US010282852B1

(12) **United States Patent**
Buibas et al.

(10) **Patent No.: US 10,282,852 B1**
(45) **Date of Patent: May 7, 2019**

(54) **AUTONOMOUS STORE TRACKING SYSTEM**

(71) Applicant: **ACCEL ROBOTICS CORPORATION**, La Jolla, CA (US)

(72) Inventors: **Marius Buibas**, San Diego, CA (US); **John Quinn**, San Diego, CA (US); **Kaylee Feigum**, San Diego, CA (US); **Csaba Petre**, San Diego, CA (US); **Michael Brandon Maseda**, San Diego, CA (US); **Martin Alan Cseh**, San Diego, CA (US)

(73) Assignee: **ACCEL ROBOTICS CORPORATION**, La Jolla, CA (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) Appl. No.: **16/254,776**

(22) Filed: **Jan. 23, 2019**

Related U.S. Application Data

(63) Continuation-in-part of application No. 16/138,278, filed on Sep. 21, 2018, which is a continuation-in-part of application No. 16/036,754, filed on Jul. 16, 2018.

(51) **Int. Cl.**
G06K 9/00 (2006.01)
G06T 7/80 (2017.01)
(Continued)

(52) **U.S. Cl.**
CPC ***G06T 7/248*** (2017.01); ***G06K 9/00201*** (2013.01); ***G06K 9/00369*** (2013.01);
(Continued)

(58) **Field of Classification Search**
CPC G06T 7/248; G06T 7/292; G06T 7/80; G06T 2207/20081; G06T 2207/20084; G06K 9/00201; G06K 9/00369
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

6,193,154 B1 2/2001 Phillips et al.
6,364,206 B1 4/2002 Keohane
(Continued)

OTHER PUBLICATIONS

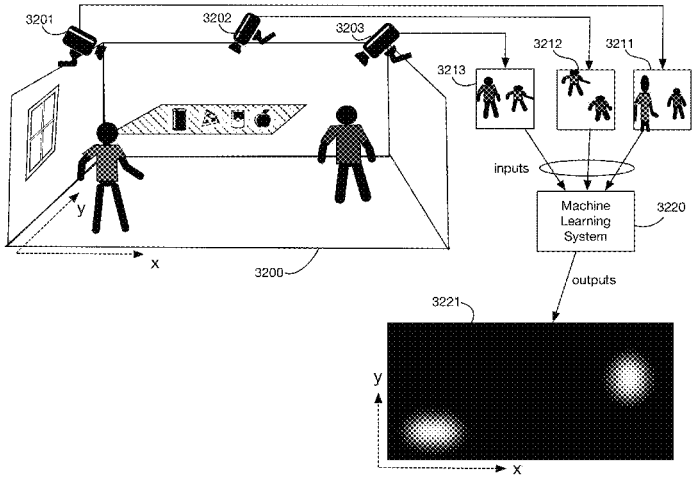
Swift Local Solutions, "What is frictionless shopping and how will it impact your retail business", Apr. 25. 2017 (6 pages).
(Continued)

Primary Examiner — Katrina R Fujita
(74) *Attorney, Agent, or Firm* — Arc IP Law, PC; Joseph J. Mayo

(57) **ABSTRACT**

A system that analyzes camera images to track a person in an autonomous store, and to determine when a tracked person takes or moves items in the store. The system may associate a field of influence volume around a person's location; intersection of this volume with an item storage area, such as a shelf, may trigger the system to look for changes in the items on the shelf. Items that are taken from, placed on, or moved on a shelf may be determined by a neural network that processes before and after images of the shelf. Person tracking may be performed by analyzing images from fisheye ceiling cameras projected onto a plane horizontal to the floor. Projected ceiling camera images may be analyzed using a neural network trained to recognize shopper locations. The autonomous store may include modular ceiling and shelving fixtures that contain cameras, lights, processors, and networking.

20 Claims, 52 Drawing Sheets
(25 of 52 Drawing Sheet(s) Filed in Color)



(51) **Int. Cl.**
G06T 7/246 (2017.01)
G06T 7/292 (2017.01)

(52) **U.S. Cl.**
CPC ***G06T 7/292*** (2017.01); ***G06T 7/80*** (2017.01); *G06T 2207/20081* (2013.01); *G06T 2207/20084* (2013.01)

(56) **References Cited**

U.S. PATENT DOCUMENTS

6,507,279 B2 * 1/2003 Loof G06Q 20/20 235/375
6,882,900 B1 4/2005 Terranova
7,436,887 B2 * 10/2008 Yeredor G06K 9/00771 375/240.01
8,300,890 B1 10/2012 Gaikwad et al.
9,064,161 B1 * 6/2015 Boman G06K 7/01
9,805,266 B2 * 10/2017 Zhang H04N 7/18
9,911,290 B1 3/2018 Zalewski
9,996,818 B1 6/2018 Ren et al.
1,004,492 A1 8/2018 Bradski et al.
10,127,438 B1 * 11/2018 Fisher G06K 9/00335
10,133,933 B1 * 11/2018 Fisher G06K 9/00718
2006/0187305 A1 * 8/2006 Trivedi G06K 9/00234 348/169
2006/0279630 A1 12/2006 Aggarwal et al.
2007/0282665 A1 12/2007 Buehler et al.
2008/0181507 A1 7/2008 Gope et al.
2009/0057068 A1 3/2009 Lin et al.
2009/0324010 A1 * 12/2009 Hou G06K 9/00771 382/103
2011/0317016 A1 12/2011 Saeki et al.
2012/0113294 A1 5/2012 Oyabu et al.
2014/0132728 A1 * 5/2014 Verano G06K 9/00771 348/46
2015/0039458 A1 2/2015 Reid
2015/0199890 A1 * 7/2015 Hewett G08B 13/2451 340/572.1
2015/0294483 A1 * 10/2015 Wells H04N 7/181 382/103
2015/0294496 A1 * 10/2015 Medasani H04N 7/181 348/159
2015/0347846 A1 * 12/2015 Guzm N-Rivera G06K 9/6256 382/103
2016/0092739 A1 * 3/2016 Oami G06K 9/00771 348/159
2016/0358145 A1 12/2016 Montgomery
2017/0309136 A1 10/2017 Schoner
2017/0323376 A1 * 11/2017 Glaser G06Q 30/0643
2018/0052228 A1 * 2/2018 Markison A61B 5/02055
2018/0150685 A1 * 5/2018 Ebrom G06K 9/00335

OTHER PUBLICATIONS

Stores NRF's Magazine, "Standard Cognition makes frictionless checkout a reality", Jan. 23, 2018 (3 pages).
Porikli, Fatih, "Inter-Camera Color Calibration by Correlation Model Function", Mitsubishi Electric Research Lab, Cambridge, MA, 2003, 4 pages.

* cited by examiner

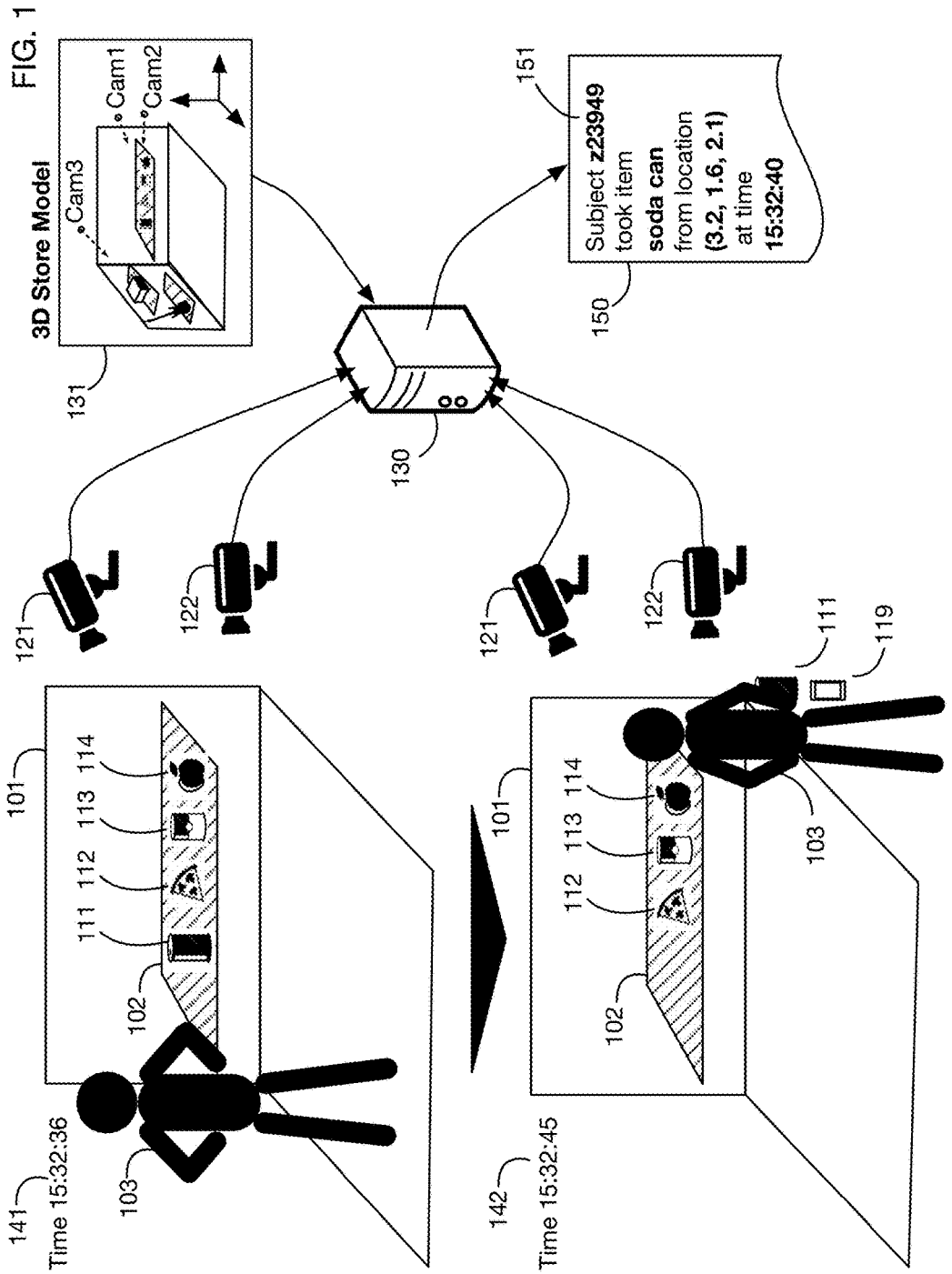

3D Store Model
Cam1
Cam2
Cam3
131
130
121
122
121
122
Subject z23949
took item
soda can
from location
(3.2, 1.6, 2.1)
at time
15:32:40
150
151
101
102
103
111
112
113
114
119
141
Time 15:32:36
142
Time 15:32:45

FIG. 1

FIG. 2

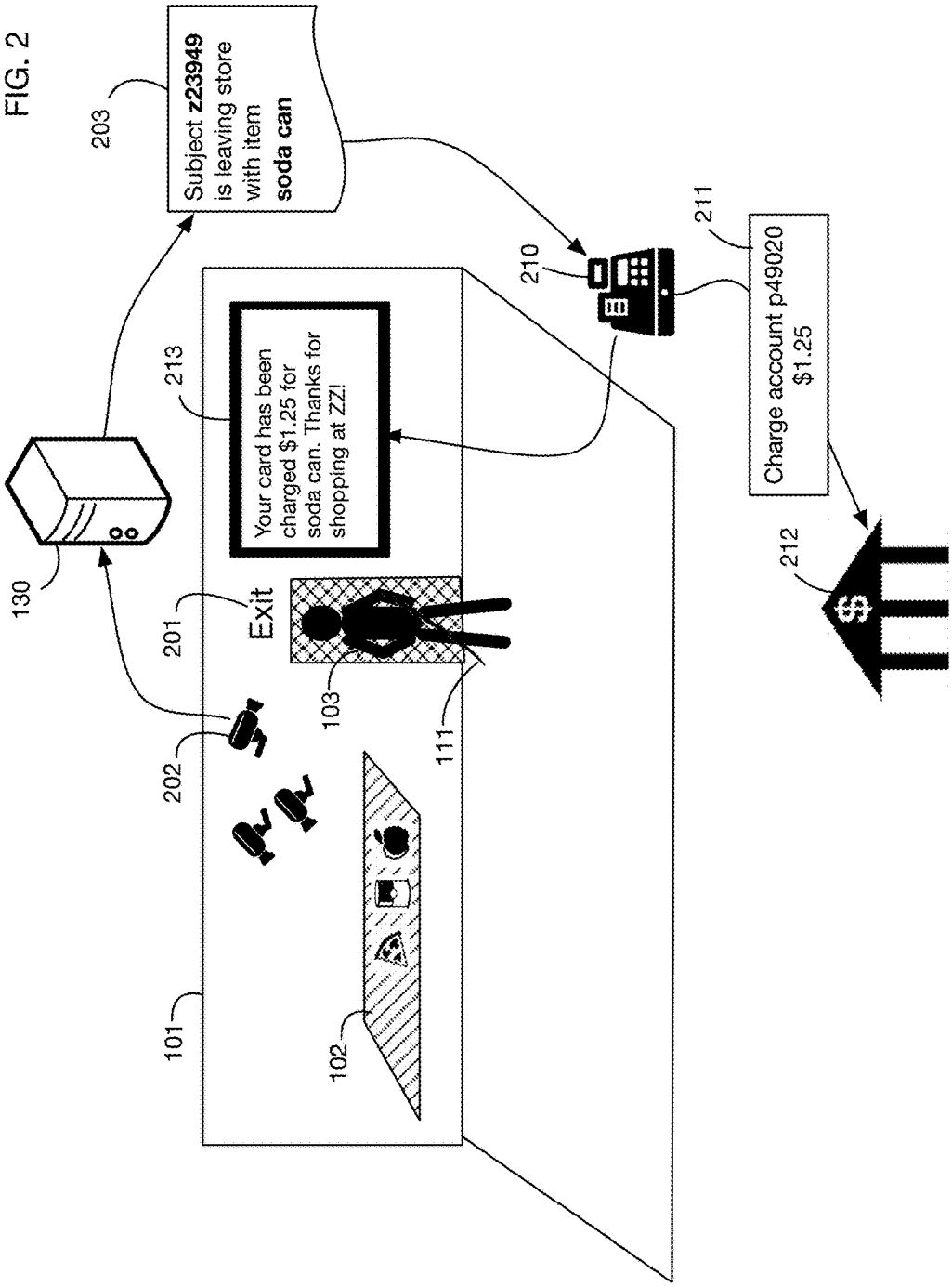
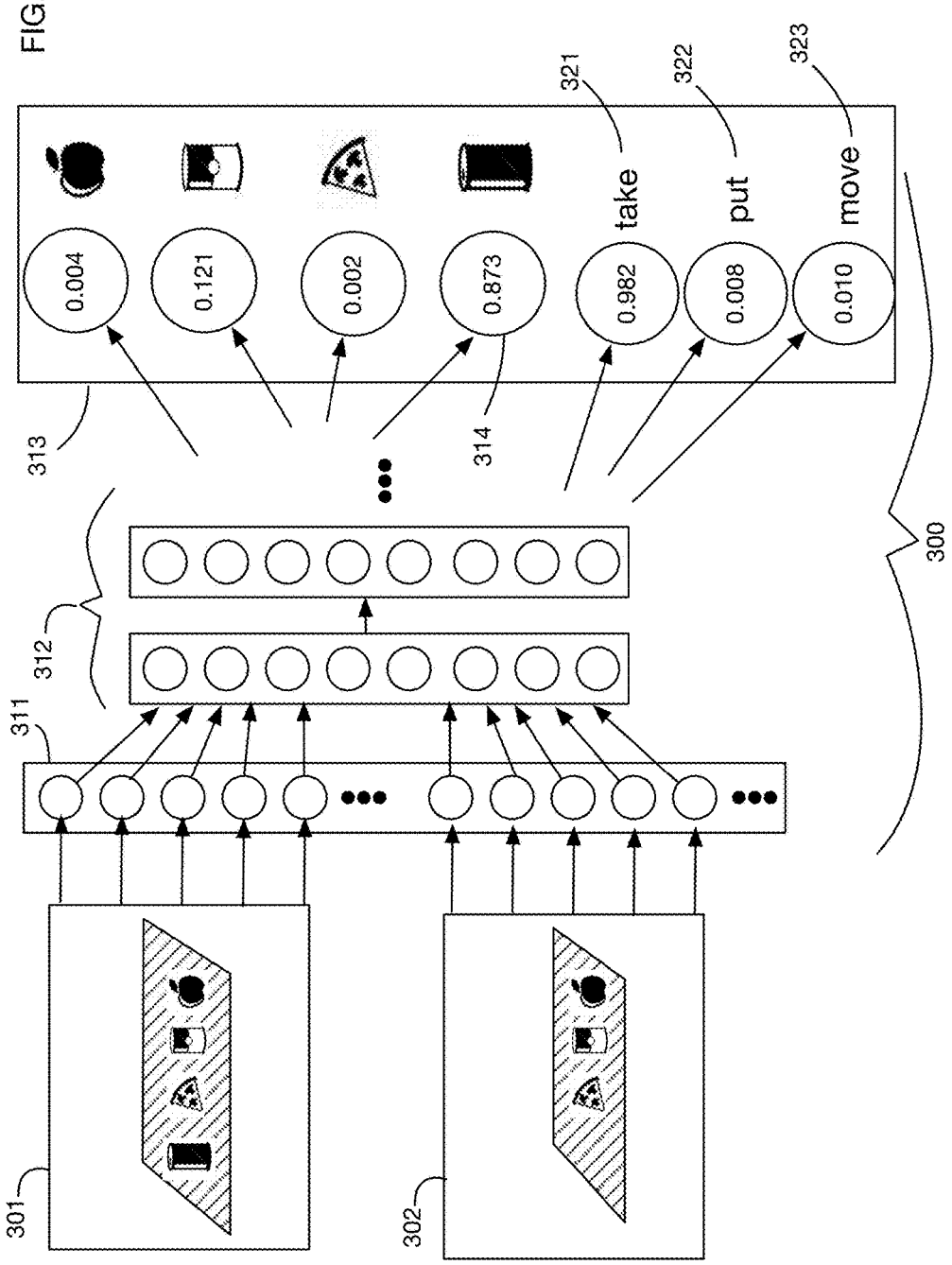

Subject **z23949** is leaving store with item **soda can**
203
130
Your card has been charged $1.25 for soda can. Thanks for shopping at ZZ!
213
210
Charge account p49020 $1.25
211
212
Exit
201
202
103
111
101
102

FIG.3


0.004
0.121
0.002
0.873
0.982
0.008
0.010
take
put
move
313
314
321
322
323
312
311
300
301
302

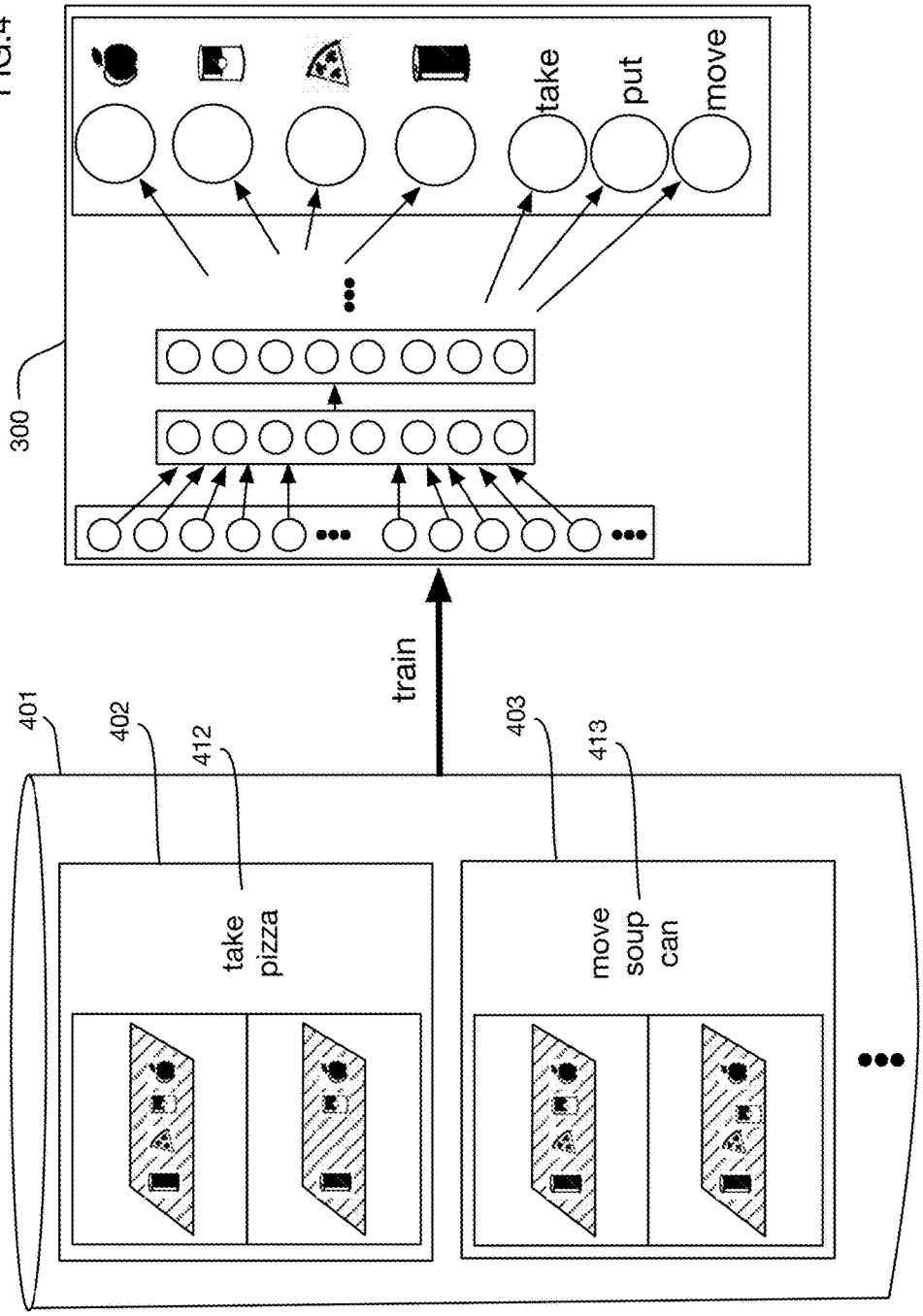

300
take
put
move
train
401
402
412
take pizza
403
413
move soup can

FIG.4

FIG.4A

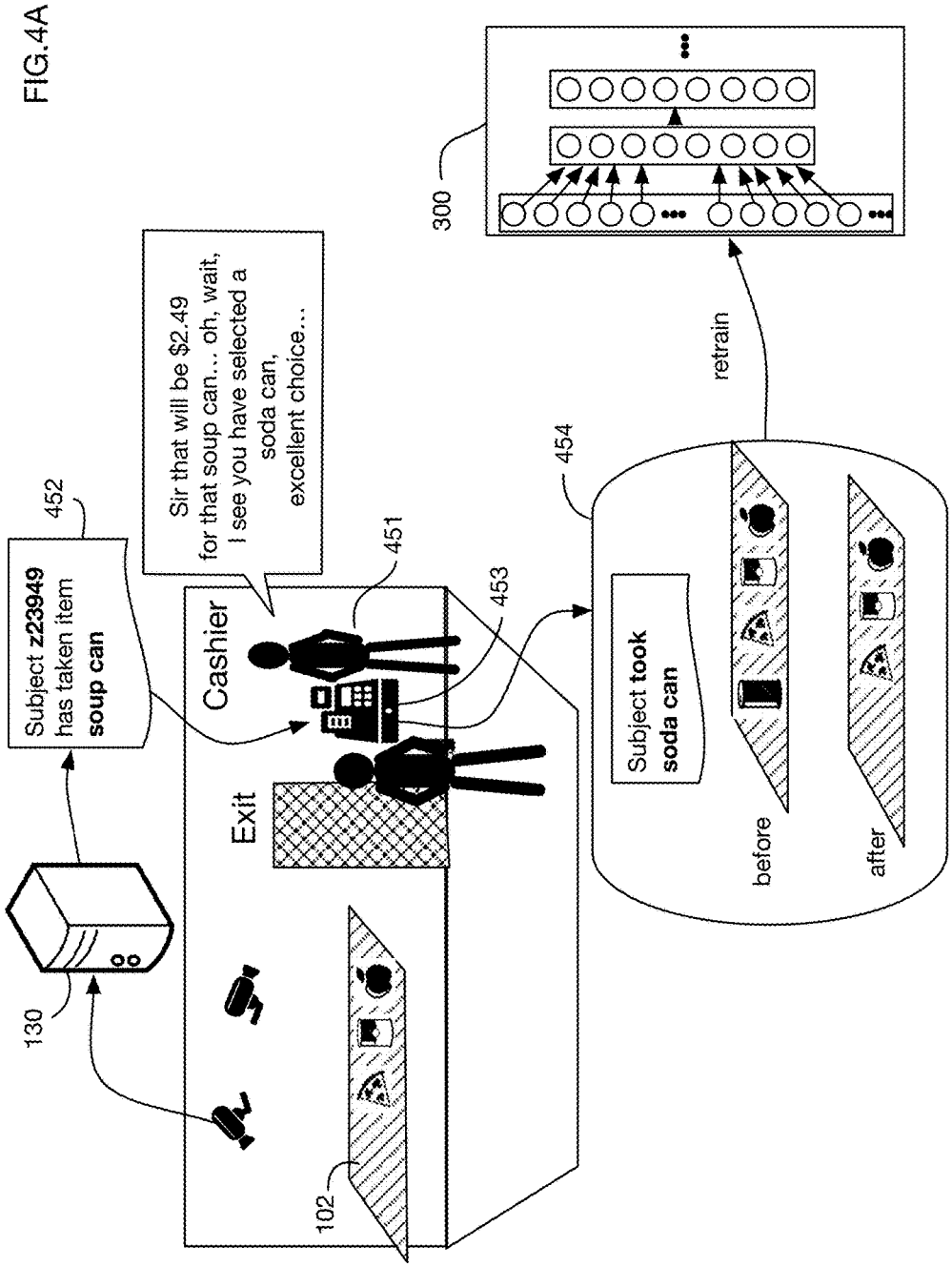
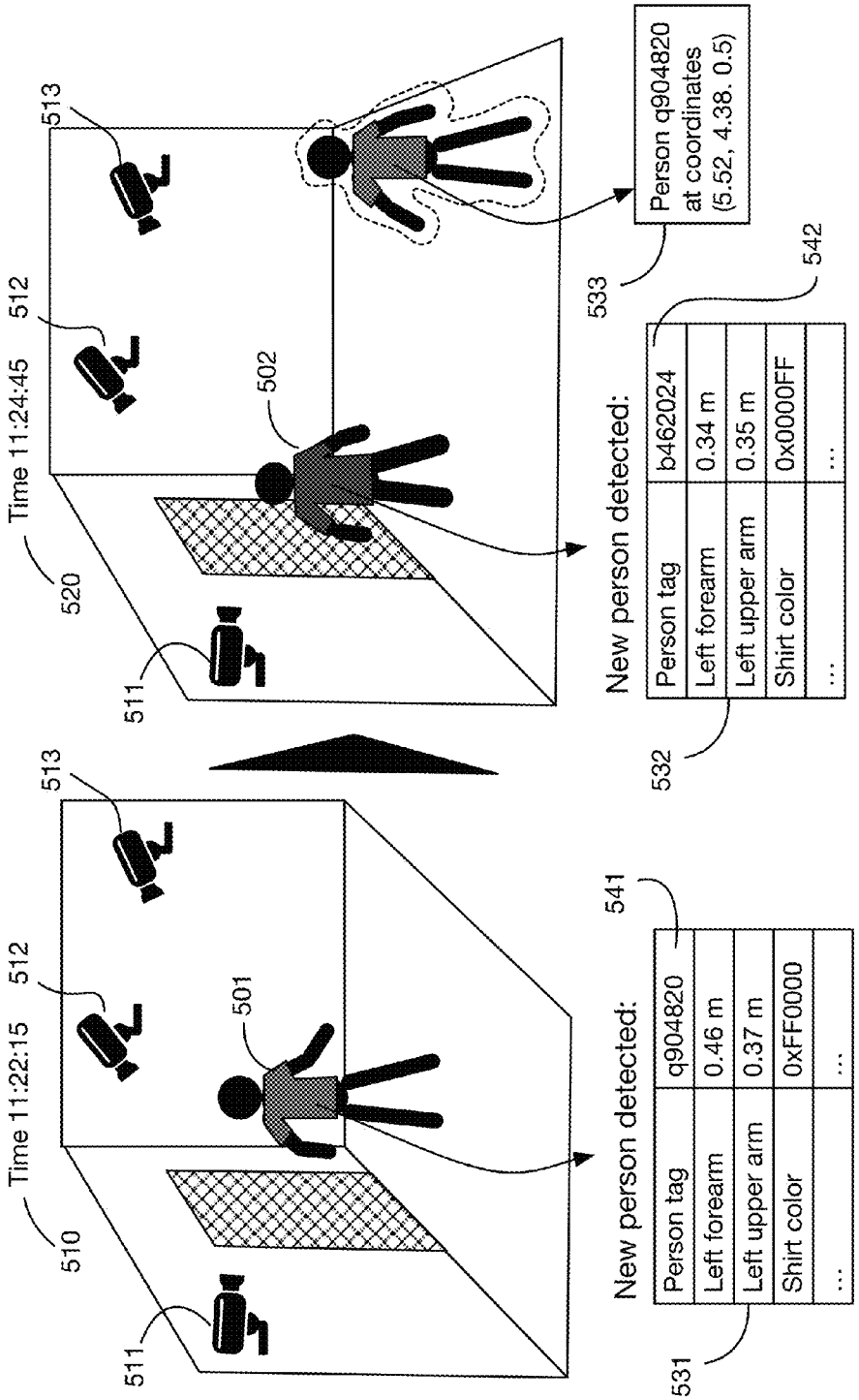

300
retrain
454
Subject took soda can
before
after
Sir that will be $2.49 for that soup can… oh, wait, I see you have selected a soda can, excellent choice…
451
453
Cashier
Exit
452
Subject z23949 has taken item soup can
130
102

FIG. 5


Time 11:22:15
510
511
512
513
501
New person detected:
Person tag q904820
Left forearm 0.46 m
Left upper arm 0.37 m
Shirt color 0xFF0000
...
531
541
Time 11:24:45
520
511
512
513
502
New person detected:
Person tag b462024
Left forearm 0.34 m
Left upper arm 0.35 m
Shirt color 0x0000FF
...
532
542
Person q904820 at coordinates (5.52, 4.38. 0.5)
533

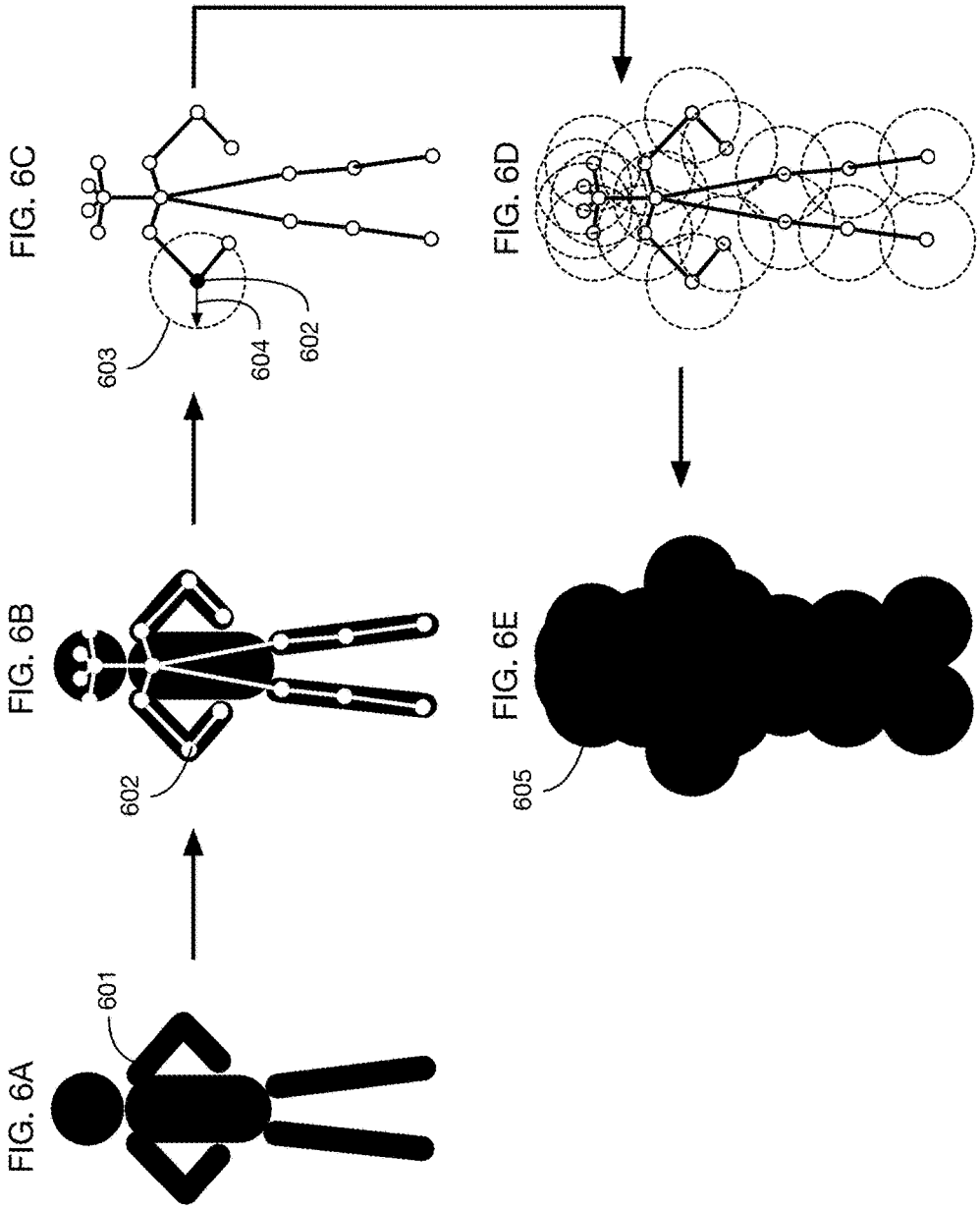


FIG. 6A
601
FIG. 6B
602
FIG. 6C
603
604
602
FIG. 6D
FIG. 6E
605

FIG. 7B

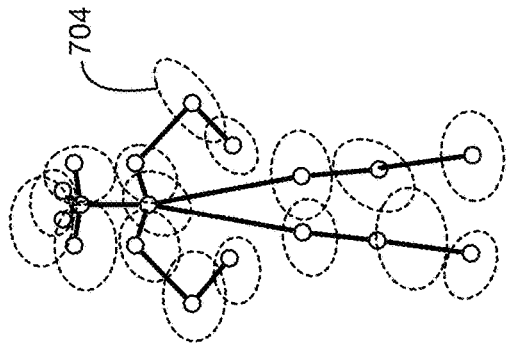


FIG. 7A

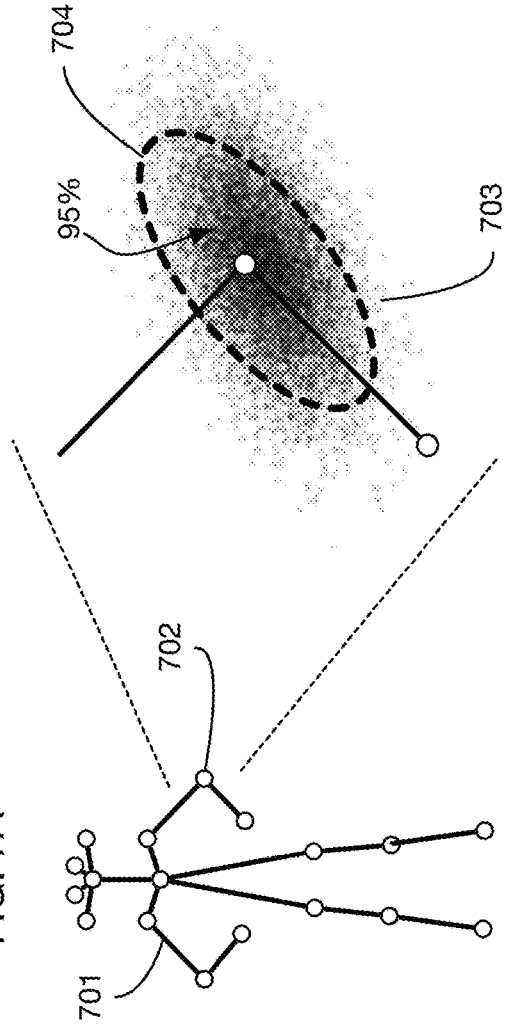


FIG. 8

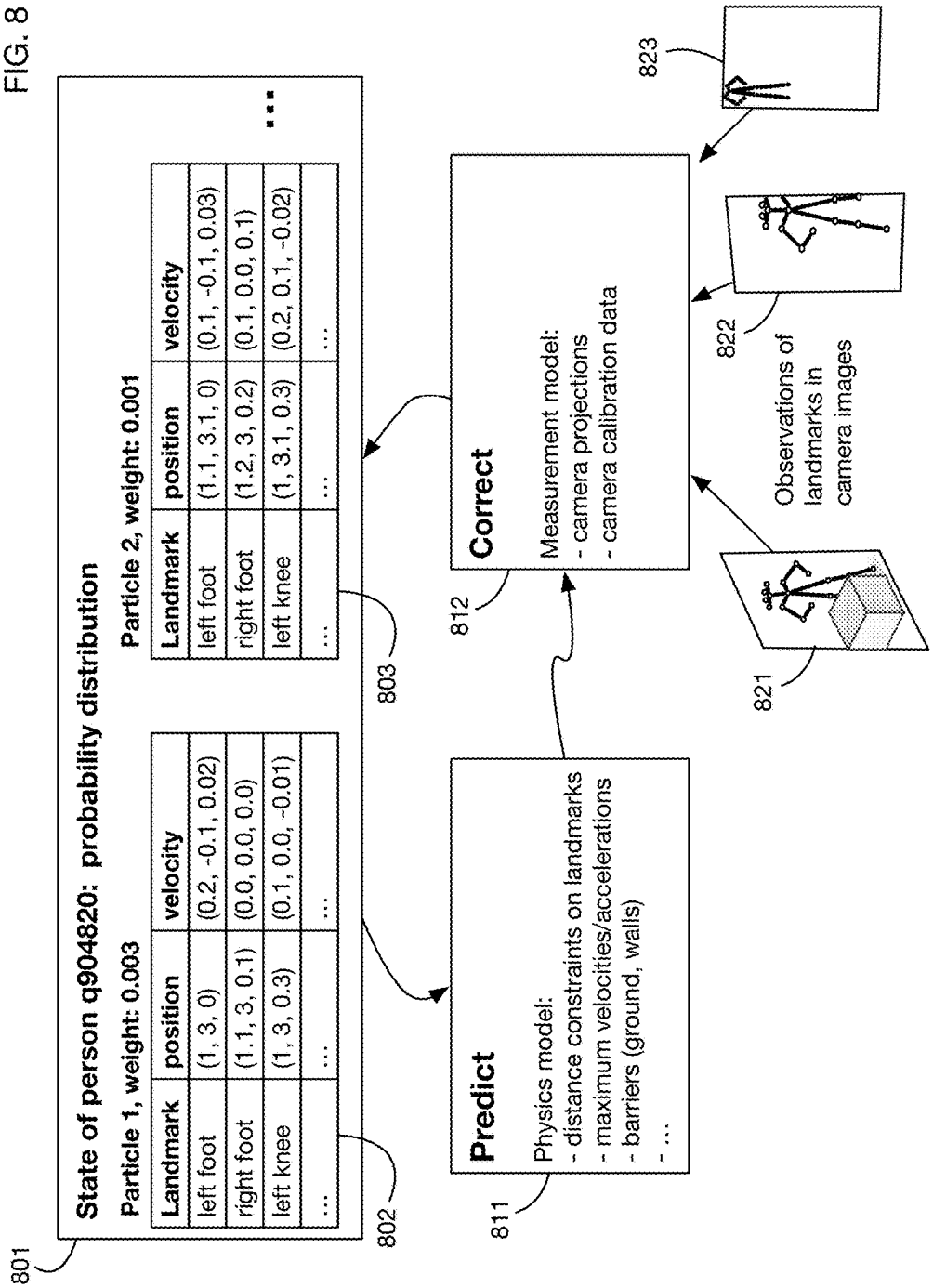

801
State of person q904820: probability distribution
Particle 1, weight: 0.003
Landmark | position | velocity
left foot | (1, 3, 0) | (0.2, -0.1, 0.02)
right foot | (1.1, 3, 0.1) | (0.0, 0.0, 0.0)
left knee | (1, 3, 0.3) | (0.1, 0.0, -0.01)
... | ... | ...
802
Particle 2, weight: 0.001
Landmark | position | velocity
left foot | (1.1, 3.1, 0) | (0.1, -0.1, 0.03)
right foot | (1.2, 3, 0.2) | (0.1, 0.0, 0.1)
left knee | (1, 3.1, 0.3) | (0.2, 0.1, -0.02)
... | ... | ...
803
...
Predict
Physics model:
- distance constraints on landmarks
- maximum velocities/accelerations
- barriers (ground, walls)
- ...
811
Correct
Measurement model:
- camera projections
- camera calibration data
812
821
822
Observations of landmarks in camera images
823

