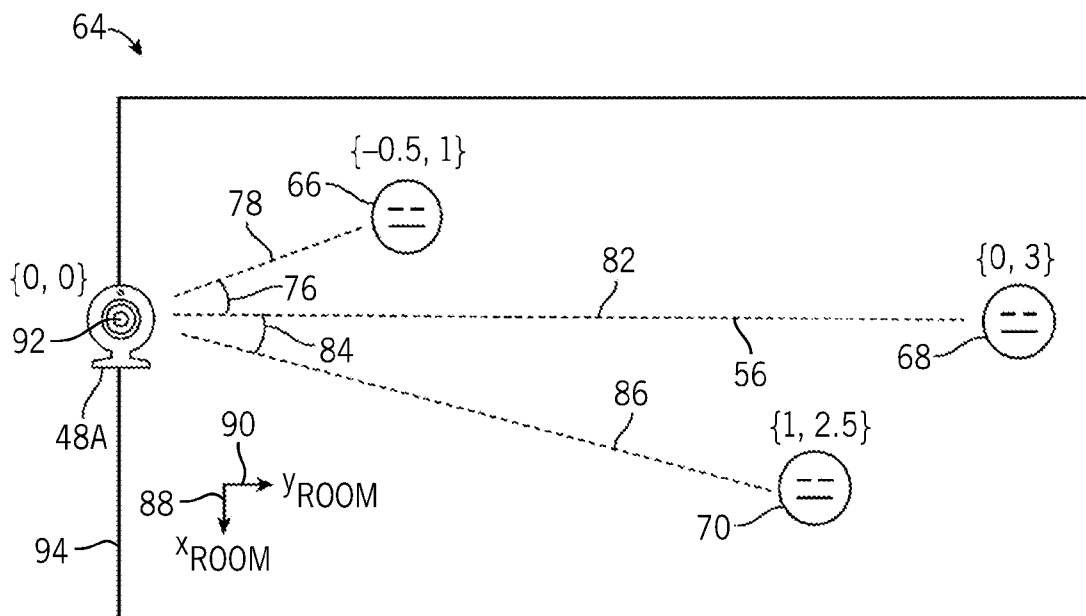


FIG. 3

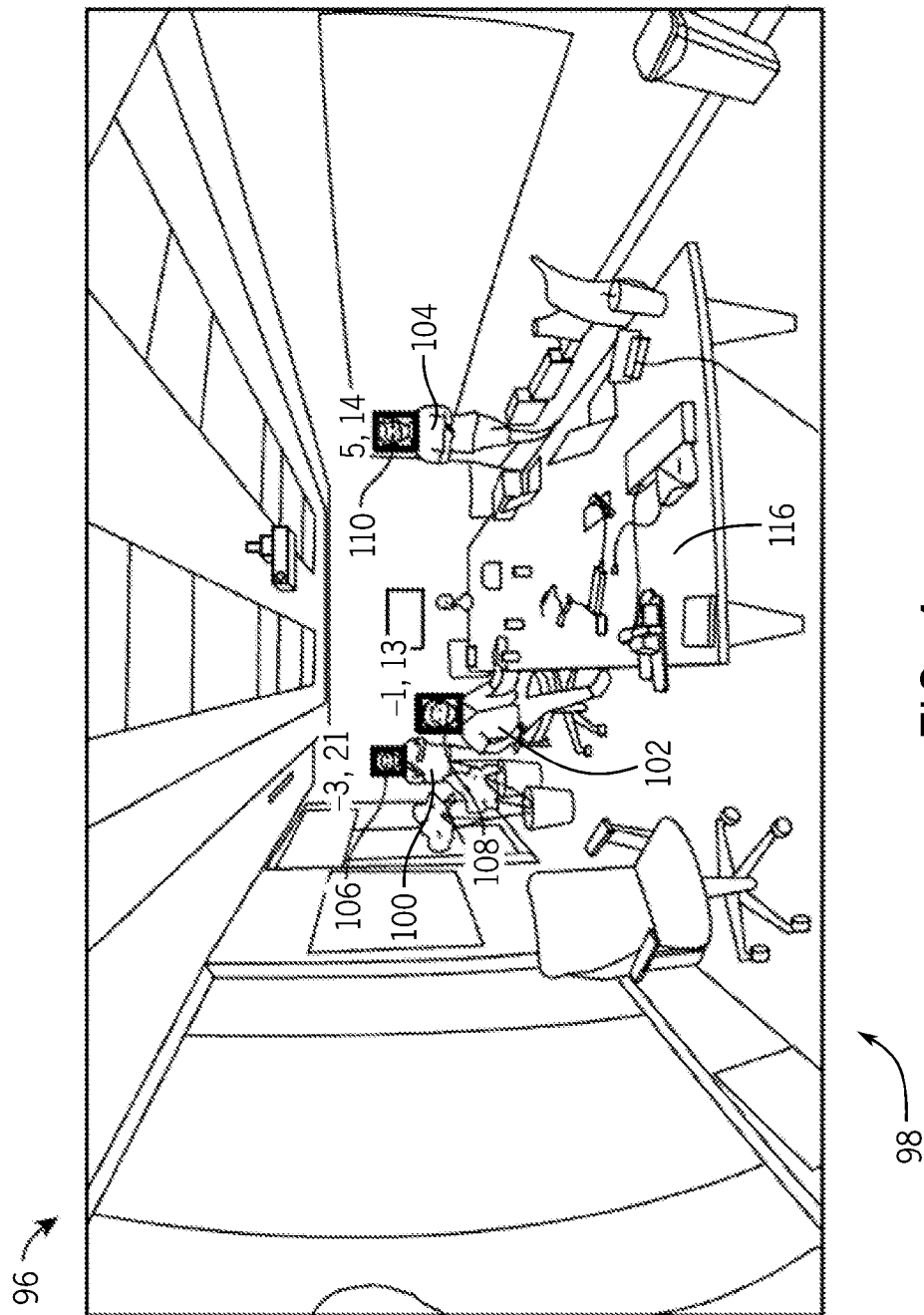


FIG. 4

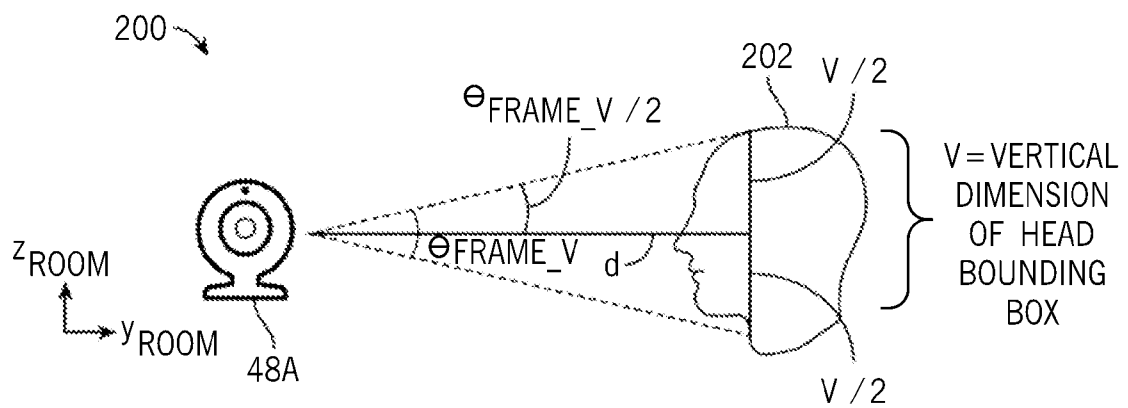


FIG. 5

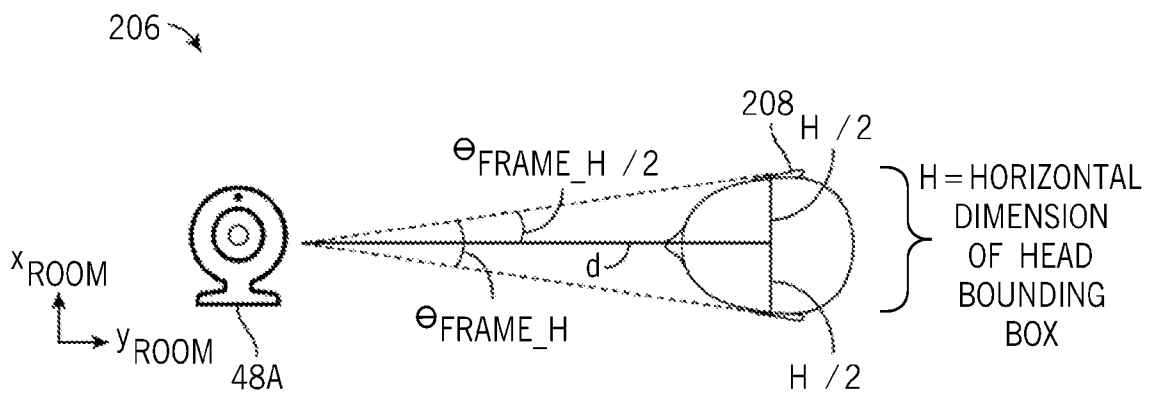
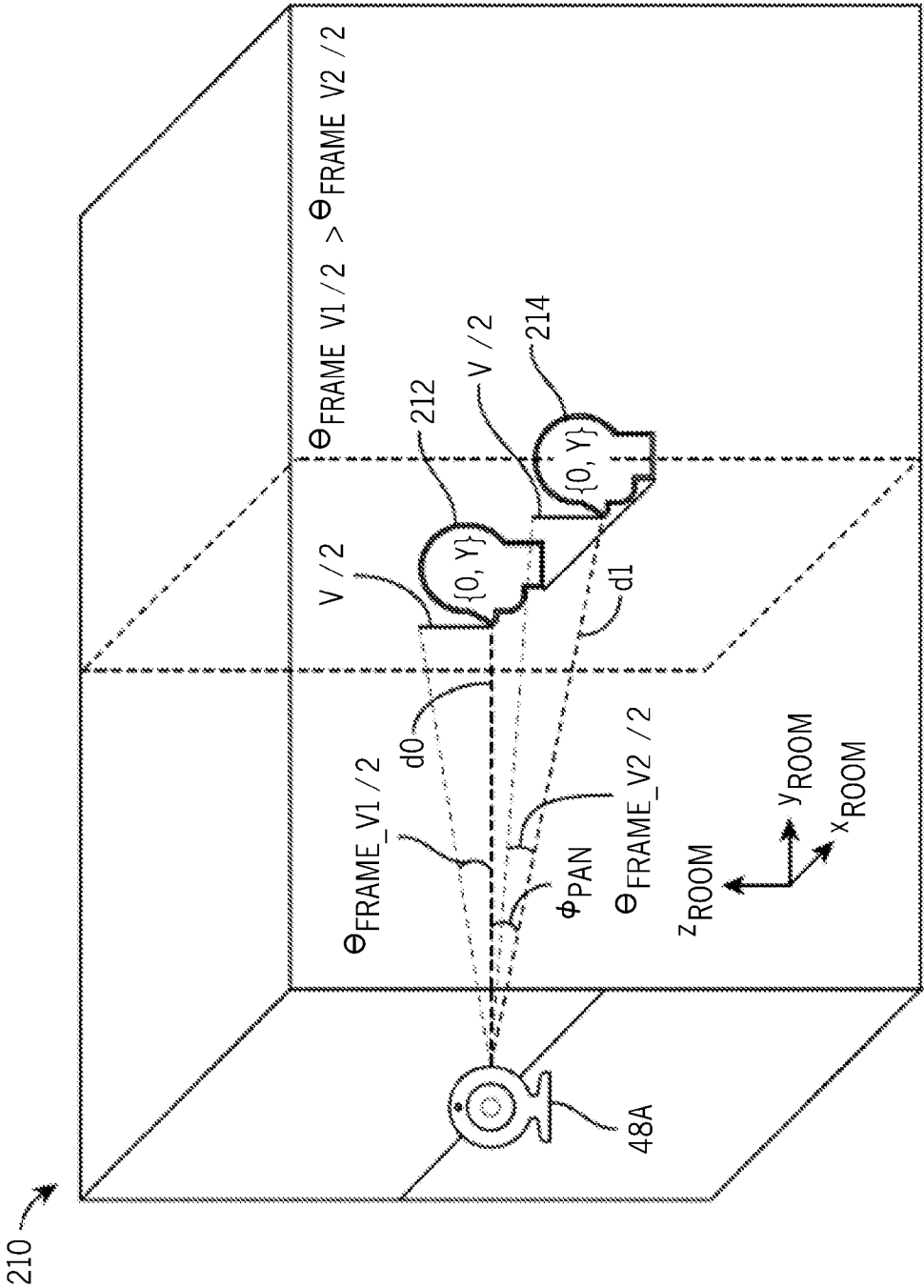


