(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
27 December 2012 (27.12.2012)

WIPO | PCT

(10) International Publication Number
WO 2012/177336 A2

(51) International Patent Classification:
G06T 17/05 (2011.01) *G06T 7/00* (2006.01)

(21) International Application Number:
PCT/US2012/037673

(22) International Filing Date:
12 May 2012 (12.05.2012)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/484,904 11 May 2011 (11.05.2011) US

(71) Applicant *(for all designated States except US)*: **UNIVERSITY OF FLORIDA RESEARCH FOUNDATION, INC.** [US/US]; 223 Grinter Hall, Gainesville, FL 32611 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **KIM, Hakjae** [US/US]; 2600 Sw Williston Rd. #1722, Gainesville, FL 32608 (US). **DIXON, Warren** [US/US]; 8681 Sw 89th Lane, Gainesville, FL 32608 (US).

(74) Agent: **RISLEY, David, R.**; THOMAS, KAYDEN, HORSTEMEYER & RISLEY, LLP, 400 Interstate North Parkway, Suite 1500, Atlanta, GA 30339 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

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

Published:

— *without international search report and to be republished upon receipt of that report (Rule 48.2(g))*

— *with information concerning request for restoration of the right of priority in respect of one or more priority claims (Rules 26bis.3 and 48.2(b)(vii))*

(54) **Title**: SYSTEMS AND METHODS FOR ESTIMATING THE GEOGRAPHIC LOCATION AT WHICH IMAGE DATA WAS CAPTURED

(57) **Abstract**: In some embodiments, a system and method for estimating the geographical location at which image data was captured with a camera identifies matching feature points between the captured images, estimates a pose of the camera during the image capture from the feature points, performs geometric reconstruction of a scene in the images using the estimated pose of the camera to obtain a reconstructed scene, and compares the reconstructed scene to overhead images of known geographical origin to identify potential matches

WO 2012/177336 A2

SYSTEMS AND METHODS FOR ESTIMATING THE GEOGRAPHIC LOCATION AT WHICH IMAGE DATA WAS CAPTURED

Notice of Government-Sponsored Research

This invention was made with Government support under contract number 00078739, awarded by the U.S. Department of Defense. The Government has rights in the claimed inventions.

Cross-Reference to Related Application(s)

This application claims priority to co-pending U.S. Provisional Application serial number 61/484,904, filed May 11, 2011, which is hereby incorporated by reference herein in its entirety.

Background

In certain situations, it is desirable to be able to determine the geographic location at which arbitrary image data, such as still images or video, was captured. That location can in some cases be determined from a description that is presented along with the image data or from metadata that is embedded in the image data. When such information is not provided, image matching algorithms can sometimes be used to determine the location at which the image data was captured. Specifically, the scenes

within the image data can be compared to scenes within images of one or more databases. If the scene or a part of the scene in the image data matches a scene or part of a scene of an image of known origin stored within the database, it can be assumed that the image data was captured at the same location as was the database image. Unfortunately, the effectiveness of a technique is limited by the content of the database. Although it may be relatively easy to identify a location when it is a location that is frequently photographed, such as tourist destinations, it may be more difficult to find matches for image data that was captured at other locations.

From the above discussion, it can be appreciated that it would be desirable to be able to estimate the geographic location at which image data was captured without having to rely on conventional image matching.

Brief Description of the Drawings

The present disclosure may be better understood with reference to the following figures. Matching reference numerals designate corresponding parts throughout the figures, which are not necessarily drawn to scale.

Fig. 1A is an overhead map of an example geographical area of interest.

Fig. 1B is a ground-level image captured a particular location identified on the map of Fig. 1A

Figs. 2A and 2B are close-ups of portions of the image shown in Fig. 1B and identify feature points on first and second buildings captured in the image.

Fig. 3 is a projection of translation to a unit sphere that is performed during pose estimation by gridding of unit spheres (PEGUS).

Fig. 4 comprises an uncropped version of the image shown in Fig. 1B and illustrates a reconstructed scene that has been generated from the image using geometric reconstruction.

Fig. 5 is an occupancy grid map generated from a satellite image of the geographical area of interest shown in Fig. 1A.

Figs. 6A-6D are example segmented maps that show only particular types of white spaces contained within a satellite image.

Fig. 7 is a flow diagram that describes of an embodiment of a method for estimating the geographic location at which image data was captured.

Fig. 8 shows an example result of estimating geographic location in which a highest probability match is highlighted within the occupancy grid map of Fig. 5, the match identifying the location at which the image in Fig. 1B and Fig. 4 was captured.

Fig. 9 is a block diagram of a computing device that can be used to estimate the geographic location at which image data was captured.

Detailed Description

As described above, it would be desirable to be able to estimate the geographic location at which image data, such as a still image or a video clip, was captured without relying on conventional image matching. Described herein are systems and methods with which the location at which "ground-level" image data (e.g., video) of a scene was captured can be estimated by determining the geometry of the scene and comparing it with overhead images of known geographical origin. With such a process, the ground-level viewpoint of the location can be correlated with an overhead viewpoint of the

location. In some embodiments, feature points are matched between captured images (frames), the rotation and translation of the camera are estimated, and Euclidean reconstruction of the scene is performed to obtain a reconstructed scene process. The geometry within the reconstructed scene can then be compared with overhead images, such as satellite images, to find a possible match. In further embodiments, "white spaces" that exist between objects in the scene and the camera are also used in the matching process.