FIG. 9

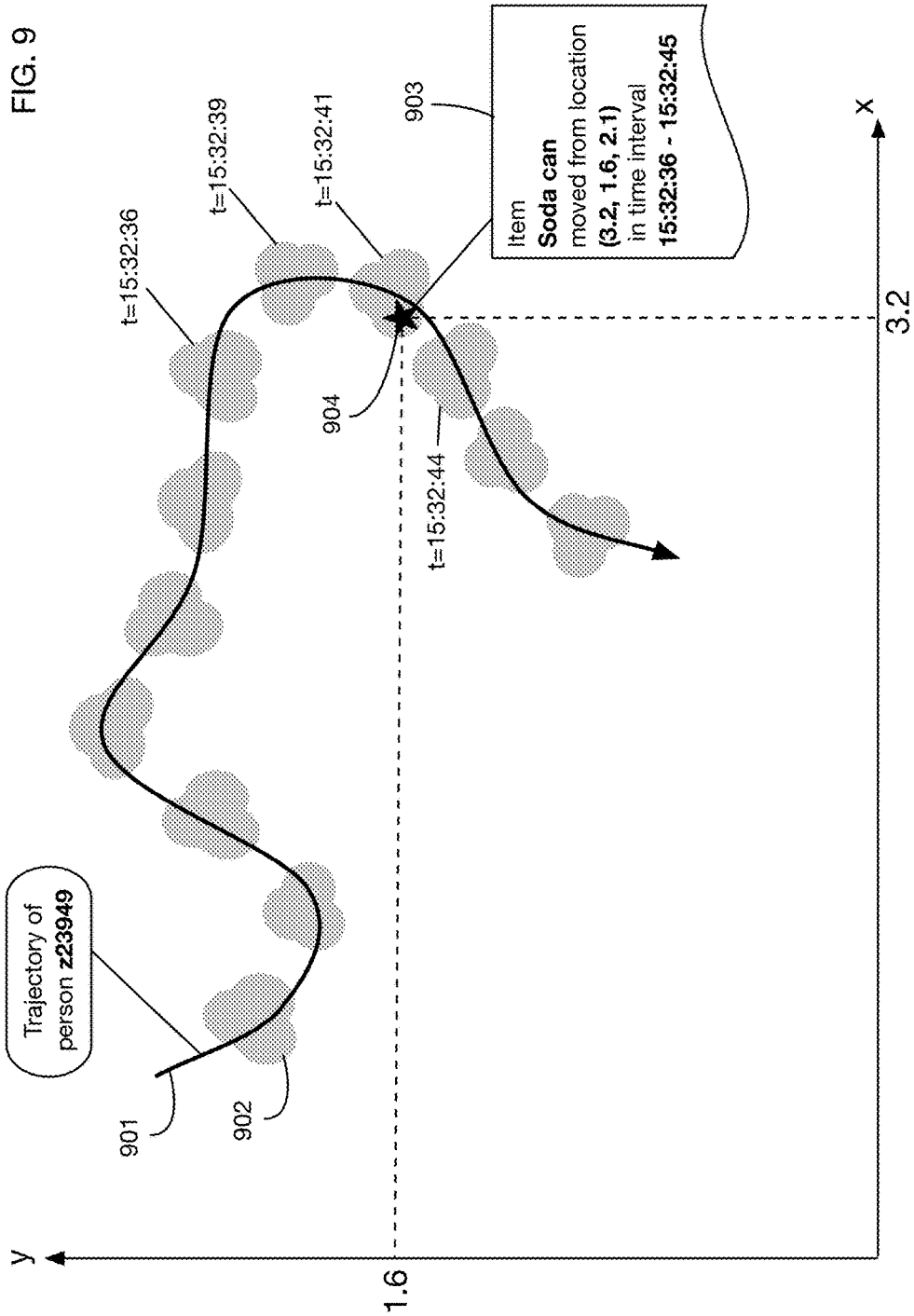

Trajectory of person **z23949**
901
902
903
904
t=15:32:36
t=15:32:39
t=15:32:41
t=15:32:44
Item
Soda can
moved from location
(3.2, 1.6, 2.1)
in time interval
15:32:36 - 15:32:45
y
x
1.6
3.2

FIG. 10

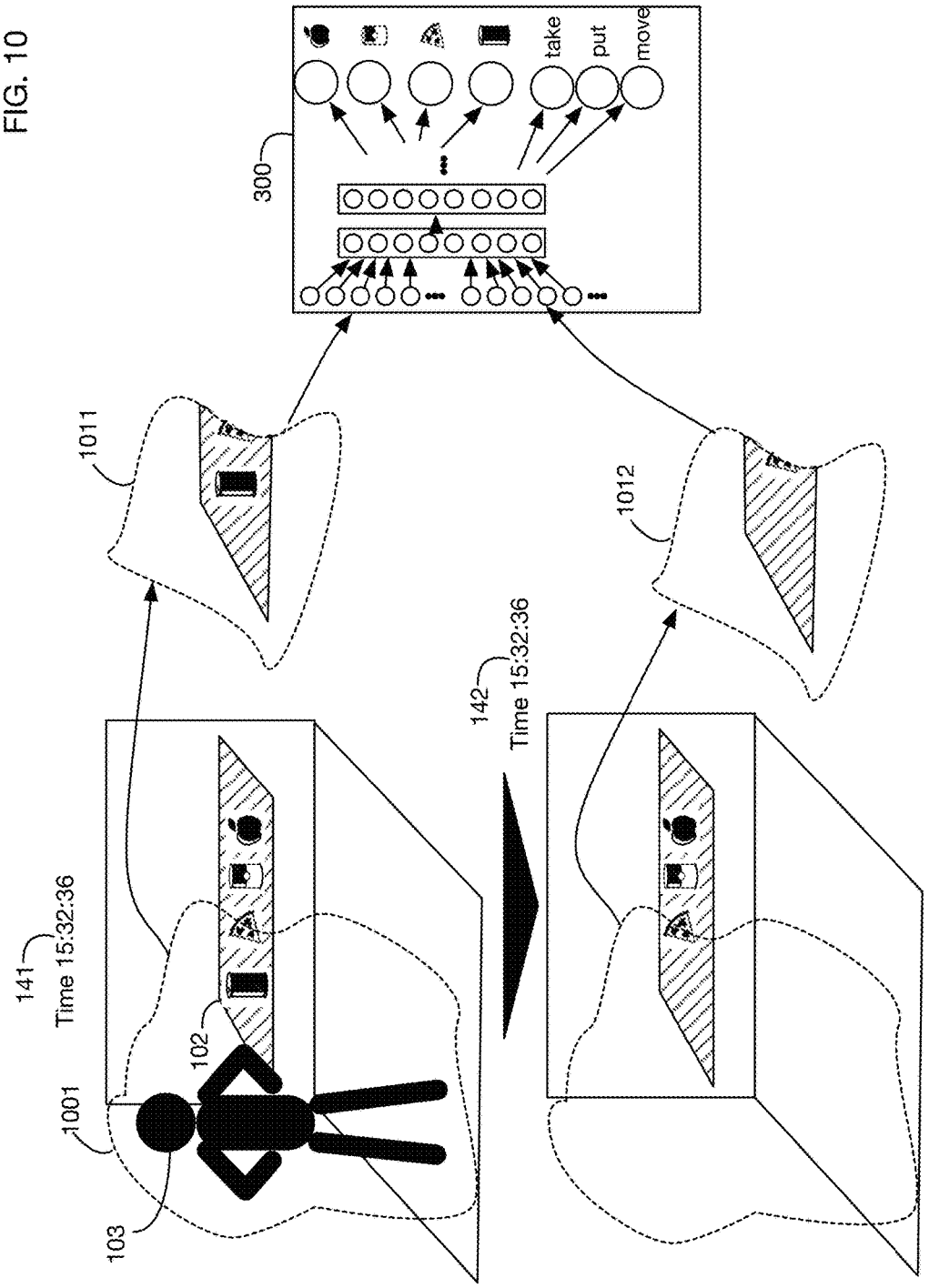

300
take
put
move
1011
1012
141
Time 15:32:36
142
Time 15:32:36
1001
102
103

FIG. 11

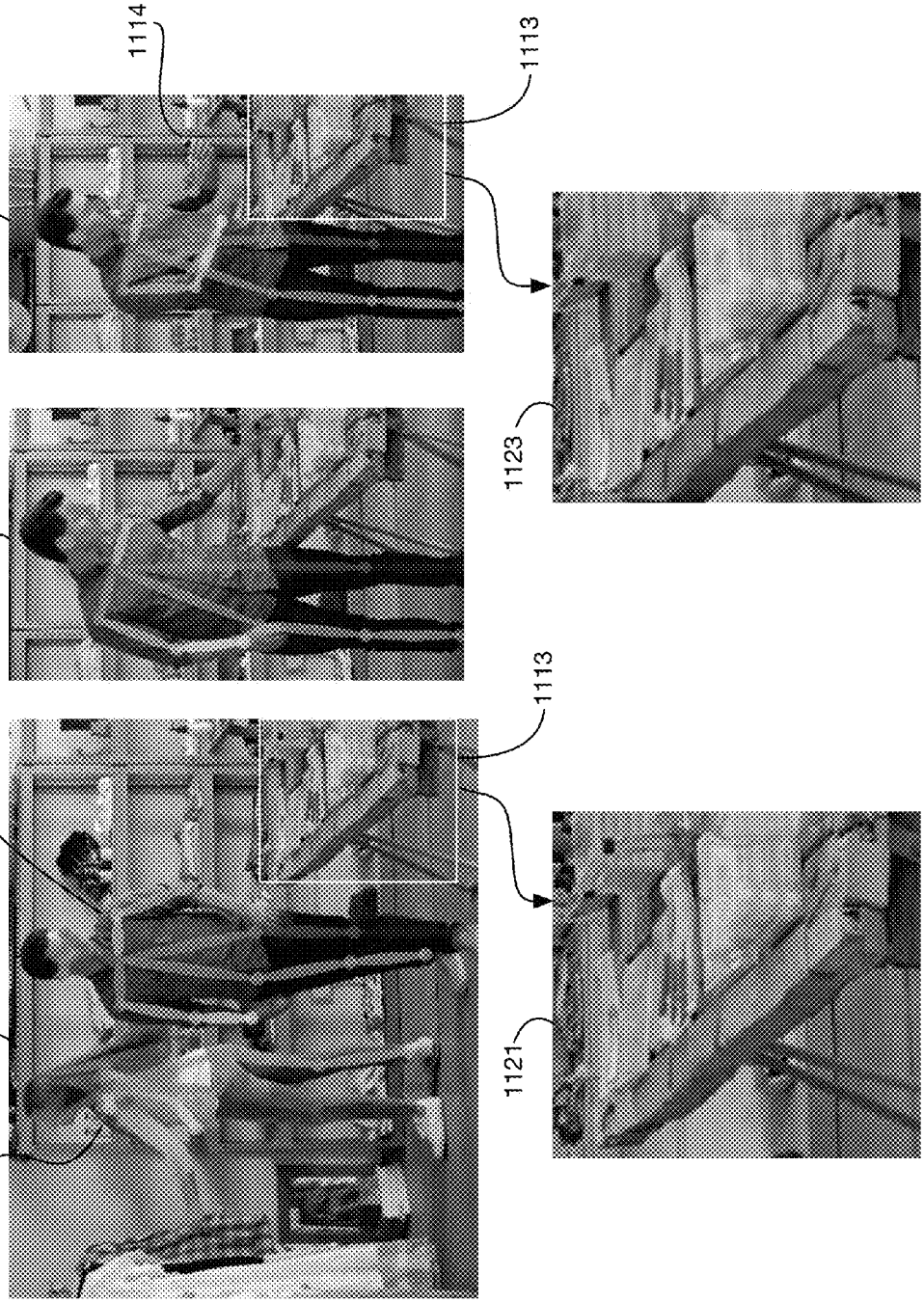


FIG. 12

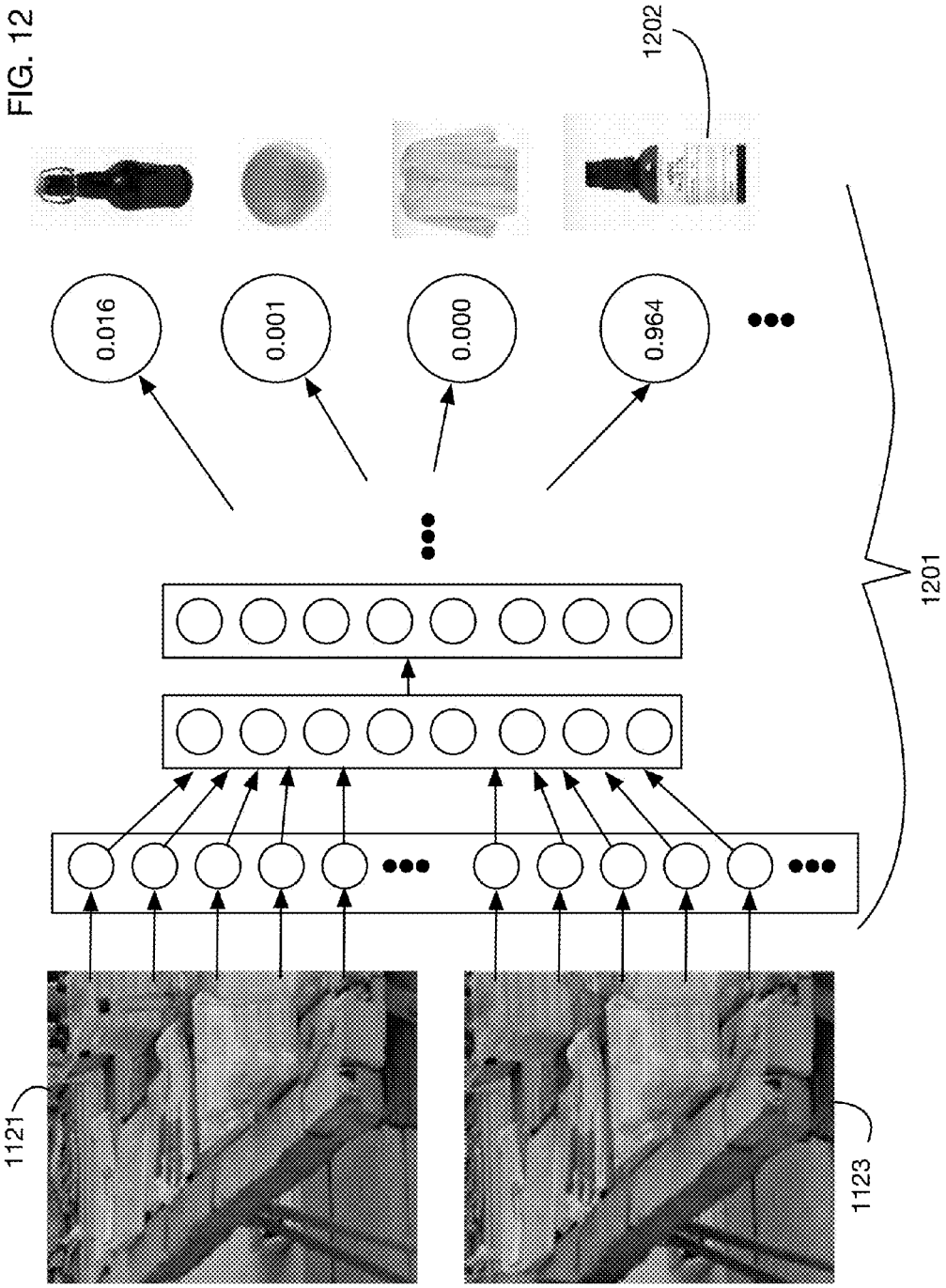


FIG. 13

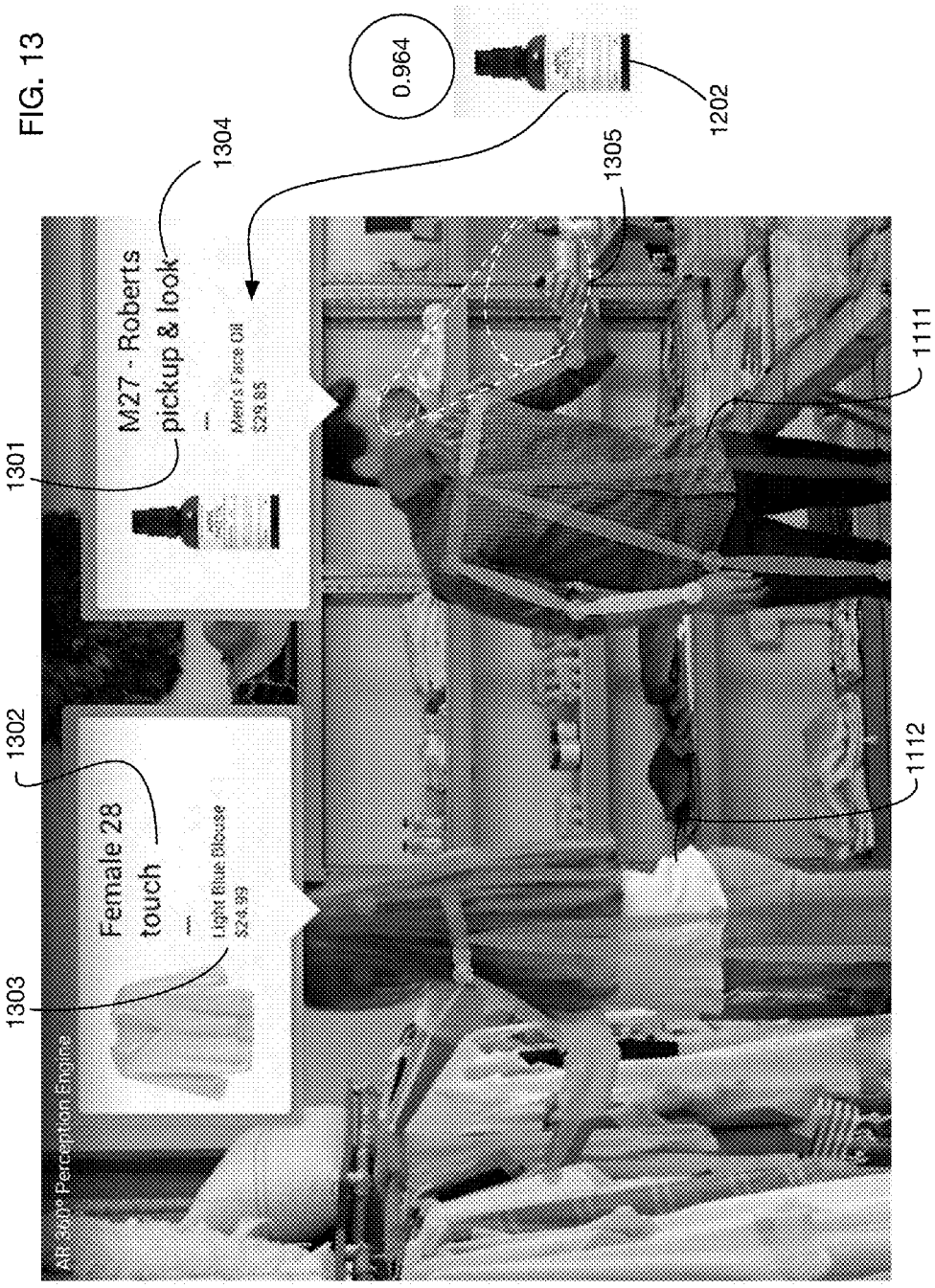

1301
M27 - Roberts pickup & look
Men's Face Oil
$29.85
1302
Female 28 touch
Light Blue Blouse
$24.99
1303
1304
0.964
1202
1305
1111
1112

FIG. 14

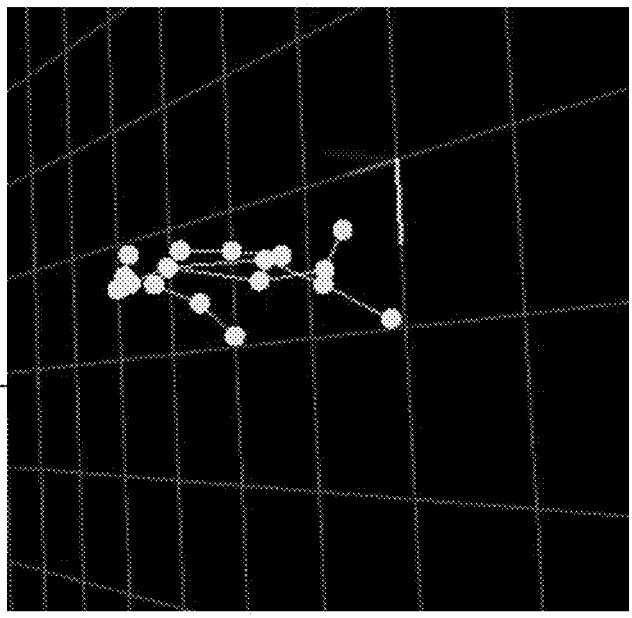


FIG. 15

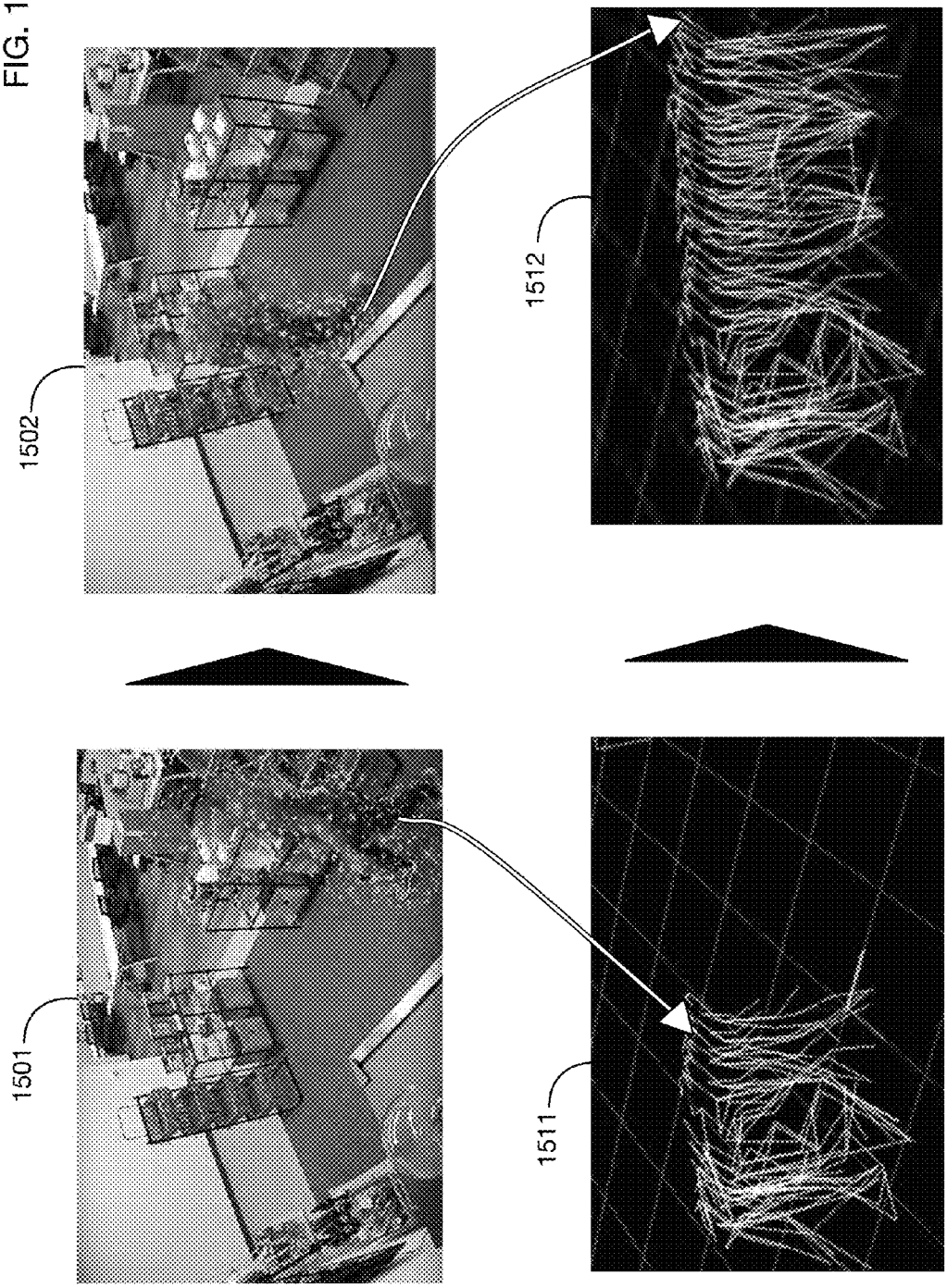


FIG. 16

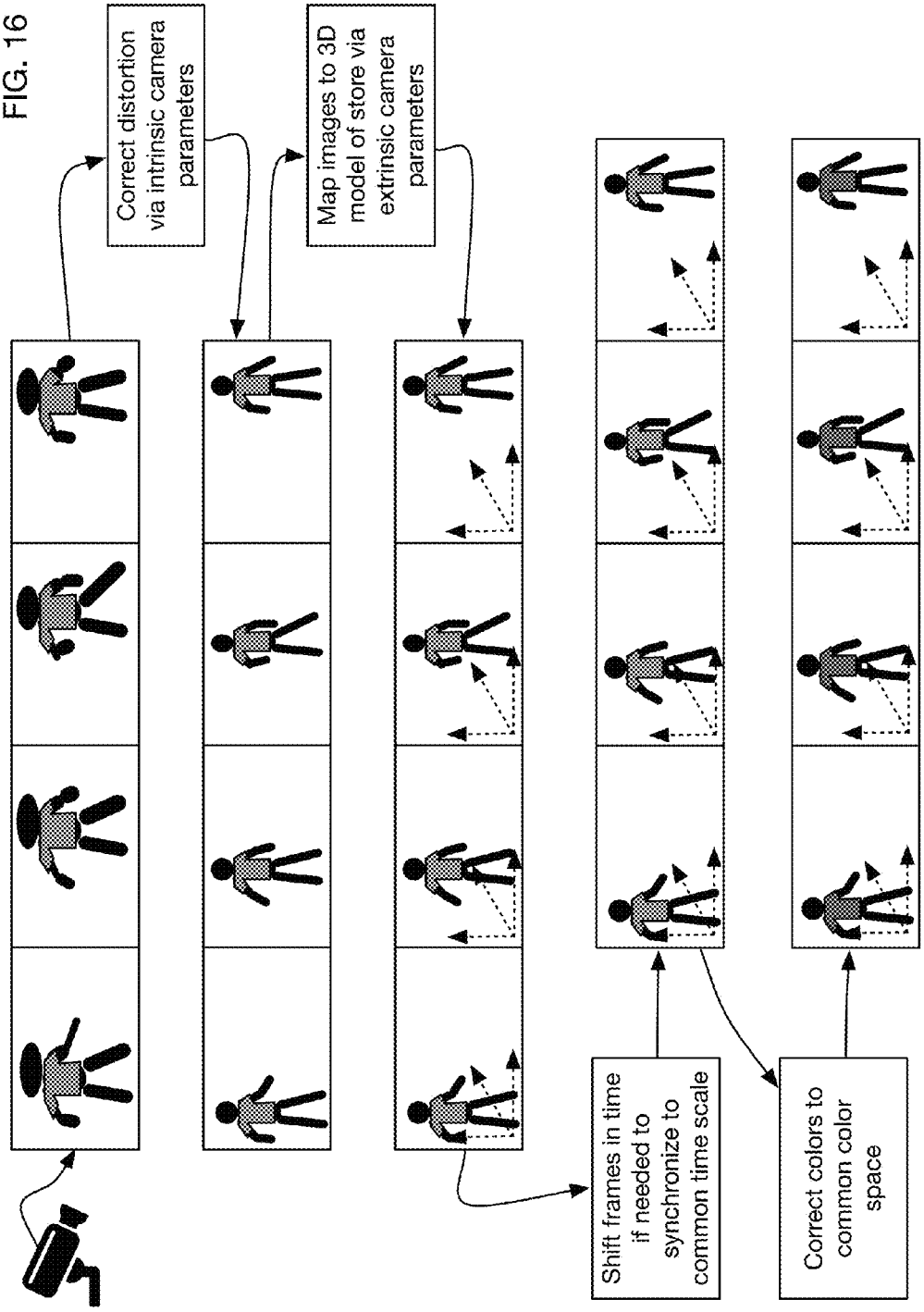

Correct distortion via intrinsic camera parameters
Map images to 3D model of store via extrinsic camera parameters
Shift frames in time if needed to synchronize to common time scale
Correct colors to common color space

FIG. 17

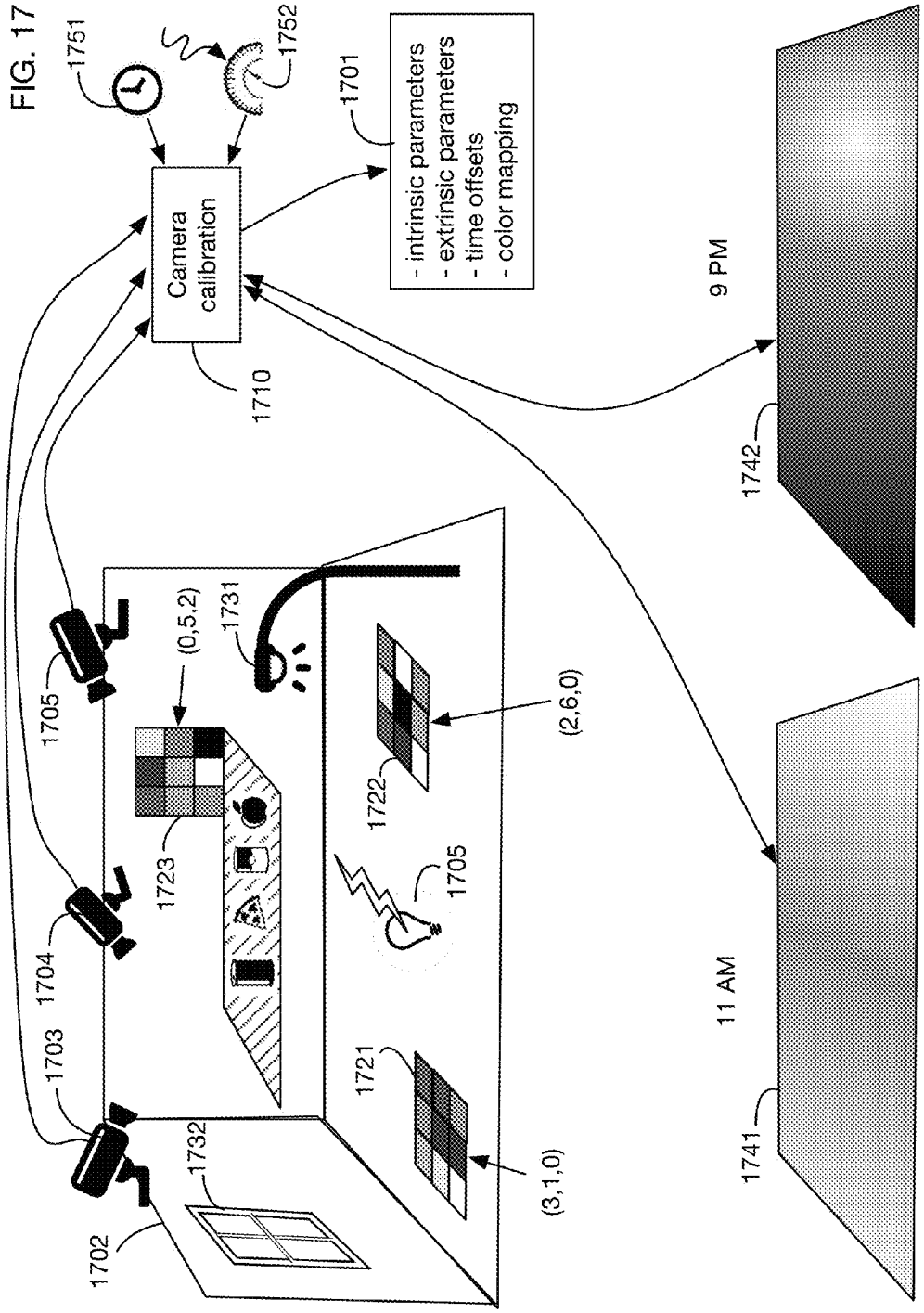
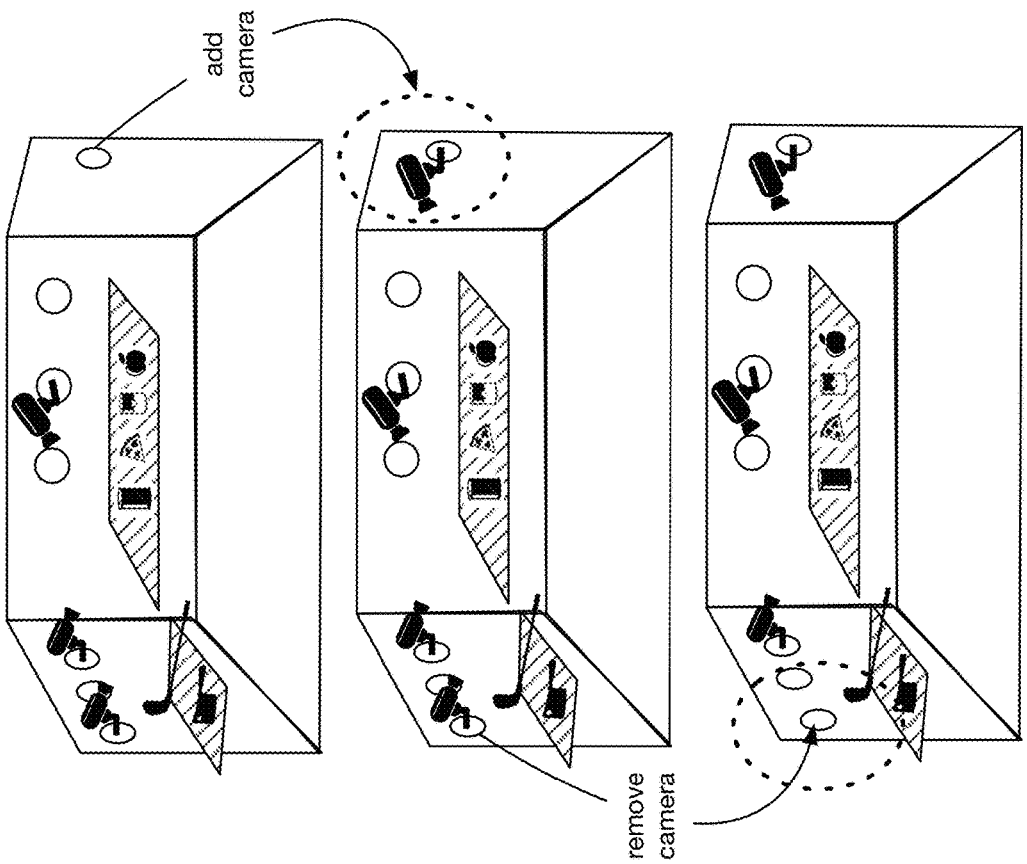

1751
1752
Camera calibration
1710
1701
- intrinsic parameters
- extrinsic parameters
- time offsets
- color mapping
9 PM
1742
11 AM
1741
1705
1704
1703
1702
(0,5,2)
1731
1723
1722
(2,6,0)
1705
1721
(3,1,0)
1732

FIG. 18



Item	Views	Cost
Pan	1	1000
Golf club	1	1000
Soda can	2	0
Pizza	2	0
Soup can	2	0
Apple	2	0
Cameras		3
Total cost		**2003**

Item	Views	Cost
Pan	2	0
Golf club	2	0
Soda can	3	0
Pizza	3	0
Soup can	3	0
Apple	3	0
Cameras		4
Total cost		**4**

Item	Views	Cost
Pan	2	0
Golf club	2	0
Soda can	2	0
Pizza	2	0
Soup can	2	0
Apple	2	0
Cameras		3
Total cost		**3**

FIG. 19

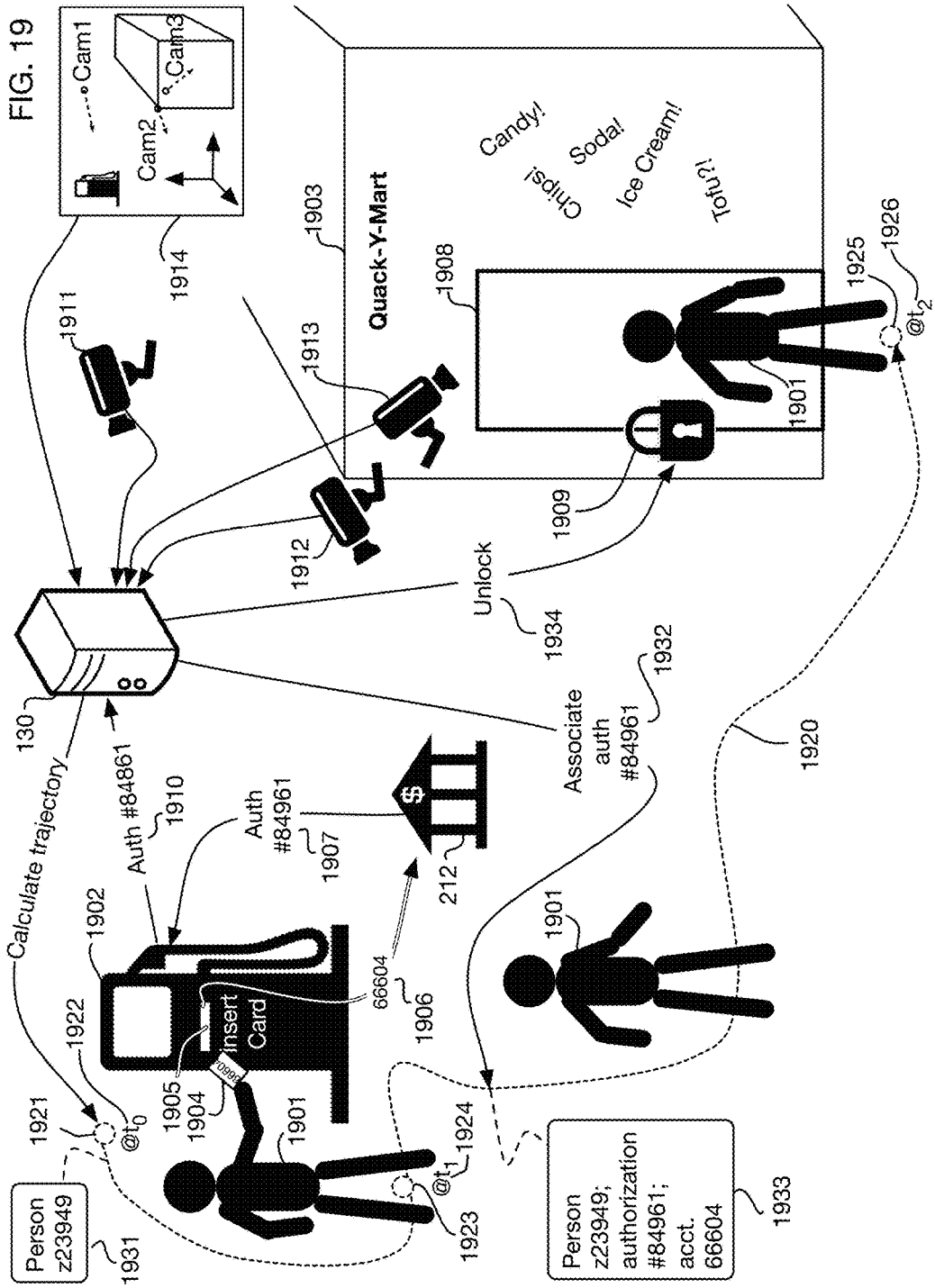

Cam1
Cam2
Cam3
1914
1911
1913
1912
1903
Quack-Y-Mart
1908
Candy!
Chips!
Soda!
Ice Cream!
Tofu?!
1909
1901
1925
1926
@t2
Unlock
1934
130
Calculate trajectory
Auth #84861
1910
Auth #84961
1907
212
Associate auth #84961
1932
1920
1902
1922
Insert Card
66604
1906
1905
1904
1901
1921
@t0
1924
@t1
1923
Person z23949
1931
Person z23949; authorization #84961; acct. 66604
1933