FIG. 6



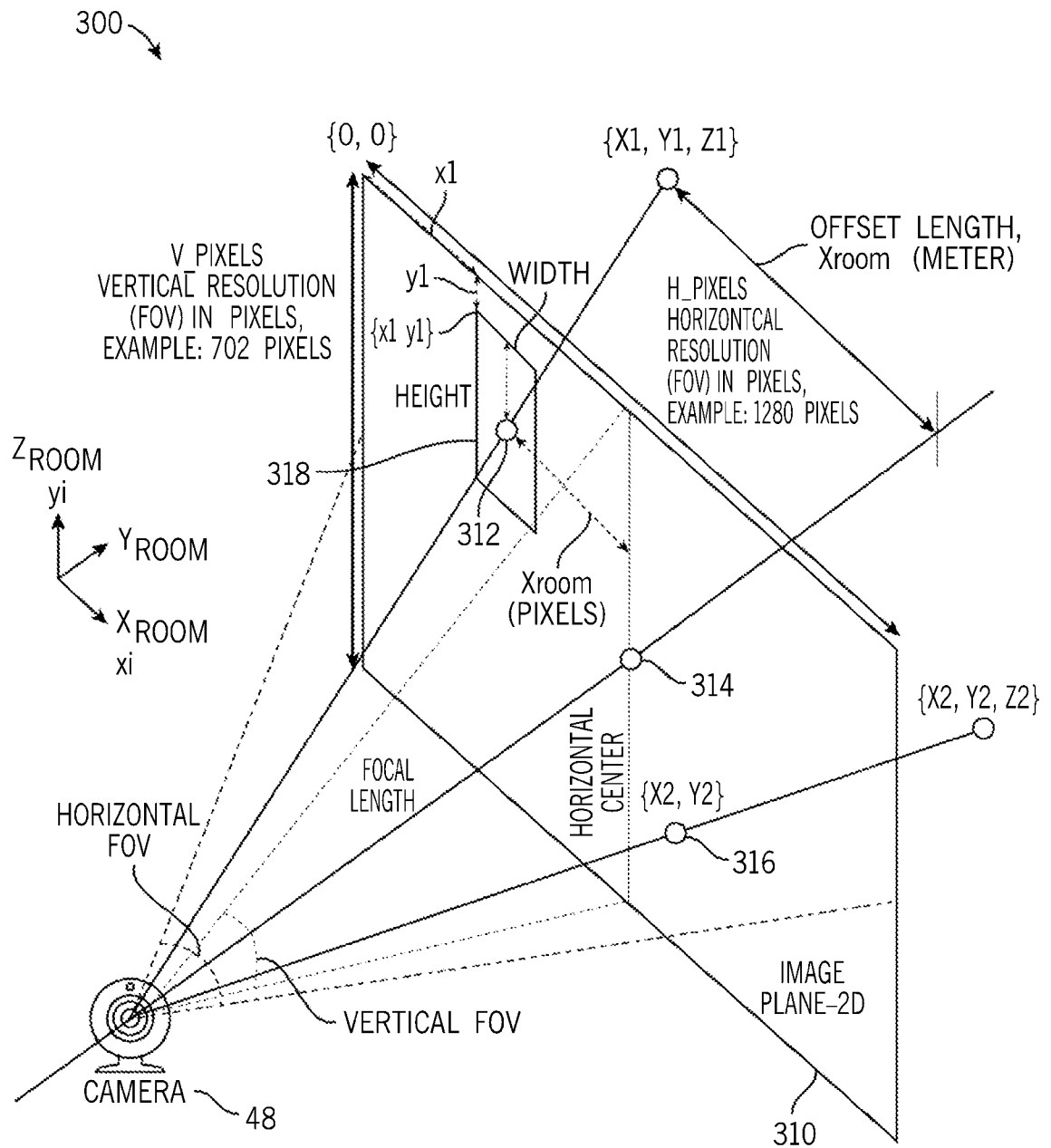


FIG. 8

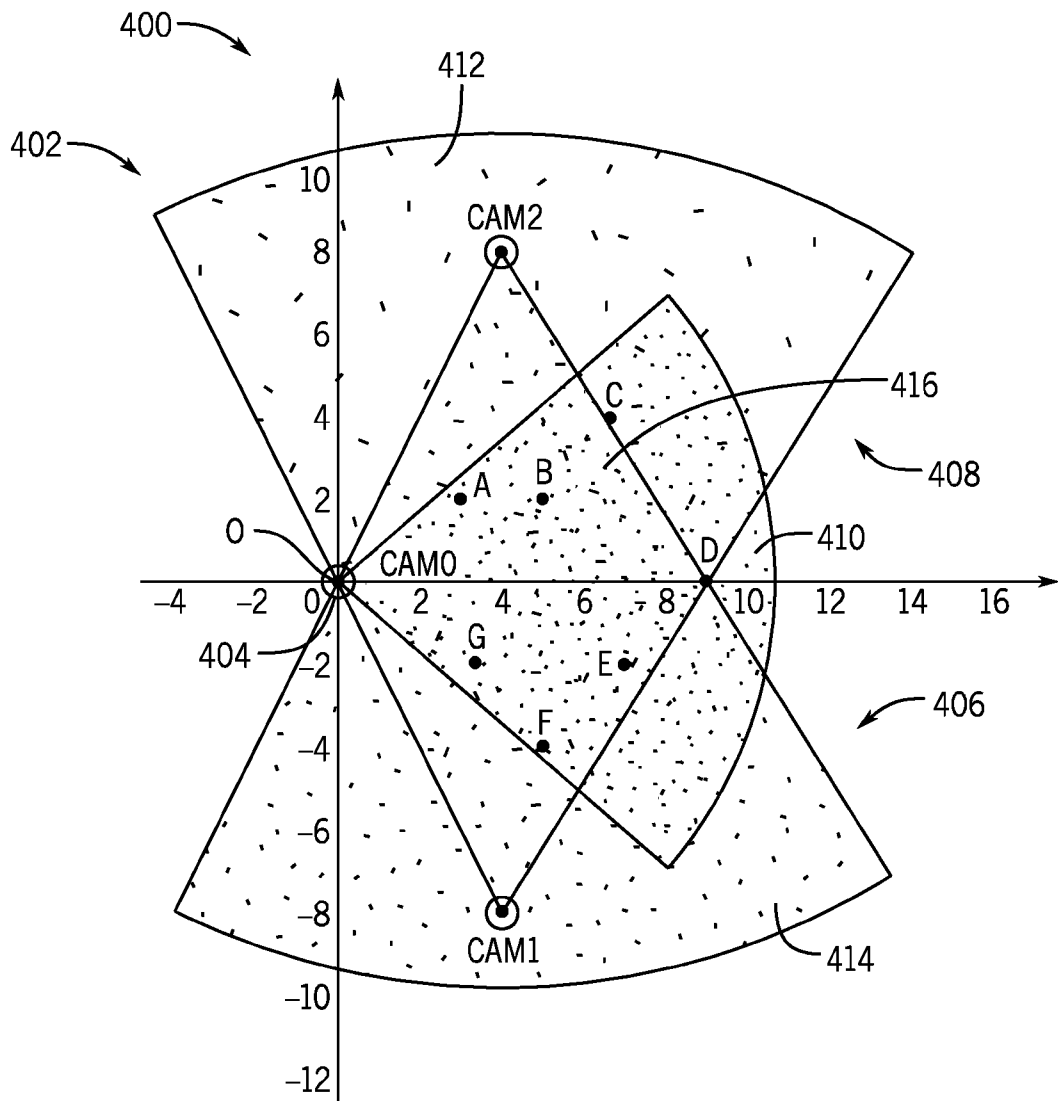


FIG. 9

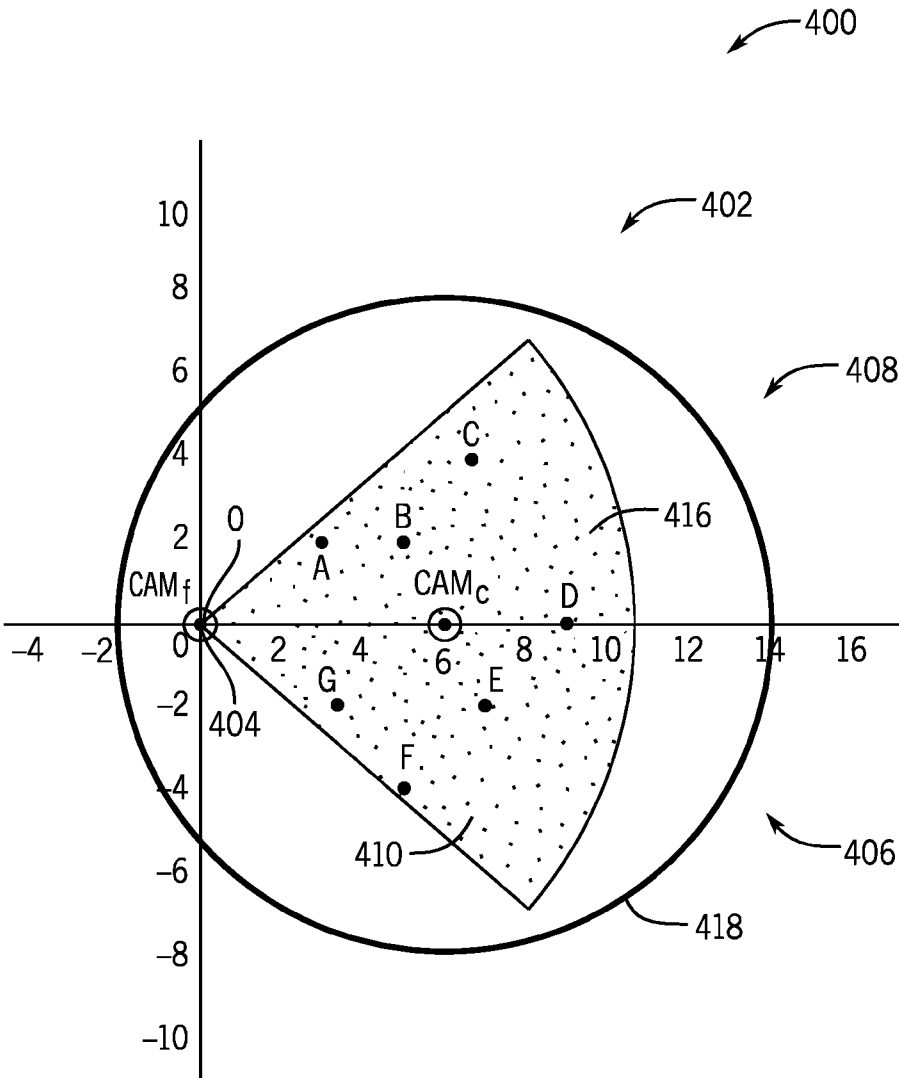


FIG. 10

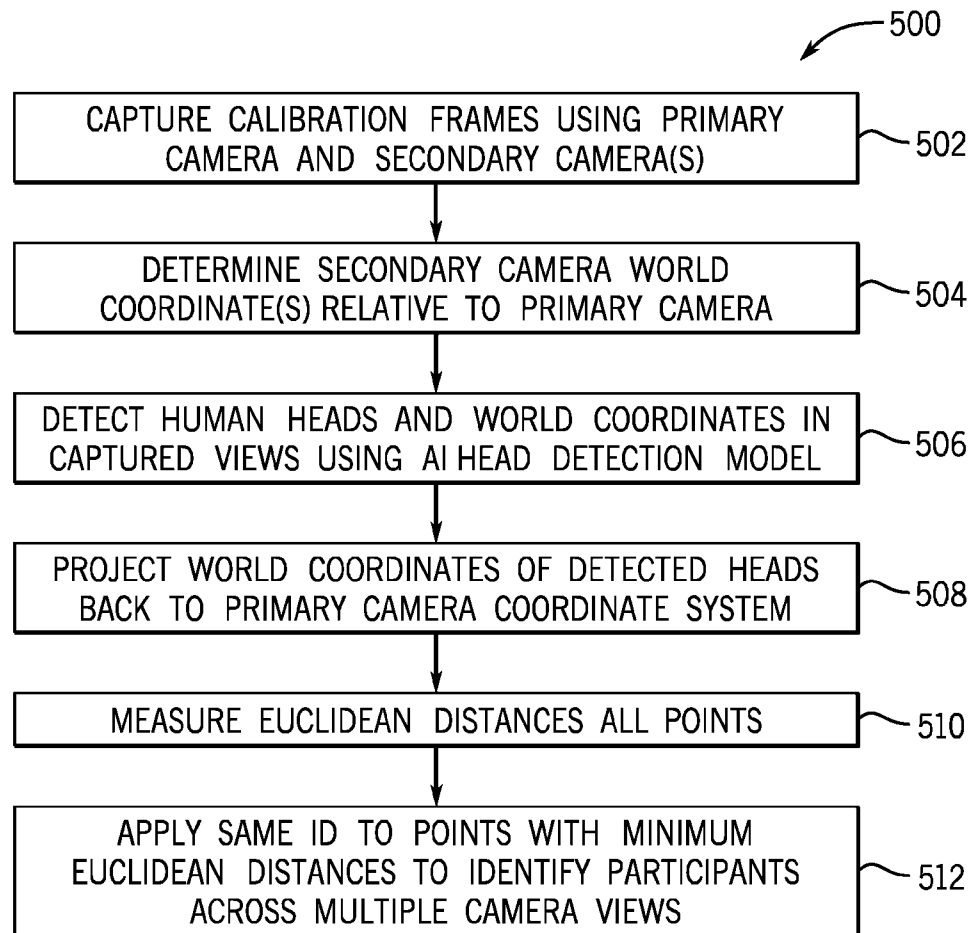


FIG. 11

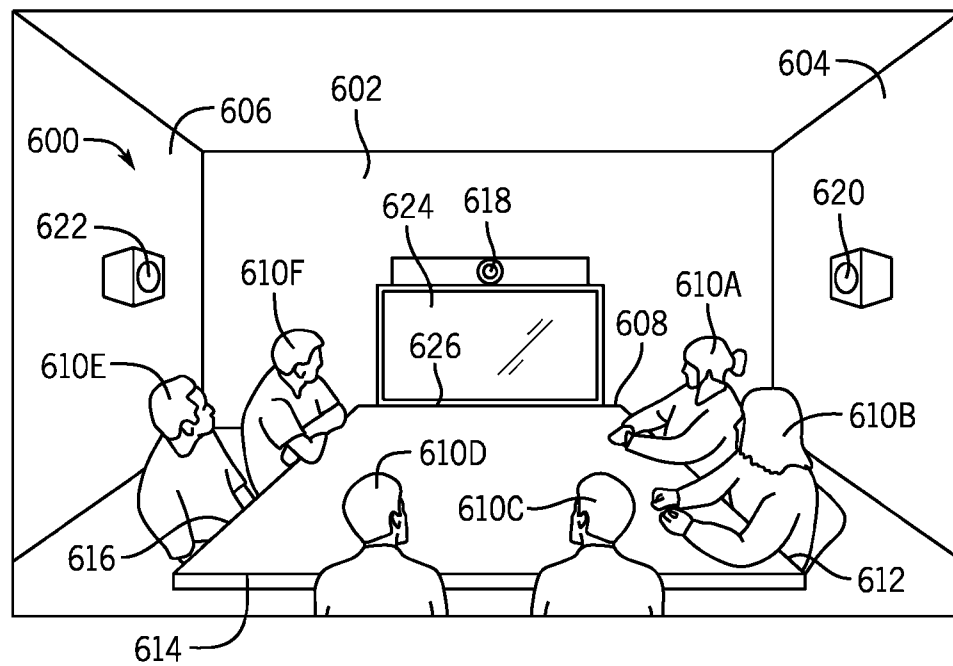


FIG. 12A

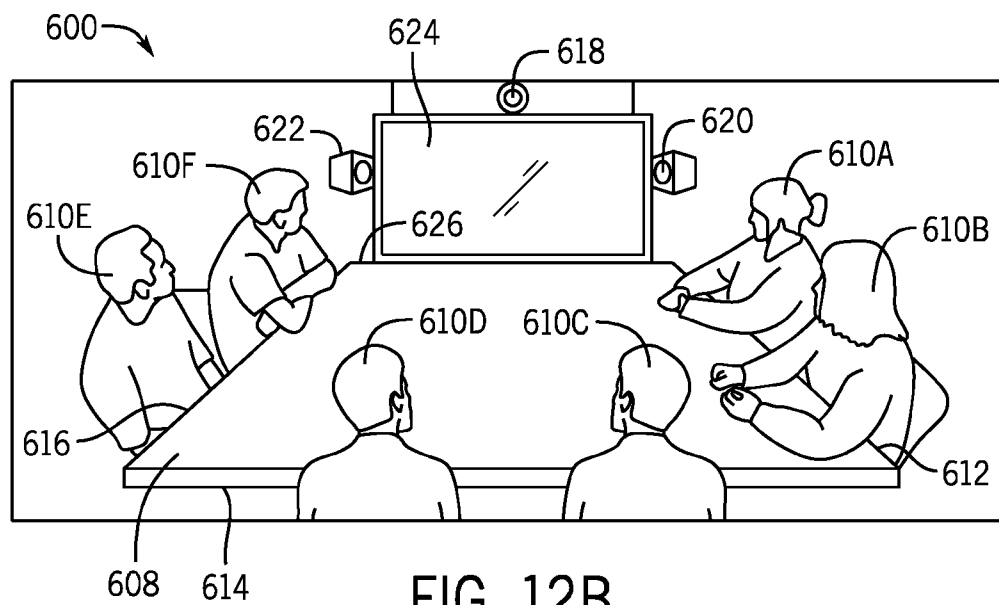


FIG. 12B

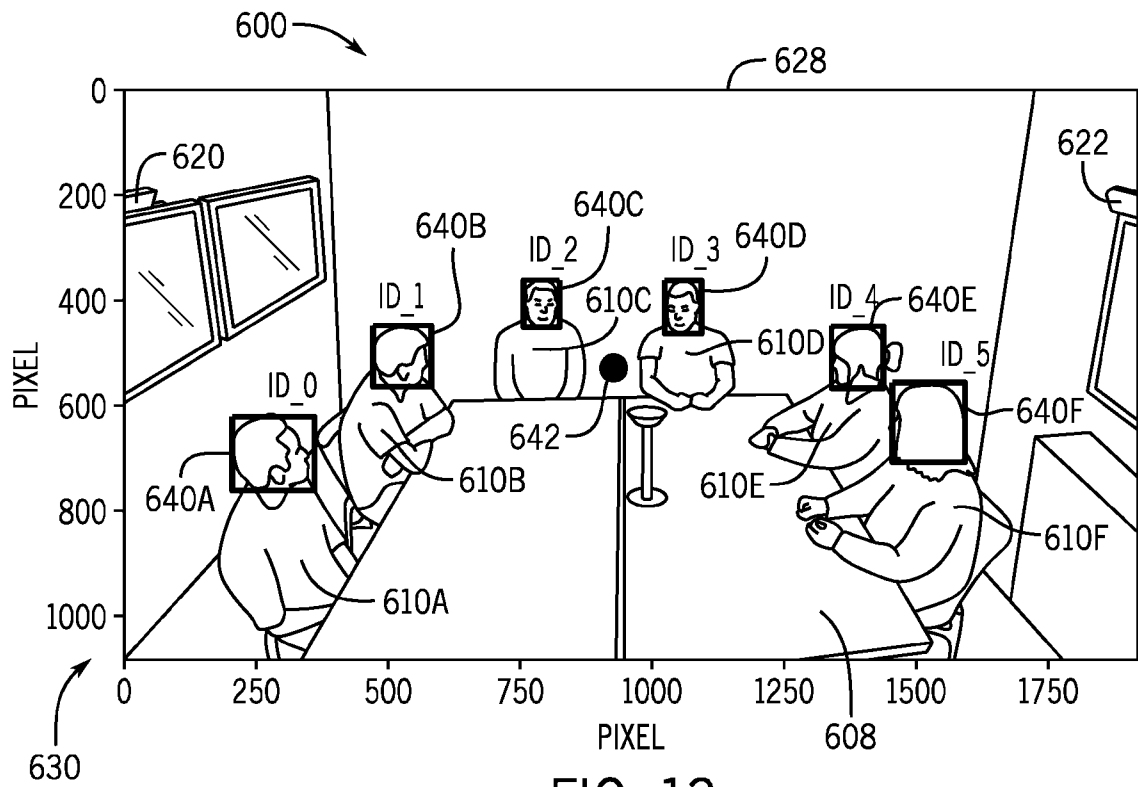


FIG. 13

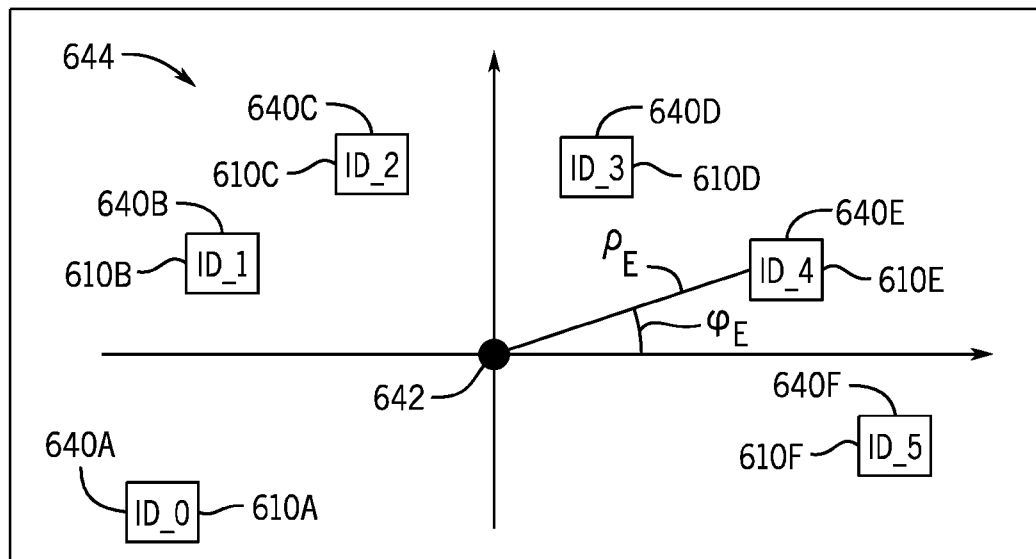


FIG. 14

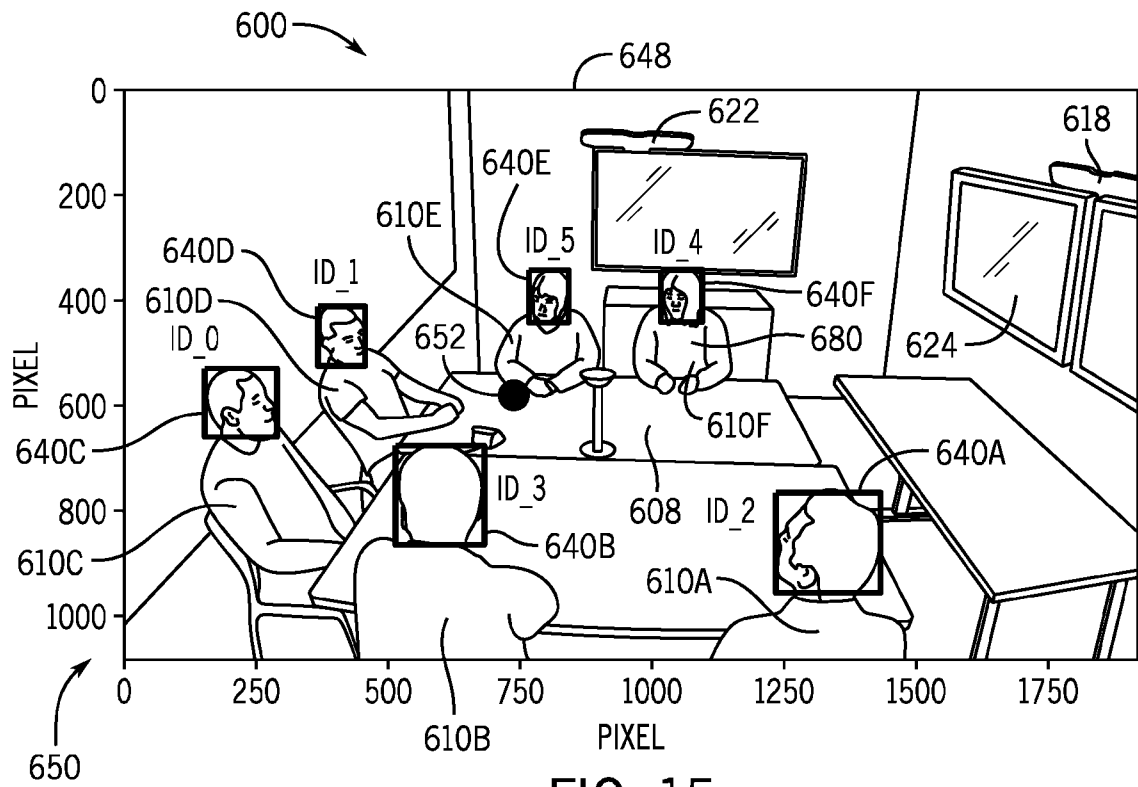


FIG. 15

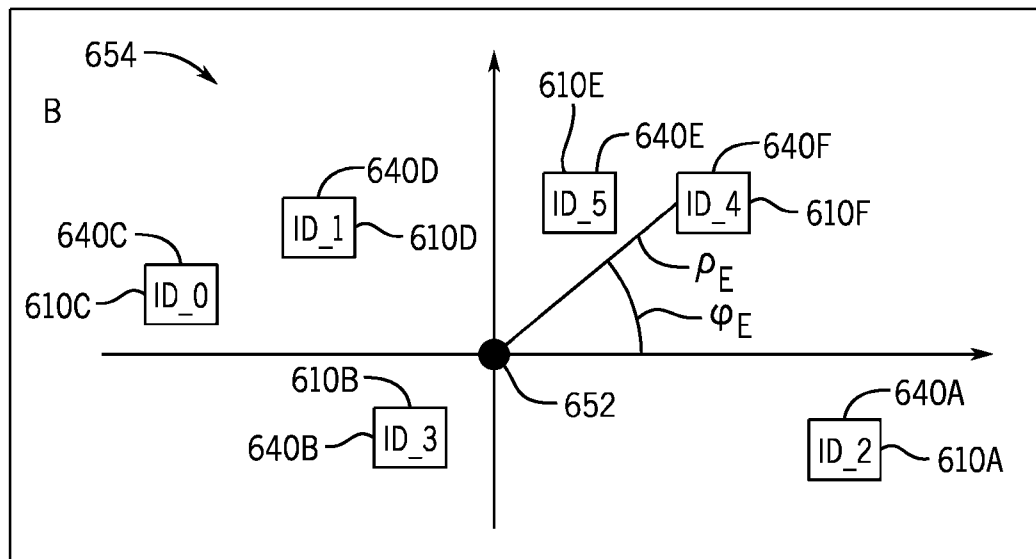


FIG. 16

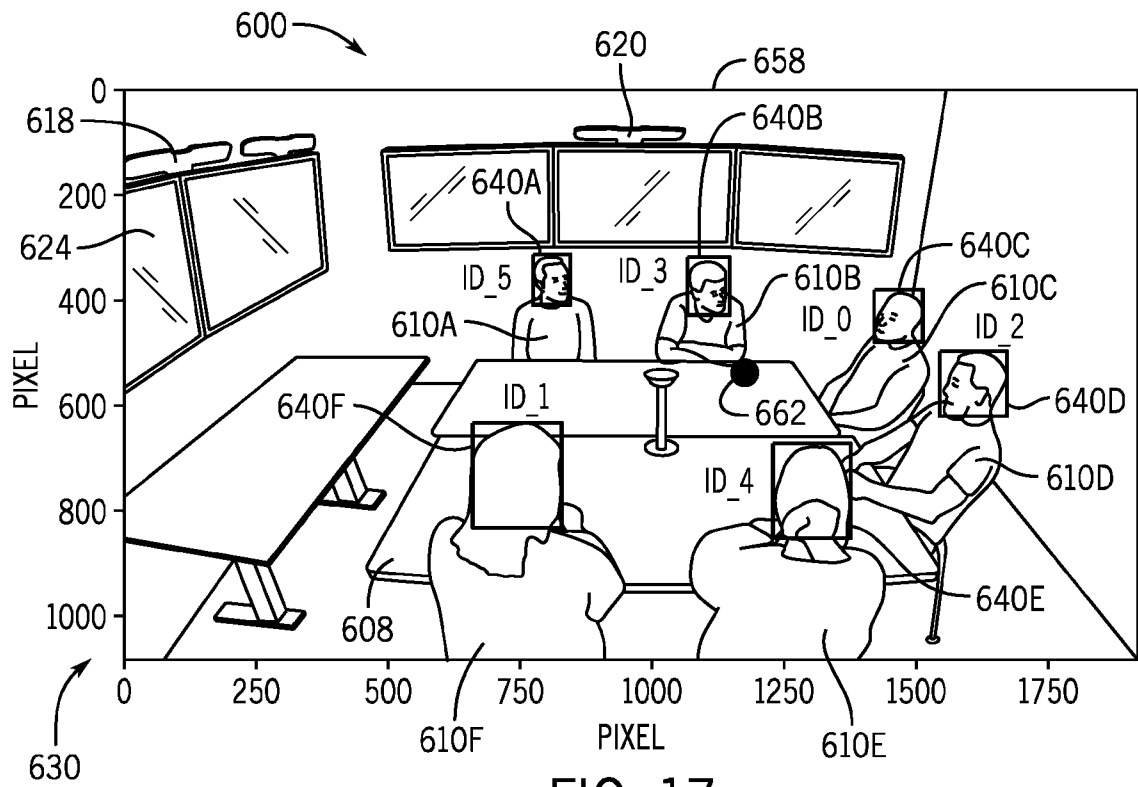


FIG. 17

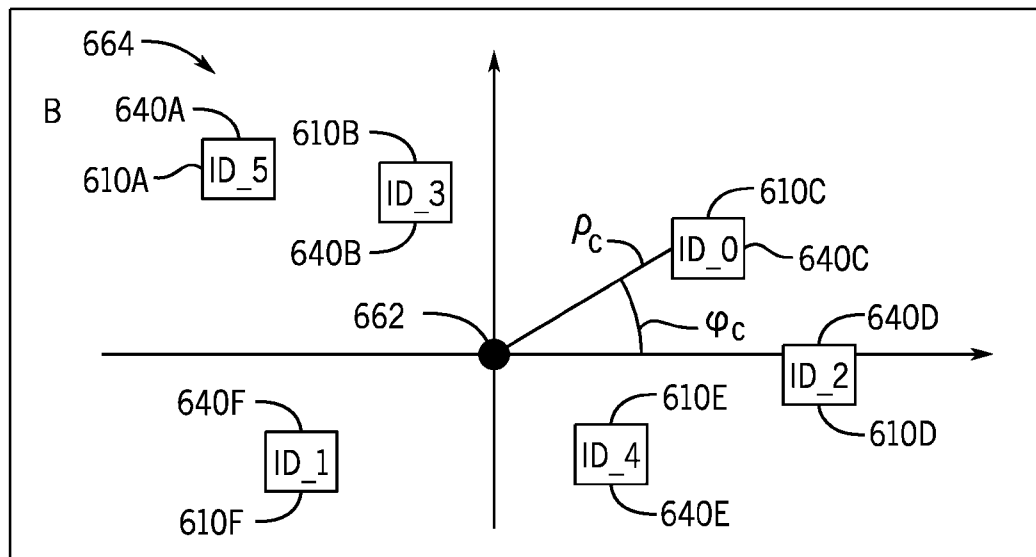


FIG. 18