In the following disclosure, multiple embodiments are described. It is noted that those embodiments are merely example implementations of the disclosed inventions and that other embodiments are possible. All such embodiments are intended to fall within the scope of this disclosure.

Described herein are systems and methods for determining the geographical location at which image data, such as arbitrary hand-held video captured by a user on the ground, was captured by transforming the image data into a reconstructed scene and comparing reconstructed scene with an overhead image of the area. The systems and methods overcome major difficulties of other geolocation approaches, such as the limitations of image databases and computational expense.

Because the ground-level perspective is substantially orthogonal to the overhead perspective, images taken from the two perspectives do not share much information. This, of course, makes comparison of ground-level images and overhead images difficult. This difficulty can be overcome, however, by conducting viewpoint transformation using the Euclidian geometry reconstruction of the captured scene and using spatial information of the reconstructed scene to localize the scene within the

overhead image. Such reconstruction can be performed by first matching feature points across the images (frames) and estimating the pose of the camera that captured the images. These steps are described below.

Feature Detection and Matching

Fig. 1B shows an example ground-level image I that can, for example, comprise one or many frames of a video sequence that was captured with a camera. The image was captured at a particular location identified in an overhead map M of a geographical area of interest (the campus of the University of Florida in this example), which is shown in Fig. 1A. The scene captured in the image of Fig. 1B includes two buildings: building S1 on the left and building S2 on the right. The locations of those buildings are identified in the map of Fig. 1A, as is the perspective from which the image was captured (see the diverging arrows in Fig. 1A).

As noted above, one step in the reconstruction process identifying matching feature points in the images. Figs. 2A and 2B are portions of the image shown in Fig. 1B (portions identified in Fig. 1B with rectangles labeled S1 and S2) and illustrate example feature points associated with buildings S1 and S2, respectively. Feature point matching is performed to see how the feature points in the scene "move" from image to image (frame to frame) and therefore provides an indication of the relative movement between the scene and the camera. Assuming that the scene is stationary, the relative movement arises from movement of the camera in terms of rotation and translation. In some cases, this movement can be the result of panning of the camera during image capture. In other cases, the movement can be the result of movement of an object, such

as an unmanned autonomous vehicle (UAV), to which the camera is mounted. Regardless, understanding how the camera is moving relative to the scene enables determination of the geometry and position of objects (and voids) within the scene. The determination of such that geometry and position is a structure-from-motion problem. In other words, the geometry of the scene can be reconstructed from an estimated motion of the camera.

A number of different techniques can be used to match local image features. One such method is scale invariant feature transformation (SIFT). The SIFT process has many properties that are important in matching different images of an object or scene. Specifically, the process invariant to image scaling and rotation and is partially invariant to change in illumination and three-dimensional camera viewpoint. In SIFT, potential key points are selected by scanning over entire scales and image locations. This can be efficiently implemented by constructing a Gaussian pyramid and searching for the local maxima in a collection of different Gaussian images.

Next, candidate key-points are localized to sub-pixel accuracy and are selected based upon their stability. The dominant orientations for each key point are identified based on their local image patch. Assigned orientation, scale, and location for each key point are used as bases to transform all image data, which provides invariance to similarity transforms. The last stage of SIFT involves building a local image descriptor that is compact, distinctive, and robust to illumination and camera viewpoint. The key points from the images are extracted and saved from each image, and then the descriptors are compared to compute corresponding feature matches between images.

Significantly, feature point matching can be performed using other than SIFT methods. For example, feature point matching can be achieved using one or more of Harris points, mutual information points, or Kanade-Lucas-Tomasi points. Regardless of how the feature point matching is performed, the result is point correspondence between the images and a set of feature points is obtained, the set comprising a plurality of feature point pairs, each pair comprising a first point within a first image associated with a given feature in the scene and a second point within a second image also associated with the given feature.

Camera Pose Estimation

The feature points obtained from the feature point matching described above can be used to estimate the pose (i.e., rotation and translation) of the camera as the images were captured. As was previously noted, the pose of the camera may change because of panning of the camera because of movement of an object to which the camera is mounted. Regardless, the movement of the camera facilitates determination of the geometry of objects in the captured scene.

An important aspect of effective pose estimation is rejecting outliers prior to estimation because point correspondences returned by the feature tracker often contain gross mismatches or large errors in feature point locations. A number of algorithms exist to solve this problem. Example solutions include random sample consensus (RANSAC), RANSAC + Least Squares, and nonlinear mean shift, which are hypothesize-and-test methods. Another solution is a new pose estimation algorithm developed by the inventors called pose estimation by gridding of unit spheres (PEGUS), which has been

shown to have superior performance as compared to that of RANSAC + Least Squares and the nonlinear mean shift.

PEGUS involves three major stages. The first stage is the hypothesis generation stage, which is used to reduce the number of feature points that will be used to estimate the camera pose. In this stage, a sampling-with-replacement strategy is used to generate n number of hypotheses that have small "correlation." The total number of pose hypotheses, $N_{max} = (M/P)$, is typically an extremely large number, where M is the number of corresponding points and P is minimum number required for pose estimation. Because it is computationally expensive to consider all of the pose hypotheses, only a relatively small number of sample hypotheses, n, is used. In some embodiments, n= 4 or 8.