FIG. 20

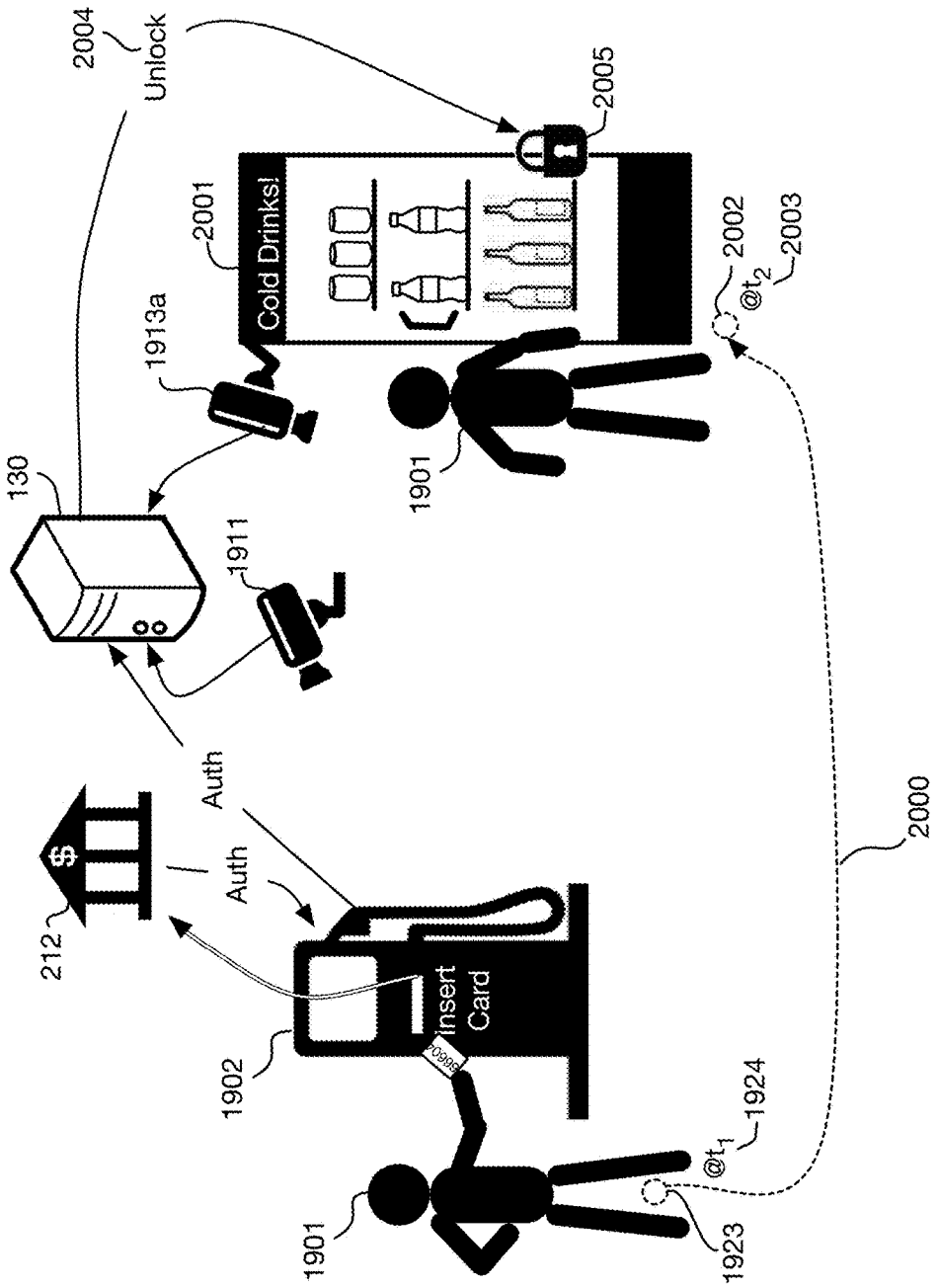

2004
Unlock
2005
2001
Cold Drinks!
1913a
130
1911
1901
2002
2003
@t2
Auth
Auth
212
1902
Insert
Card
1901
1924
@t1
1923
2000

FIG. 21

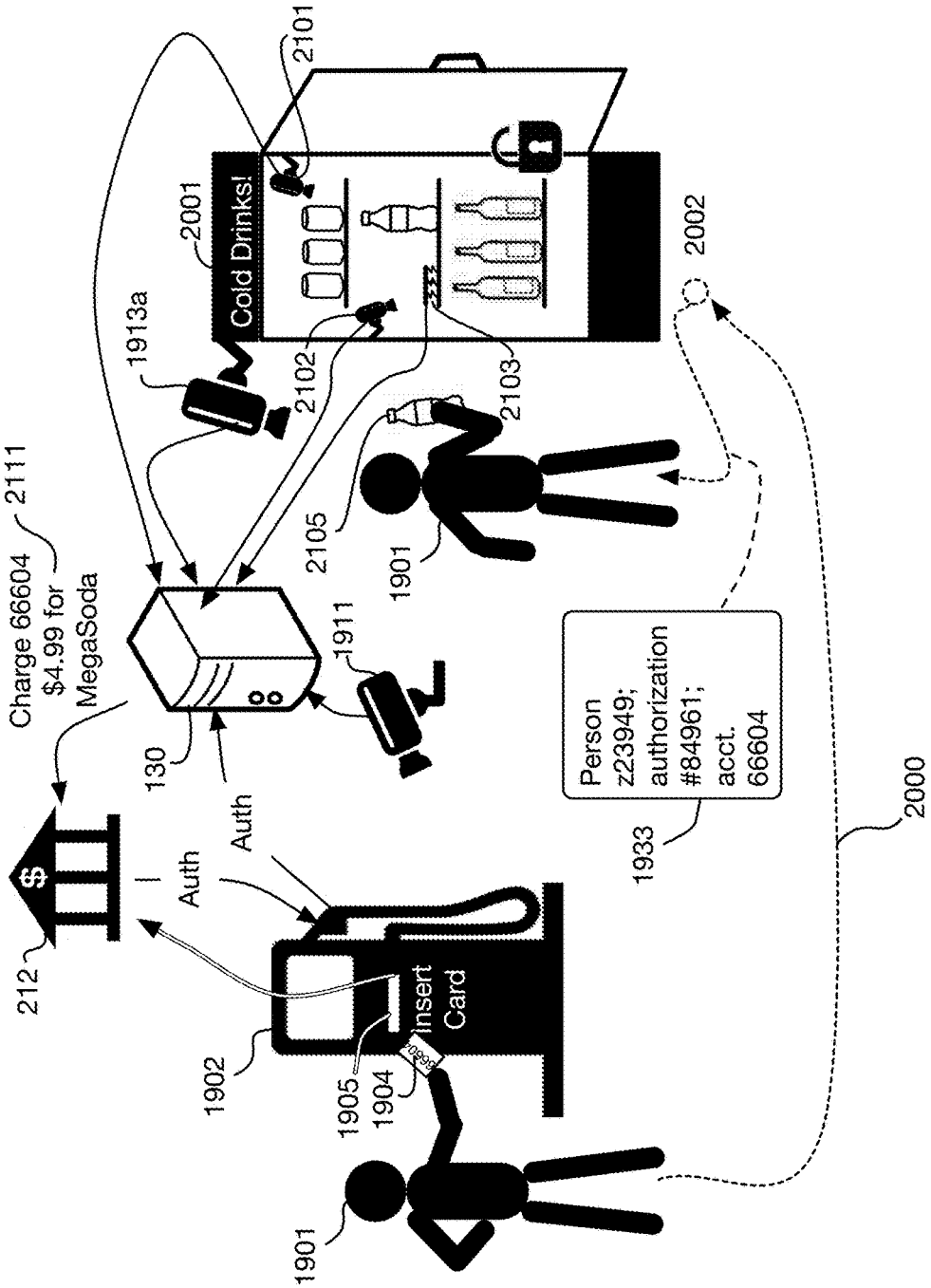

2101
2001
Cold Drinks!
2002
1913a
2102
2103
2105
1901
2111
Charge 66604
$4.99 for
MegaSoda
1911
130
Auth
Auth
Person
z23949;
authorization
#84961;
acct.
66604
1933
2000
212
1902
1905
1904
Insert
Card
1901

FIG. 22

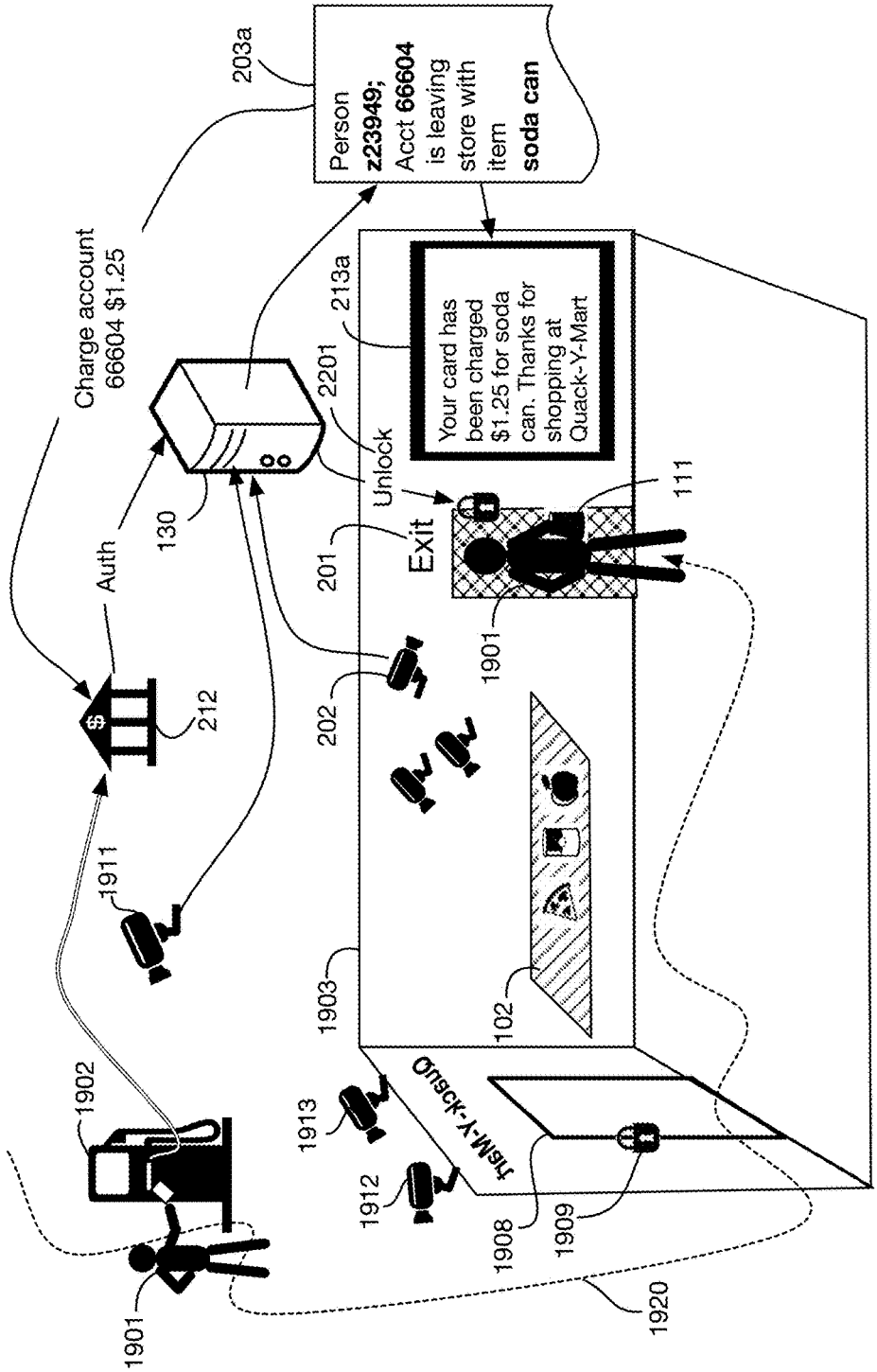

Charge account
66604 $1.25
203a
Person
z23949;
Acct 66604
is leaving
store with
item
soda can
213a
Your card has
been charged
$1.25 for soda
can. Thanks for
shopping at
Quack-Y-Mart
2201
Unlock
Exit
111
201
1901
130
Auth
212
202
1911
1903
102
1902
1913
1912
1908
1909
1920
1901

FIG. 23

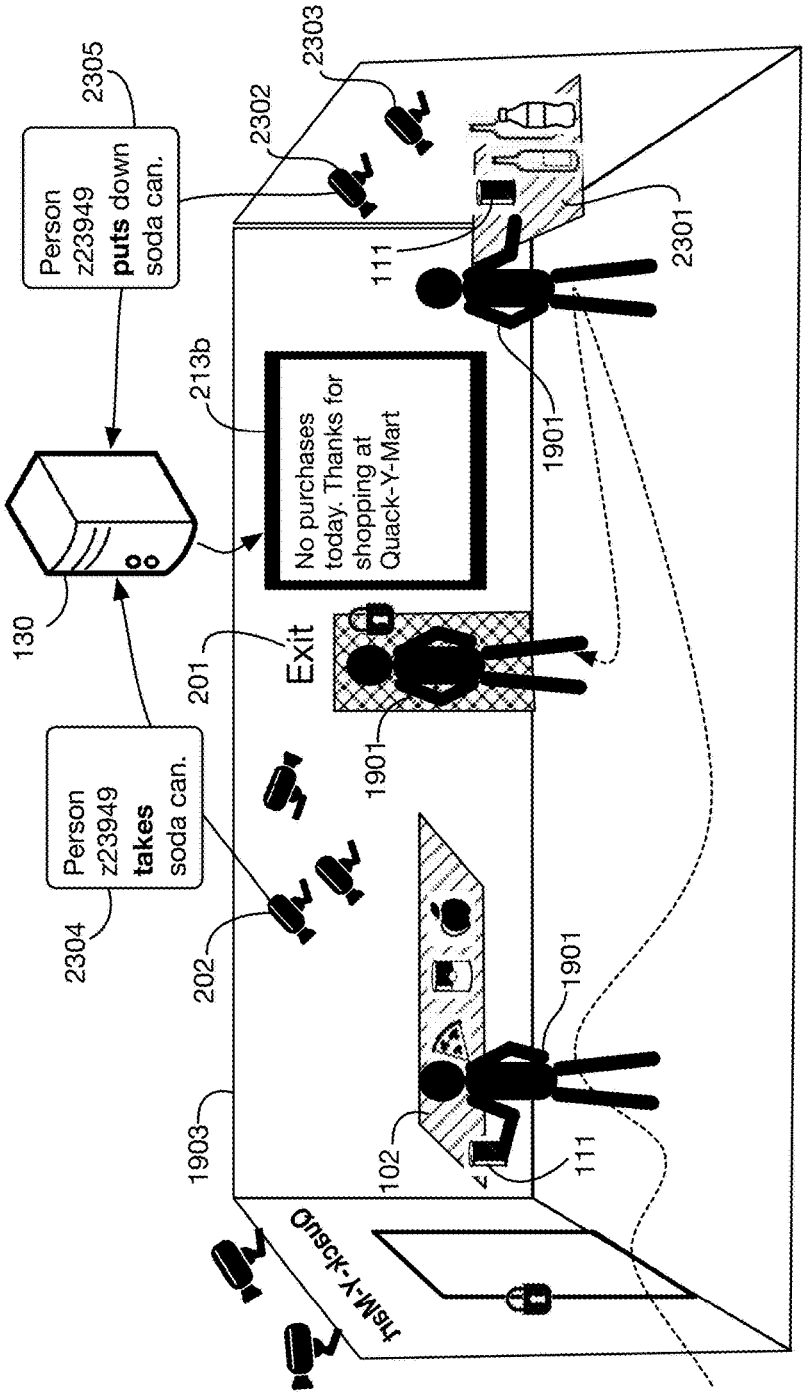
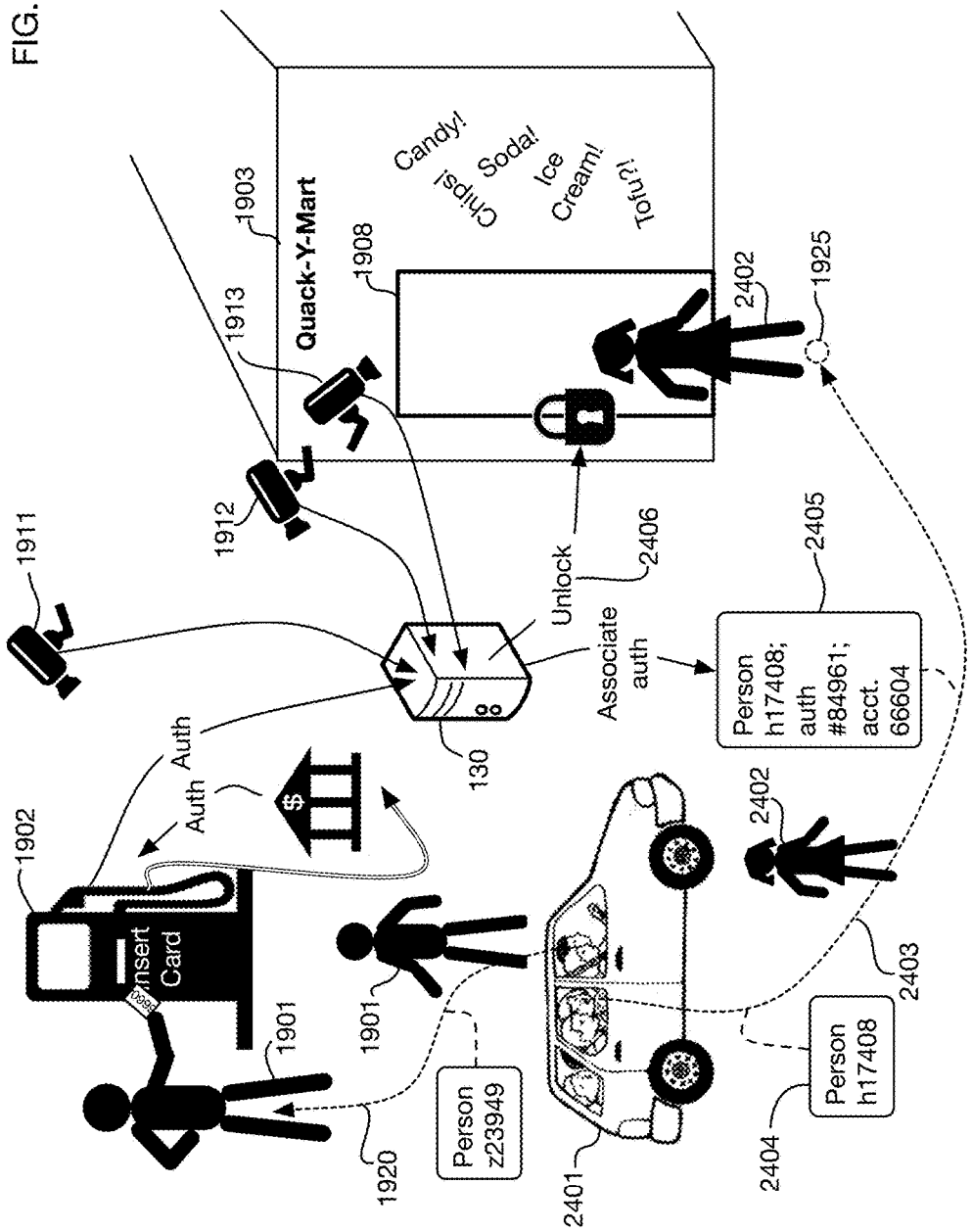

Person z23949 puts down soda can.
2305
2302
2303
111
2301
213b
No purchases today. Thanks for shopping at Quack-Y-Mart
1901
130
201
Exit
1901
Person z23949 takes soda can.
2304
202
1901
1903
102
111
Quack-Y-Mart

FIG. 24


1911
1902
Insert Card
Auth
Auth
1901
1901
1920
Person z23949
1912
1913
1903
Quack-Y-Mart
1908
Candy!
Chips!
Soda!
Ice Cream!
Tofu?!
130
Unlock
2406
Associate auth
2405
Person h17408; auth #84961; acct. 66604
2402
2402
1925
2401
2404
Person h17408
2403

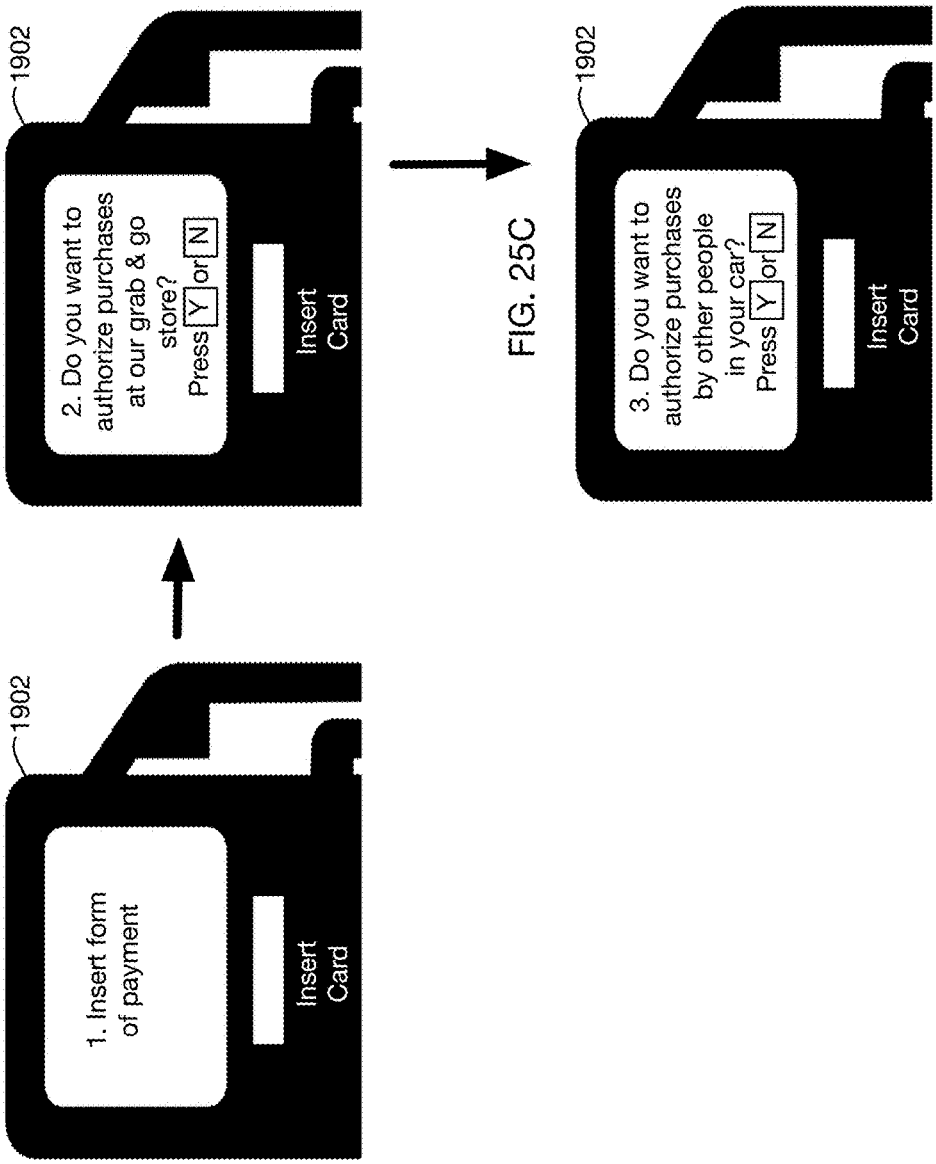


FIG. 25A

FIG. 25B

FIG. 25C

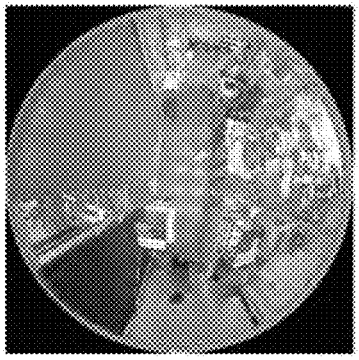


FIG. 26A

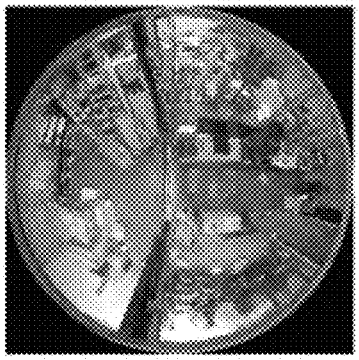


FIG. 26B

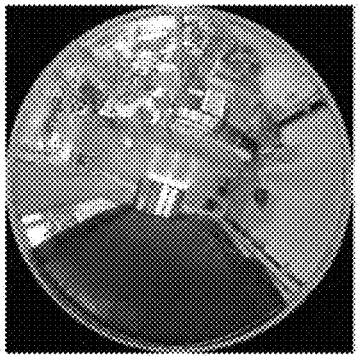


FIG. 26C

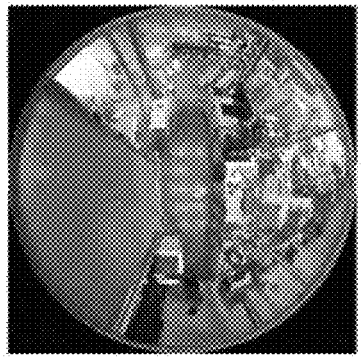


FIG. 26D

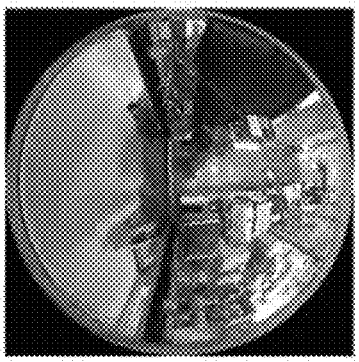


FIG. 26E

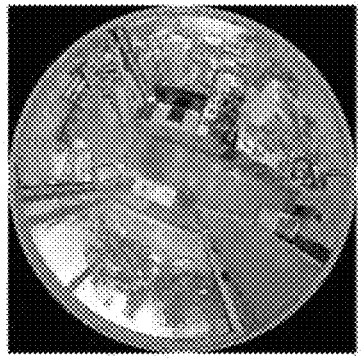


FIG. 26F

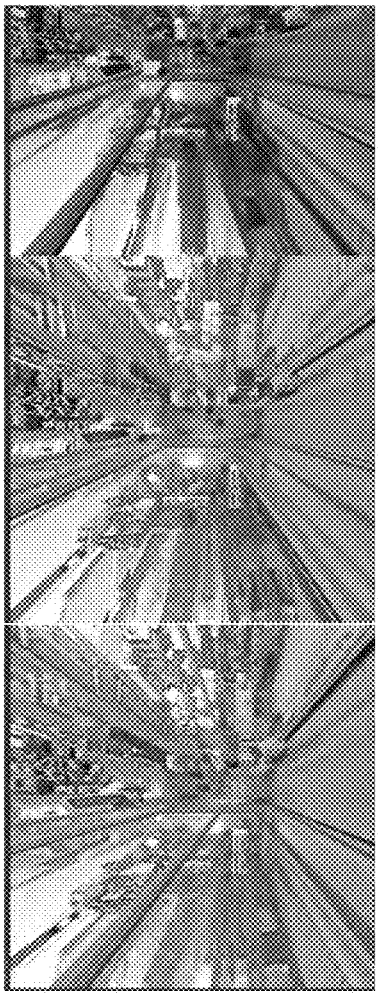


FIG. 27A FIG. 27B FIG. 27C

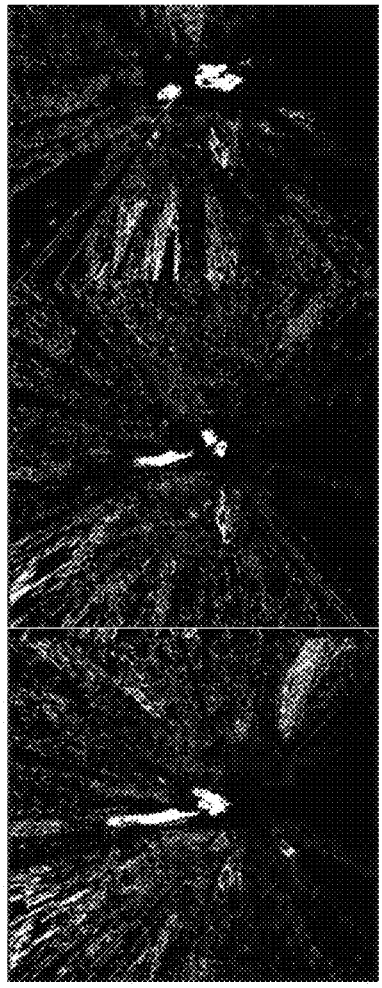


FIG. 28A FIG. 28B FIG. 28C

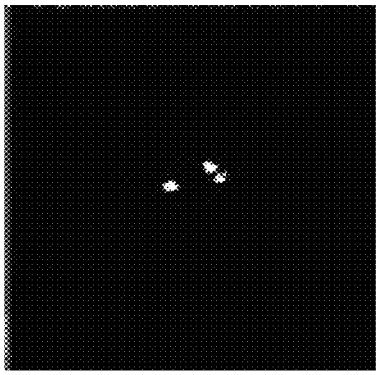


FIG. 28D

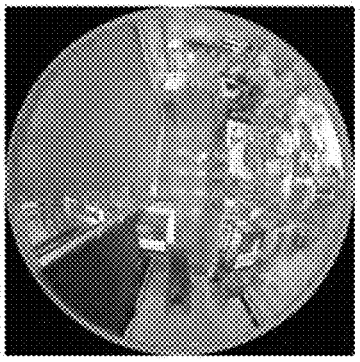


FIG. 29A

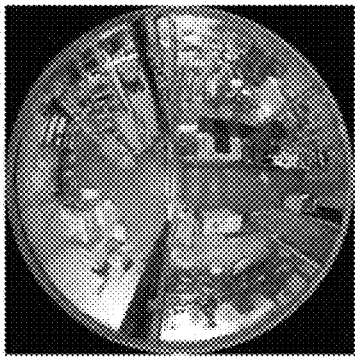


FIG. 29B

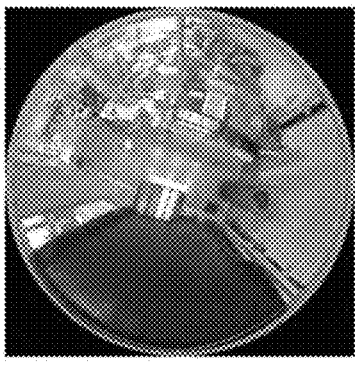


FIG. 29C

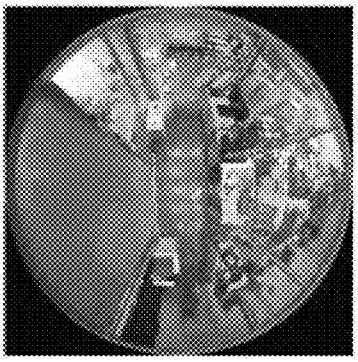


FIG. 29D

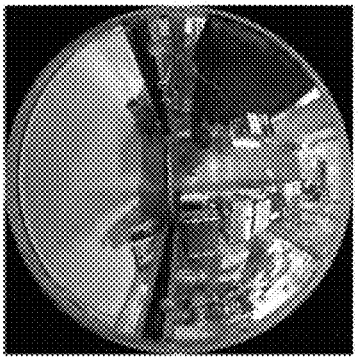


FIG. 29E

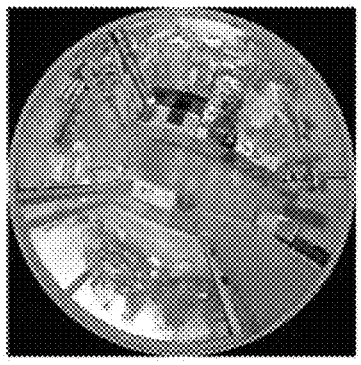


FIG. 29F

FIG. 30A

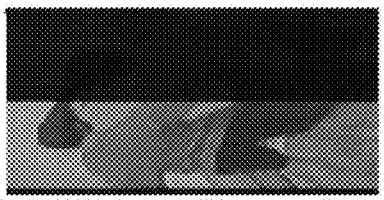


FIG. 30B

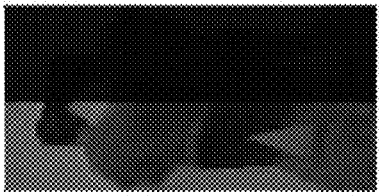


FIG. 30C

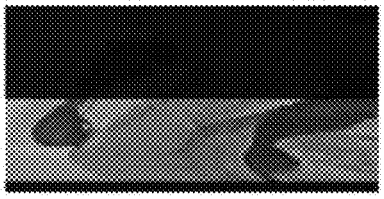


FIG. 30D

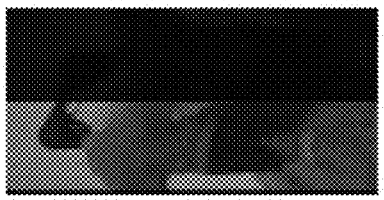


FIG. 30E

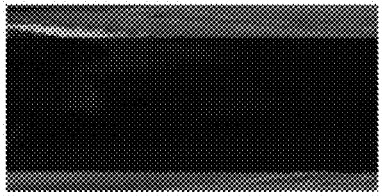


FIG. 30F

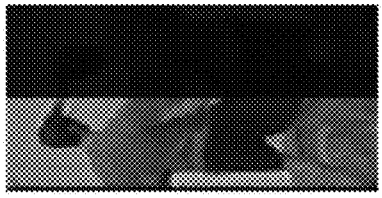
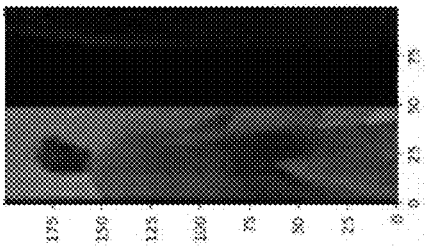