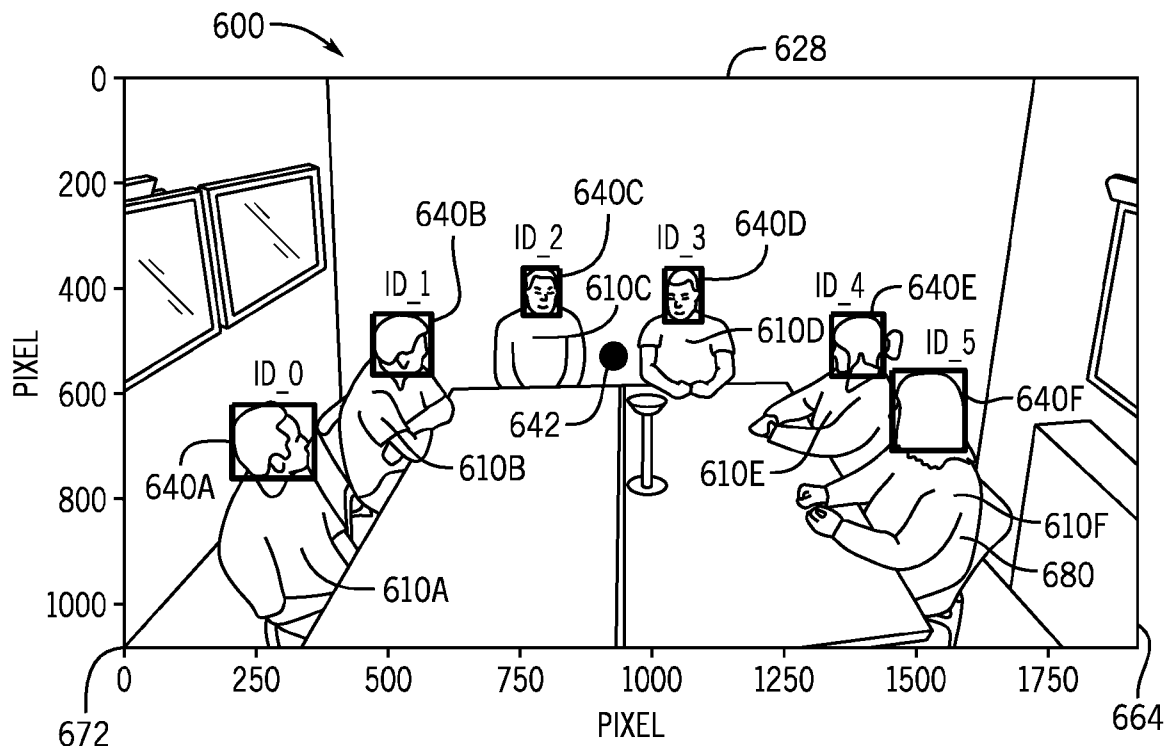


FIG. 19

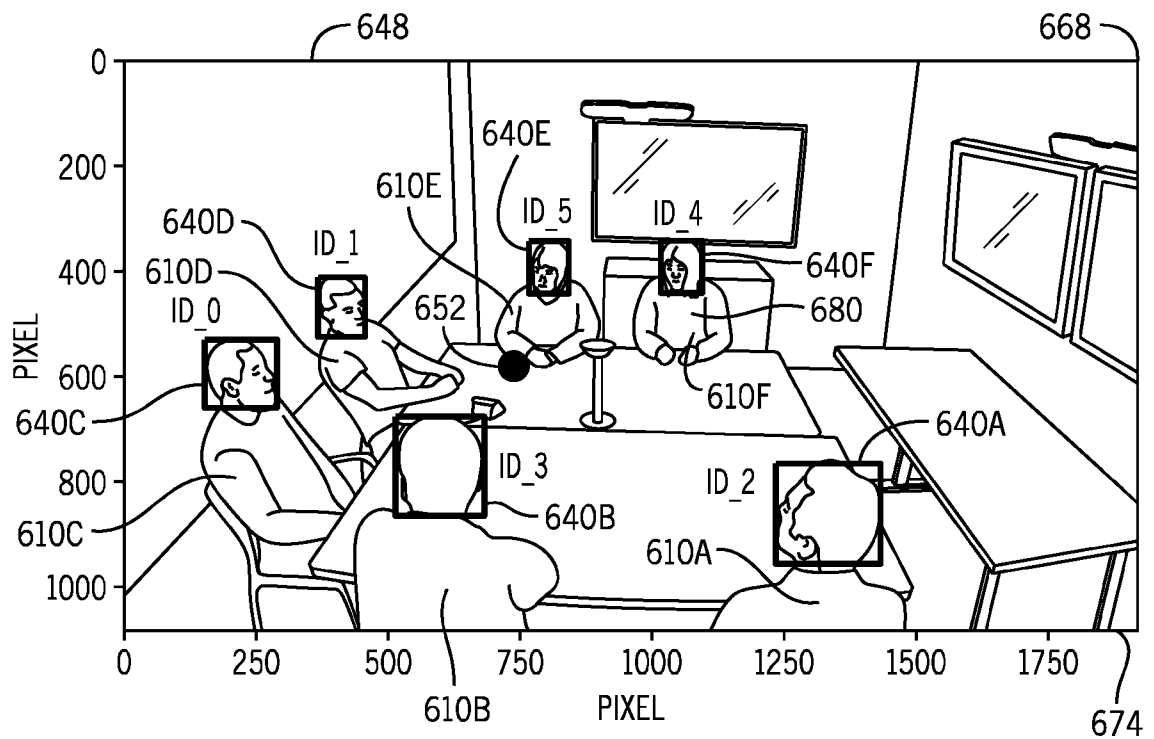


FIG. 20

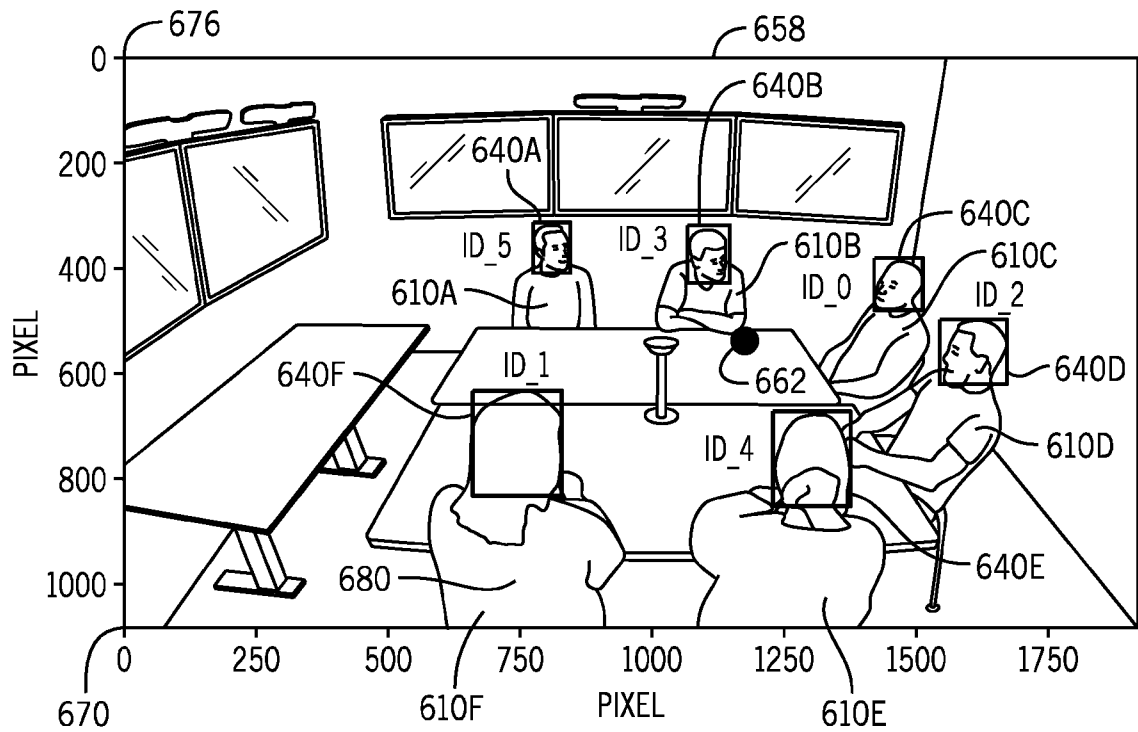


FIG. 21

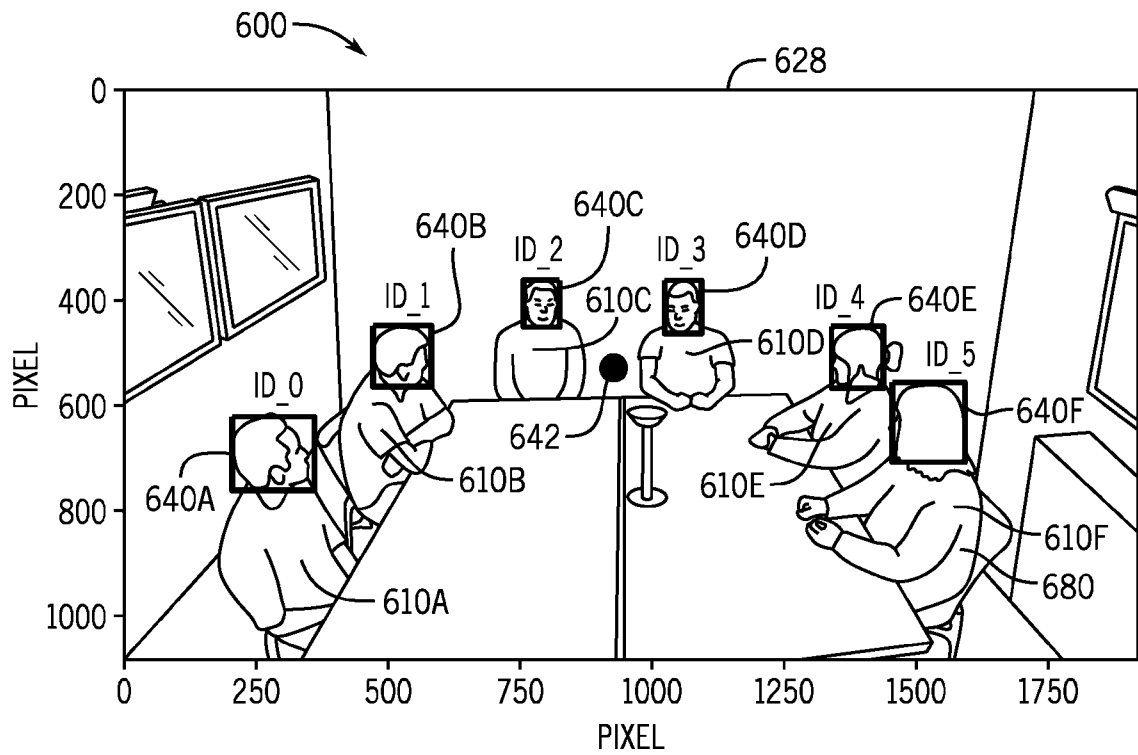


FIG. 22

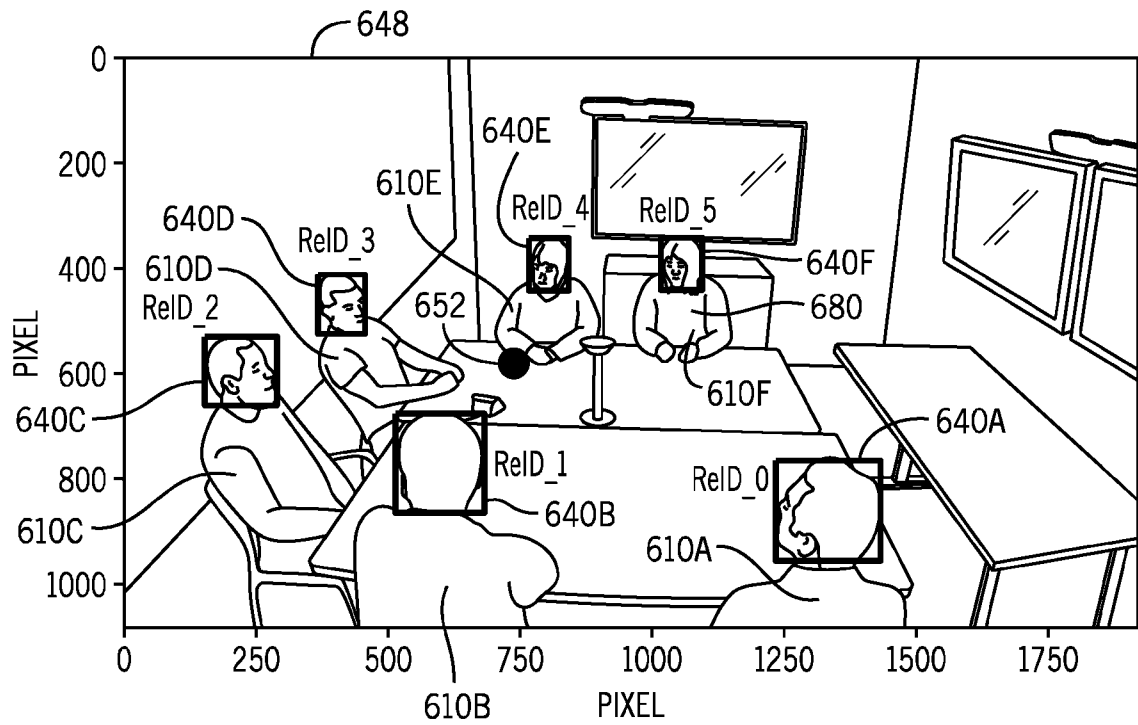


FIG. 23

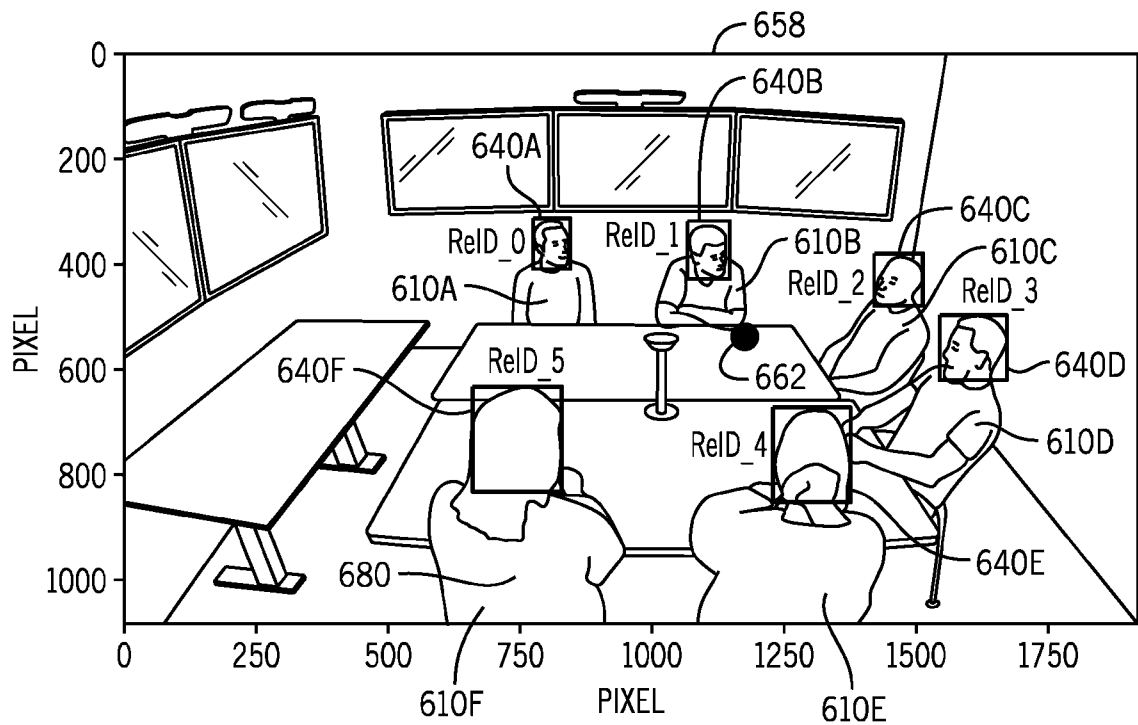


FIG. 24

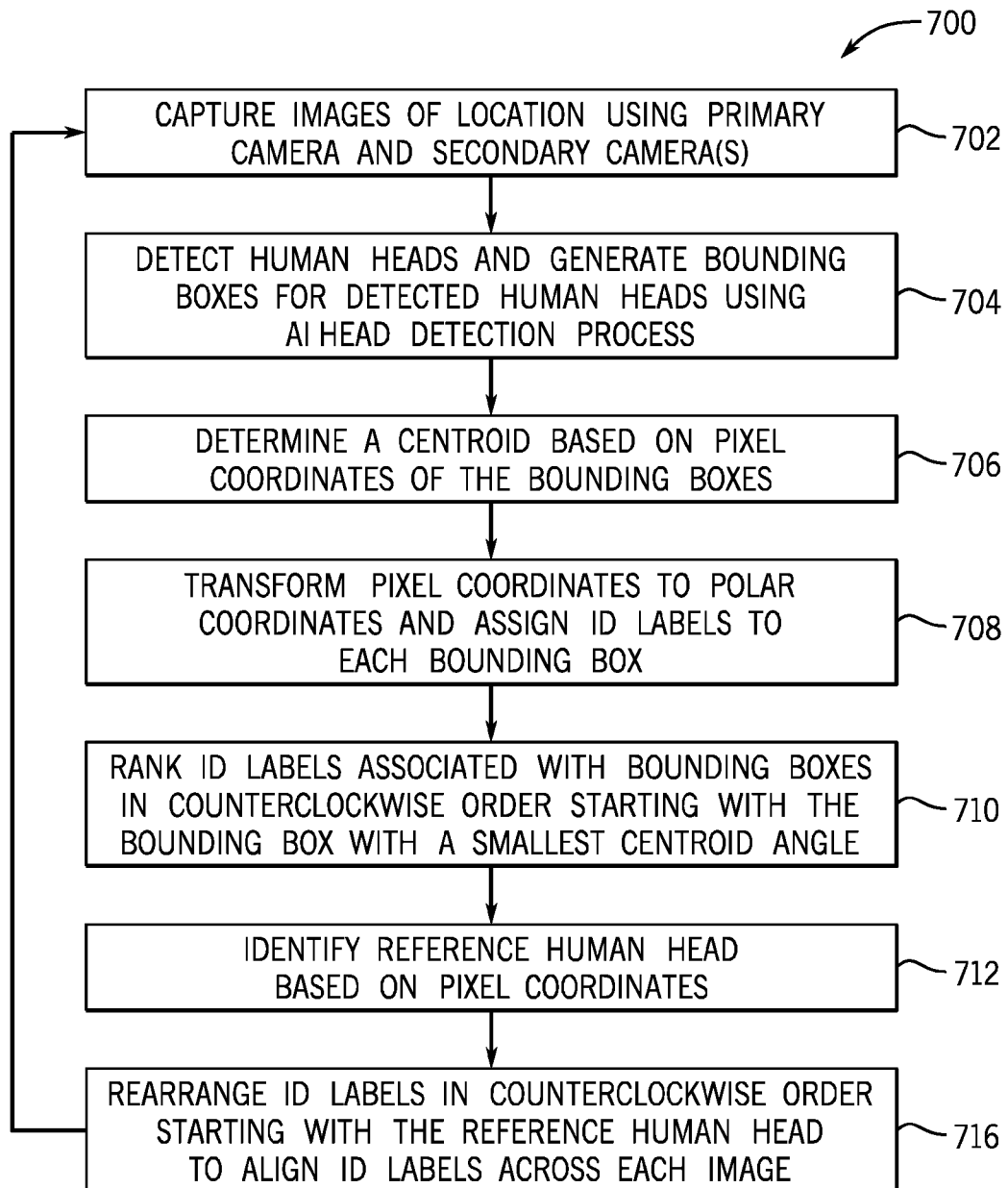


FIG. 25

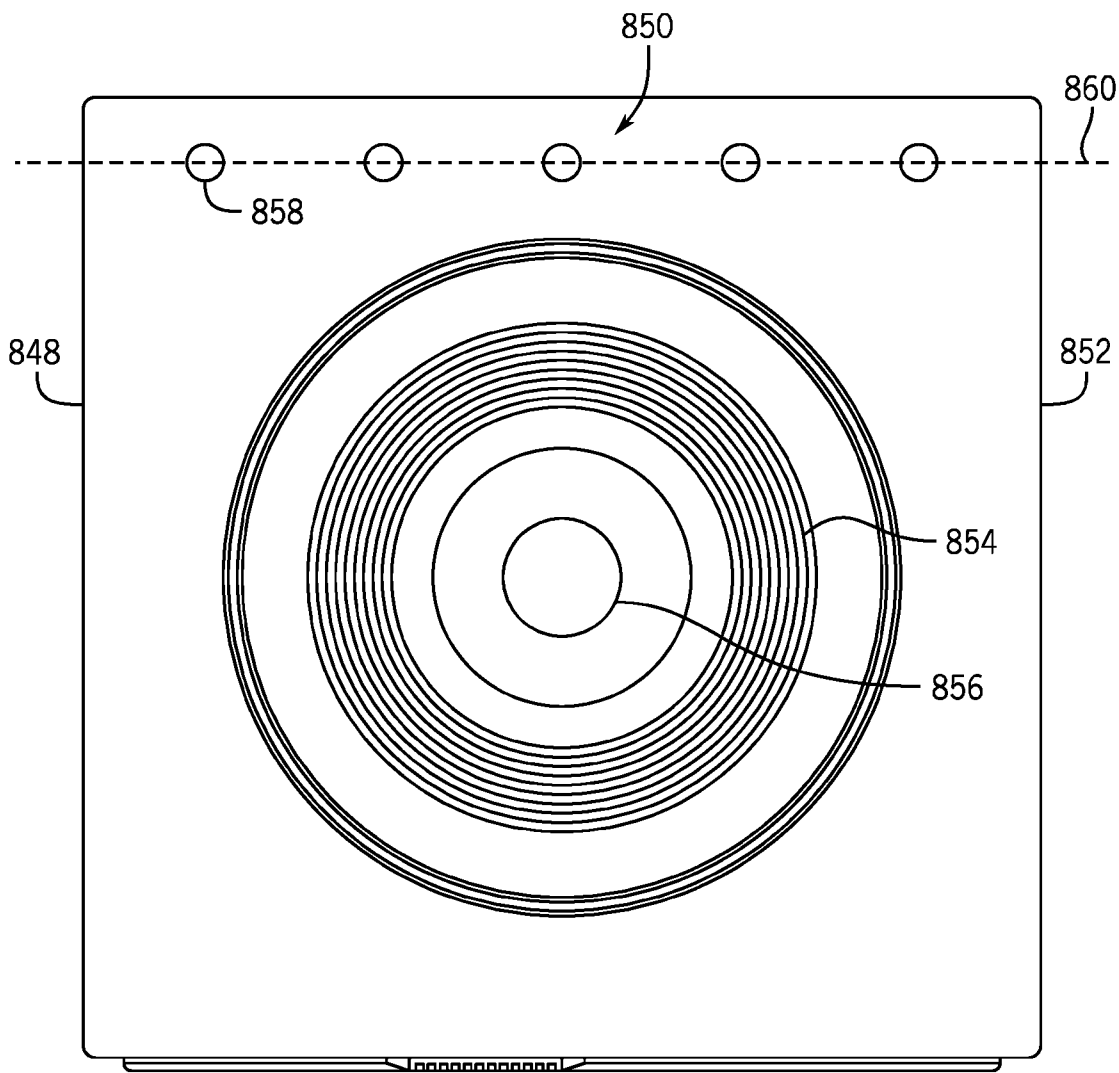


FIG. 26

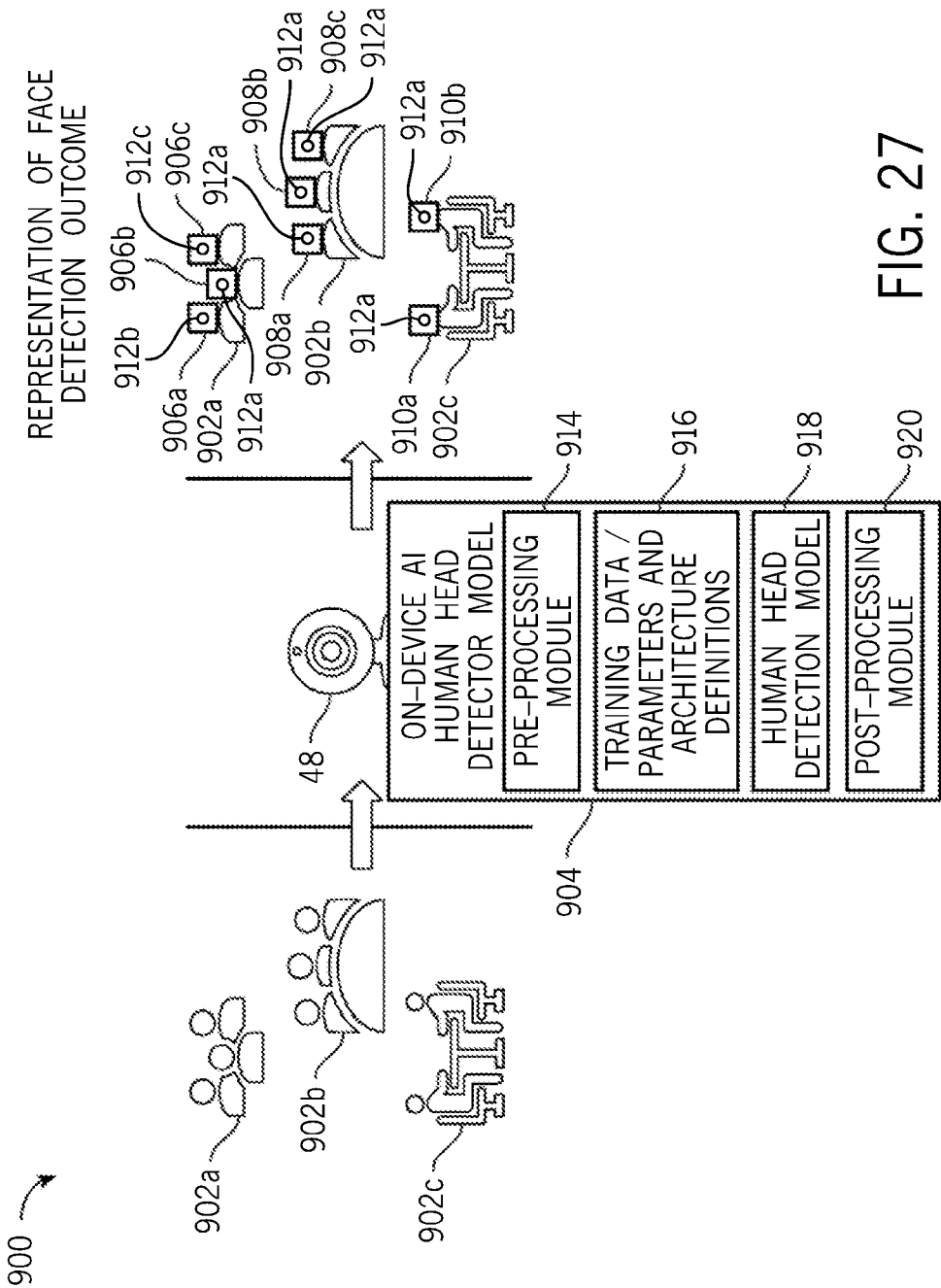


FIG. 27

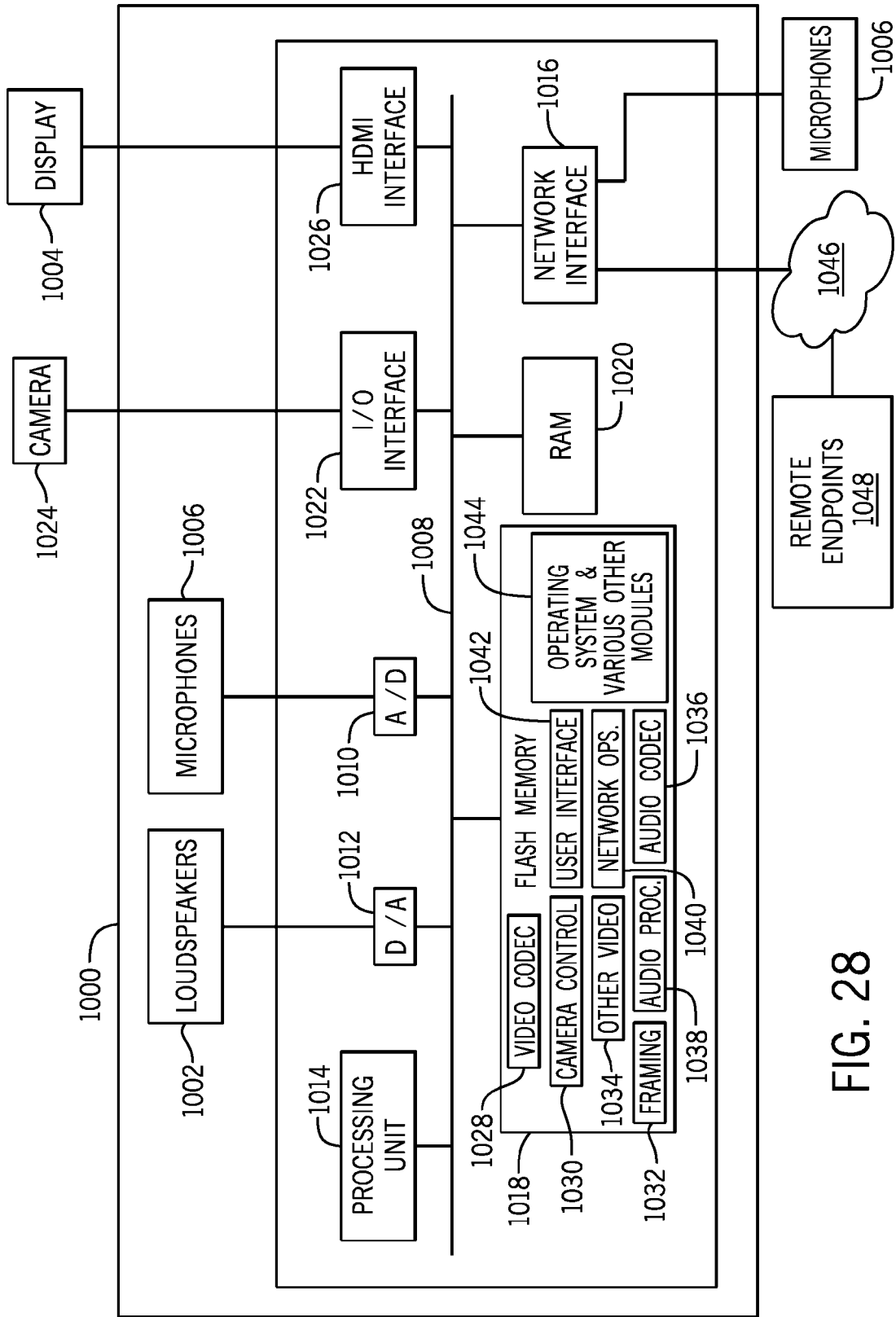


FIG. 28

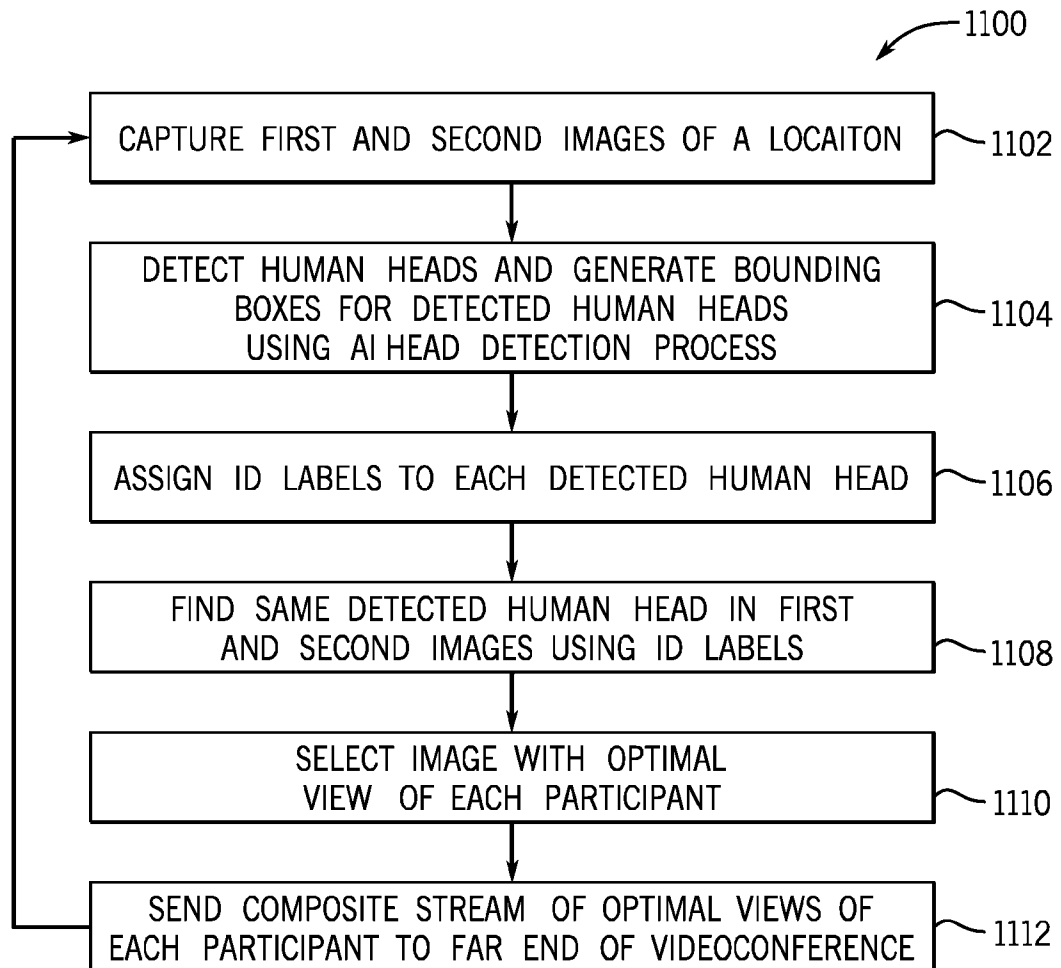


FIG. 29

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2023/075906