The second stage of the PEGUS process is rotation estimation. The first step in this stage is estimating the mode of the rotation hypotheses. Each rotation hypotheses matrix is transformed into unit-quaternions, q. Each unit-quaternion q_i is plotted onto the surface of 3-sphere $\mathbb{S}3$ and the surface of sphere is divided into a number of regions of equal area. The probability mask function (PMF) of the random variable q can be estimated by counting the number of rotation hypotheses that lie within each region. Once the region with the greatest number of unit-quaternions is identified, the mode of the PMF, q^*, can be estimated by finding the point that occurs most frequently in the region. The next step involves is extracting low-noise measurements. The rotation hypotheses that is within the distance of εq form the mode q^* is selected such that following equation is satisfied

$$d_q\left(q^*, q_i\right) < \varepsilon_p \qquad \text{[Equation 1]}$$

where the distance function $d_p(\cdot,\cdot)$ is the Riemannian distance. The final step in the rotation estimation is averaging low-noise data. In a Euclidian sense, the optimal average, $\hat{R}$, of the rotation matrices can be computed by

$$\hat{R} = \underset{R \in SO(3)}{\arg\min} \sum_{i=1}^{N_1} \left\| R_i - R \right\|^2 \qquad \text{[Equation 2]}$$

where N_1 is the number of elements in the low noise rotation hypotheses and R_i denotes the rotation matrix corresponding to q_i. The optimal average $\hat{R}$ can be computed by performing an orthogonal projection of the arithmetic average on to the special orthogonal group $SO(3)$.

The third stage of the PEGUS process is translation estimation. Unit translation estimation, which provides the direction of the camera, is very similar to that rotation estimation. A mode of translation hypotheses is first identified. Each translation hypothesis is plotted onto the surface of 2-sphere $\mathbb{S}^2$, and the surface of sphere is divided into number of regions of equal area as shown in Fig. 3. The PMF of the random variable t can be estimated by counting the number of rotation hypotheses that lie within each region. Once the region with greatest number of unit-quaternions has been identified, the mode of the PMF, t^*, is estimated by finding the point that occurs most frequently in the region. Next, low-noise measurements are extracted. A low-noise set

of translation hypotheses is then selected by taking pre-defined small positive number ε_t, and collecting the points around the mode t^* that satisfies

$$d_t\left(t^*, t_i\right) < \varepsilon_t \qquad \text{[Equation 3]}$$

where $d_t\left(t^*, t_i\right)$ is the geodesic distance between the unit translation vectors t^*, and t_i. Because the hypotheses of a unit translation are elements of $\mathbb{S}^2$, the optimal average is achieved by taking the normalized arithmetic mean of the low-noise set, which is given by

$$t = \frac{\sum_{i=1}^{N_2} \frac{t_i}{N_2}}{\left\| \sum_{i=1}^{N_2} \frac{t_i}{N_2} \right\|} \qquad \text{[Equation 4]}$$

where N_2 is the number of elements in the low-noise data set of unit translation. When PEGUS is used with a homography matrix that includes scaled depth information, the translation between the two views are elements of $\mathbb{R}^3$, not $\mathbb{S}^2$. In this case, histogram construction, mode estimation, and hypotheses extraction are performed by dividing a particular volume of $\mathbb{R}^3$ into K_t bins of equal volume, where each bin is a cube with equal sides. The volume to grid is chosen to include all the hypotheses. The remainder of the process is substantially the same as the unit translation case. In some

embodiments, the rotation and translation hypotheses can be coupled together to represent complete pose of the camera between two images.

Further details regarding PEGUS are provided in PCT/US06/004469, which is hereby incorporated by reference into the present disclosure. Although PEGUS has been explicitly identified as a suitable method for determining camera pose, it is noted that other pose estimation techniques, such as one or more of those identified above, can be used.

Euclidian Reconstruction with a Single Known Length

The structure-from-motion problem can be solved to attach reference frames to piecewise planar objects in a scene using the methods described above. Geometric reconstruction can then be performed using the Euclidian homography relationship. Although the reconstructed scene could be iteratively rescaled and compared to overhead images, the matching process can be conducted far more quickly when a length of an object in the scene is known. In some embodiments, a reference length can be manually input by a user. For example, if the scene includes an object that has a typical length, such as a door, the user can communicate to the program performing the location estimation that the door in the scene is of a particular length (e.g., approximately 7 feet). Alternatively, one or more image analysis algorithms can be used to automatically determine the reference length. For example, an image analysis algorithm can identify an object in the scene having a typical length and provide that length to the location estimation program. In other embodiments, the known velocity of the camera can be used in lieu of a known length.

The case considered here is one in which a large number of feature points P are grouped into k sets of coplanar points $P_h \subset P, \forall h \in \{1...k\}$, where all points in P_h lie in a plane π_h. The coordinate of the j^{th} point in P_h is given as

$$\begin{aligned} \bar{m}_{hj}^{*} &= \left[x_{hj}^{*}, y_{hj}^{*}, z_{hj}^{*}\right]^{T}, \\ \bar{m}_{hj}(t) &= \left[x_{hj}, y_{hj}, z_{hj}\right]^{T}, \\ & \forall j \in \{1...N_h\}, \forall h \in \{1...k\} \end{aligned} \qquad \text{[Equation 5]}$$

in the frames F_c^{*} and $F_c(t)$, respectively. The normalized coordinates of $\bar{m}_{hj}^{*}$ and $\bar{m}_{hj}(t)$ projected onto the image plane π_i are given as

$$\begin{aligned} m_{hj}^{*} &= \left[\frac{x_{hj}^{*}}{z_{hj}^{*}}, \frac{y_{hj}}{z_{hj}^{*}}, 1\right]^{T}, \forall j \in \{1....k\} \\ m_{hj}(t) &= \left[\frac{x_{hj}(t)}{z_{hj}(t)}, \frac{y_{hj}(t)}{z_{hj}(t)}, 1\right]^{T}, \forall j \in \{1....k\} \end{aligned} \qquad \text{[Equation 6]}$$