FIG. 30G



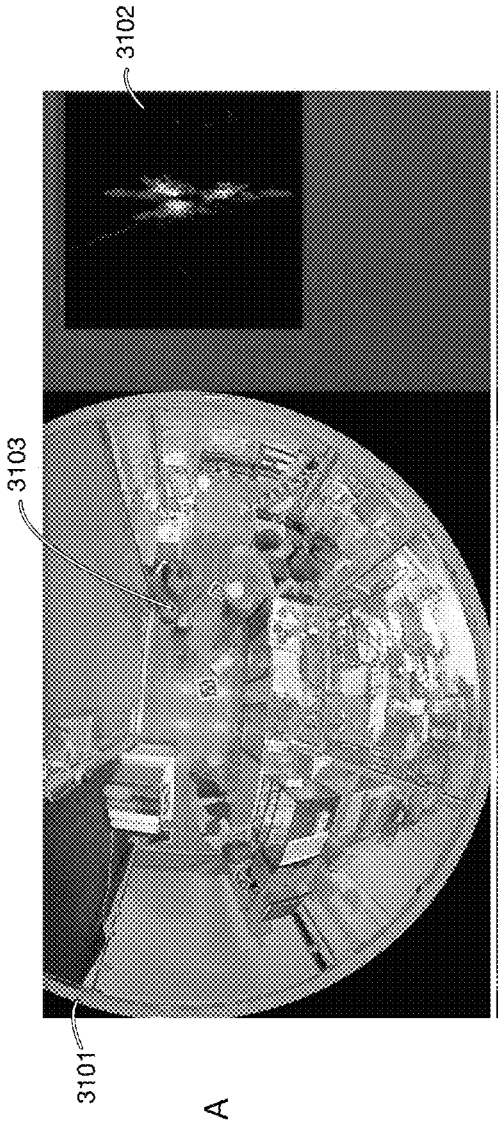


FIG. 31A

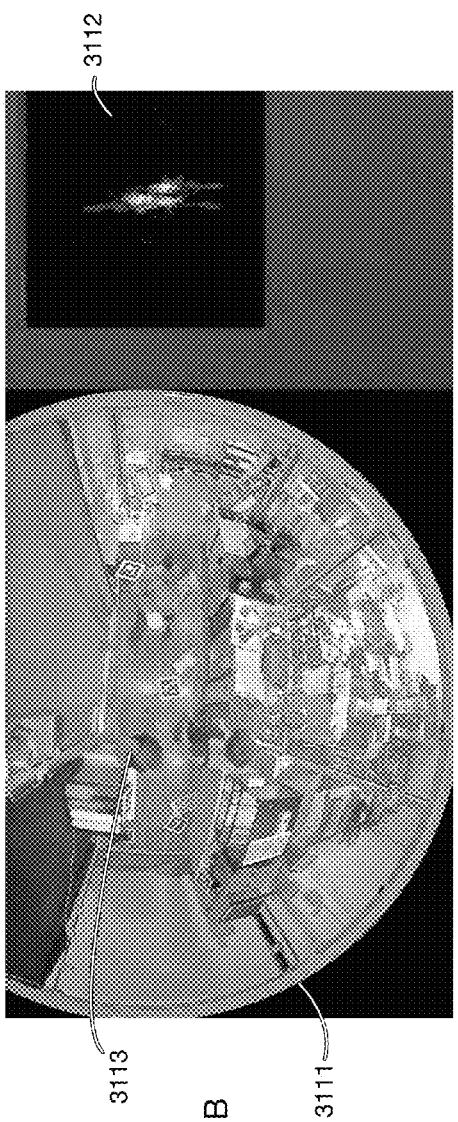


FIG. 31B

FIG. 32

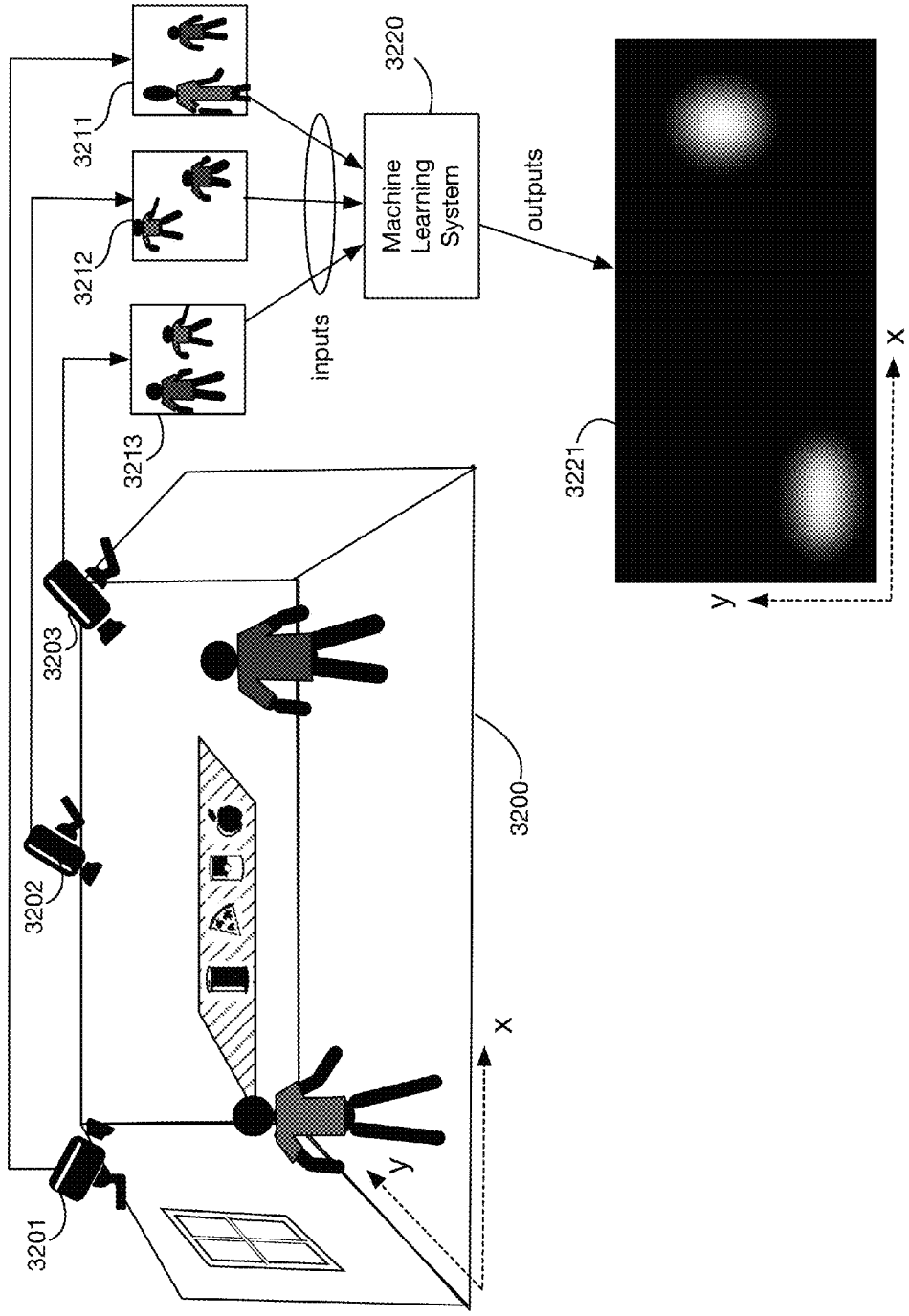
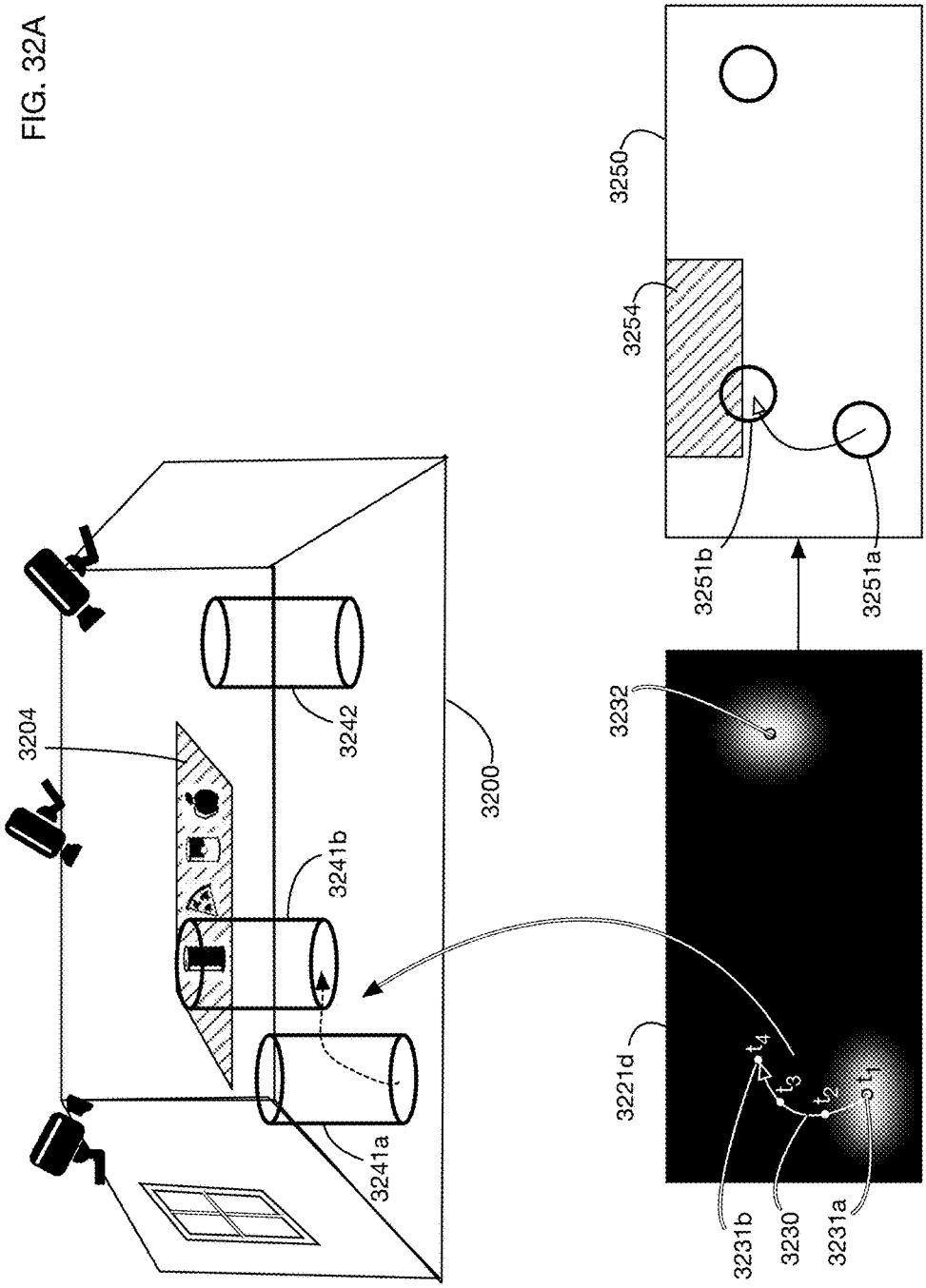

3211
3212
3213
3220
Machine Learning System
inputs
outputs
3221
x
y
3203
3202
3201
3200
x
y

FIG. 32A


3204
3242
3241b
3241a
3200
3250
3254
3251b
3251a
3232
3221d
t_4
t_3
t_2
t_1
3231b
3230
3231a

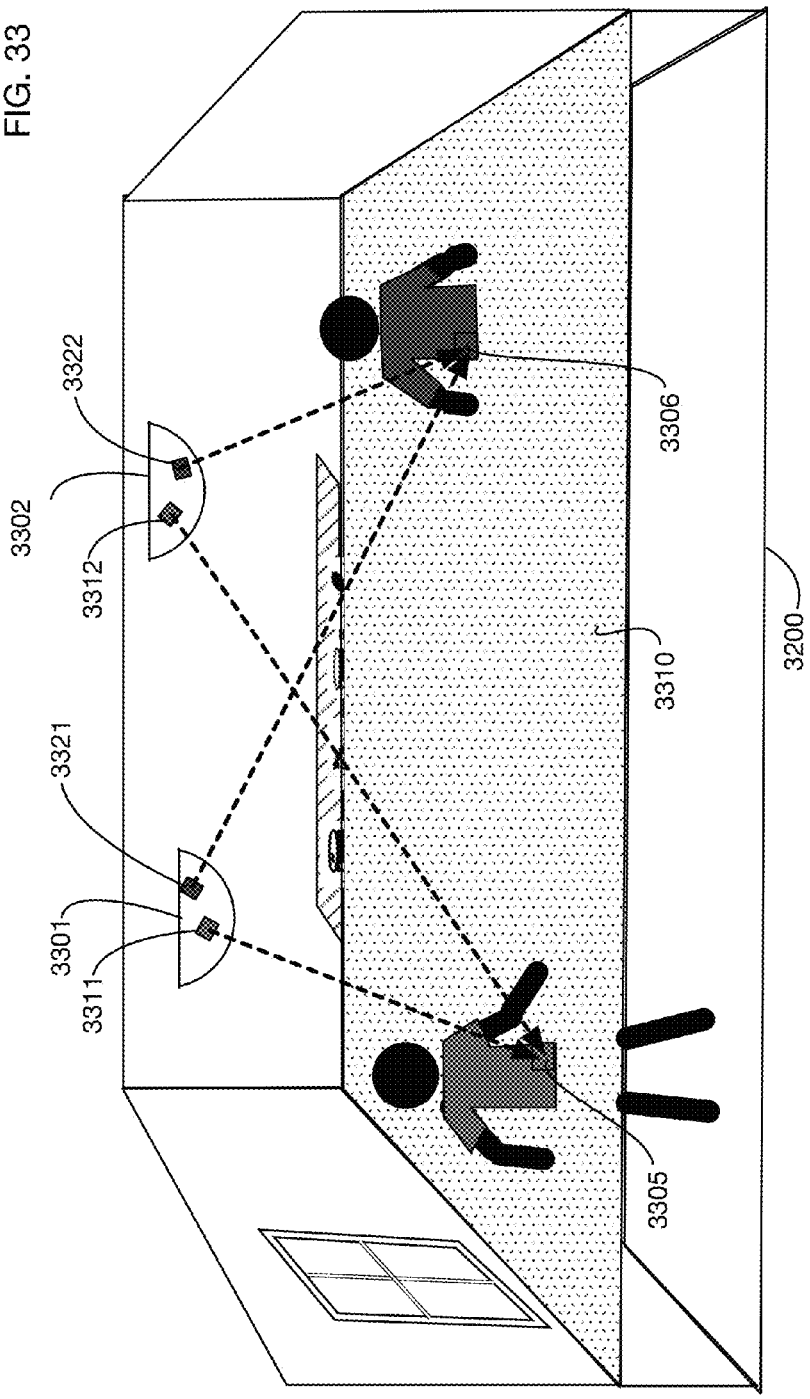


FIG. 33

FIG. 34B

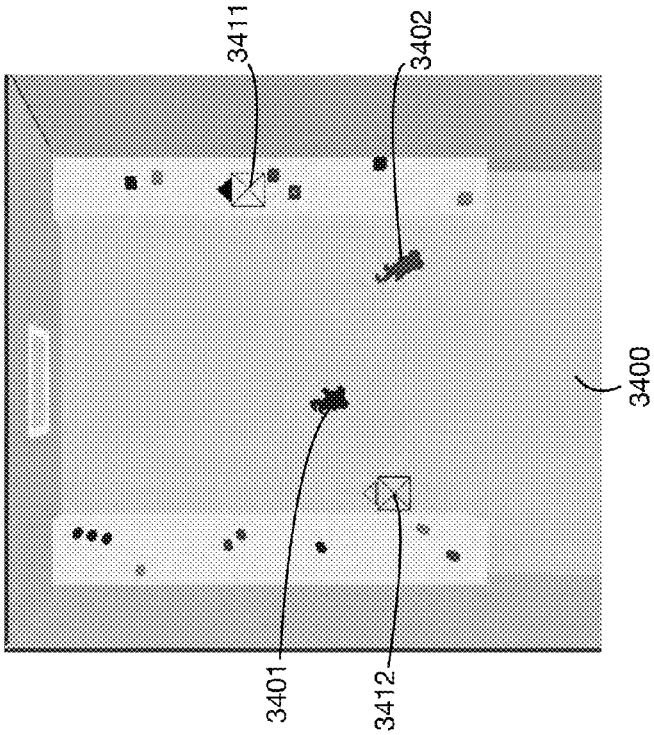


FIG. 34A

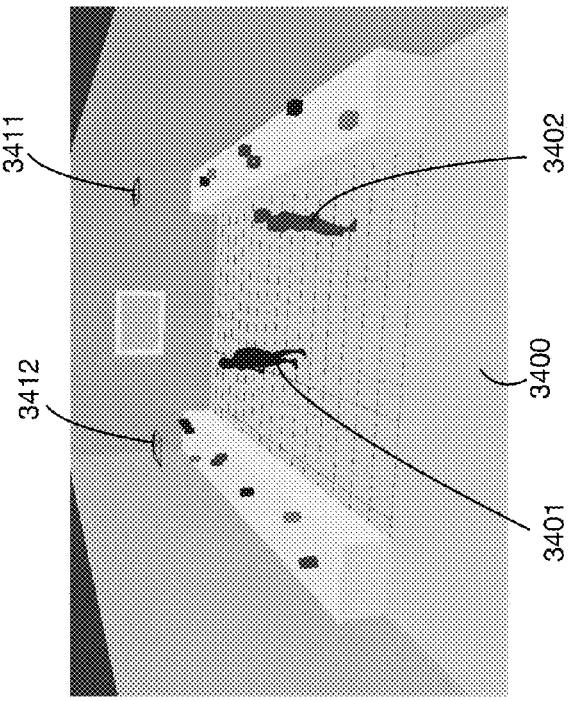


FIG. 35

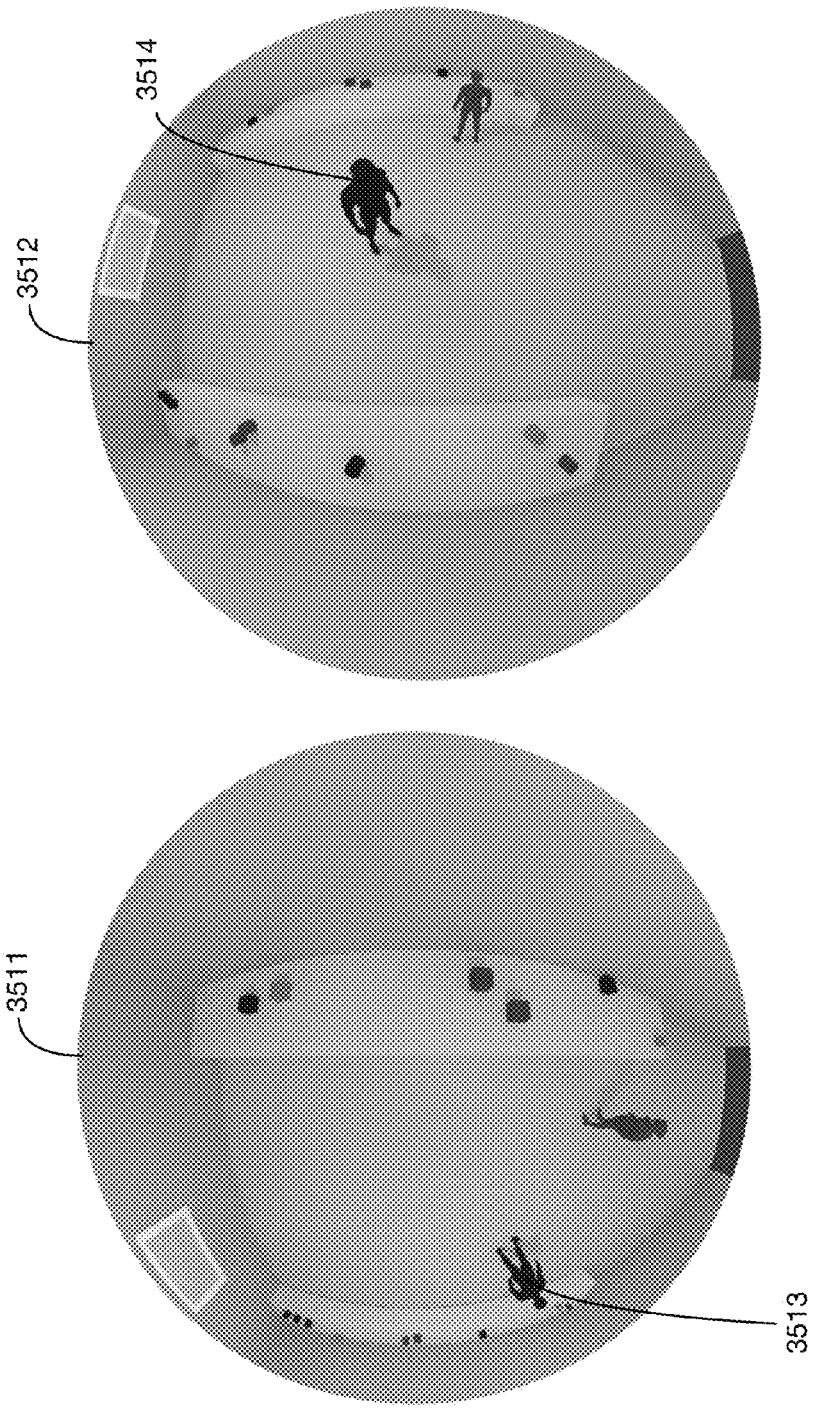


FIG. 36

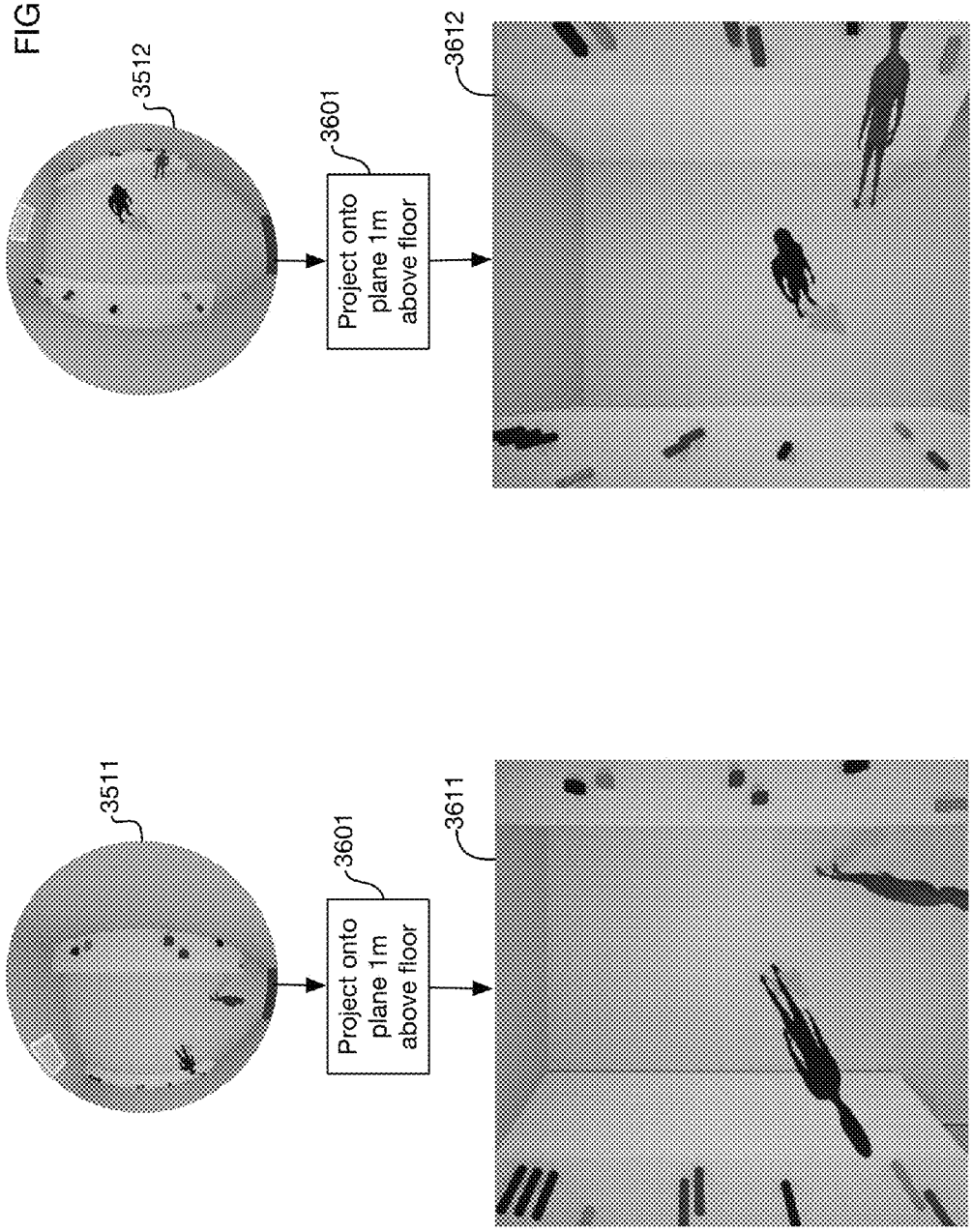

3512
3601
Project onto plane 1m above floor
3612
3511
3601
Project onto plane 1m above floor
3611

FIG. 37

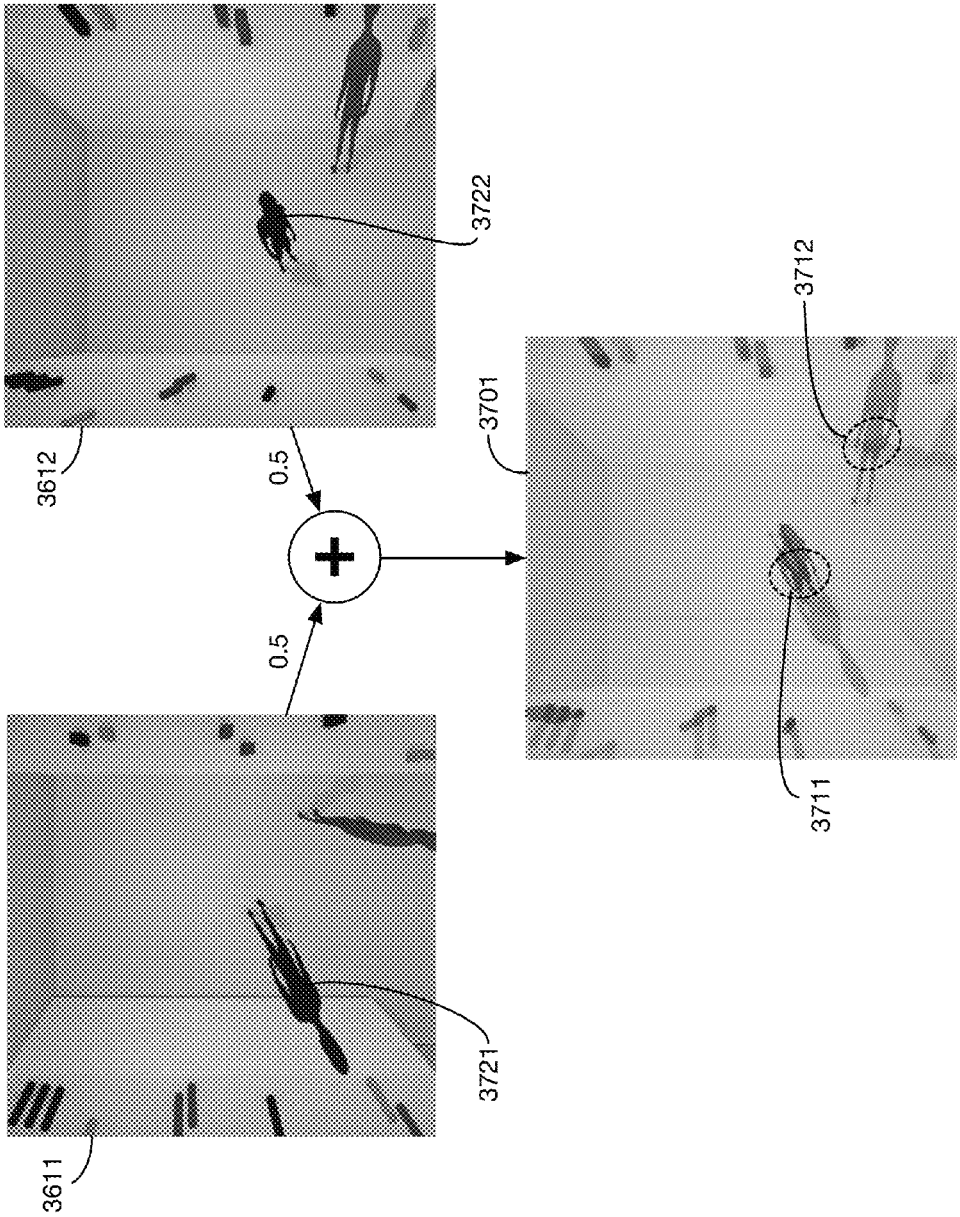


FIG. 38

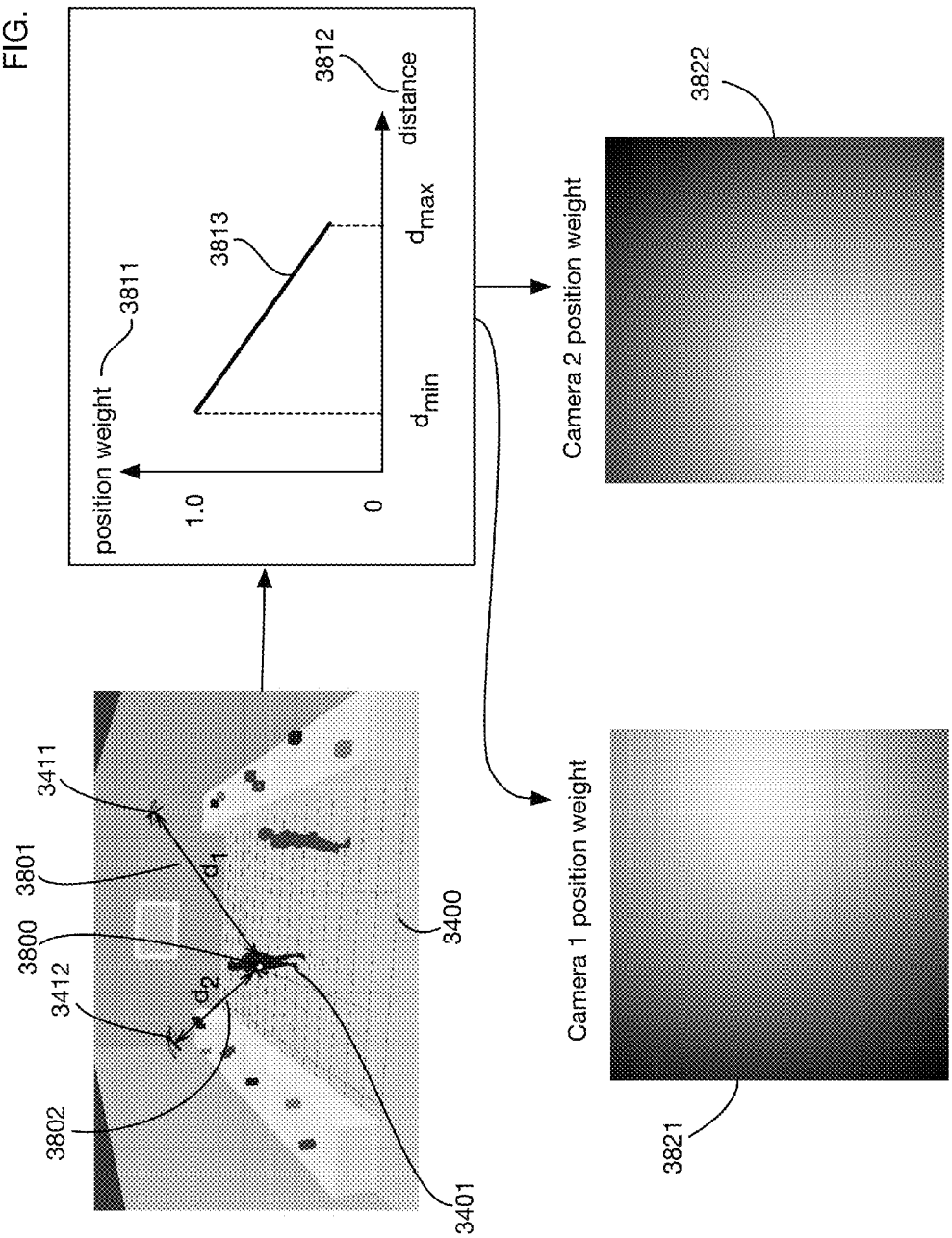

position weight
3811
1.0
0
3813
d_{min}
d_{max}
3812
distance
3411
3801
3800
3412
3802
d_1
d_2
3400
3401
Camera 1 position weight
3821
Camera 2 position weight
3822

FIG. 39

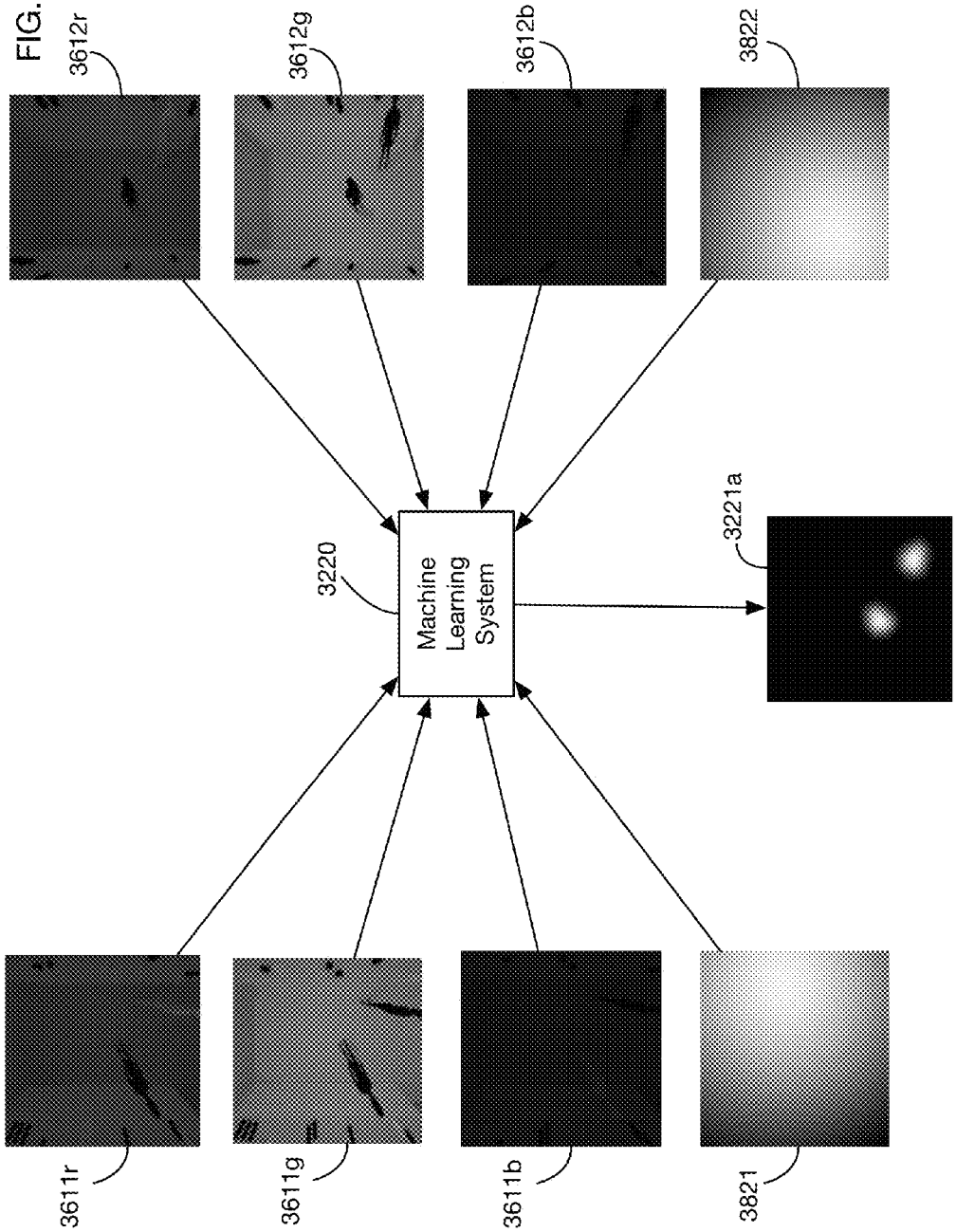

3612r
3612g
3612b
3822
3220
Machine Learning System
3221a
3611r
3611g
3611b
3821

FIG. 40

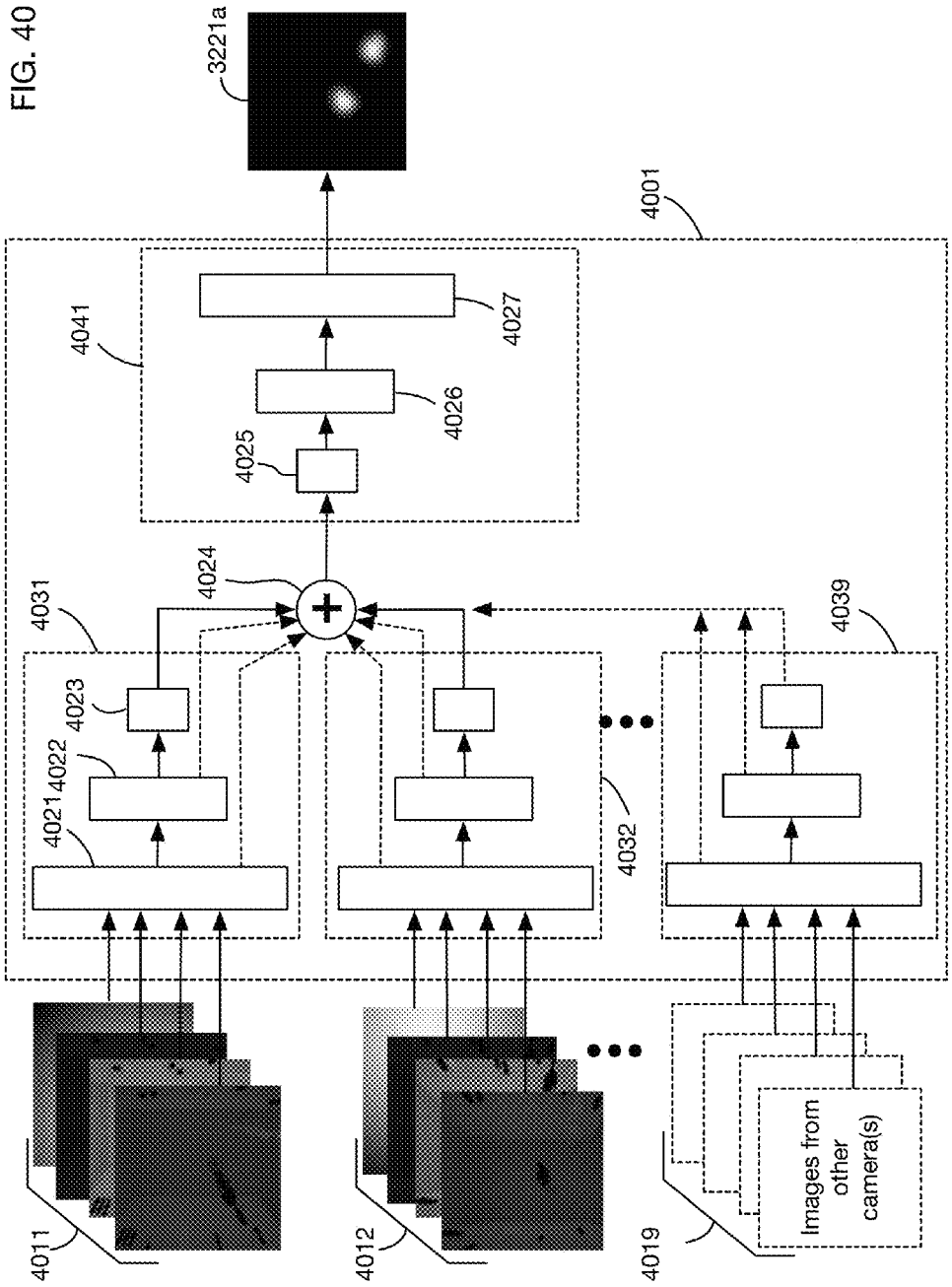

3221a
4001
4041
4027
4026
4025
4024
4031
4039
4023
4022
4021
4032
4011
4012
4019
Images from other camera(s)

FIG. 41

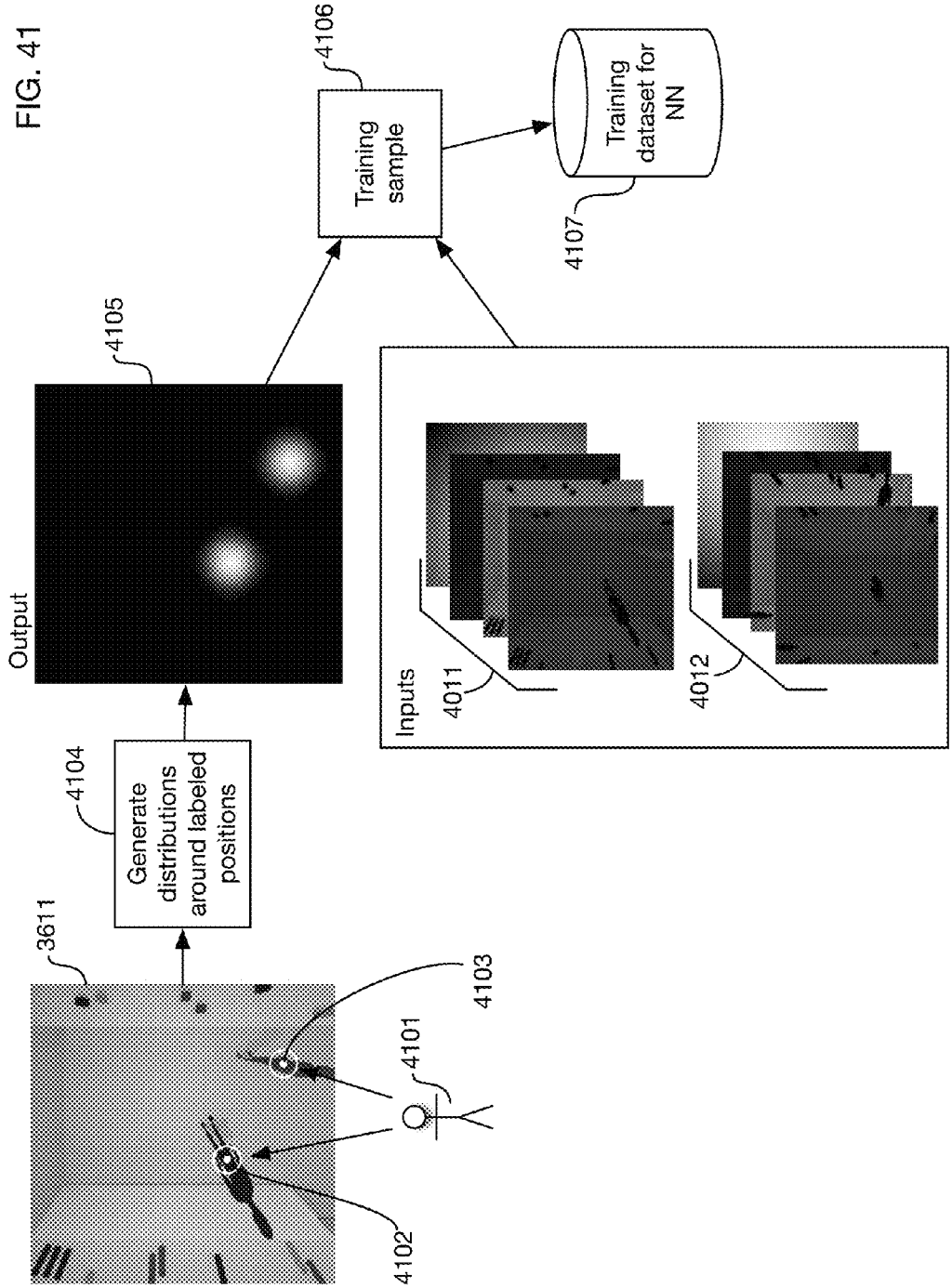

3611
4104
Generate distributions around labeled positions
Output
4105
4106
Training sample
Training dataset for NN
4107
Inputs
4011
4012
4101
4102
4103