A. CLASSIFICATION OF SUBJECT MATTER INV. G06V20/52 G06V40/16 H04N7/15 H04N23/90 ADD.		
According to International Patent Classification (IPC) or to both national classification and IPC		
B. FIELDS SEARCHED Minimum documentation searched (classification system followed by classification symbols) G06V H04N		
Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched		
Electronic data base consulted during the international search (name of data base and, where practicable, search terms used) EPO-Internal, WPI Data		
C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 11 777 754 B1 (ZHANG KUI [US] ET AL) 3 October 2023 (2023-10-03) column 2, line 25 - line 59 column 3, line 48 - line 58 -----	1-7
☐ Further documents are listed in the continuation of Box C. ☒ See patent family annex.		
* Special categories of cited documents : "A" document defining the general state of the art which is not considered to be of particular relevance "E" earlier application or patent but published on or after the international filing date "L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified) "O" document referring to an oral disclosure, use, exhibition or other means "P" document published prior to the international filing date but later than the priority date claimed "T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention "X" document of particular relevance;; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone "Y" document of particular relevance;; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art "&" document member of the same patent family		
Date of the actual completion of the international search 29 February 2024		Date of mailing of the international search report 06/05/2024
Name and mailing address of the ISA/ European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016		Authorized officer Martinez, Francis

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2023/075906

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

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

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:
2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:
3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

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

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

see additional sheet

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.
2. ☐ As all searchable claims could be searched without effort justifying an additional fees, this Authority did not invite payment of additional fees.
3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:
4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims;; it is covered by claims Nos.:

1-7

Remark on Protest

- ☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.
- ☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.
- ☐ No protest accompanied the payment of additional search fees.

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/US2023/075906

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 11777754	B1	03-10-2023	NONE

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

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

1. claims: 1-7

reidentify human heads detected across different camera views

2. claims: 8-15

identify human heads detected across different camera views