For each set of points, there exists a homography $H_h(t) \in \mathbb{R}^{3x3}$ such that relationship between $\bar{m}_{hj}^{*}$ and $\bar{m}_{hj}(t)$ is given by

$$m_{hj} = \alpha H_h m_{hj}^{*} \qquad \text{[Equation 7]}$$

$$= \alpha \left(R(t) + \frac{x(t)}{d_h^*} n_h^{*T} \right) m_{hj}^* \qquad \text{[Equation 8]}$$

where $\alpha = \frac{z_{hj}^*}{z_{hj}}$ is a scalar depth ratio, $R(t)$ is rotation matrix, $x(t)$ is true translation vector, d_h^* is depth, n_h^* is normal to the plane π_h. $R(t)$ and $x(t)$ are the same for all point sets because all coordinate changes are due to the motion of the camera. However, each P_h has a different d_h^* and n_h^*. Therefore, $H_h(t)$ is also distinct. Each $H_h(t)$ is decomposed into $R(t), \frac{x(t)}{d_h^*}$, and n_h^*. Note that translation is only recovered up to a scaled factor $x_h(t) = \frac{x(t)}{d_h^*}$ and the depth d_h^* is generally unknown.

Based on the assumption that the constant, scalar length

$$\bar{s}_{h1} = \left\| \bar{m}_{h1}^* - \bar{m}_{h2}^* \right\| \qquad \text{[Equation 9]}$$

is known, d_h^* can be recovered. Without loss of generality, the length is assumed to be known in set P_1. The translation $x(t)$ is then recovered from d_1^* as

$$x(t) = d_1^* x_1(t) \qquad \text{[Equation 10]}$$

Given $x(t)$, each d_h^* can be recovered by

$$d_h^* = \frac{x_h^T x(t)}{x_h^T x_h} \qquad \text{[Equation 11]}$$

At this point, all of the information that is needed to compute Euclidian coordinates of all points P visible in the image are possessed. Euclidian coordinates of the j^{th} point in plane P_h is given by

$$\overline{m}_{hj}^* = \frac{d_h^* m_{hj}}{n_h^* m_{hj}}, \forall j \in \{1....N\} \qquad \text{[Equation 12]}$$

Geographic Location Estimation

The three-dimensional scene reconstructed from the sequence of captured images will comprise the information about the geometry of various objects in the scene as well as the "white spaces" in the scene, which can be defined as the areas between the objects and the camera that are substantially devoid of vertically-oriented objects. Examples of white space can include "voids" such as lawns or fields, forests, bodies of water, parking lots, etc. The object geometries and the white spaces can be compared with the overhead images in an effort to find a match. Because the geometry and the white spaces extracted from the reconstructed scene cannot be compared directly with a raw satellite image, the satellite images are first processed such that both the reconstructed scene and satellite image contain similar information that can be easily

compared. One way of achieving this is to generate an occupancy-grid map (OGM) representation in which each pixel represents actual scaled physical space in real world.

The OGM comprises a two-dimensional array of cells corresponding to a horizontal grid imposed on the area to be mapped. The grid has $n \times m$ cells, and each cell has size of $s \times s$. Occupancy status with an associated certainty factor are assigned to every cell in the OGM using "0" for empty and "1" for occupied. Probabilistic representation can alternatively be used in which case the probability of a cell being occupied is represented with values between "0" to "1". The OGM representation is simple to construct, even in large-scale environments. Because the intrinsic geometry of a grid corresponds directly to the geometry of the environment, the location estimation of the reconstructed scene can be determined by its pose (position and orientation) in real world. In some embodiments, the environment is represented with two OGMs. The first OGM is a local model of the environment, which represents three-dimensional virtual space reconstructed from the image data, and the second OGM is a global model of the environment, which represents virtual space of a satellite image that is segmented into different layers, such as the contours of objects (e.g., buildings) and different types of white spaces.

Fig. 4 illustrates an example of a reconstructed scene (lower right) associated with an image (upper right), which is an uncropped version of the image I shown in Fig. 1B. The parallel lines L in the reconstructed scene represent the faces of the buildings S1 and S2 in the scene (see Fig. 1B), and the triangles T represent the white spaces between camera and the building faces. A satellite image can be segmented in a similar manner and the outlines of objects (e.g., buildings) and white spaces can be

represented in an OGM. Fig. 5 shows an example OGM that was created from a satellite image of the geographical area represented by the map of Fig. 1A. As can be appreciated from Fig. 5 when compared to Fig. 1A, the OGM identifies the boundaries of the buildings in the geographical area that is the subject of the map.

Standard and hierarchical brute-force searching can be used to search the OGM for the best match. A sufficient number of points is sampled from the outlines of the objects and white spaces in the OGM to represent a probabilistic space of the reconstructed scene. For a hierarchical search, a minimum number of sample points from the white space is used to search along the matching white spaces to ensure time is not wasted searching through an unwanted area. Conditional probability of a match *X* given an ith orientation local OGM is calculated by counting the number of matching points then using mismatched points to penalize the function.