FIG. 42

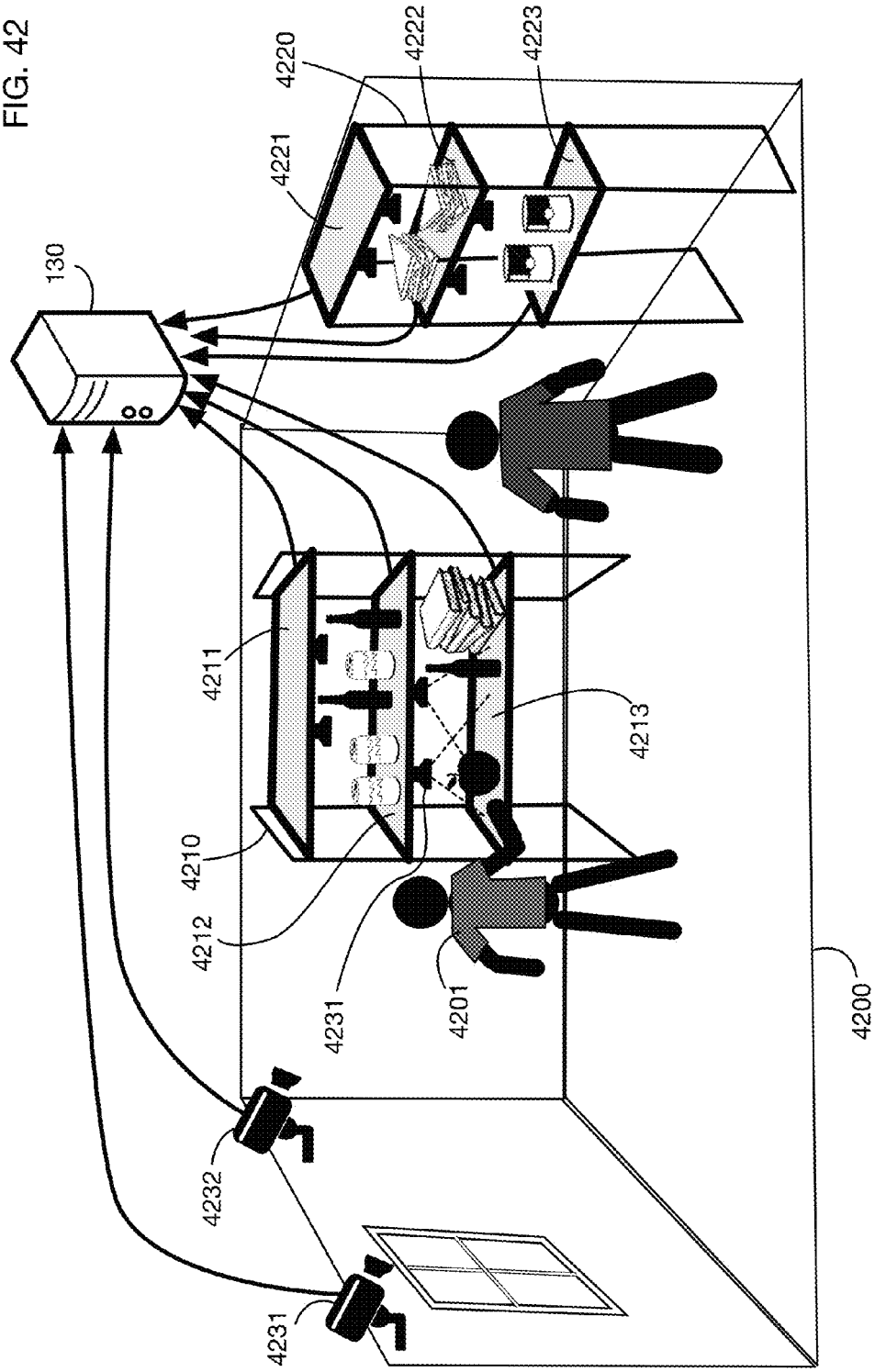
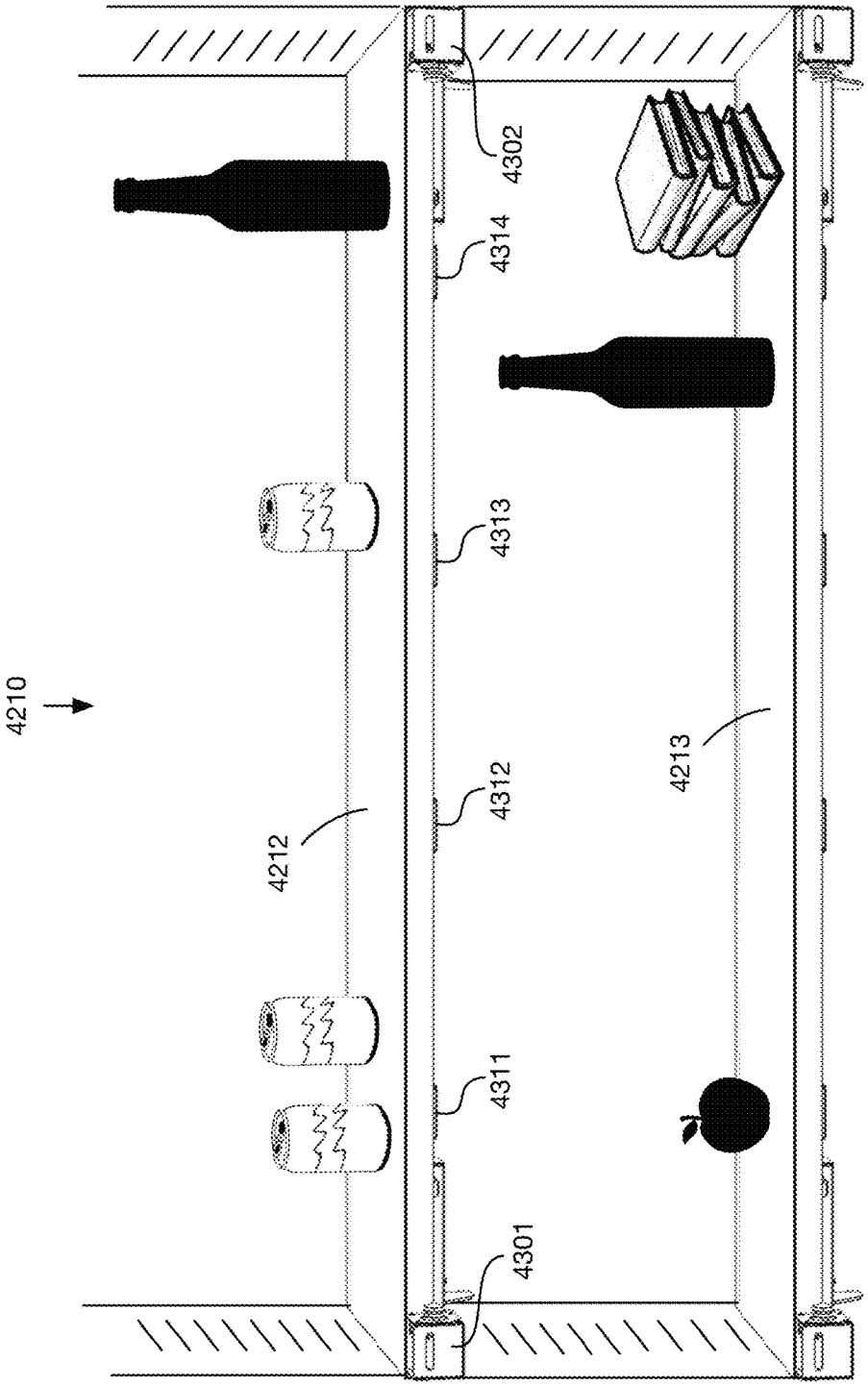

130
4220
4221
4222
4223
4211
4210
4212
4213
4231
4201
4200
4232
4231

FIG. 43



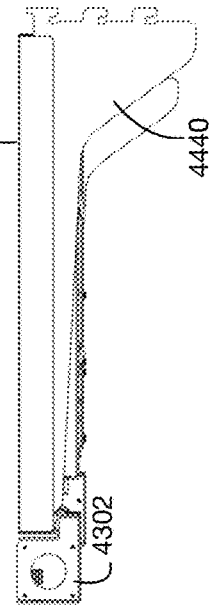


FIG. 44B

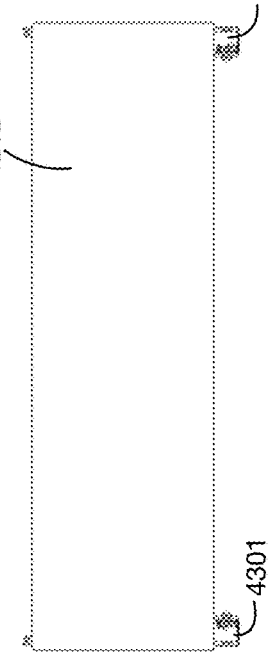


FIG. 44A

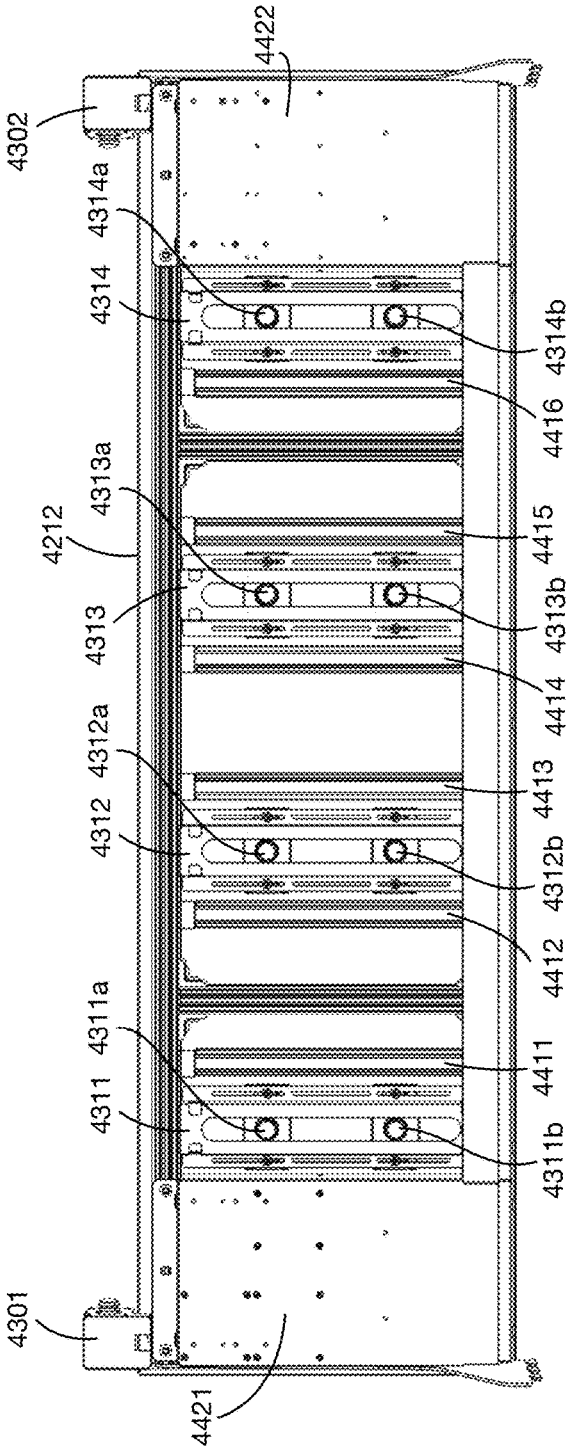


FIG. 44C

FIG. 45

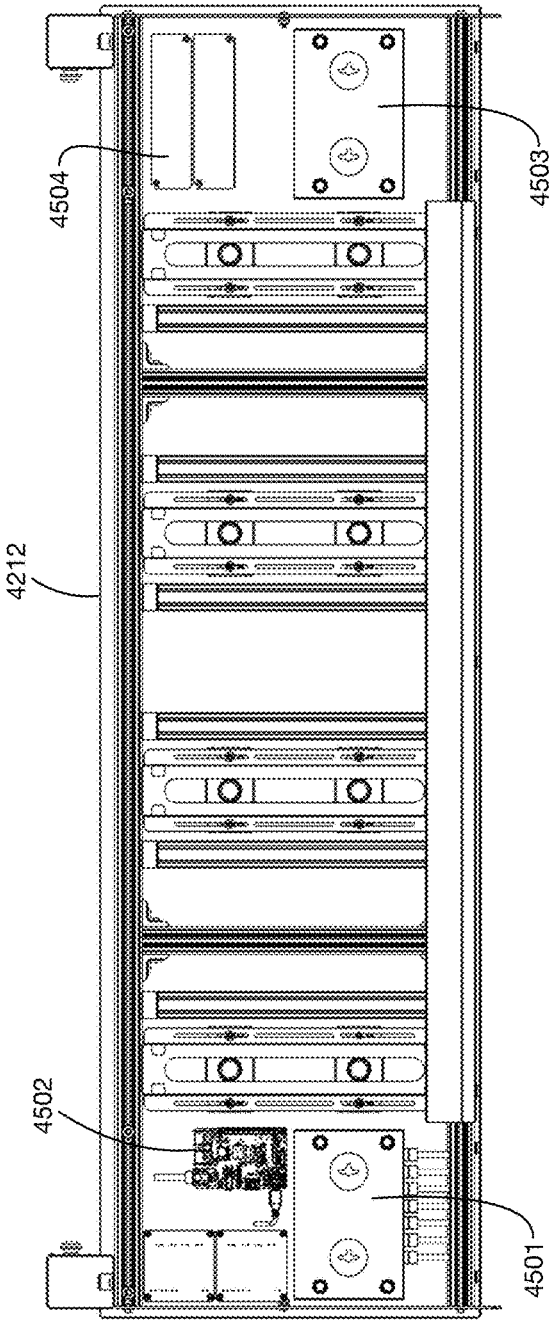


FIG. 46B

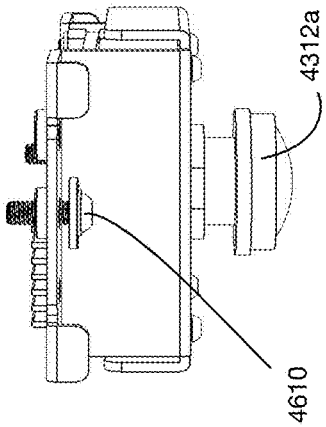


FIG. 46A

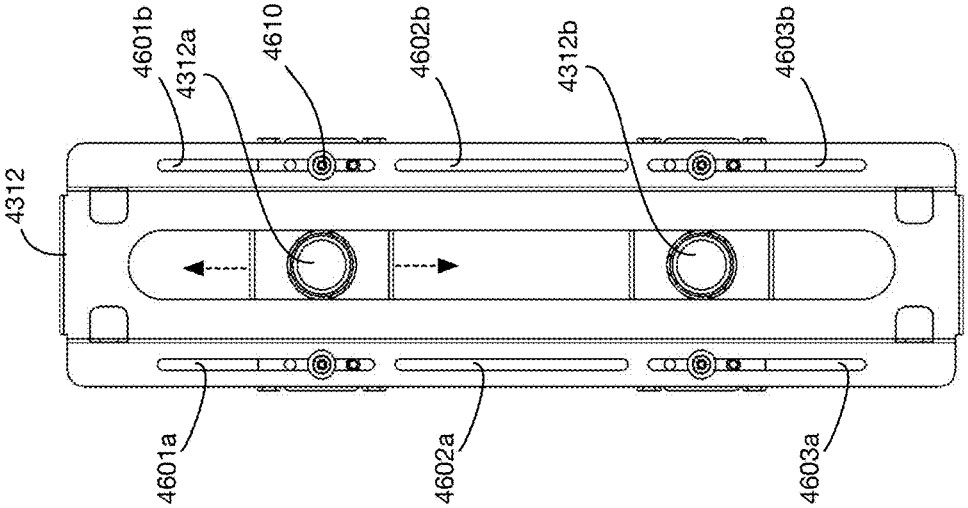


FIG. 47

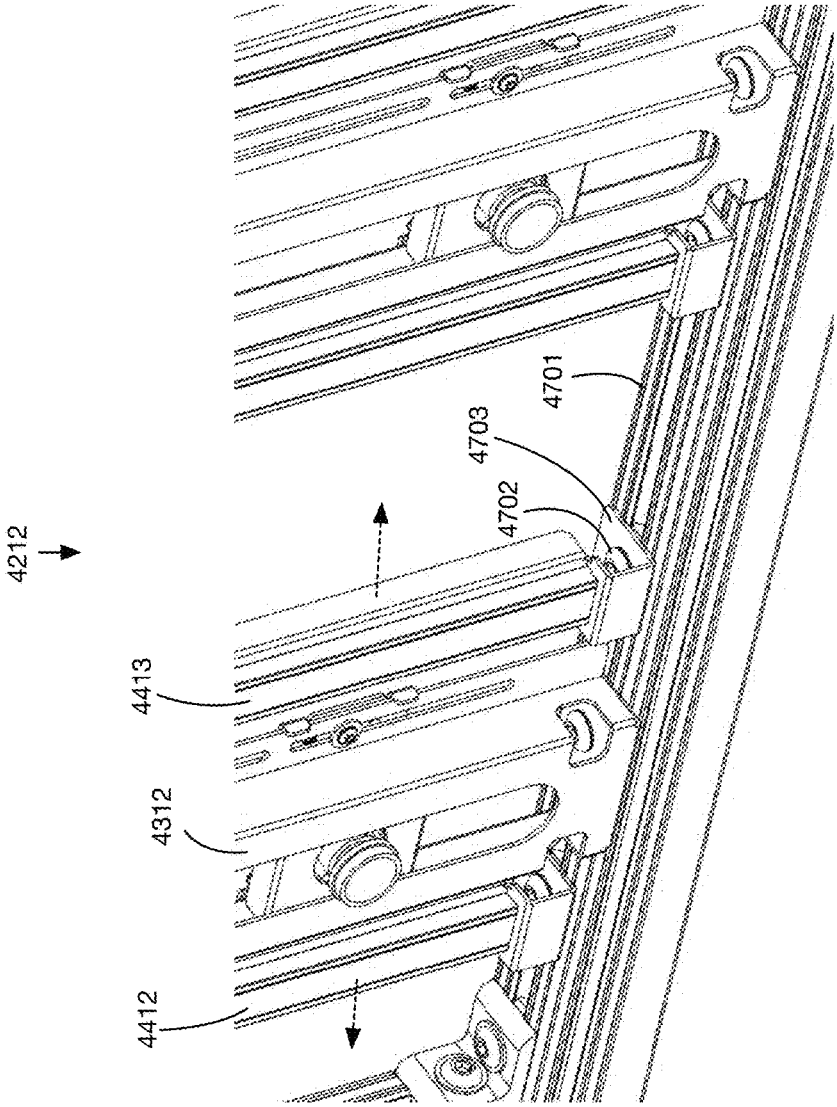


FIG. 48

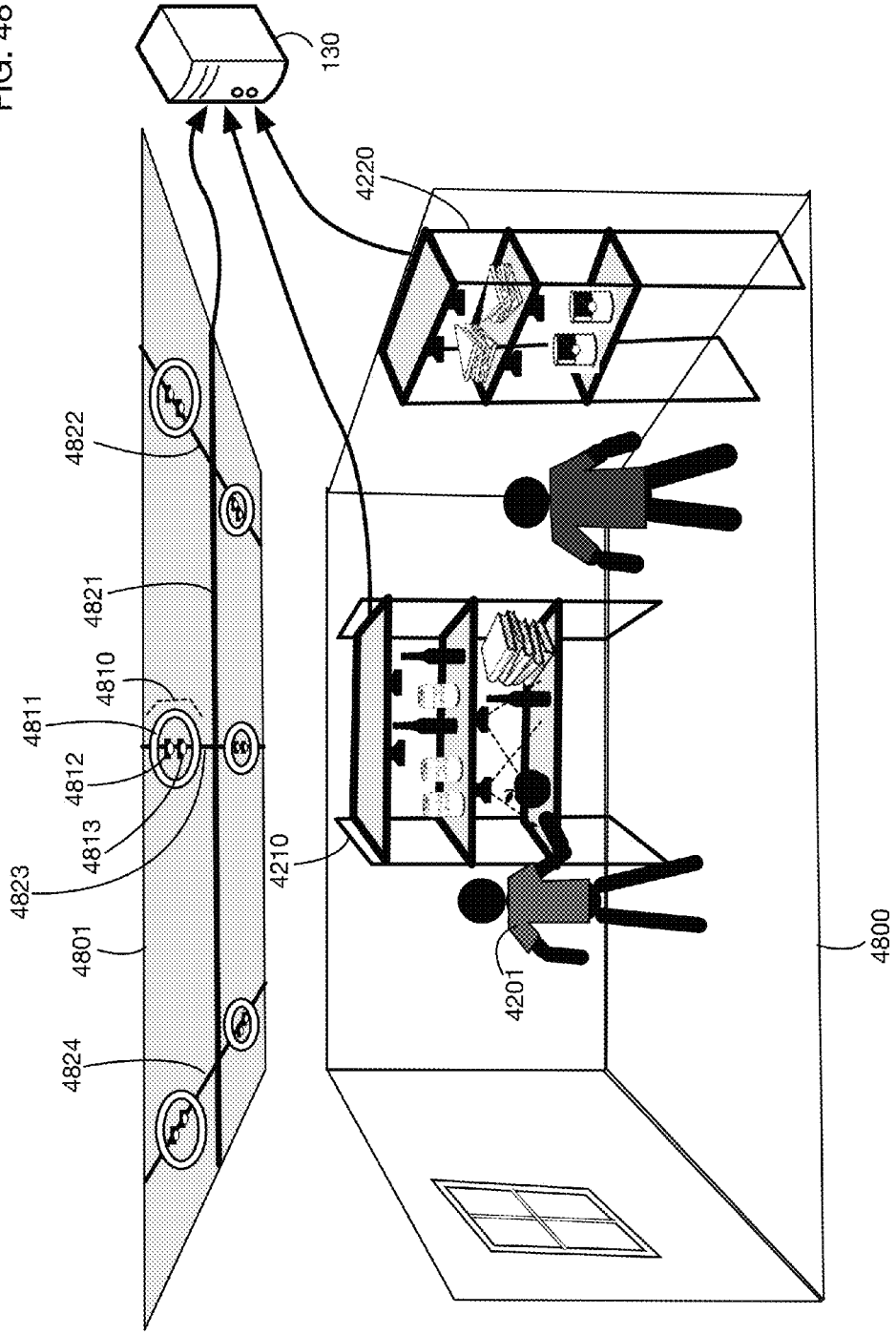

130
4220
4822
4821
4810
4811
4812
4813
4823
4801
4824
4210
4201
4800

FIG. 49

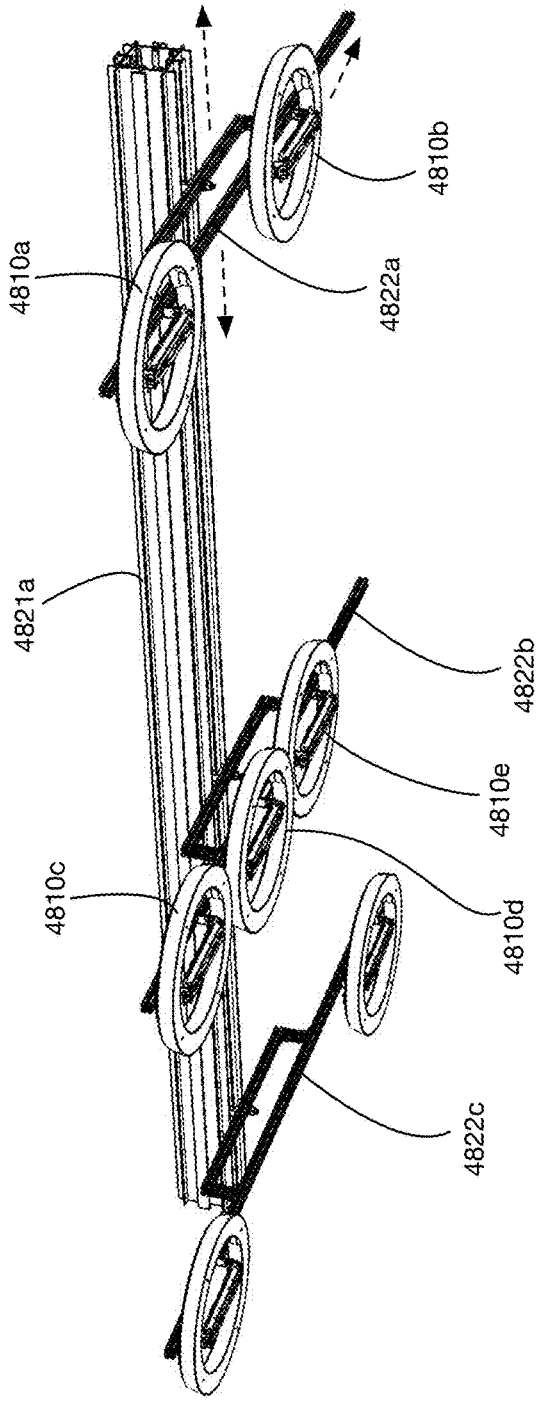


FIG. 50

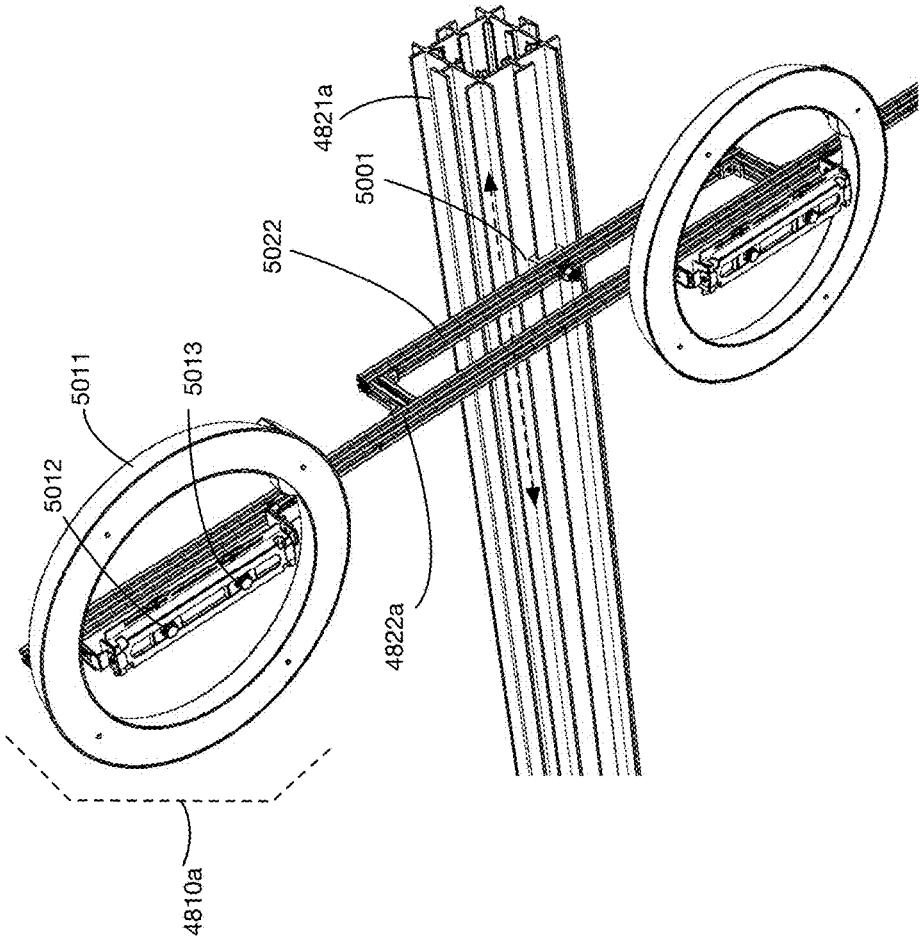
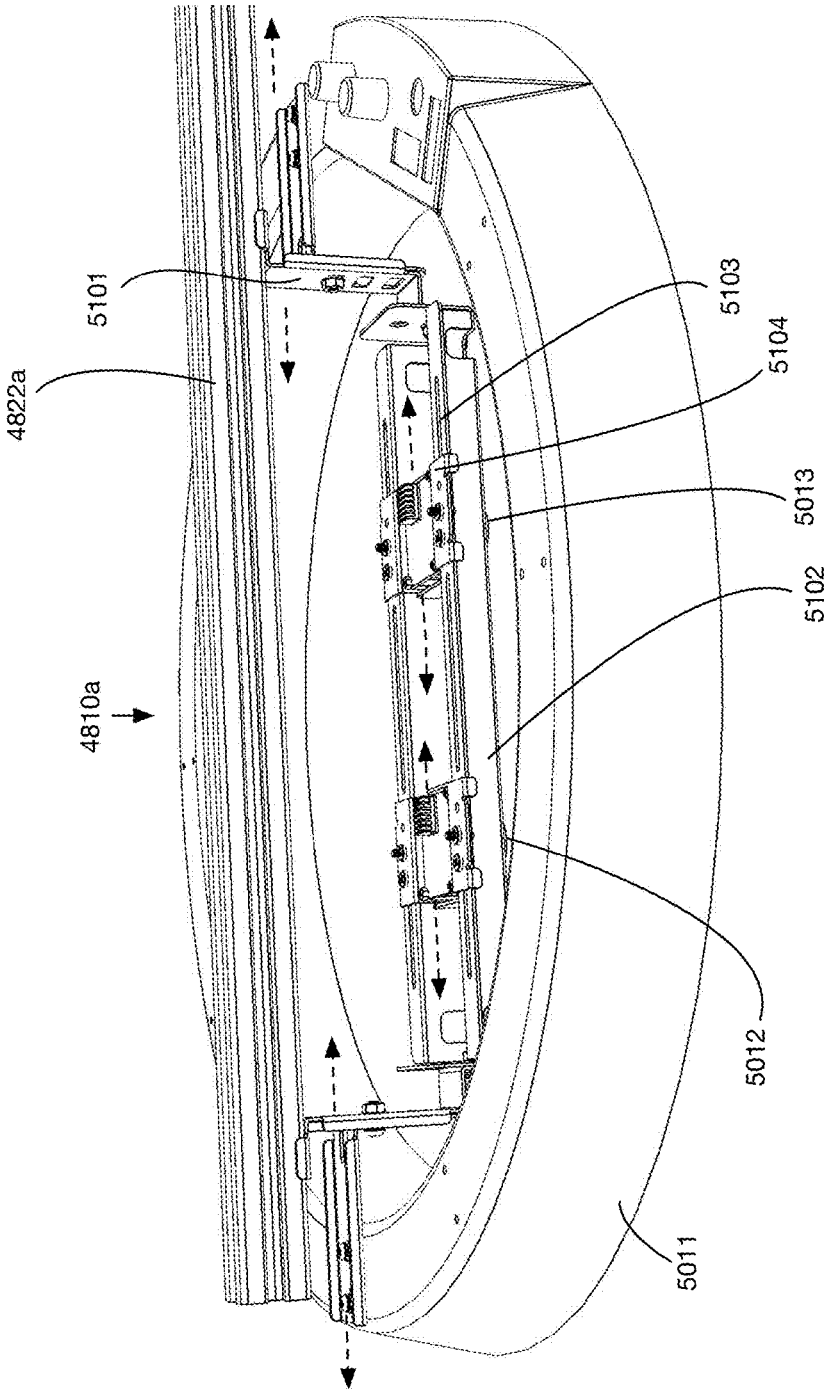


FIG. 51



AUTONOMOUS STORE TRACKING SYSTEM

This application is a continuation-in-part of U.S. Utility patent application Ser. No. 16/138,278, filed 21 Sep. 2018, which is a continuation-in-part of U.S. Utility patent application Ser. No. 16/036,754, filed 16 Jul. 2018, the specifications of which are hereby incorporated herein by reference.

BACKGROUND OF THE INVENTION

Field of the Invention

One or more embodiments of the invention are related to the fields of image analysis, artificial intelligence, automation, camera calibration, camera placement optimization and computer interaction with a point of sale system. More particularly, but not by way of limitation, one or more embodiments of the invention enable a camera-based system that extends an authorization, for example from one location to another and an autonomous store system that analyzes images from cameras to track people in the store and to detect interactions of these people with items in the store such as products on store shelves.

Description of the Related Art

Previous systems involving security cameras have had relatively limited people tracking, counting, loiter detection and object tampering analytics. These systems employ relatively simple algorithms that have been utilized in cameras and NVRs (network video recorders).

Other systems such as retail analytics solutions utilize additional cameras and sensors in retail spaces to track people in relatively simple ways, typically involving counting and loiter detection.

Currently there are initial "grab-n-go" systems that are in the initial prototyping phase. These systems are directed at tracking people that walk into a store, take what they want, put back what they don't want and get charged for what they leave with. These solutions generally use additional sensors and/or radio waves for perception, while other solutions appear to be using potentially uncalibrated cameras or non-optimized camera placement. For example, some solutions may use weight sensors on shelves to determine what products are taken from a shelf; however, these weight sensors alone are not sufficient to attribute the taking of a product with a particular shopper. To date all known camera-based grab-n-go companies utilize algorithms that employ the same basic software and hardware building blocks, drawing from academic papers that address parts of the overall problem of people tracking, action detection, object recognition.

Academic building blocks utilized by entities in the automated retail sector include a vast body of work around computer vision algorithms and open source software in this space. The basic available toolkits utilize deep learning, convolutional neural networks, object detection, camera calibration, action detection, video annotation, particle filtering and model-based estimation.

To date, none of the known solutions or systems enable a truly automated store and require additional sensors, use more cameras than are necessary, do not integrate with existing cameras within a store, for example security cameras, thus requiring more initial capital outlay. In addition, known solutions may not calibrate the cameras, allow for

heterogenous camera types to be utilized or determine optimal placement for cameras, thus limiting their accuracy.

For an automated store or similar applications, it may be valuable to allow a customer to obtain an authorization at an entry point or at another convenient location, and then extend this authorization automatically to other locations in the store or site. For example, a customer of an automated gas station may provide a credit card at a gas pump to purchase gas, and then enter an automated convenience store at the gas station to purchase products; ideally the credit card authorization obtained at the gas pump would be extended to the convenience store, so that the customer could enter the store (possibly through a locked door that is automatically unlocked for this customer), and take products and have them charged to the same card.

Authorization systems integrated into entry control systems are known in the art. Examples include building entry control systems that require a person to present a key card or to enter an access code. However, these systems do not extend the authorization obtained at one point (the entry location) to another location. Known solutions to extend authorization from one location to additional locations generally require that the user present a credential at each additional location where authorization is needed. For example, guests at events or on cruise ships may be given smart wristbands that are linked to a credit card or account; these wristbands may be used to purchase additional products or to enter locked areas. Another example is the system disclosed in U.S. Pat. No. 6,193,154, "Method and apparatus for vending goods in conjunction with a credit card accepting fuel dispensing pump," which allows a user to be authorized at a gas pump (using a credit card), and to obtain a code printed on a receipt that can then be used at a different location to obtain goods from a vending machine. A potential limitation of all of these known systems is that additional devices or actions by the user are required to extend authorization from one point to another. There are no known systems that automatically extend authorization from one point (such as a gas pump) to another point (such as a store or vending machine) using only tracking of a user from the first point to the second via cameras. Since cameras are widely available and often are already installed in sites or stores, tracking users with cameras to extend authorization from one location to another would add significant convenience and automation without burdening the user with codes or wristbands and without requiring additional sensors or input devices.

Another limitation of existing systems for automated stores is the complexity of the person tracking approaches. These systems typically use complex algorithms that attempt to track joints or landmarks of a person based on multiple camera views from arbitrary camera locations. This approach may be error-prone, and it requires significant processing capacity to support real-time tracking. A simpler person tracking approach may improve robustness and efficiency of the tracking process.

For at least the limitations described above there is a need for an autonomous store tracking system.

BRIEF SUMMARY OF THE INVENTION

One or more embodiments described in the specification are related to an automated store system that analyzes camera images to track people and their interactions with items. One or more embodiments include a processor that is configured to obtain a 3D model of a store that contains items and item storage areas. The processor receives a

respective time sequence of images from cameras in the store, wherein the time sequence of images is captured over a time period and analyzes the time sequence of images from each camera and the 3D model of the store to detect a person in the store based on the time sequence of images, calculate a trajectory of the person across the time period, identify an item storage area of the item storage areas that is proximal to the trajectory of the person during an interaction time period within the time period, analyze two or more images of the time sequence of images to identify an item of the items within the item storage area that moves during the interaction time period, wherein the two or more images are captured within or proximal in time to the interaction time period and the two or more images contain views of the item storage area and attribute motion of the item to the person. One or more embodiments of the system rely on images for tracking and do not utilize item tags, for example RFID tags or other identifiers on the items that are manipulated and thus do not require identifier scanners. In addition, one or more embodiments of the invention enable a "virtual door" where entry and exit of users triggers a start or stop of the tracker, i.e., via images and computer vision. Other embodiments may utilize physical gates or electronic check-in and check-out, e.g., using QR codes or Bluetooth, but these solutions add complexity that other embodiments of the invention do not require.

At least one embodiment of the processor is further configured to interface with a point of sale computer and charge an amount associated with the item to the person without a cashier. Optionally, a description of the item is sent to a mobile device associated with the person and wherein the processor or point of sale computer is configured to accept a confirmation from the mobile device that the item is correct or in dispute. In one or more embodiments, a list of the items associated with a particular user, for example a shopping cart list associated with the shopper, may be sent to a display near the shopper or that is closest to the shopper.

In one or more embodiments, each image of the time sequence of images is a 2D image and the processor calculates a trajectory of the person consisting of a 3D location and orientation of the person and at least one body landmark from two or more 2D projections of the person in the time sequence of images.

In one or more embodiments, the processor is further configured to calculate a 3D field of influence volume around the person at points of time during the time period.

In one or more embodiments, the processor identifies an item storage area that is proximal to the trajectory of the person during an interaction time period utilizes a 3D location of the storage area that intersects the 3D field of influence volume around the person during the interaction time period. In one or more embodiments, the processor calculates the 3D field of influence volume around the person utilizing a spatial probability distribution for multiple landmarks on the person at the points of time during the time period, wherein each landmark of the multiple landmarks corresponds to a location on a body part of the person. In one or more embodiments, the 3D field of influence volume around the person comprises points having a distance to a closest landmark of the multiple landmarks that is less than or equal to a threshold distance. In one or more embodiments, the 3D field of influence volume around the person comprises a union of probable zones for each landmark of the multiple landmarks, wherein each probable zone of the probable zones contains a threshold probability of the spatial probability distribution for a corresponding landmark. In one or more embodiments, the processor calculates the spatial probability distribution for multiple landmarks on the person at the points of time during the time period through calculation of a predicated spatial probability distribution for the multiple landmarks at one or more points of time during the time period based on a physics model and calculation of a corrected spatial probability distribution at one or more points of time during the time period based on observations of one or more of the multiple landmarks in the time sequence of images. In one or more embodiments, the physics model includes the locations and velocities of the landmarks and thus the calculated field of influence. This information can be used to predict a state of landmarks associated with a field at a time and a space not directly observed and thus may be utilized to interpolate or augment the observed landmarks.

In one or more embodiments, the processor is further configured to analyze the two or more images of the time sequence of images to classify the motion of the item as a type of motion comprising taking, putting or moving.

In one or more embodiments, the processor analyzes two or more images of the time sequence of images to identify an item within the item storage area that moves during the interaction time period. Specifically, the processor uses or obtains a neural network trained to recognize items from changes across images, sets an input layer of the neural network to the two or more images and calculates a probability associated with the item based on an output layer of the neural network. In one or more embodiments, the neural network is further trained to classify an action performed on an item into classes comprising taking, putting, or moving. In one or more embodiments, the system includes a verification system configured to accept input confirming or denying that the person is associated with motion of the item. In one or more embodiments, the system includes a machine learning system configured to receive the input confirming or denying that the person is associated with the motion of the item and updates the neural network based on the input. Embodiments of the invention may utilize a neural network or more generally, any type of generic function approximator. By definition the function to map inputs of before-after image pairs, or before-during-after image pairs to output actions, then the neural network can be trained to be any such function map, not just traditional convolutional neural networks, but also simpler histogram or feature based classifiers. Embodiments of the invention also enable training of the neural network, which typically involves feeding labeled data to an optimizer that modifies the network's weights and/or structure to correctly predict the labels (outputs) of the data (inputs). Embodiments of the invention may be configured to collect this data from customer's acceptance or correction of the presented shopping cart. Alternatively, or in combination, embodiments of the system may also collect human cashier corrections from traditional stores. After a user accepts a shopping cart or makes a correction, a ground truth labeled data point may be generated and that point may be added to the training set and used for future improvements.

In one or more embodiments, the processor is further configured to identify one or more distinguishing characteristics of the person by analyzing a first subset of the time sequence of images and recognizes the person in a second subset of the time sequence of images using the distinguishing characteristics. In one or more embodiments, the processor recognizes the person in the second subset without determination of an identity of the person. In one or more embodiments, the second subset of the time sequence of images contains images of the person and images of a

second person. In one or more embodiments, the one or distinguishing characteristics comprise one or more of shape or size of one or more body segments of the person, shape, size, color, or texture of one or more articles of clothing worn by the person and gait pattern of the person.

In one or more embodiments of the system, the processor is further configured to obtain camera calibration data for each camera of the cameras in the store and analyze the time sequence of images from each camera of the cameras using the camera calibration data. In one or more embodiments, the processor configured to obtain calibration images from each camera of the cameras and calculate the camera calibration data from the calibration images. In one or more embodiments, the calibration images comprise images captured of one or more synchronization events and the camera calibration data comprises temporal offsets among the cameras. In one or more embodiments, the calibration images comprise images captured of one or markers placed in the store at locations defined relative to the 3D model and the camera calibration data comprises position and orientation of the cameras with respect to the 3D model. In one or more embodiments, the calibration images comprise images captured of one or more color calibration targets located in the store, the camera calibration data comprises color mapping data between each camera of the cameras and a standard color space. In one or more embodiments, the camera calibration processor is further configured to recalculate the color mapping data when lighting conditions change in the store. For example, in one or more embodiments, different camera calibration data may be utilized by the system based on the time of day, day of year, current light levels or light colors (hue, saturation or luminance) in an area or entire image, such as occur at dusk or dawn color shift periods. By utilizing different camera calibration data, for example for a given camera or cameras or portions of images from a camera or camera, more accurate determinations of items and their manipulations may be achieved.

In one or more embodiments, any processor in the system, such as a camera placement optimization processor is configured to obtain the 3D model of the store and calculate a recommended number of the cameras in the store and a recommended location and orientation of each camera of the cameras in the store. In one or more embodiments, the processor calculates a recommended number of the cameras in the store and a recommended location and orientation of each camera of the cameras in the store. Specifically, the processor obtains a set of potential camera locations and orientations in the store, obtains a set of item locations in the item storage areas and iteratively updates a proposed number of cameras and a proposed set of camera locations and orientations to obtain a minimum number of cameras and a location and orientation for each camera of the minimum number of cameras such that each item location of the set of item locations is visible to at least two of the minimum number of cameras.

In one or more embodiments, the system comprises the cameras, wherein the cameras are coupled with the processor. In other embodiments, the system includes any subcomponent described herein.

In one or more embodiments, processor is further configured to detect shoplifting when the person leaves the store without paying for the item. Specifically, the person's list of items on hand (e.g., in the shopping cart list) may be displayed or otherwise observed by a human cashier at the traditional cash register screen. The human cashier may utilize this information to verify that the shopper has either not taken anything or is paying/showing for all items taken from the store. For example, if the customer has taken two items from the store, the customer should pay for two items from the store. Thus, embodiments of the invention enable detection of customers that for example take two items but only show and pay for one when reaching the register.

In one or more embodiments, the computer is further configured to detect that the person is looking at an item.

In one or more embodiments, the landmarks utilized by the system comprise eyes of the person or other landmarks on the person's head, and wherein the computer is further configured to calculate a field of view of the person based on a location of the eyes or other head landmarks of the person, and to detect that the person is looking at an item when the item is in the field of view.

One or more embodiments of the system may extend an authorization obtained at one place and time to a different place or a different time. The authorization may be extended by tracking a person from the point of authorization to a second point where the authorization is used. The authorization may be used for entry to a secured environment, and to purchase items within this secured environment.

To extend an authorization, a processor in the system may analyze images from cameras installed in or around an area in order to track a person in the area. Tracking may also use a 3D model of the area, which may for example describe the location and orientation of the cameras. The processor may calculate the trajectory of the person in the area from the camera images. Tracking and calculation of the trajectory may use any of the methods described above or described in detail below.

The person may present a credential, such as a credit card, to a credential receiver, such as a card reader, at a first location and at a first time, and may then receive an authorization; the authorization may also be received by the processor. The person may then move to a second location at a second time. At this second location, an entry to a secured environment may be located, and the entry may be secured by a controllable barrier such as a lock. The processor may associate the authorization with the person by relating the time that the credential was presented, or the authorization was received, with the time that the person was at the first location where the credential receiver is located. The processor may then allow the person to enter the secured environment by transmitting an allow entry command to the controllable barrier when the person is at the entry point of the secured environment.

The credential presented by the person to obtain an authorization may include for example, without limitation, one or more of a credit card, a debit card, a bank card, an RFID tag, a mobile payment device, a mobile wallet device, an identity card, a mobile phone, a smart phone, a smart watch, smart glasses or goggles, a key fob, a driver's license, a passport, a password, a PIN, a code, a phone number, or a biometric identifier.

In one or more embodiments the secured environment may be all or portion of a building, and the controllable barrier may include a door to the building or to a portion of the building. In one or more embodiments the secured environment may be a case that contains one or more items (such as a display case with products for sale), and the controllable barrier may include a door to the case.

In one or more embodiments, the area may be a gas station, and the credential receiver may be a payment mechanism at or near a gas pump. The secured environment may be for example a convenience store at the gas station or a case (such as a vending machine for example) at the gas station that contains one or more items. A person may for

example pay at the pump and obtain an authorization for pumping gas and for entering the convenience store or the product case to obtain other products.

In one or more embodiments, the credential may be or may include a form of payment that is linked to an account of the person with the credential, and the authorization received by the system may be an authorization to charge purchases by the person to this account. In one or more embodiments, the secured environment may contain sensors that detect when one or more items are taken by the person. Signals from the sensors may be received by the system's processor and the processor may then charge the person's account for the item or items taken. In one or more embodiments the person may provide input at the location where he or she presents the credential that indicates whether to authorize purchases of items in the secured environment.

In one or more embodiments, tracking of the person may also occur in the secured environment, using cameras in the secured environment. As described above with respect to an automated store, tracking may determine when the person is near an item storage area, and analysis of two or more images of the item storage area may determine that an item has moved. Combining these analyses allows the system to attribute motion of an item to the person, and to charge the item to the person's account if the authorization is linked to a payment account. Again as described with respect to an automated store, tracking and determining when a person is at or near an item storage area may include calculating a 3D field of influence volume around the person; determining when an item is moved or taken may use a neural network that inputs two or more images (such as before and after images) of the item storage area and outputs a probability that an item is moved.

In one or more embodiments, an authorization may be extended from one person to another person, such as another person who is in the same vehicle as the person with the credential. The processor may analyze camera images to determine that one person exits a vehicle and then presents a credential, resulting in an authorization. If a second person exits the same vehicle, that second person may also be authorized to perform certain actions, such as entering a secured area or taking items that will be charge to the account associated with the credential. Tracking the second person and determining what items that person takes may be performed as described above for the person who presents the credential.

In one or more embodiments, extension of an authorization may enable a person who provides a credential to take items and have them charged to an account associated with the credential; the items may or may not be in a secured environment having an entry with a controllable barrier. Tracking of the person may be performed using cameras, for example as described above. The system may determine what item or items the person takes by analyzing camera images, for example as described above. The processor associated with the system may also analyze camera images to determine when a person takes and item and then puts the item down prior to leaving an area; in this case the processor may determine that the person should not be charged for the item when leaving the area.

One or more embodiments of the invention may analyze camera images to locate a person in the store, and may then calculate a field of influence volume around the person. This field of influence volume may be simple or detailed. It may be a simple shape, such as a cylinder for example, around a single point estimate of a person's location. Tracking of landmarks or joints on the person's body may not be needed in one or more embodiments. When the field of influence volume intersects an item storage area during an interaction period, the system may analyze images captured at the beginning of this period or before, and images captured at the end of this period or afterwards. This analysis may determine whether an item on the shelf has moved, in which case this movement may be attributed to the person whose field of influence volume intersected the item storage area. Analysis of before and after images may be done for example using a neural network that takes these two images as input. The output of the neural network may include probabilities that each item has moved, and probabilities associated with each action of a set of possible actions that a person may have taken (such as for example taking, putting, or moving an item). The item and action with the highest probabilities may be selected and may be attributed to the person that interacted with the item storage area.

In one or more embodiments the cameras in a store may include ceiling cameras mounted on the store's ceiling. These ceiling cameras may be fisheye cameras, for example. Tracking people in the store may include projecting images from ceiling cameras onto a plane parallel to the floor, and analyzing the projected images.

In one or more embodiments the projected images may be analyzed by subtracting a store background image from each, and combining the differences to form a combined mask. Person locations may be identified as high intensity locations in the combined mask.

In one or more embodiments the projected images may be analyzed by inputting them into a machine learning system that outputs an intensity map that contains a likelihood that a person is at each location. The machine learning system may be a convolutional neural network, for example. An illustrative neural network architecture that may be used in one or more embodiments is a first half subnetwork consisting of copies of a feature extraction network, one copy for each projected image, a feature merging layer that combines outputs from the copies of the feature extraction network, and a second half subnetwork that maps combined features into the intensity map.

In one or more embodiments, additional position map inputs may be provided to the machine learning system. Each position map may correspond to a ceiling camera. The value of the position map at each location may a function of the distance between the location and the ceiling camera. Position maps may be input into a convolutional neural network, for example as an additional channel associated with each projected image.

In one or more embodiments the tracked location of a person may be a single point. It may be a point on a plane, such as the plane parallel to the floor onto which ceiling camera images are projected. In one or more embodiments the field of influence volume around a person may be a translated copy of a standardized shape, such as a cylinder for example.

One or more embodiments may include one or more modular shelves. Each modular shelf may contain at least one camera module on the bottom of the shelf, at least one lighting module on the bottom of the shelf, a right-facing camera on or near the left edge of the shelf, a left-facing camera on or near the right edge of the shelf, a processor, and a network switch. The camera module may contain two or more downward-facing cameras.

Modular shelves may function as item storage areas. The downward-facing cameras in a shelf may view items on the shelf below.

The position of camera modules and lighting modules in a modular shelf may be adjustable. The modular shelf may have a front rail and back rail onto which the camera and lighting modules may be mounted and adjusted. The camera modules may have one or more slots into which the downward-facing cameras are attached. The position of the downward-facing cameras in the slots may be adjustable.

One or more embodiments may include a modular ceiling. The modular ceiling may have a longitudinal rail mounted to the store's ceiling, and one or more transverse rails mounted to the longitudinal rail. The position of each transverse rail along the longitudinal rail may be adjustable. One or more integrated lighting-camera modules may be mounted to each transverse rail. The position of each integrated lighting-camera module may be adjustable along the transverse rail. An integrated lighting-camera module may include a lighting element surrounding a center area, and two or more ceiling cameras mounted in the center area. The ceiling cameras may be mounted to a camera module in the center area with one or more slots into which the cameras are mounted; the positions of the cameras in the slots may be adjustable.

BRIEF DESCRIPTION OF THE DRAWINGS

The patent or application file contains at least one drawing executed in color. Copies of this patent or patent application publication with color drawing(s) will be provided by the Office upon request and payment of the necessary fee.

The above and other aspects, features and advantages of the invention will be more apparent from the following more particular description thereof, presented in conjunction with the following drawings wherein:

FIG. **1** illustrates operation of an embodiment of the invention that analyzes images from cameras in a store to detect that a person has removed a product from a shelf.

FIG. **2** continues the example shown in FIG. **1** to show automated checkout when the person leaves the store with an item.

FIG. **3** shows an illustrative method of determining that an item has been removed from a shelf by feeding before and after images of the shelf to a neural network to detect what item has been taken, moved, or put back wherein the neural network may be implemented in one or more embodiments of the invention through a Siamese neural network with two image inputs for example.

FIG. **4** illustrates training the neural network shown in FIG. **3**.

FIG. **4**A illustrates an embodiment that allows manual review and correction of a detection of an item taken by a shopper and retraining of the neural network with the corrected example.

FIG. **5** shows an illustrative embodiment that identifies people in a store based on distinguishing characteristics such as body measurements and clothing color.

FIGS. **6**A through **6**E illustrate how one or more embodiments of the invention may determine a field of influence volume around a person by finding landmarks on the person's body and calculating an offset distance from these landmarks.

FIGS. **7**A and **7**B illustrate a different method of determining a field of influence volume around a person by calculating a probability distribution for the location of landmarks on a person's body and setting the volume to include a specified amount of the probability distribution.

FIG. **8** shows an illustrative method for tracking a person's movements through a store, which uses a particle filter for a probability distribution of the person's state, along with a physics model for motion prediction and a measurement model based on camera image projection observations.

FIG. **9** shows a conceptual model for how one or more embodiments may combine tracking of a person's field of influence with detection of item motion to attribute the motion to a person.

FIG. **10** illustrates an embodiment that attributes item movement to a person by intersecting the person's field of influence volume with an item storage area, such as a shelf and feeding images of the intersected region to a neural network for item detection.

FIG. **11** shows screenshots of an embodiment of the system that tracks two people in a store and detects when one of the tracked people picks up an item.

FIG. **12** shows screenshots of the item storage area of FIG. **11**, illustrating how two different images of the item storage area may be input into a neural network for detection of the item that was moved by the person in the store.

FIG. **13** shows the results of the neural network classification in FIG. **12**, which tags the people in the store with the items that they move or touch.

FIG. **14** shows a screenshot of an embodiment that identifies a person in a store and builds a 3D field of influence volume around the identified landmarks on the person.

FIG. **15** shows tracking of the person of FIG. **14** as he moves through the store.

FIG. **16** illustrates an embodiment that applies multiple types of camera calibration corrections to images.

FIG. **17** illustrates an embodiment that generates camera calibration data by capturing images of markers placed throughout a store and also corrects for color variations due to hue, saturation or luminance changes across the store and across time.

FIG. **18** illustrates an embodiment that calculates an optimal camera configuration for a store by iteratively optimizing a cost function that measures the number of cameras and the coverage of items by camera fields of view.

FIG. **19** illustrates an embodiment installed at a gas station that extends an authorization from a card reader at a gas pump to provide automated access to a store where a person may take products and have them charged automatically to the card account.

FIG. **20** shows a variation of the embodiment of FIG. **19**, where a locked case containing products is automatically unlocked when the person who paid at a pump is at the case.

FIG. **21** continues the example of FIG. **20**, showing that the products taken by the person from the case may be tracked using cameras or other sensors and may be charged to the card account used at the pump.

FIG. **22** continues the example of FIG. **19**, illustrating tracking the person once he or she enters the store, analyzing images to determine what products the person has taken and charging the account associated with the card entered at the pump.

FIG. **23** shows a variation of the example of FIG. **22**, illustrating tracking that the person picks up and then later puts down an item, so that the item is not charged to the person.

FIG. **24** shows another variation of the example of FIG. **19**, where the authorization obtained at the pump may apply to a group of people in a car.

FIGS. **25**A, **25**B and **25**C illustrate an embodiment that queries a user as to whether to extend authorization from the pump to purchases at a store for the user and also for other occupants of the car.

FIGS. **26**A through **26**F show illustrative camera images from six ceiling-mounted fisheye cameras that may be used for tracking people through a store.

FIGS. **27**A, **27**B, and **27**C show projections of three of the fisheye camera images from FIGS. **26**A through **26**F onto a horizontal plane one meter above the floor.

FIGS. **28**A, **28**B, and **28**C show binary masks of the foreground objects in FIGS. **27**A, **27**B, and **27**C, respectively, as determined for example by background subtraction or motion filtering. FIG. **28**D shows a composite foreground mask that combines all camera image projections to determine the position of people in the store.

FIGS. **29**A through **29**F show a cylinder generated around one of the persons in the store, as viewed from each of the six fisheye cameras.

FIGS. **30**A through **30**F show projections of the six fisheye camera views onto the cylinders shown in FIGS. **29**A through **29**F, respectively. FIG. **30**G shows a composite of the six projections of FIGS. **30**A through **30**F.

FIGS. **31**A and **31**B show screenshots at two different points in time of an embodiment of a people tracking system using the fisheye cameras described above.

FIG. **32** shows an illustrative embodiment that uses a machine learning system to detect person locations from camera images.

FIG. **32**A shows generation of 3D or 2D fields of influence around person locations generated by a machine learning system.

FIG. **33** illustrates projection of ceiling camera images onto a plane parallel to the floor, so that pixels corresponding to the same person location on this plane are aligned in the projected images.

FIGS. **34**A and **34**B show an artificial 3D scene that is used in FIGS. **35** through **41** to illustrate embodiments of the invention that use projected images and machine learning for person detection.

FIG. **35** shows fisheye camera images captured by the ceiling cameras in the scene.

FIG. **36** shows the fisheye camera images of FIG. **35** projected onto a common plane.

FIG. **37** shows the overlap of the projected images of FIG. **36**, illustrating the coincidence of pixels for persons at the intersection of the projected plane.

FIG. **38** shows an illustrative embodiment that augments projected images with a position weight map that reflects the distance of each point from the camera that captures each image.

FIG. **39** shows an illustrative machine learning system with inputs from each camera in a store, where each input has four channels representing three color channels augmented with a position weight channel.

FIG. **40** shows an illustrative neural network architecture that may be used in one or more embodiments to detect persons from camera images.

FIG. **41** shows an illustrative process of generating training data for a machine learning person detection system.

FIG. **42** shows an illustrative store with modular "smart" shelves that integrate cameras, lighting, processing, and communication to detect movement of items on the shelves.

FIG. **43** shows a front view of an illustrative embodiment of a smart shelf.

FIGS. **44**A, **44**B, and **44**C show top, side, and bottom views of the smart shelf of FIG. **43**.

FIG. **45** shows a bottom view of the smart shelf of FIG. **44**C with the electronics covers removed to show the components.

FIGS. **46**A and **46**B show bottom and side views, respectively, of a camera module that may be installed into the smart shelf of FIG. **45**.

FIG. **47** shows a rail mounting system that may be used on the smart shelf of FIG. **45**, which allows lighting and camera modules to be installed at any desired positions along the shelf.

FIG. **48** shows an illustrative store with a modular, "smart" ceiling system into which camera and lighting modules may be installed at any desired positions and spacings.

FIG. **49** shows an illustrative smart ceiling system that supports installation of integrated lighting-camera modules at any desired horizontal positions.

FIG. **50** shows a closeup view of a portion of the smart ceiling system of FIG. **49**, showing the main longitudinal rail, and a moveable transverse rail onto which integrated lighting-camera modules are mounted.

FIG. **51** shows a closeup view of an integrated lighting-camera module of FIG. **50**.

DETAILED DESCRIPTION OF THE INVENTION

An autonomous store tracking system will now be described. Embodiments may track a person by analyzing camera images and may therefore extend an authorization obtained by this person at one point in time and space to a different point in time or space. Embodiments may also enable an autonomous store system that analyzes camera images to track people and their interactions with items and may also enable camera calibration, optimal camera placement and computer interaction with a point of sale system. The computer interaction may involve a mobile device and a point of sale system for example. In the following exemplary description, numerous specific details are set forth in order to provide a more thorough understanding of embodiments of the invention. It will be apparent, however, to an artisan of ordinary skill that the present invention may be practiced without incorporating all aspects of the specific details described herein. In other instances, specific features, quantities, or measurements well known to those of ordinary skill in the art have not been described in detail so as not to obscure the invention. Readers should note that although examples of the invention are set forth herein, the claims and the full scope of any equivalents, are what define the metes and bounds of the invention.

FIG. **1** shows an embodiment of an automated store. A store may be any location, building, room, area, region, or site in which items of any kind are located, stored, sold, or displayed, or through which people move. For example, without limitation, a store may be a retail store, a warehouse, a museum, a gallery, a mall, a display room, an educational facility, a public area, a lobby, an office, a home, an apartment, a dormitory, or a hospital or other health facility. Items located in the store may be of any type, including but not limited to products that are for sale or rent.

In the illustrative embodiment shown in FIG. **1**, store **101** has an item storage area **102**, which in this example is a shelf. Item storage areas may be of any type, size, shape and location. They may be of fixed dimensions or they may be of variable size, shape, or location. Item storage areas may include for example, without limitation, shelves, bins, floors, racks, refrigerators, freezers, closets, hangers, carts, containers, boards, hooks, or dispensers. In the example of FIG. **1**, items **111**, **112**, **113** and **114** are located on item storage area **102**. Cameras **121** and **122** are located in the store and they

are positioned to observe all or portions of the store and the item storage area. Images from the cameras are analyzed to determine the presence and actions of people in the store, such as person **103** and in particular to determine the interactions of these people with items **111-114** in the store. In one or more embodiments, camera images may be the only input required or used to track people and their interactions with items. In one or more embodiments, camera image data may be augmented with other information to track people and their interactions with items. One or more embodiments of the system may utilize images to track people and their interactions with items for example without the use of any identification tags, such as RFID tags or any other non-image based identifiers associated with each item.

FIG. **1** illustrates two cameras, camera **121** and camera **122**. In one or more embodiments, any number of cameras may be employed to track people and items. Cameras may be of any type; for example, cameras may be 2D, 3D, or 4D. 3D cameras may be stereo cameras, or they may use other technologies such as rangefinders to obtain depth information. One or more embodiments may use only 2D cameras and may for example determine 3D locations by triangulating views of people and items from multiple 2D cameras. 4D cameras may include any type of camera that can also gather or calculate depth over time, e.g., 3D video cameras.

Cameras **121** and **122** observe the item storage area **102** and the region or regions of store **101** through which people may move. Different cameras may observe different item storage areas or different regions of the store. Cameras may have overlapping views in one or more embodiments. Tracking of a person moving through the store may involve multiple cameras, since in some embodiments no single camera may have a view of the entire store.

Camera images are input into processor **130**, which analyzes the images to track people and items in the store. Processor **130** may be any type or types of computer or other device. In one or more embodiments, processor **130** may be a network of multiple processors. When processor **130** is a network of processors, different processors in the network may analyze images from different cameras. Processors in the network may share information and cooperate to analyze images in any desired manner. The processor or processors **130** may be onsite in the store **101**, or offsite, or a combination of onsite and offsite processing may be employed. Cameras **121** and **122** may transfer data to the processor over any type or types of network or link, including wired or wireless connections. Processor **130** includes or couples with memory, RAM or disk and may be utilized as a non-transitory data storage computer-readable media that embodiments of the invention may utilize or otherwise include to implement all functionality detailed herein.

Processor or processors **130** may also access or receive a 3D model **131** of the store and may use this 3D model to analyze camera images. The model **131** may for example describe the store dimensions, the locations of item storage areas and items and the location and orientation of the cameras. The model may for example include the floorplan of the store, as well as models of item storage areas such as shelves and displays. This model may for example be derived from a store's planogram, which details the location of all shelving units, their height, as well as which items are placed on them. Planograms are common in retail spaces, so should be available for most stores. Using this planogram, measurements may for example be converted into a 3D model using a 3D CAD package.

If no planogram is available, other techniques may be used to obtain the item storage locations. One illustrative technique is to measure the locations, shapes and sizes of all shelves and displays within the store. These measurements can then be directly converted into a planogram or 3D CAD model. A second illustrative technique involves taking a series of images of all surfaces within the store including the walls, floors and ceilings. Enough images may be taken so that each surface can be seen in at least two images. Images can be either still images or video frames. Using these images, standard 3D reconstruction techniques can be used to reconstruct a complete model of the store in 3D.

In one or more embodiments, a 3D model **131** used for analyzing camera images may describe only a portion of a site, or it may describe only selected features of the site. For example, it may describe only the location and orientation of one or more cameras in the site; this information may be obtained for example from extrinsic calibration of camera parameters. A basic, minimal 3D model may contain only this camera information. In one or more embodiments, geometry describing all or part of a store may be added to the 3D model for certain applications, such as associating the location of people in the store with specific product storage areas. A 3D model may also be used to determine occlusions, which may affect the analysis of camera images. For example, a 3D model may determine that a person is behind a cabinet and is therefore occluded by the cabinet from the viewpoint of a camera; tracking of the person or extraction of the person's appearance may therefore not use images from that camera while the person is occluded.

Cameras **121** and **122** (and other cameras in store **101** if available) may observe item storage areas such as area **102**, as well as areas of the store where people enter, leave and circulate. By analyzing camera images over time, the processor **130** may track people as they move through the store. For example, person **103** is observed at time **141** standing near item storage area **102** and at a later time **142** after he has moved away from the item storage area. Using possibly multiple cameras to triangulate the person's position and the 3D store model **131**, the processor **130** may detect that person **103** is close enough to item storage area **102** at time **141** to move items on the shelf. By comparing images of storage area **102** at times **141** and **142**, the system may detect that item **111** has been moved and may attribute this motion to person **103** since that person was proximal to the item in the time range between **141** and **142**. Therefore, the system derives information **150** that the person **103** took item **111** from shelf **102**. This information may be used for example for automated checkout, for shoplifting detection, for analytics of shopper behavior or store organization, or for any other purposes. In this illustrative example, person **103** is given an anonymous tag **151** for tracking purposes. This tag may or may not be cross referenced to other information such as for example a shopper's credit card information; in one or more embodiments the tag may be completely anonymous and may be used only to track a person through the store. This enables association of a person with products without require identification of who that particular user is. This is important in locales where people typically wear masks when sick, or other garments which cover the face for example. Also shown is electronic device **119** that generally includes a display that the system may utilize to show the person's list of items, i.e., shopping cart list and with which the person may pay for the items for example.

In one or more embodiments, camera images may be supplemented with other sensor data to determine which products are removed or the quantity of a product that is taken or dispensed. For example, a product shelf such as shelf **102** may have weight sensors or motion sensors that

assist in detecting that products are taken, moved, or replaced on the shelf. One or more embodiments may receive and process data indicating the quantity of a product that is taken or dispensed, and may attribute this quantity to a person, for example to charge this quantity to the person's account. For example, a dispenser of a liquid such as a beverage may have a flow sensor that measures the amount of liquid dispensed; data from the flow sensor may be transmitted to the system to attribute this amount to a person proximal to the dispenser at the time of dispensing. A person may also press a button or provide other input to determine what products or quantities should be dispensed; data from the button or other input device may be transmitted to the system to determine what items and quantities to attribute to a person.

FIG. **2** continues the example of FIG. **1** to show an automated checkout. In one or more embodiments, processor **130** or another linked system may detect that a person **103** is leaving a store or is entering an automated checkout area. For example, a camera or cameras such as camera **202** may track person **103** as he or she exits the store. If the system **130** has determined that person **103** has an item, such as item **111** and if the system is configured to support automated checkout, then it may transmit a message **203** or otherwise interface with a checkout system such as a point of sale system **210**. This message may for example trigger an automated charge **211** for the item (or items) believed to be taken by person **103**, which may for example be sent to financial institution or system **212**. In one or more embodiments a message **213** may also be displayed or otherwise transmitted to person **103** confirming the charge, e.g., on the person's electronic device **119** shown in FIG. **1**. The message **213** may for example be displayed on a display visible to the person exiting or in the checkout area, or it may be transmitted for example via a text message or email to the person, for example to a computer or mobile device **119** (see FIG. **1**) associated with the user. In one or more embodiments the message **213** may be translated to a spoken message. The fully automated charge **211** may for example require that the identity of person **103** be associated with financial information, such as a credit card for example. One or more embodiments may support other forms of checkout that may for example not require a human cashier but may ask person **103** to provide a form of payment upon checkout or exit. A potential benefit of an automated checkout system such as that shown in FIG. **2** is that the labor required for the store may be eliminated or greatly reduced. In one or more embodiments, the list of items that the store believes the user has taken may be sent to a mobile device associated with the user for the user's review or approval.

As illustrated in FIG. **1**, in one or more embodiments analysis of a sequence of two or more camera images may be used to determine that a person in a store has interacted with an item in an item storage area. FIG. **3** shows an illustrative embodiment that uses an artificial neural network **300** to identify an item that has been moved from a pair of images, e.g., an image **301** obtained prior to the move of the item and an image **302** obtained after the move of the item. One or more embodiments may analyze any number of images, including but not limited to two images. These images **301** and **302** may be fed as inputs into input layer **311** of a neural network **300**, for example. (Each color channel of each pixel of each image may for example be set as the value of an input neuron in input layer **311** of the neural network.) The neural network **300** may then have any number of additional layers **312**, connected and organized in any desired fashion. For example, without limitation, the neural network may employ any number of fully connected layers, convolutional layers, recurrent layers, or any other type of neurons or connections. In one or more embodiments the neural network **300** may be a Siamese neural network organized to compare the two images **301** and **302**. In one or more embodiments, neural network **300** may be a generative adversarial network, or any other type of network that performs input-output mapping.

The output layer **313** of the neural network **300** may for example contain probabilities that each item was moved. One or more embodiments may select the item with the highest probability, in this case output neuron **313** and associate movement of this item with the person near the item storage area at the time of the movement of the item. In one or more embodiments there may be an output indicating no item was moved.

The neural network **300** of FIG. **3** also has outputs classifying the type of movement of the item. In this illustrative example there are three types of motions: a take action **321**, which indicates for example that the item appeared in image **301** but not in image **302**; a put action **322**, which indicates for example that the item appears in image **302** but not in image **301**; and a move action **323**, which indicates for example that the item appears in both images but in a different location. These actions are illustrative; one or more embodiments may classify movement or rearrangement of items into any desired classes and may for example assign a probability to each class. In one or more embodiments, separate neural networks may be used to determine the item probabilities and the action class probabilities. In the example of FIG. **3**, the take class **321** has the highest calculated probability, indicating that the system most likely detects that the person near the image storage area has taken the item away from the storage area.

The neural network analysis as indicated in FIG. **3** to determine which item or items have been moved and the types of movement actions performed is an illustrative technique for image analysis that may be used in one or more embodiments. One or more embodiments may use any desired technique or algorithm to analyze images to determine items that have moved and the actions that have been performed. For example, one or more embodiments may perform simple frame differences on images **301** and **302** to identify movement of items. One or more embodiments may preprocess images **301** and **302** in any desired manner prior to feeding them to a neural network or other analysis system. For example, without limitation, preprocessing may align images, remove shadows, equalize lighting, correct color differences, or perform any other modifications. Images may be processed with any classical image processing algorithms such as color space transformation, edge detection, smoothing or sharpening, application of morphological operators, or convolution with filters.

One or more embodiments may use machine learning techniques to derive classification algorithms such as the neural network algorithm applied in FIG. **3**. FIG. **4** shows an illustrative process for learning the weights of the neural network **300** of FIG. **3**. A training set **401** of examples may be collected or generated and used to train network **300**. Training examples such as examples **402** and **403** may for example include before and after images of an item storage area and output labels **412** and **413** that indicate the item moved and the type of action applied to the item. These examples may be constructed manually, or in one or more embodiments there may be an automated training process that captures images and then uses checkout data that associates items with persons to build training examples.

FIG. **4**A shows an example of augmenting the training data with examples that correct misclassifications by the system. In this example, the store checkout is not fully automated; instead, a cashier **451** assists the customer with checkout. The system **130** has analyzed camera images and has sent message **452** to the cashier's point of sale system **453**. The message contains the system's determination of the item that the customer has removed from the item storage area **102**. However, in this case the system has made an error. Cashier **451** notices the error and enters a correction into the point of sale system with the correct item. The corrected item and the images from the camera may then be transmitted as a new training example **454** that may be used to retrain neural network **300**. In time, the cashier may be eliminated when the error rate converges to an acceptable predefined level. In one or more embodiments, the user may show the erroneous item to the neural network via a camera and train the system without cashier **451**. In other embodiments, cashier **451** may be remote and accessed via any communication method including video or image and audio-based systems.

In one or more embodiments, people in the store may be tracked as they move through the store. Since multiple people may be moving in the store simultaneously, it may be beneficial to distinguish between persons using image analysis, so that people can be correctly tracked. FIG. **5** shows an illustrative method that may be used to distinguish among different persons. As a new person **501** enters a store or enters a specified area or areas of the store at time **510**, images of the person from cameras such as cameras **511**, **512** and **513** may be analyzed to determine certain characteristics **531** of the person's appearance that can be used to distinguish that person from other people in the store. These distinguishing characteristics may include for example, without limitation: the size or shape of certain body parts; the color, shape, style, or size of the person's hair; distances between selected landmarks on the person's body or clothing; the color, texture, materials, style, size, or type of the person's clothing, jewelry, accessories, or possessions; the type of gait the person uses when walking or moving; the speed or motion the person makes with any part of their body such as hands, arms, legs, or head; and gestures the person makes. One or more embodiments may use high resolution camera images to observe biometric information such as a person's fingerprints or handprints, retina, or other features.

In the example shown in FIG. **5**, at time **520** a person **502** enters the store and is detected to be a new person. New distinguishing characteristics **532** are measured and observed for this person. The original person **501** has been tracked and is now observed to be at a new location **533**. The observations of the person at location **533** are matched to the distinguishing characteristics **531** to identify the person as person **501**.

In the example of FIG. **5**, although distinguishing characteristics are identified for persons **501** and **502**, the identities of these individuals remain anonymous. Tags **541** and **542** are assigned to these individuals for internal tracking purposes, but the persons' actual identities are not known. This anonymous tracking may be beneficial in environments where individuals do not want their identities to be known to the autonomous store system. Moreover, sensitive identifying information, such as for example images of a person's face, need not be used for tracking; one or more embodiments may track people based on other less sensitive information such as the distinguishing characteristics **531** and **532**. As previously described, in some areas, people wear masks when sick or otherwise wear face garments, making identification based on a user's face impossible.

The distinguishing characteristics **531** and **532** of persons **501** and **502** may or may not be saved over time to recognize return visitors to the store. In some situations, a store may want to track return visitors. For example, shopper behavior may be tracked over multiple visits if the distinguishing characteristics are saved and retrieved for each visitor. Saving this information may also be useful to identify shoplifters who have previously stolen from the store, so that the store personnel or authorities can be alerted when a shoplifter or potential shoplifter returns to the store. In other situations, a store may want to delete distinguishing information when a shopper leaves the store, for example if there are potential concern that the store may be collecting information that the shopper's do not want saved over time.

In one or more embodiments, the system may calculate a 3D field of influence volume around a person as it tracks the person's movement through the store. This 3D field of influence volume may for example indicate a region in which the person can potentially touch or move items. A detection of an item that has moved may for example be associated with a person being tracked only if the 3D field of influence volume for that person is near the item at the time of the item's movement.

Various methods may be used to calculate a 3D field of influence volume around a person. FIGS. **6**A through **6**E illustrate a method that may be used in one or more embodiments. (These figures illustrate the construction of a field of influence volume using 2D figures, for ease of illustration, but the method may be applied in three dimensions to build a 3D volume around the person.) Based on an image or images **601** of a person, image analysis may be used to identify landmarks on the person's body. For example, landmark **602** may be the left elbow of the person. FIG. **6**B illustrates an analysis process that identifies **18** different landmarks on the person's body. One or more embodiments may identify any number of landmarks on a body, at any desired level of detail. Landmarks may be connected in a skeleton in order to track the movement of the person's joints. Once landmark locations are identified in the 3D space associated with the store, one method for constructing a 3D field of influence volume is to calculate a sphere around each landmark with a radius of a specified threshold distance. For example, one or more embodiments may use a threshold distance of 25 cm offset from each landmark. FIG. **6**C shows sphere **603** with radius **604** around landmark **602**. These spheres may be constructed around each landmark, as illustrated in FIG. **6**D. The 3D field of influence volume may then be calculated as the union of these spheres around the landmarks, as illustrated with 3D field of influence volume **605** in FIG. **6**E.

Another method of calculating a 3D field of influence volume around a person is to calculate a probability distribution for the location of each landmark and to define the 3D field of influence volume around a landmark as a region in space that contains a specified threshold amount of probability from this probability distribution. This method is illustrated in FIGS. **7**A and **7**B. Images of a person are used to calculate landmark positions **701**, as described with respect to FIG. **6**B. As the person is tracked through the store, uncertainty in the tracking process results in a probability distribution for the 3D location of each landmark. This probability distribution may be calculated and tracked using various methods, including a particle filter as described below with respect to FIG. **8**. For example, for the right elbow landmark **702** in FIG. **7**A, a probability density

703 may be calculated for the position of the landmark. (This density is shown in FIG. 7A as a 2D figure for ease of illustration, but in tracking it will generally be a 3D spatial probability distribution.) A volume may be determined that contains a specified threshold probability amount of this probability density for each landmark. For example, the volume enclosed by surface may enclose 95% (or any other desired amount) of the probability distribution 703. The 3D field of influence volume around a person may then be calculated as the union of these volumes 704 around each landmark, as illustrated in FIG. 7B. The shape and size of the volumes around each landmark may differ, reflecting differences in the uncertainties for tracking the different landmarks.

FIG. 8 illustrates a technique that may be used in one or more embodiments to track a person over time as he or she moves through a store. The state of a person at any point in time may for example be represented as a probability distribution of certain state variables such as the position and velocity (in three dimensions) of specific landmarks on the person's body. One approach to representing this probability distribution is to use a particle filter, where a set of particles is propagated over time to represent weighted samples from the distribution. In the example of FIG. 8, two particles 802 and 803 are shown for illustration; in practice the probability distribution at any point in time may be represented by hundreds or thousands of particles. To propagate state 801 to a subsequent point in time, one or more embodiments may employ an iterative prediction/correction loop. State 801 is first propagated through a prediction step 811, which may for example use a physics model to estimate for each particle what the next state of the particle is. The physics model may include for example, without limitation, constraints on the relative location of landmarks (for example, a constraint that the distance between the left foot and the left knee is fixed), maximum velocities or accelerations at which body parts can move and constraints from barriers in the store, such as floors, walls, fixtures, or other persons. These physics model components are illustrative; one or more embodiments may use any type of physics model or other model to propagate tracking state from one time period to another. The predict step 811 may also reflect uncertainties in movements, so that the spread of the probability distribution may increase over time in each predict step, for example. The particles after the prediction step 811 are then propagated through a correction step 812, which incorporates information obtained from measurements in camera images, as well as other information if available. The correction step uses camera images such as images 821, 822, 823 and information on the camera projections of each camera as well as other camera calibration data if available. As illustrated in images 821, 822 and 823, camera images may provide only partial information due to occlusion of the person or to images that capture only a portion of the person's body. The information that is available is used to correct the predictions, which may for example reduce the uncertainty in the probability distribution of the person's state. This prediction/correction loop may be repeated at any desired interval to track the person through the store.

By tracking a person as he or she moves through the store, one or more embodiments of the system may generate a 3D trajectory of the person through the store. This 3D trajectory may be combined with information on movement of items in item storage areas to associate people with the items they interact with. If the person's trajectory is proximal to the item at a time when the item is moved, then the movement of the item may be attributed to that person, for example. FIG. 9 illustrates this process. For ease of illustration, the person's trajectory and the item position are shown in two dimensions; one or more embodiments may perform a similar analysis in three dimensions using the 3D model of the store, for example. A trajectory 901 of a person is tracked over time, using a tracking process such as the one illustrated in FIG. 8, for example. For each person, a 3D field of influence volume 902 may be calculated at each point in time, based for example on the location or probability distribution of landmarks on the person's body. (Again, for ease of illustration the field of influence volume shown in FIG. 9 is in the two dimension, although in implementation this volume may be three dimensional.) The system calculates the trajectory of the 3D influence volume through the store. Using camera image analysis such as the analysis illustrated in FIG. 3, motion 903 of an item is detected at a location 904. Since there may be multiple people tracked in a store, the motion may be attributed to the person whose field of influence volume was at or near this location at the time of motion. Trajectory 901 shows that the field of influence volume of this tracked person intersected the location of the moved item during a time interval proximal in time to this motion; hence the item movement may be attributed to this person.

In one or more embodiments the system may optimize the analysis described above with respect to FIG. 9 by looking for item movements only in item storage areas that intersect a person's 3D field of influence volume. FIG. 10 illustrates this process. At a point in time 141 or over a time interval, the tracked 3D field of influence volume 1001 of person 103 is calculated to be near item storage area 102. The system therefore calculates an intersection 1011 of the item storage area 102 and the 3D field of influence volume 1001 around person 1032 and locates camera images that contain views of this region, such as image 1011. At a subsequent time 142, for example when person 103 is determined to have moved away from item storage area 102, an image 1012 (or multiple such images) is obtained of the same intersected region. These two images are then fed as inputs to neural network 300, which may for example detect whether any item was moved, which item was moved (if any) and the type of action that was performed. The detected item motion is attributed to person 103 because this is the person whose field of influence volume intersected the item storage area at the time of motion. By applying the classification analysis of neural network 300 only to images that represent intersections of person's field of influence volume with item storage areas, processing resources may be used efficiently and focused only on item movement that may be attributed to a tracked person.