In some embodiments, the probability calculation is performed as follows:

% Define variables and sample from work space

So = sampled points from contour of objects

Sw = sampled points from whitespace

$[xo_{id}, yo_{id}]$ = find index of So

$[xw_{id}, yw_{id}]$ = find the index of Sw

N_{index} = size of (xo_{id})

C_v = cells in local OGM representing reconstructed video

C_s = cells in global OGM representing satellite map

m_i = i^{th} orientation of local OGM

% Calculate probability of match X

$$\Pr(X|m_i) = \frac{\left(C_v(xo_{id}, yo_{id})C_s(xo_{id}, yo_{id})^T\right)}{N_{index}} - \lambda\left(C_v(xw_{id}, yw_{id})C_s(xw_{id}, yw_{id})^T\right)$$

The calculated probability is stored and the process is repeated until the entire search space is covered for all m_i for $i = (1, 2, ..., n)$.

In some embodiments, the white spaces within the satellite image can be identified and categorized using an appropriate image analysis algorithm to distinguish between multiple different types of white spaces. Figs. 6A-6D illustrate pavement, flat vegetation, water, and dense vegetation white spaces, respectively, that were obtained from an example satellite image (not shown). When similar categorization of the white spaces of the reconstructed scene is performed (e.g., again using an appropriate image analysis algorithm), such categorization can be used to eliminate possible match candidates, and therefore can increase speed of the matching process. For example, if the white space within the captured scene is determined to comprise water and a satellite image is determined to contain no water, the satellite image and its OGM can be eliminated as a potential match for the reconstructed scene.

Experimental Results

Experiments were conducted to evaluate the performance of the above-described geolocation method. Test videos were taken across the University of Florida campus and reconstructed scenes were compared with an OGM of a satellite map of the campus. The range finder of the camera was used to measure distances between the camera and each object in the capture scene, and the measured distances were

used as ground truth for scene reconstruction. Both standard search and hierarchical search results were determined and compared in terms of search duration for a single orientation. For each experiment, a probabilistic measure of match between a reconstructed scene and satellite map was calculated for an entire space of interest and then the best possible candidates (within 5% from the highest probability) were chosen based on their probability score. The location with highest probability was identified in the OGM with red star and other possible locations were marked with green circles for each case.

Both the standard and hierarchical search schemes were found to be effective at identifying the location at which the video was captured. The standard search was also found to generate more false positives than the hierarchical search and suffered from longer computation times. A search duration time comparison showed that the hierarchical search improves the speed of search by an average of 76%.

Example Systems and Methods

Fig. 7 is a flow diagram of an example method for estimating geographic location that is consistent with the above discussion. Beginning with block 10, multiple sequential images of a scene are captured with a camera. In some embodiments, the images can be individual images that are captured with a still camera. In other embodiments, the images are video frames that are captured with the camera. Regardless, there is relative motion between the camera and the scene as the images are captured. As was noted above, the relative motion can be the result of, for instance, panning of the camera or movement of an object to which the camera is mounted. The

number of frames that are required for acceptable results may depend upon several factors, including the quality of video data. Generally speaking, however, the greater the number of frames that are selected, the better the estimate will be.

Referring next to block 12, matching feature points between the captured images are identified to obtain a set of feature points. In this process, feature points are extracted from the images and those feature points are matched with corresponding feature points of the other images. By identifying the feature points, one can track how the feature points in the scene move from image to image (frame to frame). This provides an indication of the relative movement between the scene and the camera, which in turn enables determination of the geometry and position of objects and voids within the scene. In some embodiments, the matching can be performed using SIFT. Generally speaking, the feature points are associated with distinctive aspects of the objects within the scene (e.g., distinctive elements of buildings) that can be identified across multiple images. Generally speaking, the greater the number of matching feature points that are identified, the better the estimation.

In some embodiments, only a small number of feature points are needed to determine camera pose. Therefore, it may be desirable to select a subset of the feature points that will be used to estimate the pose of the camera, as indicated in block 14. As noted above, the number of feature points can be reduced by using a sampling-with-replacement strategy to generate a number of feature point hypotheses that have small correlation.

Once the desired number of feature points have been selected (e.g., 4 or 8), the rotation and translation (pose) of the camera during image capture can be estimated

from the feature points, as indicated in block 16. In some embodiments, the rotation and the translation of the camera can be determined using PEGUS as described above. At this point, a known length within the captured scene or a known velocity of the camera can be identified, as indicated in block 18. As was described above, the length can either be manually input by a user or can be automatically determined using an image analysis algorithm. The velocity of the camera can be determined, for example, from speed and direction measurement devices provided on the object to which the camera is mounted.

Referring next to block 20, geometric reconstruction of the scene is performed using the estimated pose and either the known length or known velocity to obtain a reconstructed scene. Again, Fig. 4 shows an example reconstructed scene. In that example, the reconstructed scene comprises two lines L that represent the walls of buildings in the scene that faced the camera, and the triangles T that extend to the lines represent the white space between the buildings and the camera. As can be appreciated from that reconstructed scene, performing geometric reconstruction in essence converts the perspective of the video camera (i.e., a ground-level perspective) to an overhead perspective that can be compared to overhead images. Although the reconstructed scene shown in Fig. 4 is two-dimensional, it is noted that the geometric reconstruction can be three-dimensional, in which case other information, such as the height of objects in the scene, are determined. Such information can be obtained from LIDAR images that not only comprise a two-dimensional image of a geographical area but further comprise height information about objects in the image. In such a case, the

height of the objects determined when geometric reconstruction is performed can be used as a further means to improve the estimation.