FIGS. 11 through 15 show screenshots of an embodiment of the system in operation in a typical store environment. FIG. 11 shows three camera images 1101, 1102 and 1103 taken of shoppers moving through the store. In image 1101, two shoppers 1111 and 1112 have been identified and tracked. Image 1101 shows landmarks identified on each shopper that are used for tracking and for generating a 3D field of influence volume around each shopper. Distances between landmarks and other features such as clothing may be used to distinguish between shoppers 1111 and 1112 and to track them individually as they move through the store. Images 1102 and 1103 show views of shopper 1111 as he approaches item storage area 1113 and picks up an item 114 from the item storage area. Images 1121 and 1123 show close up views from images 1101 and 1103, respectively, of item storage area 1113 before and after shopper 1111 picks up the item.

FIG. **12** continues the example shown in FIG. **11** to show how images **1121** and **1123** of the item storage area are fed as inputs into a neural network **1201** to determine what item, if any, has been moved by shopper **1111**. The network assigns the highest probability to item **1202**. FIG. **13** shows how the system attributes motion of this item **1202** to shopper **1111** and assigns an action **1301** to indicate that the shopper picked up the item. This action **1301** may also be detected by neural network **1201**, or by a similar neural network. Similarly, the system has detected that item **1303** has been moved by shopper **1112** and it assigns action **1302** to this item movement.

FIG. **13** also illustrates that the system has detected a "look at" action **1304** by shopper **1111** with respect to item **1202** that the shopper picked up. In one or more embodiments, the system may detect that a person is looking at an item by tracking the eyes of the person (as landmarks, for example) and by projecting a field of view from the eyes towards items. If an item is within the field of view of the eyes, then the person may be identified as looking at the item. For example, in FIG. **13** the field of view projected from the eyes landmarks of shopper **1111** is region **1305** and the system may recognize that item **1202** is within this region. One or more embodiments may detect that a person is looking at an item whether or not that item is moved by the person; for example, a person may look at an item in an item storage area while browsing and may subsequently choose not to touch the item.

In one or more embodiments, other head landmarks instead of or in addition to the eyes may be used to compute head orientation relative to the store reference frame to determine what a person is looking at. Head orientation may be computed for example via 3D triangulated head landmarks. One or more embodiments may estimate head orientation from 2D landmarks using for example a neural network that is trained to estimate gaze in 3D from 2D landmarks.

FIG. **14** shows a screenshot **1400** of the system creating a 3D field of influence volume around a shopper. The surface of the 3D field of influence volume **1401** is represented in this image overlay as a set of dots on the surface. The surface **1401** may be generated as an offset from landmarks identified on the person, such as landmark **1402** for the person's right foot for example. Screenshot **1410** shows the location of the landmarks associated with the person in the 3D model of the store.

FIG. **15** continues the example of FIG. **14** to show tracking of the person and his 3D field of influence volume as he moves through the store in camera images **1501** and **1502** and generation of a trajectory of the person's landmarks in the 3D model of the store in screenshots **1511** and **1512**.

In one or more embodiments, the system may use camera calibration data to transform images obtained from cameras in the store. Calibration data may include for example, without limitation, intrinsic camera parameters, extrinsic camera parameters, temporal calibration data to align camera image feeds to a common time scale and color calibration data to align camera images to a common color scale. FIG. **16** illustrates the process of using camera calibration data to transform images. A sequence of raw images **1601** is obtained from camera **121** in the store. A correction **1602** for intrinsic camera parameters is applied to these raw images, resulting in corrected sequence **1603**. Intrinsic camera parameters may include for example the focal length of the camera, the shape and orientation of the imaging sensor, or lens distortion characteristics. Corrected images **1603** are then transformed in step **1604** to map the images to the 3D store model, using extrinsic camera parameters that describe the camera projection transformation based on the location and orientation of the camera in the store. The resulting transformed images **1605** are projections aligned with respect to a coordinate system **1606** of the store. These transformed images **1605** may then be shifted in time to account for possible time offsets among different cameras in the store. This shifting **1607** synchronizes the frames from the different cameras in the store to a common time scale. In the last transformation **1609**, the color of pixels in the time corrected frames **1608** may be modified to map colors to a common color space across the cameras in the store, resulting in final calibrated frames **1610**. Colors may vary across cameras because of differences in camera hardware or firmware, or because of lighting conditions that vary across the store; color correction **1609** ensures that all cameras view the same object as having the same color, regardless of where the object is in the store. This mapping to a common color space may for example facilitate the tracking of a person or an item selected by a person as the person or item moves from the field of view of one camera to another camera, since tracking may rely in part on the color of the person or item.

The camera calibration data illustrated in FIG. **16** may be obtained from any desired source. One or more embodiments may also include systems, processes, or methods to generate any or all of this camera calibration data. FIG. **17** illustrates an embodiment that generates camera calibration data **1701**, including for example any or all of intrinsic camera parameters, extrinsic camera parameter, time offsets for temporal synchronization and color mapping from each camera to a common color space. Store **1702** contains for this example three cameras, **1703**, **1704** and **1705**. Images from these cameras are captured during calibration procedures and are analyzed by camera calibration system **1710**. This system may be the same as or different from the system or systems used to track persons and items during store operations. Calibration system **1710** may include or communicate with one or more processors. For calibration of intrinsic camera parameters, standard camera calibration grids for example may be placed in the store **1702**. For calibration of extrinsic camera parameters, markers of a known size and shape may for example be placed in known locations in the store, so that the position and orientation of cameras **1703**, **1704** and **1705** may be derived from the images of the markers. Alternatively, an iterative procedure may be used that simultaneously solves for marker positions and for camera positions and orientations.

A temporal calibration procedure that may be used in one or more embodiments is to place a source of light **1705** in the store and to pulse a flash of light from the source **1705**. The time that each camera observes the flash may be used to derive the time offset of each camera from a common time scale. The light flashed from source **1705** may be visible, infrared, or of any desired wavelength or wavelengths. If all cameras cannot observe a single source, then either multiple synchronized light sources may be used, or cameras may be iteratively synchronized in overlapping groups to a common time scale.

A color calibration procedure that may be used in one or more embodiments is to place one or more markers of known colors into the store and to generate color mappings from each camera into a known color space based on the images of these markers observed by the cameras. For example, color markers **1721**, **1722** and **1723** may be placed in the store; each marker may for example have a grid of

standard color squares. In one or more embodiments the color markers may also be used for calibration of extrinsic parameters; for example, they may be placed in known locations as shown in FIG. **17**. In one or more embodiments items in the store may be used for color calibration if for example they are of a known color.

Based on the observed colors of the markers **1721**, **1722** and **1723** in a specific camera, a mapping may be derived to transform the observed colors of the camera to a standard color space. This mapping may be linear or nonlinear. The mapping may be derived for example using a regression or using any desired functional approximation methodology.

The observed color of any object in the store, even in a camera that is color calibrated to a standard color space, depends on the lighting at the location of the object in the store. For example, in store **1702** an object near light **1731** or near window **1732** may appear brighter than objects at other locations in the store. To correct for the effect of lighting variations on color, one or more embodiments may create and/or use a map of the luminance or other lighting characteristics across the store. This luminance map may be generated based on observations of lighting intensity from cameras or from light sensors, on models of the store lighting, or on a combination thereof. In the example of FIG. **17**, illustrative luminance map **1741** may be generated during or prior to camera calibration and it may be used in mapping camera colors to a standard color space. Since lighting conditions may change at different times of day, one or more embodiments may generate different luminance maps for different times or time periods. For example, luminance map **1742** may be used for nighttime operation, when light from window **1732** is diminished but store light **1731** continues to operate.

In one or more embodiments, filters may be added to light sources or to cameras, or both, to improve tracking and detection. For example, point lights may cause glare in camera images from shiny products. Polarizing filters on light may reduce this glare, since polarized light generates less glare. Polarizing filters on light sources may be combined with polarizers on cameras to further reduce glare.

In addition to or instead of using different luminance maps at different times to account for changes in lighting conditions, one or more embodiments may recalibrate cameras as needed to account for the effects of changing lighting conditions on camera color maps. For example, a timer **1751** may trigger camera calibration procedure **1710**, so that for example camera colors are recalibrated at different times of day. Alternatively, or in addition, light sensors **1752** located in store **1702** may trigger camera calibration procedure **1710** when the sensor or sensors detect that lighting conditions have changed or may have changed. Embodiments of the system may also sub-map calibration to specific areas of images, for example if window **1732** allows sunlight in to a portion of the store. In other words, the calibration data may also be based on area and time to provide even more accurate results.

In one or more embodiments, camera placement optimization may be utilized in the system. For example, in a 2D camera scenario, one method that can be utilized is to assign a cost function to camera positions to optimize the placement and number of cameras for a particular store. In one embodiment, assigning a penalty of 1000 to any item that is only found in one image from the cameras results in a large penalty for any item only viewable by one camera. Assigning a penalty of 1 to the number of cameras results in a slight penalty for additional cameras required for the store. By penalizing camera placements that do not produce at least two images or a stereoscopic image of each item, then the number of items for which 3D locations cannot be obtained is heavily penalized so that the final camera placement is under a predefined cost. One or more embodiments thus converge on a set of camera placements where two different viewpoints to all items is eliminated given enough cameras. By placing a cost function on the number of cameras, the iterative solution according to this embodiment thus is employed to find at least one solution with a minimal number of cameras for the store. As shown in the upper row of FIG. **18**, the items on the left side of the store only have one camera, the middle camera pointing towards them. Thus, those items in the upper right table incur a penalty of 1000 each. Since there are 3 cameras in this iteration, the total cost is 2003. In the next iteration, a camera is added as shown in the middle row of the figure. Since all items can now be seen by at least two cameras, the cost drops to zero for items, while another camera has been added so that the total cost is 4. In the bottom row as shown for this iteration, a camera is removed, for example by determining that certain items are viewed by more than 2 cameras as shown in the middle column of the middle row table, showing 3 views for 4 items. After removing the far-left camera in the bottom row store, the cost decreases by 1, thus the total cost is 3. Any number of camera positions, orientations and types may be utilized in embodiments of the system. One or more embodiments of the system may optimize the number of cameras by using existing security cameras in a store and by moving those cameras if needed or augmenting the number of cameras for the store to leverage existing video infrastructure in a store, for example in accordance with the camera calibration previously described. Any other method of placing and orienting cameras, for example equal spacing and a predefined angle to set an initial scenario may be utilized.

In one or more embodiments, one or more of the techniques described above to track people and their interactions with an environment may be applied to extend an authorization obtained by a person at one point in time and space to another point in time or space. For example, an authorization may be obtained by a person at an entry point to an area or a check point in the area and at an initial point in time. The authorization may authorize the person to perform one or more actions, such as for example to enter a secure environment such as a locked building, or to charge purchases to an account associated with the person. The system may then track this person to a second location at a subsequent point in time and may associate the previously obtained authorization with that person at the second location and at the subsequent point in time. This extension of an authorization across time and space may simplify the interaction of the person with the environment. For example, a person may need to or choose to present a credential (such as a payment card) at the entry point to obtain an authorization to perform purchases; because the system may track that person afterwards, this credential may not need to be presented again to use the previously obtained authorization. This extension of authorization may for example be useful in automated stores in conjunction with the techniques described above to determine which items a person interacts with or takes within the store; a person might for example present a card at a store entrance or at a payment kiosk or card reader associated with the store and then simply take items as desired and be charged for them automatically upon leaving the store, without performing any explicit checkout.

FIG. **19** shows an illustrative embodiment that enables authorization extension using tracking via analysis of cam-

era images. This figure and several subsequent figures illustrate one or more aspects of authorization extension using a gas station example. This example is illustrative; one or more embodiments may enable authorization extension at any type of site or area. For example, without limitation, authorization extension may be applied to or integrated into all of or any portion of a building, a multi-building complex, a store, a restaurant, a hotel, a school, a campus, a mall, a parking lot, an indoor or outdoor market, a residential building or complex, a room, a stadium, a field, an arena, a recreational area, a park, a playground, a museum, or a gallery. It may be applied or integrated into any environment where an authorization obtained at one time and place may be extended to a different time or different place. It may be applied to extend any type of authorization.

In the example shown in FIG. **19**, a person **1901** arrives at a gas station and goes to gas pump **1902**. To obtain gas (or potentially to authorize other actions without obtaining gas), person **1901** presents a credential **1904**, such as for example a credit or debit card, into credential reader **1905** on or near the pump **1902**. The credential reader **1905** transmits a message **1906** to a bank or clearinghouse **212** to obtain an authorization **1907**, which allows user **1901** to pump gas from pump **1902**.

In one or more embodiments, a person may present any type of credential to any type of credential reader to obtain an authorization. For example, without limitation, a credential may be a credit card, a debit card, a bank card, an RFID tag, a mobile payment device, a mobile wallet device, a mobile phone, a smart phone, a smart watch, smart glasses or goggles, a key fob, an identity card, a driver's license, a passport, a password, a PIN, a code, a phone number, or a biometric identifier. A credential may be integrated into or attached to any device carried by a person, such as a mobile phone, smart phone, smart watch, smart glasses, key fob, smart goggles, tablet, or computer. A credential may be worn by a person or integrated into an item of clothing or an accessory worn by a person. A credential may be passive or active. A credential may or may not be linked to a payment mechanism or an account. In one or more embodiments a credential may be a password, PIN, code, phone number, or other data typed or spoken or otherwise entered by a person into a credential reader. A credential reader may be any device or combination of devices that can read or accept a presented credential. A credential reader may or may not be linked to a remote authorization system like bank **212**. In one or more embodiments a credential reader may have local information to authorize a user based on a presented credential without communicating with other systems. A credential reader may read, recognize, accept, authenticate, or otherwise process a credential using any type of technology. For example, without limitation, a credential reader may have a magnetic stripe reader, a chip card reader, an RFID tag reader, an optical reader or scanner, a biometric reader such as a fingerprint scanner, a near field communication receiver, a Bluetooth receiver, a Wi-Fi receiver, a keyboard or touchscreen for typed input, or a microphone for audio input. A credential reader may receive signals, transmit signals, or both.

In one or more embodiments, an authorization obtained by a person may be associated with any action or actions the person is authorized to perform. These actions may include, but are not limited to, financial transactions such as purchases. Actions that may be authorized may include for example, without limitation, entry to or exit from a building, room, or area; purchasing or renting of items, products, or services; use of items, products, or services; or access to controlled information or materials.

In one or more embodiments, a credential reader need not be integrated into a gas pump or into any other device. It may be standalone, attached to or integrated into any device, or distributed across an area. A credential reader may be located in any location in an area, including for example, without limitation, at an entrance, exit, check-in point, checkpoint, control point, gate, door, or other barrier. In one or more embodiments, several credential readers may be located in an area; multiple credential readers may be used simultaneously by different persons.

The embodiment illustrated in FIG. **19** extends the authorization for pumping gas obtained by person **1901** to authorize one or more other actions by this person, without requiring the person to re-present credential **1904**. In this illustrative example, the gas station has an associated convenience store **1903** where customers can purchase products. The authorization extension embodiment may enable the convenience store to be automated, for example without staff. Because the store **1903** may be unmanned, the door **1908** to the store may be locked, for example with a controllable lock **1909**, thereby preventing entry to the store by unauthorized persons. The embodiment described below extends the authorization of person **1901** obtained by presenting credential **1904** at the pump **1902** to enable the person **1901** to enter store **1903** through locked door **1908**.

One or more embodiments may enable authorization extension to allow a user to enter a secured environment of any kind, including but not limited to a store such as convenience store **1903** in FIG. **19**. The secured environment may have an entry that is secured by a barrier, such as for example, without limitation, a door, gate, fence, grate, or window. The barrier may not be a physical device preventing entry; it may be for example an alarm that must be disabled to enter the secured environment without sounding the alarm. In one or more embodiments the barrier may be controllable by the system so that for example commands may be sent to the barrier to allow (or to disallow) entry. For example, without limitation, an electronically controlled lock to a door or gate may provide a controllable barrier to entry.

In FIG. **19**, authorization extension may be enabled by tracking the person **1901** from the point of authorization to the point of entry to the convenience store **1903**. Tracking may be performed using one or more cameras in the area. In the gas station example of FIG. **19**, cameras **1911**, **1912** and **1913** are installed in or around the area of the gas station. Images from the cameras are transmitted to processor **130**, which processes these images to recognize people and to track them over a time period as they move through the gas station area. Processor **130** may also access and use a 3D model **1914**. The 3D model **1914** may for example describe the location and orientation of one or more cameras in the site; this data may be obtained for example from extrinsic camera calibration. In one or more embodiments, the 3D model **1914** may also describe the location of one or more objects or zones in the site, such as the pump and the convenience store in the gasoline station site of FIG. **19**. The 3D model **1914** need not be a complete model of the entire site; a minimal model may for example contain only enough information on one or more cameras to support tracking of persons in locations or regions of the site that are relevant to the application.

Recognition, tracking and calculation of a trajectory associated with a person may be performed for example as described above with respect to FIGS. **1** through **10** and as

illustrated in FIG. **15**. Processor **130** may calculate a trajectory **1920** for person **1901**, beginning for example at a point **1921** at time **1922** when the person enters the area of the gas station or is first observed by one or more cameras. The trajectory may be continuously updated as the person moves through the area. The starting point **1921** may or may not coincide with the point **1923** at which the person presents credential **1904**. On beginning tracking of a person, the system may for example associate a tag **1931** with the person **1901** and with the trajectory **1920** that is calculated over a period of time for this person as the person is tracked through the area. This tag **1931** may be associated with distinguishing characteristics of the person (for example as described above with respect to FIG. **5**). In one or more embodiments it may be an anonymous tag that is an internal identifier used by processor **130**.

The trajectory **1920** calculated by processor **130**, which may be updated as the person **1901** moves through the area, may associate locations with times. For example, person **1901** is at location **1921** at time **1922**. In one or more embodiments the locations and the times may be ranges rather than specific points in space and time. These ranges may for example reflect uncertainties or limitations in measurement, or the effects of discrete sampling. For example, if a camera captures images every second, then a time associated with a location obtained from one camera image may be a time range with a width of two seconds. Sampling and extension of a trajectory with a new point may also occur in response to an event, such as a person entering a zone or triggering a sensor, instead of or in addition to sampling at a fixed frequency. Ranges for location may also reflect that a person occupies a volume in space, rather than a single point. This volume may for example be or be related to the 3D field of influence volume described above with respect to FIGS. **6**A through **7**B.

The processor **130** tracks person **1901** to location **1923** at time **1924**, where credential reader **1905** is located. In one or more embodiments location **1923** may be the same as location **1921** where tracking begins; however, in one or more embodiments the person may be tracked in an area upon entering the area and may provide a credential at another time, such as upon entering or exiting a store. In one or more embodiments, multiple credential readers may be present; for example, the gas station in FIG. **19** may have several pay-at-the-pump stations at which customers can enter credentials. Using analysis of camera images, processor **130** may determine which credential reader a person uses to enter a credential, which allows the processor to associate an authorization with the person, as described below.

As a result of entering credential **1904** into credential reader **1905**, an authorization **1907** is provided to gas pump **1902**. This authorization, or related data, may also be transmitted to processor **130**. The authorization may for example be sent as a message **1910** from the pump or credential reader, or directly from bank or payment processor (or another authorization service) **212**. Processor **130** may associate this authorization with person **1901** by determining that the trajectory **1920** of the person is at or near the location of the credential reader **1904** at or near the time that the authorization message is received or the time that the credential is presented to the credential reader **1905**. In embodiments with multiple credential readers in an area, the processor **130** may associate a particular authorization with a particular person by determining which credential reader that authorization is associated with and by correlating the time of that authorization and the location of that credential reader with the trajectories of one or more people to determine which person is at or near that credential reader at that time. In some situations, the person **1901** may wait at the credential reader **1905** until the authorization is received; therefore processor **130** may use either the time that the credential is presented or the time that the authorization is received to determine which person is associated with the authorization.

By determining that person **1901** is at or near location **1923** at or near time **1924**, determining that location **1923** is the location of credential reader **1905** (or within a zone near the credential reader) and determining that authorization **1910** is associated with credential reader **1905** and is received at or near time **1924** (or is associated with presentation of a credential at or near time **1924**), processor **130** may associate the authorization with the trajectory **1920** of person **1901** after time **1924**. This association **1932** may for example add an extended tag **1933** to the trajectory that includes authorization information and may include account or credential information associated with the authorization. Processor **130** may also associate certain allowed actions with the authorization; these allowed actions may be specific to the application and may also be specific to the particular authorization obtained for each person or each credential.

Processor **130** then continues to track the trajectory **1920** of person **1901** to the location **1925** at time **1926**. This location **1925** as at the entry **1908** to the convenience store **1903**, which is locked by lock **1909**. Because in this example the authorization obtained at the pump also allows entry into the store, processor **130** transmits command **1934** to the controllable lock **1909**, which unlocks door **1908** to allow entry to the store. (Lock **1909** is shown symbolically as a padlock; in practice it may be integrated into door **1908** or any barrier, along with electronic controls to actuate the barrier to allow or deny entry.) The command **1934** to unlock the barrier is issued automatically at or near time **1926** when person **1901** arrives at the door, because camera images are processed to recognize the person, to determine that the person is at the door at location **1925** and to associate this person with the authorization obtained previously as a result of presenting the credential **1904** at previous time **1924**.

One or more embodiments may extend authorization obtained at one point in time to allow entry to any type of secure environment at a subsequent point in time. The secure environment may be for example a store or building as in FIG. **19**, or a case or similar enclosed container as illustrated in FIG. **20**. FIG. **20** illustrates a gas station example that is similar to the example shown in FIG. **19**; however, in FIG. **20**, products are available in an enclosed and locked case as opposed to (or in addition to) in a convenience store. For example, a gas station may have cases with products for sale next to or near gas pumps, with authorization to open the cases obtained by extending authorization obtained at a pump. In the example of FIG. **20**, person **1901** inserts a credential into pump **1902** at location **1923** and time **1924**, as described with respect to FIG. **19**. Processor **130** associates the resulting authorization with the person and with the trajectory **2000** of the person after time **1924**. Person **1901** then walks to case **2001** that contains products for sale. The processor tracks the path of the person to location **2002** at time **2003**, by analyzing images from cameras **1911** and **1913***a*. It then issues command **2004** to unlock the controllable lock **2005** that locks the door of case **2001**, thereby opening the door so that the person can take products.

In one or more embodiments, a trajectory of a person may be tracked and updated at any desired time intervals. Depending for example on the placement and availability of cameras in the area, a person may pass through one or more

locations where cameras do not observe the person; therefore, the trajectory may not be updated in these "blind spots". However, because for example distinguishing characteristics of the person being tracked may be generated during one or more initial observations, it may be possible to pick up the track of the person after he or she leaves these blind spots. For example, in FIG. **20**, camera **1911** may provide a good view of location **1924** at the pump and camera **1913***a* may provide a good view of location **2002** at case **2001**, but there may be no views or limited views between these two points. Nevertheless, processor **130** may recognize that person **1901** is the person at location **2002** at time **2003** and is therefore authorized to open the case **2001**, because the distinguishing characteristics viewed by camera **1913***a* at time **2003** match those viewed by camera **1911** at time **1924**.

FIG. **21** continues the example of FIG. **20**. Case **2001** is opened when person **1901** is at location **2002**. The person then reaches into the case and removes item **2105**. Processor **130** analyzes data from cameras or other sensors that detect removal of item **2105** from the case. In the example in FIG. **21**, these sensors include camera **2101**, camera **2102** and weight sensor **2103**. Cameras **2101** and **2102** may for example be installed inside case **2001** and positioned and oriented to observe the removal of an item from a shelf. Processor **130** may determine that person **1901** has taken a specific item using for example techniques described above with respect to FIGS. **3** and **4**. In addition, or alternatively, one or more other sensors may detect removal of a product. For example, a weight sensor may be placed under each item in the case to detect when the item is removed and data from the weight sensor may be transmitted to processor **130**. Any type or types of sensors may be used to detect or confirm that a user takes an item. Detection of removal of a product, using any type of sensor, may be combined with tracking of a person using cameras in order to attribute the taking of a product to a specific user.

In the scenario illustrated in FIG. **21**, person **1901** removes product **2105** from case **2001**. Processor **130** analyzes data from one or more of cameras **2102**, **2101**, **1913***a* and sensor **2103**, to determine the item that was taken and to associate that item with person **1901** (based for example on the 3D influence volume of the person being located near the item at the time the item was moved). Because authorization information **1933** is also associated with the person at the time the item is taken, processor **130** may transmit message **2111** to charge the account associated with the user for the item. This charge may be pre-authorized by the person **1901** by previously presenting credential **1904** to credential reader **1905**.

FIG. **22** extends the example of FIG. **19** to illustrate the person entering the convenience store and taking an item. This example is similar in some respects to the previous example of FIG. **21**, in that the person takes an item from within a secure environment (a case in FIG. **21**, a convenience store in FIG. **22**) and a charge is issued for the item based on a previously obtained authorization. This example is also similar to the example illustrated in FIG. **2**, with the addition that an authorization is obtained by person **1901** at pump **1902**, prior to entering the convenience store **1903**. External cameras **1911**, **1912** and **1913** track person **1901** to the entrance **1908** and processor **130** unlocks lock **1909** so that person **1901** may enter the store. Afterwards images from internal cameras such as camera **202** track the person inside the store and the processor analyzes these images to determine that the person takes item **111** from shelf **102**. At exit **201**, message **203***a* is generated to automatically charge the account of the person for the item; the message may also be sent to a display in the store (or for example on the person's mobile phone) indicating what item or items are to be charged. In one or more embodiments the person may be able to enter a confirmation or to make modifications before the charge is transmitted. In one or more embodiments the processor **130** may also transmit an unlock message **2201** to unlock the exit door; this barrier at the exit may for example force unauthorized persons in the store to provide a payment mechanism prior to exiting.

In a variation of the example of FIG. **22**, in one or more embodiments a credential may be presented by a person at entrance **1908** to the store, rather than at a different location such as at pump **1902**. For example, a credential reader may be placed within or near the entrance **1908**. Alternatively, the entrance to the store may be unlocked and the credential may be presented at the exit **201**. More generally, in one or more embodiments a credential may be presented and an authorization may be obtained at any point in time and space and may then be used within a store (or at any other area) to perform one or more actions; these actions may include, but are not limited to, taking items and having them charged automatically to an authorized account. Controllable barriers, for example on entry or on exit, may or may not be integrated into the system. For example, the door locks at the store entrance **1908** and at the exit **201** may not be present in one or more embodiments. An authorization obtained at one point may authorize only entry to a secure environment through a controllable barrier, it may authorize taking and charging of items, or it may authorize both (as illustrated in FIG. **22**).

FIG. **23** shows a variation on the scenario illustrated in FIG. **22**, where a person removes and item from a shelf but then puts it down prior to leaving the store. As in FIG. **22**, person **1901** takes item **111** from shelf **102**. Prior to exiting the store, person **1901** places item **111** back onto a different shelf **2301**. Using techniques such as those described above with respect to FIGS. **3** and **4**, processor **130** initially determines take action **2304**, for example by analyzing images from cameras such as camera **202** that observe shelf **102**. Afterwards processor **130** determines put action **2305**, for example by analyzing images from cameras such as cameras **2302** and **2303** that observe shelf **2301**. The processor therefore determines that person **1901** has no items in his or her possession upon leaving the store and transmits message **213***b* to a display to confirm this for the person.

One or more embodiments may enable extending an authorization from one person to another person. For example, an authorization may apply to an entire vehicle and therefore may authorize all occupants of that vehicle to perform actions such as entering a secured area or taking and purchasing products. FIG. **24** illustrates an example that is a variation of the example of FIG. **19**. Person **1901** goes to gas pump **1902** to present a credential to obtain an authorization. Camera **1911** (possibly in conjunction with other cameras) captures images of person **1901** exiting vehicle **2401**. Processor **130** analyzes these images and associates person **1901** with vehicle **2401**. The processor analyzes subsequent images to track any other occupants of the vehicle that exit the vehicle. For example, a second person **2402** exits vehicle **2401** and is detected by the cameras in the gas station. The processor generates a new trajectory **2403** for this person and assigns a new tag **2404** to this trajectory. After the authorization of person **1901** is obtained, processor **130** associates this authorization with person **2402** (as well as with person **1901**), since both people exited the same vehicle **2401**. When person **2402** reaches location **1925** at entry

1908 to store 1903, processor 130 sends a command 2406 to allow access to the store, since person 2402 is authorized to enter by extension of the authorization obtained by person 1901.

One or more embodiments may query a person to determine whether authorization should be extended and if so to what extent. For example, a person may be able to selectively extend authorization to certain locations, for certain actions, for a certain time period, or to selected other people. FIGS. 25A, 25B and 25C show an illustrative example with queries provided at gas pump 1902 when person 1901 presents a credential for authorization. The initial screen shown in FIG. 25A asks the user to provide the credential. The next screen shown in FIG. 25B asks the user whether to extend authorization to purchases as the attached convenience store; this authorization may for example allow access to the store through the locked door and may charge items taken by the user automatically to the user's account. The next screen in FIG. 25C asks the user if he or she wants to extend authorization to other occupants of the vehicle (as in FIG. 24). These screens and queries are illustrative; one or more embodiments may provide any types of queries or receive any type of user input (proactively from the user or in response to queries) to determine how and whether authorization should be extended. Queries and responses may for example be provided via a mobile phone as opposed to on a screen associated with a credential reader, or via any other device or devices.

Returning now to the tracking technology that tracks people through a store or an area using analysis of camera images, in one or more embodiments it may be advantageous or necessary to track people using multiple ceiling-mounted cameras, such as fisheye cameras with wide fields of view (such as 180 degrees). These cameras provide potential benefits of being less obtrusive, less visible to people, and less accessible to people for tampering. Ceiling-mounted cameras also usually provide unoccluded views of people moving through an area, unlike wall cameras that may lose views of people as they move behind fixtures or behind other people. Ceiling-mounted fisheye cameras are also frequently already installed, and they are widely available.

One or more embodiments may simultaneously track multiple people through an area using multiple ceiling-mounted cameras using the technology described below. This technology provides potential benefits of being highly scalable to arbitrarily large spaces, inexpensive in terms of sensors and processing, and adaptable to various levels of detail as the area or space demands. It also offers the advantage of not needing as much training as some deep-learning detection and tracking approaches. The technology described below uses both geometric projection and appearance extraction and matching.

FIGS. 26A through 26F show views from six different ceiling-mounted fisheye cameras installed in an illustrative store. The images are captured at substantially the same time. The cameras may for example be calibrated intrinsically and extrinsically, as described above. The tracking system therefore knows where the cameras are located and oriented in the store, as described for example in a 3D model of the store. Calibration also provides a mapping from points in the store 3D space to pixels in a camera image, and vice-versa.

Tracking directly from fisheye camera images may be challenging, due for example to the distortion inherent in the fisheye lenses. Therefore, in one or more embodiments, the system may generate a flat planar projection from each camera image to a common plane. For example, in one or more embodiments the common plane may be a horizontal plane 1 meter above the floor or ground of the site. This plane has an advantage that most people walking in the store intersect this plane. FIGS. 27A, 27B, and 27C show projections of three of the fisheye images from FIGS. 26A through 26F onto this plane. Each point in the common plane 1 meter above the ground corresponds to a pixel in the planar projections at the same pixel coordinates. Thus, the pixels at the same pixel coordinates in each of the image projections onto the common plane, such as the images 27A, 27B, and 27C, all correspond to the same 3D point in space. However, since the cameras may be two-dimensional cameras that do not capture depth, the 3D point may be sampled anywhere along the ray between it and the camera.

Specifically, in one or more embodiments the planar projections 27A, 27B and 27C may be generated as follows. Each fisheye camera may be calibrated to determine the correspondence between a pixel in the fisheye image (such as image 26A for example) and a ray in space starting at the focal point of the camera. To project from a fisheye image like image 26A to a plane or any other surface in a store or site, a ray may be formed from the camera focal point to that point on the surface, and the color or other characteristics of the pixel in the fisheye image associated with that ray may be assigned to that point on the surface.

When an object is at a 1-meter height above the floor, all cameras will see roughly the same pixel intensities in their respective projective planes, and all patches on the projected 2D images will be correlated if there is an object at the 1-meter height. This is similar to the plane sweep stereo method known in the art, with the provision that the technique described here projects onto a plane that is parallel to the floor as people will be located there (not flying above the floor). Analysis of the projected 2D images may also take into account the walkable space of a store or site, and occlusions of some parts of the space in certain camera images. This information may be obtained for example from a 3D model of the store or site.

In some situations, it may be possible for points on a person that are 1-meter high from the floor to be occluded in one or more fisheye camera views by other people or other objects. The use of ceiling-mounted fisheye cameras minimizes this risk, however, since ceiling views provide relatively unobstructed views of people below. For store fixtures or features that are in fixed locations, occlusions may be pre-calculated for each camera, and pixels on the 1-meter plane projected image for that camera that are occluded by these features or fixtures may be ignored. For moving objects like people in the store, occlusions may not be pre-calculated; however, one or more embodiments may estimate these occlusions based on the position of each person in the store in a previous frame, for example.

To track moving objects, in particular people, one or more embodiments of the system may incorporate a background subtraction or motion filter algorithm, masking out the background from the foreground for each of the planar projected images. FIGS. 28A, 28B, and 28C show foreground masks for the projected planar images 27A, 27B, and 27C, respectively. A white pixel shows a moving or non-background object, and a black pixel shows a stationary or background object. (These masks may be noisy, for example because of lighting changes or camera noise.) The foreground masks may then be combined to form mask 28D. Foreground masks may be combined for example by adding the mask values or by binary AND-ing them as shown in FIG. 28D. The locations in FIG. 28D where the combined

mask is non-zero show where the people are located in the plane at 1-meter above the ground.

In one or more embodiments, the individual foreground masks for each camera may be filtered before they are combined. For example, a gaussian filter may be applied to each mask, and the filtered masks may be summed together to form the combined mask. In one or more embodiments, a thresholding step may be applied to locate pixels in the combined mask with values above a selected intensity. The threshold may be set to a value that identifies pixels associated with a person even if some cameras have occluded views of that person.

After forming a combined mask, one or more embodiments of the system may for example use a simple blob detector to localize people in pixel space. The blob detector may filter out shapes that are too large or too small to correspond to an expected cross-sectional size of a person at 1-meter above the floor. Because pixels in the selected horizontal plane correspond directly to 3D locations in the store, this process yields the location of the people in the store.

Tracking a person over time may be performed by matching detections from one time step to the next. An illustrative tracking framework that may be used in one or more embodiments is as follows:

(1) Match new detections to existing tracks, if any. This may be done via position and appearance, as described below.

(2) Update existing tracks with matched detections. Track positions may be updated based on the positions of the matched detections.

(3) Remove tracks that have left the space or have been inactive (such as false positives) for some period of time.

(4) Add unmatched detections from step (1) to new tracks. The system may optionally choose to add tracks only at entry points in the space.

The tracking algorithm outlined above thus maintains the positions in time of all tracked persons.

As described above in step (1) of the illustrative tracking framework, matching detections to tracks may be done based on either or both of position and appearance. For example, if a person detection at a next instant in time is near the previous position of only one track, this detection may be matched to that track based on position alone. However, in some situations, such as a crowded store, it may be more difficult to match detections to tracks based on position alone. In these situations, the appearance of persons may be used to assist with matching.

In one or more embodiments, an appearance for a detected person may be generated by extracting a set of images that have corresponding pixels for that person. An approach to extracting these images that may be used in one or more embodiments is to generate a surface around a person (using the person's detected position to define the location of the surface), and to sample the pixel values for the 3D points on the surface for each camera. For example, a cylindrical surface may be generated around a person's location, as illustrated in FIGS. **29**A through **29**F. These figures show the common cylinder (in red) as seen from each camera. The surface normal vectors of the cylinder (or other surface) may be used to only sample surface points that are visible from each camera. For each detected person, a cylinder may be generated around a center vertical axis through the person's location (defined for example as a center of the blob associated with that person in the combined foreground mask); the radius and height of the cylinder may be set to fixed values, or they may be adapted for the apparent size and shape of the person.

As shown in FIGS. **29**A through **29**F, a cylindrical surface is localized in each of the original camera views (FIGS. **26**A through **26**F) based on the intrinsics/extrinsics of each camera. The points on the cylinder are sampled from each image and form the projections shown in FIGS. **30**A through **30**F. Using surface normal vectors of the cylinders, the system may only sample the points that would be visible in each camera, if there was an opaque surface of the cylinder. The occluded points are darkened in FIGS. **30**A through **30**F. An advantage of this approach is that the cylindrical surface provides a corresponding view from each camera, and the views can be combined into a single view, taking into account the visibilities at each pixel. Visibility for each pixel in each cylindrical image for each camera may take into account both the front and back sides of the cylinder as viewed from the camera, and occlusion by other cylinders around other people. Occlusions may be calculated for example using a method similar to a graphics pipeline: cylinders closer to the camera may be projected first, and the pixels on the fisheye image that are mapped to those cylinders are removed (e.g., set to black) so that they are not projected onto other cylinders; this process repeats until all cylinders receive projected pixels from the fisheye image. Cylindrical projections from each camera may be combined for example as follows: back faces may be assigned a 0 weight, and visible, unoccluded pixels may be assigned a 1 weight; the combined image may be calculated as a weighted average for all projections onto the cylinder. Combining the occluded cylindrical projections creates a registered image of the tracked person that facilitates appearance extraction. The combined registered image corresponding to cylindrical projections **30**A through **30**F is shown in FIG. **30**G.

Appearance extraction from image **30**G may for example be done by histograms, or by any other dimensionality reduction method. A lower dimensional vector may be formed from the composite image of each tracked person and used to compare it with other tracked subjects. For example, a neural network may be trained to take composite cylindrical images as input, and to output a lower-dimensional vector that is close to other vectors from the same person and far from vectors from other persons. To distinguish between people, vector-to-vector distances may be computed and compared to a threshold; for example, a distance of 0.0 to 0.5 may indicate the same person, and a greater distance may indicate different people. One or more embodiments may compare tracks of people by forming distributions of appearance vectors for each track, and comparing distributions using a distribution-to-distribution measure (such as KL-divergence, for example). A discriminant between distributions may be computed to label a new vector to an existing person in a store or site.

A potential advantage of the technique described above over appearance vector and people matching approaches known in the art is that it may be more robust in a crowded space, where there are many potential occlusions of people in the space. By combining views from multiple cameras, while taking into account visibility and occlusions, this technique may succeed in generating usable appearance data even in crowded spaces, thereby providing robust tracking. This technique treats the oriented surface (cylinder in this example) as the basic sampling unit and generates projections based on visibility of 3D points from each camera. A point on a surface is not visible from a camera if the normal

to that surface points away from the camera (dot product is negative). Furthermore, in a crowded store space, sampling the camera based on physical rules (visibility and occlusion) and cylindrical projections from multiple cameras provides cleaner images of individuals without pixels from other individuals, making the task of identifying or separating people easier.

FIGS. **31**A and **31**B show screenshots at two points in time from an embodiment that incorporates the tracking techniques described above. Three people in the store are detected and tracked as they move, using both position and appearance. The screenshots show fisheye views **3101** and **3111** from one of the fisheye cameras, with the location of each person indicated with a colored dot overlaying the person's image. They also show combined masks **3102** and **3112** for the planar projections to the plane 1 meter above the ground, as discussed above with respect to FIG. **27**D. The brightest spots in combined masks **3102** and **3112** correspond to the detection locations. As an illustration of tracking, the location of one of the persons moves from location **3103** at the time corresponding to FIG. **31**A to the location **3113** at the subsequent time corresponding to FIG. **31**B.

Embodiments of the invention may utilize more complicated models, for example spherical models for heads, additional cylindrical models for upper and lower arms and/or upper and lower legs as well. These embodiments enable more detailed differentiation of users, and may be utilized in combination with gait analysis, speed of movement, any derivative of position, including velocity acceleration, jerk or any other frequencies of movement to differentiate users and their distinguishing characteristics. In one or more embodiments, the complexity of the model may be altered over time or as needed based on the number of users in a given area for example. Other embodiments may utilize simple cylindrical or other geometrical shapes per user based on the available computing power or other factors, including the acceptable error rate for example.