Once the reconstructed scene has been generated, it can be compared to overhead images to identify possible matches, as indicated in block 22 of Fig. 7. For example, the object geometries and the white spaces of the reconstructed scene can be compared to the geometries and white space of the OGM obtained from a satellite image. Possible matches can then be identified in the OGM. Such a situation is shown in Fig. 8 in which the highest probability match is identified in the lower left corner, which correlates with the location identified in the map of Fig. 1A at which the image of Fig. 1B was captured.

As was described above, the white space in both the reconstructed scene and the overhead images can be categorized to further improve the estimation. For example, that white space can be categorized as a field, a forest, a body of water, a parking lot, etc. to reduce the number of possible matches. The categorization can either be manual, in which case a user manually designates the white space as pertaining to a given category, or automatic in which case an algorithm automatically recognizes the type of white space based upon various cues, such as color or texture.

Fig. 9 illustrates an example architecture for a computing device 30 that can be used to perform at least part of the geolocation estimation described above in relation to Fig. 7. As indicated in Fig. 9, the computing device 30 at least comprises a processing device 32 and memory 34. The processing device 32 can include a central processing unit (CPU) or other processing device (e.g., microprocessor or digital signal processor)

and the memory 34 includes any one of or a combination of volatile memory elements (e.g., RAM) and nonvolatile memory elements (e.g., flash, hard disk, ROM).

The memory 34 stores various programs (i.e., logic), including an operating system 36 and a geolocation estimator 38. The operating system 36 controls the execution of other programs and provides scheduling, input-output control, file and data management, memory management, and communication control and related services. The geolocation estimator 38 comprises one or more algorithms and/or programs that are configured to receive ground-level image data and analyze it to estimate the location at which the image data was captured. Accordingly, the geolocation estimator 38 can include one or more of a feature point matching algorithm/program, a pose estimation algorithm/program, a Euclidean reconstruction algorithm/program, and a matching and localization algorithm/program.

CLAIMS

Claimed are:

1. A method for estimating the geographical location at which image data was captured with a camera, the method comprising:

identifying matching feature points between the captured images;

estimating a pose of the camera during the image capture from the feature points;

performing geometric reconstruction of a scene in the images using the estimated pose of the camera to obtain a reconstructed scene; and

comparing the reconstructed scene to overhead images of known geographical origin to identify potential matches.

2. The method of claim 1, wherein identifying matching feature points comprises identifying matching feature points between ground-level images.

3. The method of claim 1, wherein identifying matching feature points comprises identifying matching feature points between the captured images using scale invariant feature transformation (SIFT).

4. The method of claim 1, wherein estimating a pose of the camera comprises estimating rotation and translation of the camera.

5. The method of claim 1, wherein estimating a power of the camera comprises performing pose estimation by gridding unit sphere (PEGUS).

6. The method of claim 1, wherein performing geometric reconstruction comprises performing geometric reconstruction using both the estimated pose of the camera and a known length in the scene.

7. The method of claim 1, wherein performing geometric reconstruction comprises performing geometric reconstruction using both the estimated pose of the camera and a known velocity of the camera.

8. The method of claim 1, wherein performing geometric reconstruction converts the a ground-level perspective of the images to an overhead perspective.

9. The method of claim 1, wherein the reconstructed scene identifies geometries of objects in the scene and white spaces between the camera and the objects that are substantially free of vertically-oriented objects.

10. The method of claim 1, wherein comparing the reconstructed scene to overhead images comprises comparing the reconstructed scene to satellite images.

11. The method of claim 1, wherein comparing the reconstructed scene to overhead images comprises comparing the reconstructed scene to occupancy-grid maps (OGMs) generated from satellite images.

12. The method of claim 1, further comprising categorizing white spaces between the camera and the objects that are substantially free of vertically-oriented objects in both the reconstructed scene and the overhead images to increase the speed with which potential matches are identified.

13. A system for estimating the geographical location at which image data was captured with a camera, the system comprising:

a processing device; and

memory storing a geolocation estimator comprising logic configured to:

identify matching feature points between the captured images,

estimate a pose of the camera during the image capture from the feature points,

perform geometric reconstruction of a scene in the images using the estimated pose of the camera to obtain a reconstructed scene, and

compare the reconstructed scene to overhead images of known geographical origin to identify potential matches.

14. The system of claim 13, wherein the logic configured to identify matching feature points comprises logic configured to identify matching feature points between the captured images using scale invariant feature transformation (SIFT).

15. The system of claim 13, wherein the logic configured to estimate a pose of the camera comprises logic configured to estimate rotation and translation of the camera by performing pose estimation by gridding unit sphere (PEGUS).

16. The system of claim 13, wherein the logic configured to perform geometric reconstruction comprises logic configured to perform geometric reconstruction using both the estimated pose of the camera and a known length in the scene.

17. The system of claim 13, wherein the logic configured to perform geometric reconstruction comprises logic configured to perform geometric reconstruction using both the estimated pose of the camera and a known velocity of the camera.

18. The system of claim 13, wherein the logic configured to perform geometric reconstruction converts the a ground-level perspective of the images to an overhead perspective.

19. The system of claim 13, wherein the reconstructed scene identifies geometries of objects in the scene and white spaces between the camera and the objects that are substantially free of vertically-oriented objects.

20. The system of claim 13, wherein the logic configured to compare the reconstructed scene to overhead images comprises logic configured to compare the reconstructed scene to occupancy-grid maps (OGMs) generated from satellite images.

21. The system of claim 13, further comprising logic configured to categorize white spaces between the camera and the objects that are substantially free of vertically-oriented objects in both the reconstructed scene and the overhead images to increase the speed with which potential matches are identified.

22. A non-transitory computer-readable medium that stores a geolocation estimator for estimating the geographical location at which image data was captured with a camera, the computer-readable medium comprising:

logic configured to identify matching feature points between the captured images;

logic configured to estimate a pose of the camera during the image capture from the feature points;

logic configured to perform geometric reconstruction of a scene in the images using the estimated pose of the camera to obtain a reconstructed scene; and

logic configured to compare the reconstructed scene to overhead images of known geographical origin to identify potential matches.

23. The computer-readable medium of claim 22, wherein the logic configured to identify matching feature points is configured to identify matching feature points between the captured images using scale invariant feature transformation (SIFT).

24. The computer-readable medium of claim 22, wherein the logic configured to estimate a pose of the camera is configured to estimate rotation and translation of the camera by performing pose estimation by gridding unit sphere (PEGUS).

25. The computer-readable medium of claim 22, wherein the logic configured to perform geometric reconstruction is configured to perform geometric reconstruction using both the estimated pose of the camera and a known length in the scene.

26. The computer-readable medium of claim 22, wherein the logic configured to perform geometric reconstruction is configured to perform geometric reconstruction using both the estimated pose of the camera and a known velocity of the camera.

27. The computer-readable medium of claim 22, wherein the logic configured to perform geometric reconstruction converts the a ground-level perspective of the images to an overhead perspective.

28. The computer-readable medium of claim 22, wherein the reconstructed scene identifies geometries of objects in the scene and white spaces between the camera and the objects that are substantially free of vertically-oriented objects.

29. The computer-readable medium of claim 22, wherein the logic configured to compare the reconstructed scene to overhead images is configured to compare the reconstructed scene to occupancy-grid maps (OGMs) generated from satellite images.

30. The computer-readable medium of claim 22, further comprising logic configured to categorize white spaces between the camera and the objects that are substantially free of vertically-oriented objects in both the reconstructed scene and the overhead images to increase the speed with which potential matches are identified.