As an alternative to identifying people in a store by performing background subtraction on camera images and combining the resulting masks, one or more embodiments may train and use a machine learning system that processes a set of camera images directly to identify persons. The input to the system may be or may include the camera images from all cameras, or all cameras in a relevant area. The output may be or may include an intensity map with higher values indicating a greater likelihood that a person is at that location. The machine learning system may be trained for example by capturing camera images while people move around the store area, and manually labeling the people's positions to form training data. Camera images may be used as inputs directly, or in one or more embodiments they may be processed, and the processed images may be used as inputs. For example, images from ceiling fisheye cameras may be projected onto a plane parallel to the floor, as described above, and the projected images may be used as inputs to the machine learning system.

FIG. **32** illustrates an example of a machine learning system that detects person positions in a store from camera images. This illustrative embodiment has three cameras **3201**, **3202**, and **3203** in the store **3200**. At a point in time, these three cameras capture images **3211**, **3212**, and **3213**, respectively. These three images are input into a machine learning system **3220** that has learned (or is learning) to map from the collection of camera images to an intensity map **3221** of likely person positions in the store.

In the example shown in FIG. **32**, the output of system **3220** is the likely horizontal position of persons in the store. Vertical position is not tracked. Although people occupy 3D space, horizontal position is generally all that is required to determine where each person is in a store, and to associate item motion with a person. Therefore, the intensity map **3221** maps xy position along the floor of the store into an intensity that represents how likely a person's centroid (or other point or points of a person) is at that horizontal location. This intensity map may be represented as a gray-scale image, for example, with whiter pixels representing higher probability of a person at that location.

The person detection system illustrated in FIG. **32** represents a significant simplification over systems that attempt to detect landmarks on a person's body or other features of a person's geometry. A person's location is represented only by a single 2D point, possibly with a zone around this point with a falloff in probability. This simplification makes detection potentially more efficient and more robust. Processing power to perform detection may be reduced using this method, thereby reducing the cost of installation for a system and enabling real-time person tracking.

In one or more embodiments, a 3D field of influence volume may be constructed for a person around the 2D point that represents that person's horizontal position. That field of influence volume may then be used to determine which item storage areas a person interacts with and the times of these interactions. For example, the field of influence volume may be used as described above with respect to FIG. **10**. FIG. **32**A shows an example of generating a 3D field of influence volume from a 2D location of a person, as determined for example by the machine learning system **3220** of FIG. **32**. In this example, a machine learning system or other system generates 2D location data **3221***d*. This data includes and extends the intensity map data **3221** of FIG. **32**. From the intensity data, the system estimates a point 2D location for each person in the store. These points are **3231***a* for a first shopper, and **3232** for a second shopper. The 2D point may be calculated for example as the weighted average of points in a region surrounding a local maximum of intensity, with weights proportional to the intensity of each point. The first shopper moves, and the system tracks the trajectory **3230** of this shopper's 2D location. This trajectory **3230** may for example consist of a sequence of locations, each associated with a different time. For example, at time t_1 the first shopper is at location **3231***a*, and at time t_4 the shopper arrives at 2D point **3231***b*. For each 2D point location of a shopper at different points in time, the system may generate a 3D field of influence volume around that point. This field of influence volume may be a translated copy of a standard shape that is used for all shoppers and for all points in time. For example, in FIG. **32**A the system generates a cylinder of a standard height and radius, with the center axis of the cylinder passing through the 2D location of the shopper. Cylinder **3241***a* for the first shopper corresponds to the field of influence volume at point **3231***a* at time t_1, and cylinder **3242** for the second shopper corresponds to the field of influence volume at point **3232**. The cylinder is illustrative; one or more embodiments may use any type of shape for a 3D field of influence volume, including for example, without limitation, a cylinder, a sphere, a cube, a parallelepiped, an ellipsoid, or any combinations thereof. The selected shape may be used for all shoppers and for all locations of the shoppers. Use of a simple, standardized volume around a tracked 2D location provides significant efficiency benefits

compared to tracking the specific location of landmarks or other features and constructing a detailed 3D shape for each shopper.

When the first shopper reaches 2D location **3231***b* at time t_4, the 3D field of influence volume **3241***b* intersects the item storage area **3204**. This intersection implies that the shopper may interact with items on the shelf, and it may trigger the system to track the shelf to determine movement of items and to attribute those movements to the first shopper. For example, images of the shelf **3204** before the intersection occurs, or at the beginning of the intersection time period may be compared to images of the shelf after the shopper moves away and the volume no longer intersects the shelf, or at the end of the intersection time period.

One or more embodiments may further simplify detection of intersections by performing this analysis completely or partially in 2D instead of in 3D. For example, a 2D model **3250** of the store may be used, which shows the 2D location of item storage areas such as area **3254** corresponding to shelf **3204**. In 2D, the 3D field of influence cylinders become 2D field of influence areas that are circles, such as circles **3251***a* and **3251***b* corresponding to cylinders **3241***a* and **3241***b* in 3D. The intersection of 2D field of influence area **3251***b* with 2D shelf area **3254** indicates that the shopper may be interacting with the shelf, triggering the analyses described above. In one or more embodiments, analyzing fields of influence areas and intersections in 2D instead of 3D may provide additional efficiency benefits by reducing the amount of computation and modeling required.

As described above, and as illustrated in FIGS. **26** through **31**, in one or more embodiments it may be advantageous to perform person tracking and detection using ceiling-mounted cameras, such as fisheye cameras. Camera images from these cameras, such as images **26**A through **26**F, may be used as inputs to the machine learning system **3220** in FIG. **32**. Alternatively, or in addition, these fisheye images may be projected onto one or more planes, and the projected images may be inputs to machine learning system **3220**. Projecting images from multiple cameras onto a common plane may simplify person detection since unoccluded views of a person in the projected images will overlap at the points where the person intersects this plane. This technique is illustrated in FIG. **33**, which shows two dome fisheye cameras **3301** and **3302** installed on the ceiling of store **3200**. Images captured by fisheye cameras **3301** and **3302** are projected onto an imaginary plane **3310** parallel to the floor of the store, at approximately waist level for a typical shopper. The projected pixel locations on plane **3310** coincide with actual locations of objects at this height if they are not occluded by other objects. For example, pixels **3311** and **3312** in fisheye camera images from cameras **3301** and **3302**, respectively, are projected to the same position **3305** in plane **3310**, since one of the shoppers intersects plane **3310** at this location. Similarly, pixels **3321** and **3322** are projected to the same position **3306**, since the other shopper intersects plane **3310** at this location.

FIGS. **34**AB through **37** illustrate this technique of projecting fisheye images onto a common plane for an artificially generated scene. FIG. **34**A shows the scene from a perspective view, and FIG. **34**B shows the scene from a top view. Store **3400** has a floor area between two shelves; two shoppers **3401** and **3402** are currently in this area. Store **3400** has two ceiling-mounted fisheye cameras **3411** and **3412**. (The ceiling of the store is not shown to simplify illustration). FIG. **35** shows fisheye images **3511** and **3512** captured from cameras **3411** and **3412**, respectively. Although these fisheye images may be input directly into a machine learning system, the system would have to learn how to relate the position of an object in one image to the position of that object in another image. For example, shopper **3401** appears at location **3513** in image **3511** from camera **3411**, and at a different location **3514** in image **3512** from camera **3412**. While it may be possible for a machine learning system to learn these correspondences, a large amount of training data may be needed. FIG. **36** shows the projection of the two fisheye images onto a common plane, in this case a plane one meter above the floor. Image **3511** is transformed with projection **3601** into image **3611**, and image **3512** is transformed with projection **3601** into image **3612**. The height of the projection plane in this case is selected to intersect the torso of most shoppers; in one or more embodiments any plane or planes may be used for projection. One or more embodiments may project fisheye images onto multiple planes at different heights, and may use all of these projections as inputs to a machine learning system to detect people.

FIG. **37** shows images **3611** and **3612** overlaid onto one another to illustrate that locations of shoppers coincide in these two images. For illustration, the images are alpha weighted each by 0.5 and then summed. The resulting overlaid image **3701** shows location of overlap **3711** for shopper **3401**, and location of overlap **3712** for shopper **3402**. These locations correspond to the intersection of the projection plane with each shopper. As described above with respect to FIGS. **27**ABC and **28**ABCD, in one or more embodiments the intersection areas **3711** and **3712** may be used directly to detect persons, for example via thresholding of intensity and blob detection. Alternatively, or in addition, the projected images **3611** and **3612** may be input into a machine learning system, as described below.

As illustrated in FIG. **37**, the appearance of a person in a camera image, even when this image is projected onto a common plane, varies depending on the location of the camera. For example, the FIG. **3721** in image **3611** is different from the FIG. **3722** in image **3612**, although these figures overlap in region **3711** in combined image **3701**. Because of this camera location dependence for images, knowledge of the camera locations may improve the ability of a machine learning system to detect people in camera images. The inventors have discovered that an effective technique to account for camera location is to extend each projected image with an additional "channel" that reflects the distance between each associated point on the projected plane and the camera location. Unexpectedly, adding this channel as an input feature may dramatically reduce the amount of training data needed to train a machine learning system to recognize person locations. This technique of projecting camera images to a common plane and adding a channel of distance information to each image is not known in the art. Encoding distance information as an additional image channel also has the benefit that a machine learning system (such as a convolutional neural network, as described below) organized to process images may be adapted easily to accommodate this additional channel as an input.

FIG. **38** illustrates a technique that may be used in one or more embodiments to generate a camera distance channel associated with projected images. For each point on the projected plane (such as the plane one meter above the floor), a distance to each camera may be determined. These distances may be calculated based on calibrated camera positions, for example. For instance, at point **3800**, which is on the intersection of the projected plane with the torso of shopper **3401**, these distances are distance **3801** to camera

3411 and distance 3802 to camera 3412. Distances may be calculated in any desired metric, including but not limited to a Euclidean metric as shown in FIG. 38. Based on the distance between a camera and each point on the projected plane, a position weight 3811 may be calculated for each point. This position weight may for example be used by the machine learning system to adjust the importance of pixels at different positions on an image. The position weight 3811 may be any desired function of the distance 3812 between the camera and the position. The illustrative position weight curve 3813 shown in FIG. 38 is a linear, decreasing function of distance, with a maximum weight 1.0 at the minimum distance. The position weight may decrease to 0 at the maximum distance, or it may be set to some other desired minimum weight value. One or more embodiments may use position weight functions other than linear functions. In one or more embodiments the position weight may also be a function of other variables in addition to distance from the camera, such as distance from lights or obstacles, proximity to shelves or other zones of interest, presence of occlusions or shadows, or any other factors.

Illustrative position weight maps 3821 for camera 3411 and 3822 for camera 3412 are shown in FIG. 38 as grayscale images. Brighter pixels in the grayscale images correspond to higher position weights, which correspond to shorter distances between the camera and the position on the projected plane associated with that pixel.

FIG. 39 illustrates how the position weight maps generated in FIG. 38 may be used in one or more embodiments for person detection. Projected images 3611 and 3612, from cameras 3411 and 3412, respectively, may be separated into color channels. FIG. 39 illustrates separating these images into RGB color channels; these channels are illustrative, and one or more embodiments may use any desired decomposition of images into channels using any color space or any other image processing methods. The RGB channels are combined with a fourth channel representing the position weight map for the camera that captured the image. The four channels for each image are input into machine learning system 3220, which generates an output 3221*a* with detection probabilities for each pixel. Therefore image 3611 corresponds to four inputs 3611*r*, 3611*g*, 3611*b*, and 3821; and image 3612 corresponds to four inputs 3612*r*, 3612*g*, 3612*b*, and 3822. To simplify the machine learning system, in one or more embodiments the position weight maps 3821 and 3822 may be scaled to have the same size as the associated color channels.

Machine learning system 3220 may incorporate any machine learning technologies or methods. In one or more embodiments, machine learning system 3220 may be or may include a neural network. FIG. 40 shows an illustrative neural network 4001 that may be used in one or more embodiments. In this neural network, inputs are 4 channels for each projected image, with the fourth channel containing position weights as described above. Inputs 4011 represent the four channels from the first camera, inputs 4012 represent the four channels from the second camera, and there may be additional inputs 4019 from any number of additional cameras (also augmented with position weights). By scaling all image channels, including the position weights channels, to the same size, all inputs may share the same coordinate system. Thus, for a system with N cameras, and images of size H×W, the total number of input values for the network may be N*H*W*4. More generally with C channels per image (including potentially position weights), the total of number of inputs may be N*H*W*C.

The illustrative neural network 4001 may be for example a fully convolutional network with two halves: a first (left) half that is built out of N copies (for N cameras) of a feature extraction network, which may consist of layers of decreasing size; and a second (right) half that maps the extracted features into positions. In between the two halves may be a feature merging layer 4024, which may for example be an average over the N feature maps. The first half of the network may have for example N copies of a standard image classification network. The final classifier layer of this image classification network may be removed, and the network may be used as a pre-trained feature extractor. This network may be pretrained on a dataset such as the ImageNet dataset, which is a standard objects dataset with images and labels for various types of objects, including but not limited to people. The lower layers (closer to the image) in the network generally mirror the pixel statistics and primitives. Pre-trained weights may be augmented with additional weights for the position maps, which may be initialized with random values. Then the entire network may be trained with manually labeled person positions, as described below with respect to FIG. 41. All weights, including the pretrained weights, may vary during training with the labeled dataset. In the illustrative network 4001, the copies of the image classification network (which extracts image features) are 4031, 4032, and 4039. (There may be additional copies if there are additional cameras.) Each of these copies 4031, 4032, and 4039 may have identical weights.

The first half of the network 4031 (and thus also 4032 and 4039) may for example reduce the spatial size of the feature maps several times. The illustrative network 4031 reduces the size three times, with the three layers 4021, 4022, and 4023. For example, for inputs such as input 4011 of size H×W×C, the output feature maps of layers 4021, 4022, and 4023 may be of sizes H/8×W/8, H/16×W/16, and H/32×W/32, respectively. In this illustrative network, all C channels of input 4011 are input into layer 4021 and are processed together to form output features of size H/8×W/8, which are fed downstream to layer 4022. These values are illustrative; one or more embodiments may use any number of feature extraction layers with input and output sizes of each layer of any desired dimensions.

The feature merging layer 4024 may be for example an averaging over all of the feature maps that are input into this merging layer. Since inputs from all cameras are weighted equally, the number of cameras can change dynamically without changing the network weights. This flexibility is a significant benefit of this neural network architecture. It allows the system to continue to function if one or more cameras are not working. It also allows new cameras to be added at any time without requiring retraining of the system. In addition, the number of cameras used can be different during training compared to during deployment for operational person detection. In comparison, person detection systems known in the art may not be robust when cameras change or are not functioning, and they may require significant retraining whenever the camera configuration of a store is modified.

The output features from the final reduction layer 4023, and the duplicate final reduction layers for the other cameras, are input into the feature merging layer 4024. In one or more embodiments, features from one or more previous reduction layers may also be input into the feature merging layer 4024; this combination may for example provide a mixture of lower-level features from earlier layers and higher-level features from later layers. For example, lower-level features from an earlier layer (or from multiple earlier

layers) may be averaged across cameras to form a merged lower-level feature output, which may be input into the second half network **4041** along with the average of the higher-level features.

The output of the feature merging layer **4024** (which reduces N sets of feature maps to 1 set) is input into the second half network **4041**. The second half network **4041** may for example have a sequence of transposed convolution layers (also known as deconvolution layers), which increase the size of the outputs to match the size H×W of the input image. Any number of deconvolution layers may be used; the illustrative network **4041** has three deconvolution layers **4024**, **4026**, and **4027**.

The final output **3221***a* from the last deconvolution layer **4027** may be interpreted as a "heat map" of person positions. Each pixel in the output heat map **3221***a* corresponds to an x,y coordinate in the projected plane onto which all camera images are projected. The output **3221***a* is shown as a grayscale image, with brighter pixels corresponding to higher values of the outputs from neural network **4001**. These values may be scaled for example to the range 0.0 to 1.0. The "hot spots" of the heat map correspond to person detections, and the peaks of the hot spots represent the x,y locations of the centroid of each person. Because the network **4001** does not have perfect precision in detecting the position of persons, the output heat map may contain zones of higher or moderate intensity around the centroids of the hot spots.

The machine learning system such as neural network **4001** may be trained using images captured from cameras that are projected to a plane and then manually labeled to indicate person positions within the images. This process is illustrated in FIG. **41**. A camera image is captured while persons are in the store area, and it is projected onto a plane to form an image **3611**. A user **4101** reviews this image (as well as other images captured during this session or other sessions, from the same camera or from other cameras), and the user manually labels the position of the persons at the centroid of the area where they intersect the projection plane. The user **4101** picks points such as **4102** and **4103** for the person locations. The training system then generates **4104** a probability density distribution around the selected points. For example, the distribution in one or more embodiments may be a two-dimensional gaussian of some specified width centered on the selected points. The target output **4105** may be for example the sum of the distributions generated in step **4104** at each pixel. One or more embodiments may use any type of probability distribution around the point or points selected by the user to indicate person positions. The target output **4105** is then combined with camera inputs (and position weights) from all cameras used for training, such as inputs **4011** and **4012**, to form a training sample **4106**. This training sample is added to a training dataset **4107** that is used to train the neural network.

An illustrative training process that may be used in one or more embodiments is to have one or more people move through a store, and to sample projected camera images at fixed time intervals (for example every one second). The sampled images may be labeled and processed as illustrated in FIG. **41**. On each training iteration a random subset of the cameras in an area may be selected to be used as inputs. The plane projections may also be performed on randomly selected planes parallel to the floor within some height range above the store. In addition, random data augmentation may be performed to generate additional samples; for example, synthesized images may be generated to deform the shapes or colors of persons, or to move their images to different areas of the store (and to move the labeled positions accordingly).

Tracking of persons and item movements in a store or other area may use any cameras (or other sensors), including "legacy" surveillance cameras that may already be present in a store. Alternatively, or in addition, one or more embodiments of the system may include modular elements with cameras and other components that simplify installation, configuration, and operation of an automated store system. These modular components may support a turnkey installation of an automated store, potentially reducing installation and operating costs. Quality of tracking of persons and items may also be improved using modular components that are optimized for tracking.

FIG. **42** illustrates a store **4200** with modular "smart" shelves that may be used to detect taking, moving, or placing of items on a shelf. A smart shelf may for example contain cameras, lighting, processing, and communications components in an integrated module. A store may have one or more cabinets, cases, or shelving units with multiple smart shelves stacked vertically. Illustrative store **4200** has two shelving units **4210** and **4220**. Shelving unit **4210** has three smart shelves, **4211**, **4212**, and **4213**. Shelving unit **4220** has three smart shelves, **4221**, **4222**, and **4223**. Data may be transmitted from each smart shelf to computer **130**, for analysis of what item or items are moved on each shelf. Alternatively, or in addition, in one or more embodiments each shelving unit may act as a local hub, and may consolidate data from each smart shelf in the shelving unit and forward this consolidated data to computer **130**. The shelving units **4210** and **4220** may also perform local processing on data from each smart shelf. In one or more embodiments, an automated store may be structured for example as a hierarchical system with the entire store at the top level, "smart" shelving units at the second level, smart shelves at the third level, and components such as cameras or lighting at the fourth level. One or more embodiments may organize elements in hierarchical structures with any number of levels. For example, stores may be divided into regions, with local processing performed for each region and then forwarded to a top-level store processor.

The smart shelves shown in FIG. **42** have cameras mounted on the bottom of the shelf; these cameras observe items on the shelf below. For example, camera **4231** on shelf **4212** observes items on shelf **4213**. When user **4201** reaches for an item on shelf **4213**, cameras on either or both of shelves **4212** and **4213** may detect entry of the user's hand into the shelf area, and may capture images of shelf contents that may be used to determine which item or items are taken or moved. This data may be combined with images from other store cameras, such as cameras **4231** and **4232**, to track the shoppers and attribute item movements to specific shoppers.

FIG. **43** shows an illustrative embodiment of a smart shelf **4212**, viewed from the front. FIGS. **44** through **47** show additional views of this embodiment. Smart shelf **4212** has cameras **4301** and **4302** at the left and right ends, respectively, which face inward along the front edge of the shelf. Thus the left end camera **4301** is rightward-facing, and the right end camera **4302** is leftward-facing. These cameras may be used for example to detect when a user's hand moves into or out of the shelf area. These cameras **4301** and **4302** may be used in combination with similar cameras on shelves above and/or below shelf **4212** in a shelving unit (such as shelves **4211** and **4213** in FIG. **42**) to detect hand events. For example, the system may use multiple hand detection cam-

eras to triangulate the position of a hand going into a shelf. With two cameras observing a hand, the position of a hand can be determined from the two images. With multiple cameras (for example four or more) observing a shelf, the system may be able to determine the position of more than one hand at a time since the multiple views can compensate for potential occlusions. Images of the shelf just prior to a hand entry event may be compared to images of the shelf just after a hand exit event, in order to determine which item or items may have been taken, moved, or added to the shelf. In one or more embodiments other detection technologies may be used instead of or in addition to the cameras **4301** and **4302** to detect hand entry and hand exit events for the shelf; these technologies may include for example, without limitation, light curtains, sensors on a door that must be opened to access the shelf or the shelving unit, ultrasonic sensors, and motion detectors.

Smart shelf **4212** may also have one or more downward-facing camera modules mounted on the bottom side of the shelf, facing the shelf **4213** below. For example, shelf **4214** has camera modules **4311**, **4312**, **4313**, and **4314** mounted on the bottom side of the shelf. The number of camera modules and their positions and orientations may vary across installations, and also may vary across individual shelves in a store. These camera modules may capture images of the items on the shelf. Changes in these images may be analyzed by the system, by a processor on the shelf or on a shelving unit, or by both, to determine what items have been taken, moved, or added to the shelf below.

FIGS. **44**A and **44**B show a top view and a side view, respectively, of smart shelf **4212**. Brackets **4440** may be used for example to attach shelf **4212** to a shelving unit; the shape and position of mounting brackets or similar attachment mechanisms may vary across embodiments.

FIG. **44**C shows a bottom view of smart shelf **4212**. All cameras are visible in this view, including the inside-facing cameras **4301** and **4302**, and the downward-facing cameras associated with camera modules **4311**, **4312**, **4313**, and **4314**. In this illustrative embodiment, each camera module contains two cameras: cameras **4311***a* and **4311***b* in module **4311**, cameras **4312***a* and **4312***b* in module **4312**, cameras **4313***a* and **4313***b* in module **4313**, and cameras **4314***a* and **4314***b* in module **4314**. This configuration is illustrative; camera modules may contain any number of cameras. Use of two or more cameras per camera module may assist with stereo vision, for example, in order to generate a 3D view of the items on the shelf below, and a 3D representation of the changes in shelf contents when a user interacts with items on the shelf.

Shelf **4212** also contains light modules **4411**, **4412**, **4413**, **4414**, **4415**, and **4416**. These light modules may be LED light strips, for example. Embodiments of a smart shelf may contain any number of light modules, in any locations. The intensity, wavelengths, or other characteristics of the light emitted by the light modules may be controlled by a processor on the smart shelf. This control of lighting may enhance the ability of the camera modules to accurately detect item movements and to capture images that allow identification of the items that have moved. Lighting control may also be used to enhance item presentation, or to highlight certain items such as items on sale or new offerings.

Smart shelf **4212** contains integrated electronics, including a processor and network switches. In the illustrative smart shelf **4212**, these electronics are contained in areas **4421** and **4422** at the ends of the shelf. One or more embodiments may locate any components at any position on the shelf. FIG. **45** shows a bottom view smart shelf **4212** with the covers to electronics areas **4421** and **4422** removed, to show the components. Two network switches **4501** and **4503** are included; these switches may provide for example connections to each camera and to each lighting module, and a connection between the smart shelf and the store computer or computers. A processor **4502** is included; it may be for example a Raspberry Pi® or similar embedded computer. Power supplies **4504** may also be included; these power supplies may provide AC to DC power conversion for example.

FIG. **46**A shows a bottom view of a single camera module **4312**. This module provides a mounting bracket onto which multiple cameras may be mounted in any desired positions. Camera positions and numbers may be modified based on characteristics such as item size, number of items, and distance between shelves. The bracket has slots **4601***a*, **4602***a*, **4603***a* on the left, and corresponding slots **4601***b*, **4602***b*, and **4603***b* on the right. Individual cameras may be installed at any desired position in any of these slots. Positions of cameras may be adjusted after initial installation. Camera module **4312** has two cameras **4312***a* and **4312***b* installed in the top and bottom slot pairs; the center slot pair **4602***a* and **4602***b* is unoccupied in this illustrative embodiment. FIG. **46**B shows an individual camera **4312***a* from a side view. Screw **4610** is inserted through one of the slots on the bracket **4312** to install the camera; a corresponding screw on the far side of the camera attaches the camera to the opposing slot in the bracket.

FIG. **47** illustrates how camera modules and lighting modules may be installed at any desired positions in smart shelf **4212**. Additional camera modules and lighting modules may also be added in any available positions, and positions of installed components may be adjusted. These modules mount to a rail **4701** at one end of the shelf (and to a corresponding rail at the other end, which is not shown in FIG. **47**). This rail **4701** has slots into which screws are attached to hold end brackets of the modules against the rail. For example, lighting module **4413** has an end bracket **4703**, and screw **4702** attaches through this end bracket into a groove in rail **4701**. Similar attachments are used to attach other modules such as camera module **4312** and lighting module **4412**.

One or more embodiments may include a modular, "smart" ceiling that incorporates cameras, lighting, and potentially other components at configurable locations on the ceiling. FIG. **48** shows an illustrative embodiment of a store **4800** with a smart ceiling **4801**. This illustrative ceiling has a center longitudinal rail **4821** onto which transverse rails, such as rail **4822**, may be attached at any desired locations. Lighting and camera modules may be attached to the transverse rails at any desired locations. This combined longitudinal and transverse railing system provides complete two degree of freedom positioning for lights and cameras. In the configuration shown in FIG. **48**, three transverse rails **4822**, **4823**, and **4824** each hold two integrated lighting-camera modules. For example, transverse rail **4823** holds integrated lighting-camera module **4810**, which contains a circular light strip **4811**, and two cameras **4812** and **4813** in the central area inside the circular light strip. In one or more embodiments, the rails or other mounting mechanisms of the ceiling may hold any type or types of lighting or camera components, either integrated like module **4810** or standalone. The rail configuration shown in FIG. **48** is illustrative; one or more embodiments may provide any type of lighting-camera mounting mechanisms in any desired configuration. For example, mounting

rails or other mounting mechanisms may be provided in any desired geometry, not limited to the longitudinal and transverse rail configuration illustrated in FIG. **48**.

Data from ceiling **4801** may be transmitted to store computer **130** for analysis. In one or more embodiments, ceiling **4801** may contain one or more network switches, power supplies, or processors, in addition to cameras and lights. Ceiling **4801** may perform local processing of data from cameras before transmitting data to the central store computer **130**. Store computer **130** may also transmit commands or other data to ceiling **4801**, for example to control lighting or camera parameters.

The embodiment illustrated in FIG. **48** has a modular smart ceiling **4801** as well as modular shelving units **4210** and **4220** with smart shelves. Data from ceiling **4801** and from shelves in **4210** and **4220** may be transmitted to store computer **130** for analysis. For example, computer **130** may process images from ceiling **4801** to track persons in the store, such as shopper **4201**, and may process images from shelves in **4210** and **4220** to determine what items are taken, moved, or placed on the shelves. By correlating person positions with shelf events, computer **130** may determine which shoppers take items, thereby supporting a fully or partially automated store. The combination of smart ceiling and smart shelves may provide a partially or fully turnkey solution for an automated store, which may be configured based on factors such as the store's geometry, the type of items sold, and the capacity of the store.

FIG. **49** shows an embodiment of a modular ceiling similar to the ceiling of FIG. **48**. A central longitudinal rail **4821***a* provides a mounting surface for transverse rails **4822***a*, **4822***b*, and **4822***c*, which in turn provide mounting surfaces for integrating lighting-camera modules. The transverse rails may be located at any points along longitudinal rail **4821***a*. Any number of transverse rails may be attached to the longitudinal rail. Any number of integrated lighting-camera modules, or other compatible modules, may be attached to the transverse rails at any positions. Transverse rail **4822***a* has two lighting-camera modules **4810***a* and **4810***b*, and transverse rail **4822***b* has three lighting-camera modules **4810***c*, **4810***d*, and **4810***e*. The positions of the lighting-camera modules vary across the three transverse rails to illustrate the flexibility of the mounting system.

FIG. **50** shows a closeup view of transverse rail **4822***a* and lighting-camera module **4810***a*. Transverse rail **4822***a* has a crossbar **5022** with a C-shaped attachment **5001** that clamps around a corresponding protrusion on rail **4821***a*. The position of the transverse rail **4822***a* is adjustable along the longitudinal rail **4821***a*. Lighting-camera module **4810***a* has a circularly shaped annular light **5011** with a pair of cameras **5012** and **5013** in a central area surrounded by the light **5011**. The two cameras **5012** and **5013** may be used for example to provide stereo vision. Alternatively, or in addition, two or more cameras per lighting-camera module may provide redundancy so that person tracking can continue even if one camera is down. The circular shape of light **5011** provides a diffuse light that may improve tracking by reducing reflections and improving lighting consistency across a scene. This circular shape is illustrative; one or more embodiments may use lights of any size or shape, including for example, without limitation, any polygonal or curved shape. Lights may be for example triangular, square, rectangular, pentagonal, hexagonal, or shaped like any regular or irregular polygon. In one or more embodiments lights may consist of multiple segments or multiple polygons or curves. In one or more embodiments, a light may surround a central area without lighting elements, and one or more cameras may be placed in this central area.

In one or more embodiments the light elements such as light **5011** may be controllable, so that the intensity, wavelength, or other characteristics of the emitted light may be modified. Light may be modified for example to provide consistent lighting throughout the day or throughout a store area. Light may be modified to highlight certain sections of a store. Light may be modified based on camera images received by the cameras coupled to the light elements, or based on any other camera images. For example, if the store system is having difficulty tracking shoppers, modification of emitted light may improve tracking by enhancing contrast or by reducing noise.

FIG. **51** shows a closeup view of integrated lighting-camera module **4810***a*. A bracket system **5101** connects to light **5011** (at two sides) and to the two cameras **5012** and **5013** in the center of the light, and this bracket **5101** has connections to rail **4822***a* that may be positioned at any points along the rail. The center horizontal section **5102** of the bracket system **5101** provides mounting slots for the cameras, such as slot **5103** into which camera mount **5104** for camera **5013** is mounted; these slots allow the number and position of cameras to be modified as needed. In one or more embodiments this central camera mounting bracket **5102** may be similar to or identical to the shelf camera mounting bracket shown in FIG. **46**A, for example. In one or more embodiments, ceiling cameras such as camera **5013** may also be similar to or identical to the shelf cameras such as camera **4312***a* shown in FIG. **46**A. Use of similar or identical components in both smart shelves and smart ceilings may further simplify installation, operation, and maintenance of an automated store, and may reduce cost through use of common components.

While the invention herein disclosed has been described by means of specific embodiments and applications thereof, numerous modifications and variations could be made thereto by those skilled in the art without departing from the scope of the invention set forth in the claims.

What is claimed is:

1. An autonomous store tracking system, comprising

a processor configured to

obtain a 3D model of a store that contains items and item storage areas;

receive a time sequence of images from each camera of a plurality of cameras in said store, wherein said time sequence of images from each camera is captured over a time period;

analyze said time sequence of images and said 3D model of said store to

determine a sequence of locations of a person in said store during said time period; and

calculate a field of influence volume around each location of said sequence of locations;

when said field of influence volume intersects an item storage area of said item storage areas during an interaction time period within said time period,

receive a first image from a camera in said store oriented to view said item storage area, wherein said first image is captured before or at the beginning of said interaction time period;

receive a second image from said camera in said store oriented to view said item storage area, wherein said second image is captured after or at the end of said interaction time period;

set an input of a neural network to said first image and said second image,

wherein said neural network outputs

a probability that each item of said items is moved during said interaction time period, and

a probability that each action of a set of actions is performed during said interaction time period;

select an item from said items with a highest probability of being moved during said time period in an output of said neural network;

select an action from said set of actions with a highest probability of being performed during said time period in said output of said neural network, and,

attribute said action and said item to said person.

2. The autonomous store tracking system of claim **1**, wherein

said plurality of cameras in said store comprise a plurality of ceiling cameras mounted on a ceiling of said store;

said analyze said time sequence of images and said 3D model of said store comprises

project said time sequence of images from each ceiling camera onto a plane parallel to a floor of said store, to form a time sequence of projected images corresponding to each ceiling camera;

analyze said time sequence of projected images corresponding to each ceiling camera, and said 3D model of said store to

determine said sequence of locations of a person in said store during said time period; and

calculate said field of influence volume around each location of said sequence of locations.

3. The autonomous store tracking system of claim **2**, wherein said each ceiling camera of said plurality of ceiling cameras is a fisheye camera.

4. The autonomous store tracking system of claim **2**, wherein

said field of influence volume around each location of said sequence of locations is a translated copy of a standardized shape.

5. The autonomous store tracking system of claim **4**, wherein said standardized shape comprises a cylinder.

6. The autonomous store tracking system of claim **2**, wherein said each location of said sequence of locations comprises a point.

7. The autonomous store tracking system of claim **2**, wherein said determine said sequence of locations of a person in said store during said time period comprises

for each time in said time sequence of projected images corresponding to each ceiling camera,

subtract a store background image from each projected image of said projected images captured at said each time to form a corresponding plurality of masks at said each time;

combine said plurality of masks at said each time to form a combined mask; and,

identify a location of said person at said each time as a high intensity location in said combined mask.

8. The autonomous store tracking system of claim **2**, wherein said determine said sequence of locations of a person in said store during said time period comprises

for each time in said time sequence of projected images corresponding to each ceiling camera,

input said projected images captured at said each time into a machine learning system that outputs an intensity map, wherein

said intensity map comprises a likelihood at each location that a person is at said location.

9. The autonomous store tracking system of claim **8**, wherein said machine learning system comprises a neural network.

10. The autonomous store tracking system of claim **9**, wherein said neural network comprises a fully convolutional network.

11. The autonomous store tracking system of claim **10**, wherein said fully convolutional network comprises

a first half subnetwork comprising a plurality of copies of a feature extraction network, each copy of said plurality of copies corresponding to a ceiling camera of said plurality of ceiling cameras, wherein said each copy comprises an input layer comprising a corresponding projected image of said projected images;

a feature merging layer coupled to said first half subnetwork, wherein said feature merging layer averages outputs of said plurality of copies of said feature extraction network; and,

a second half subnetwork coupled to said feature merging layer, wherein an output layer of said second half subnetwork comprises said intensity map.

12. The autonomous store tracking system of claim **11**, wherein said input layer of said each copy further comprises

a position map corresponding to the ceiling camera corresponding to said copy, wherein a value of said position map at a location is a function of a distance between said location on said plane and said each ceiling camera.

13. The autonomous store tracking system of claim **8**, wherein said determine said sequence of locations of a person in said store during said time period further comprises

input into said machine learning system a position map corresponding to each ceiling camera of said plurality of ceiling cameras, wherein a value of said position map at a location is a function of a distance between said location on said plane and said each ceiling camera.

14. The autonomous store tracking system of claim **2**, further comprising

a modular ceiling comprising

a longitudinal rail mounted to said ceiling of said store;

one or more transverse rails, wherein each transverse rail of said one or more transverse rails is mounted to said longitudinal rail;

one or more integrated lighting-camera modules mounted to said each transverse rail, wherein each integrated lighting-camera module of said one or more integrated lighting-camera modules comprises a lighting element surrounding a center area; and,

two or more ceiling-mounted cameras of said plurality of ceiling-mounted cameras mounted in said center area.

15. The autonomous store tracking system of claim **14**, wherein

a position of said each transverse rail along said longitudinal rail is adjustable;

a position of said each integrated lighting-camera module along said each transverse rail is adjustable;

said center area comprises a camera module comprising at least one slot into which said two or more ceiling-mounted cameras are attached; and,

a position of each ceiling-mounted camera of said two or more ceiling-mounted cameras in said at least one slot is adjustable.

16. The autonomous store tracking system of claim 1, further comprising

one or more modular shelves, each modular shelf of said one or more modular shelves comprising

at least one camera module mounted on a bottom side of said each modular shelf, wherein each camera module of said at least one camera module comprises two or more downward-facing cameras;

at least one lighting module, wherein each lighting module of said at least one lighting module comprises a downward-facing light;

a right-facing camera mounted on or proximal to a left edge of said each modular shelf;

a left-facing camera mounted on or proximal to a right edge of said each modular shelf;

a processor; and,

a network switch.

17. The autonomous store tracking system of claim 16, wherein

said item storage areas comprise said one or more modular shelves;

said camera in said store oriented to view said item storage area comprises a downward-facing camera of said downward-facing cameras of a modular shelf of said one or more modular shelves located above said item storage area.

18. The autonomous store tracking system of claim 16, wherein

said each modular shelf comprises a front rail and a back rail onto which said each camera module and said each lighting module are attached;

a position of said each camera module along said front rail and said back rail is adjustable; and;

a position of said each lighting module along said front rail and said back rail is adjustable.

19. The autonomous store tracking system of claim 16, wherein

said each camera module comprises at least one slot into which said two or more downward-facing cameras are attached; and,

a position of each downward-facing camera of said two or more downward-facing cameras in said at least one slot is adjustable.

20. An autonomous store tracking system, comprising:

a modular ceiling in a store comprising

a longitudinal rail mounted to a ceiling of said store;

one or more transverse rails, wherein each transverse rail of said one or more transverse rails is mounted to said longitudinal rail;

one or more integrated lighting-camera modules mounted to said each transverse rail, wherein each integrated lighting-camera module of said one or more integrated lighting-camera modules comprises

a lighting element surrounding a center area; and

two or more ceiling-mounted cameras of a plurality of ceiling-mounted cameras of said store mounted in said center area;

wherein

a position of said each transverse rail along said longitudinal rail is adjustable;

a position of said each integrated lighting-camera module along said each transverse rail is adjustable;

said center area comprises a camera module comprising at least one slot into which said two or more ceiling-mounted cameras are attached; and

a position of each ceiling-mounted camera of said two or more ceiling-mounted cameras in said at least one slot is adjustable;

one or more modular shelves in said store, each modular shelf of said one or more modular shelves comprising

at least one camera module mounted on a bottom side of said each modular shelf, wherein each camera module of said at least one camera module comprises two or more downward-facing cameras;

at least one lighting module, wherein each lighting module of said at least one lighting module comprises a downward-facing light;

a right-facing camera mounted on or proximal to a left edge of said each modular shelf;

a left-facing camera mounted on or proximal to a right edge of said each modular shelf;

a processor; and

a network switch;

wherein

said each modular shelf is an item storage area for one or more items in said store;

said each modular shelf comprises a front rail and a back rail onto which said each camera module and said each lighting module are attached;

a position of said each camera module along said front rail and said back rail is adjustable;

a position of said each lighting module along said front rail and said back rail is adjustable;

said each camera module comprises at least one slot into which said two or more downward-facing cameras are attached; and

a position of each downward-facing camera of said two or more downward-facing cameras in said at least one slot is adjustable;

a processor configured to

obtain a 3D model of said store;

receive a time sequence of images from each camera of said plurality of ceiling-mounted cameras, wherein said time sequence of images from each camera is captured over a time period;

project said time sequence of images from each ceiling camera onto a plane parallel to a floor of said store, to form a time sequence of projected images corresponding to each ceiling camera;

analyze said time sequence of projected images corresponding to each ceiling camera, and said 3D model of said store to

determine a sequence of locations of a person in said store during said time period; and

calculate a field of influence volume around each location of said sequence of locations;

when said field of influence volume intersects an item storage area of said item storage areas during an interaction time period within said time period,

receive a first image from a camera in said store oriented to view said item storage area, wherein said first image is captured before or at the beginning of said interaction time period;

receive a second image from said camera in said store oriented to view said item storage area, wherein said second image is captured after or at the end of said interaction time period;

set an input of a neural network to said first image and said second image,

wherein said neural network outputs

a probability that each item of said items is moved during said interaction time period, and

a probability that each action of a set of actions is performed during said interaction time period;

select an item from said items with a highest probability of being moved during said time period in an output of said neural network;

select an action from said set of actions with a highest probability of being performed during said time period in said output of said neural network, and,

attribute said action and said item to said person.

* * * * *