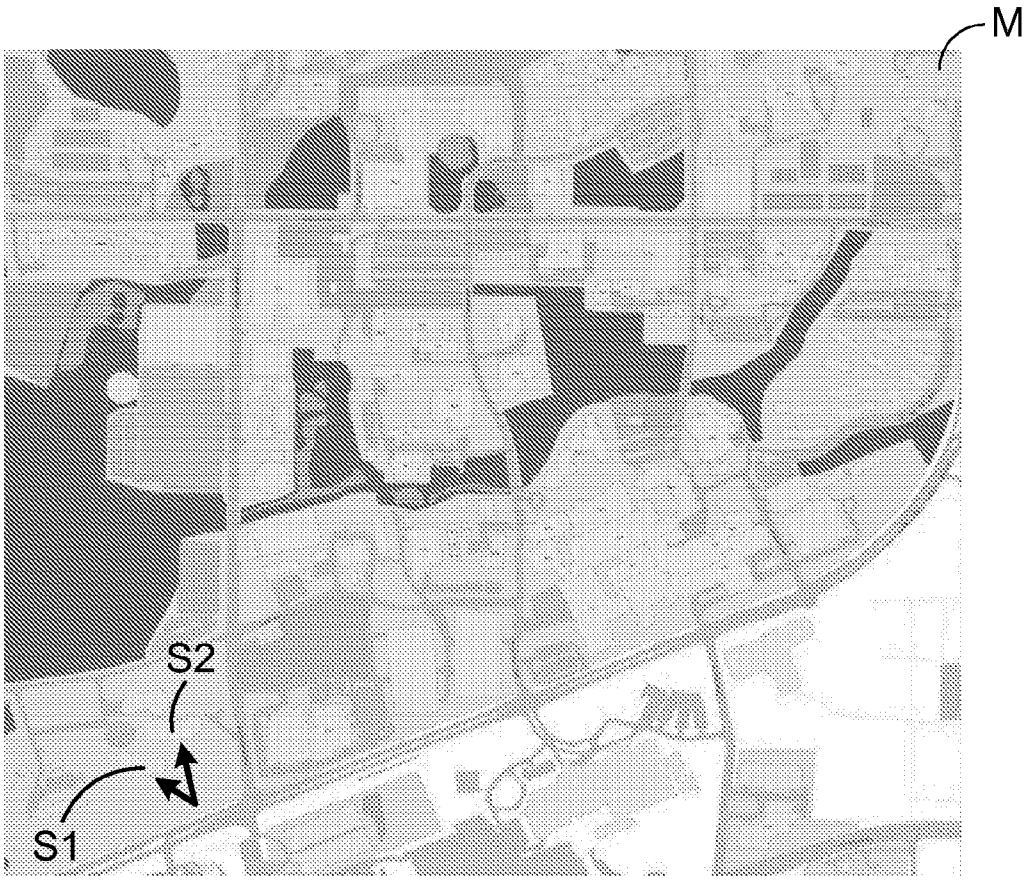

FIG. 1A

FIG. 1B

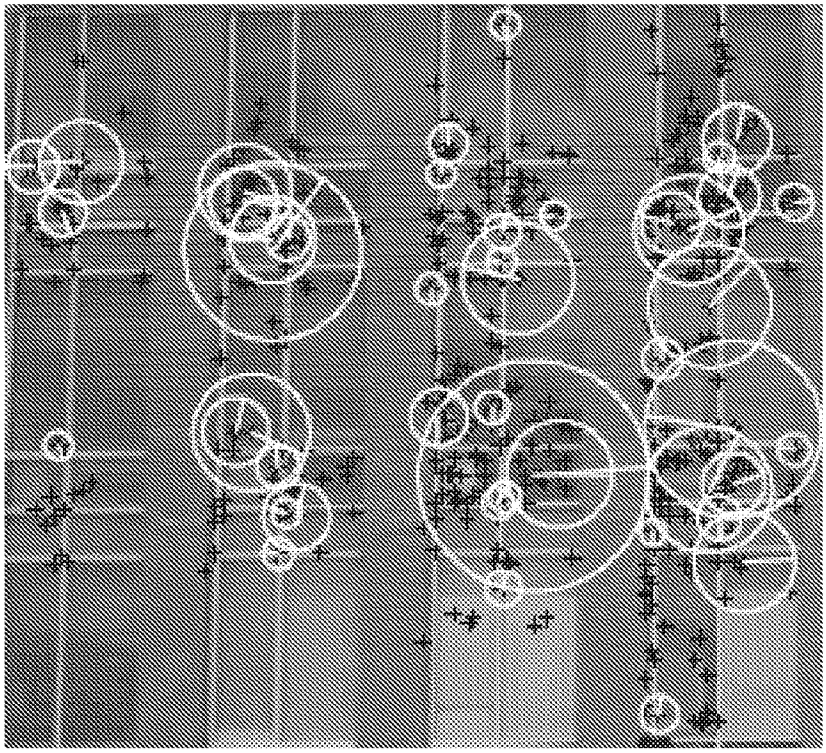

FIG. 2A

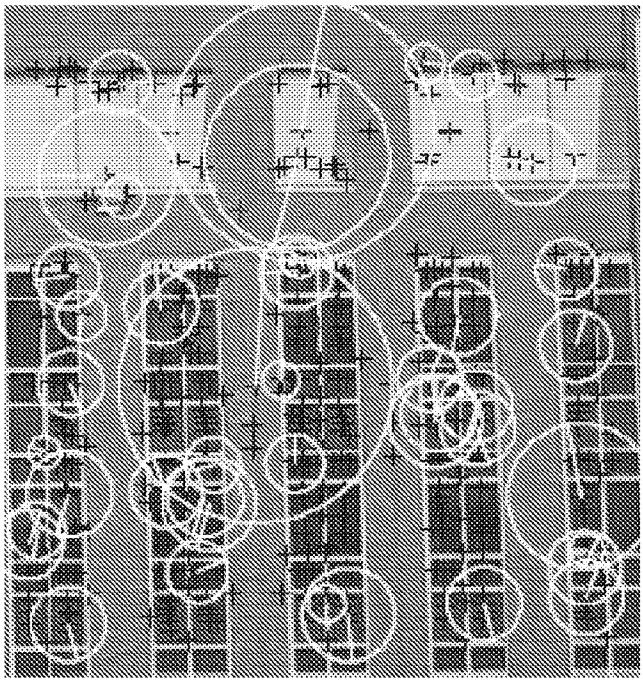

FIG. 2B

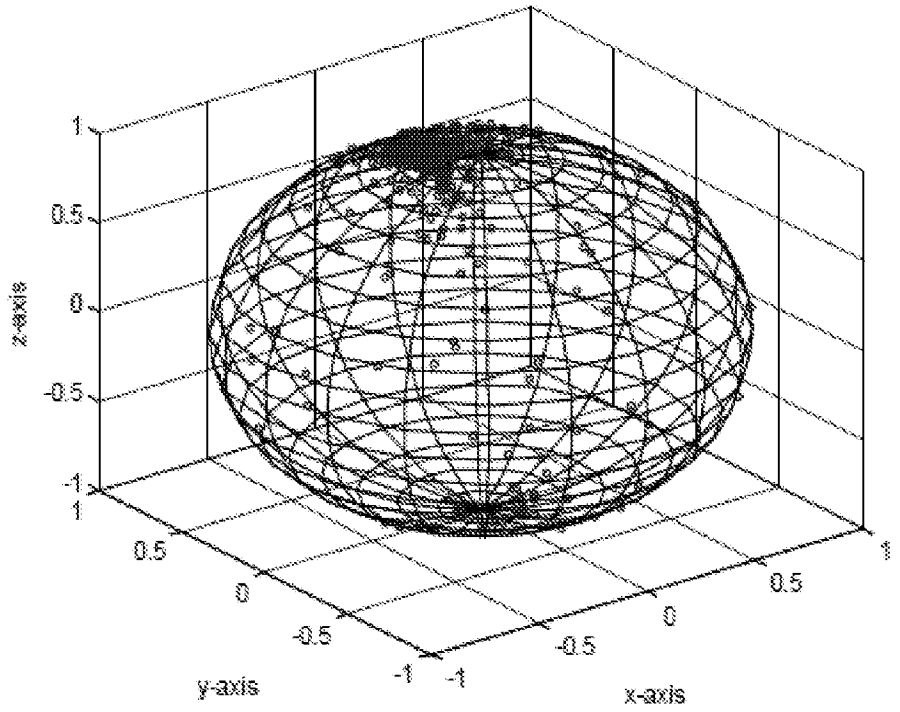

FIG. 3

FIG. 5

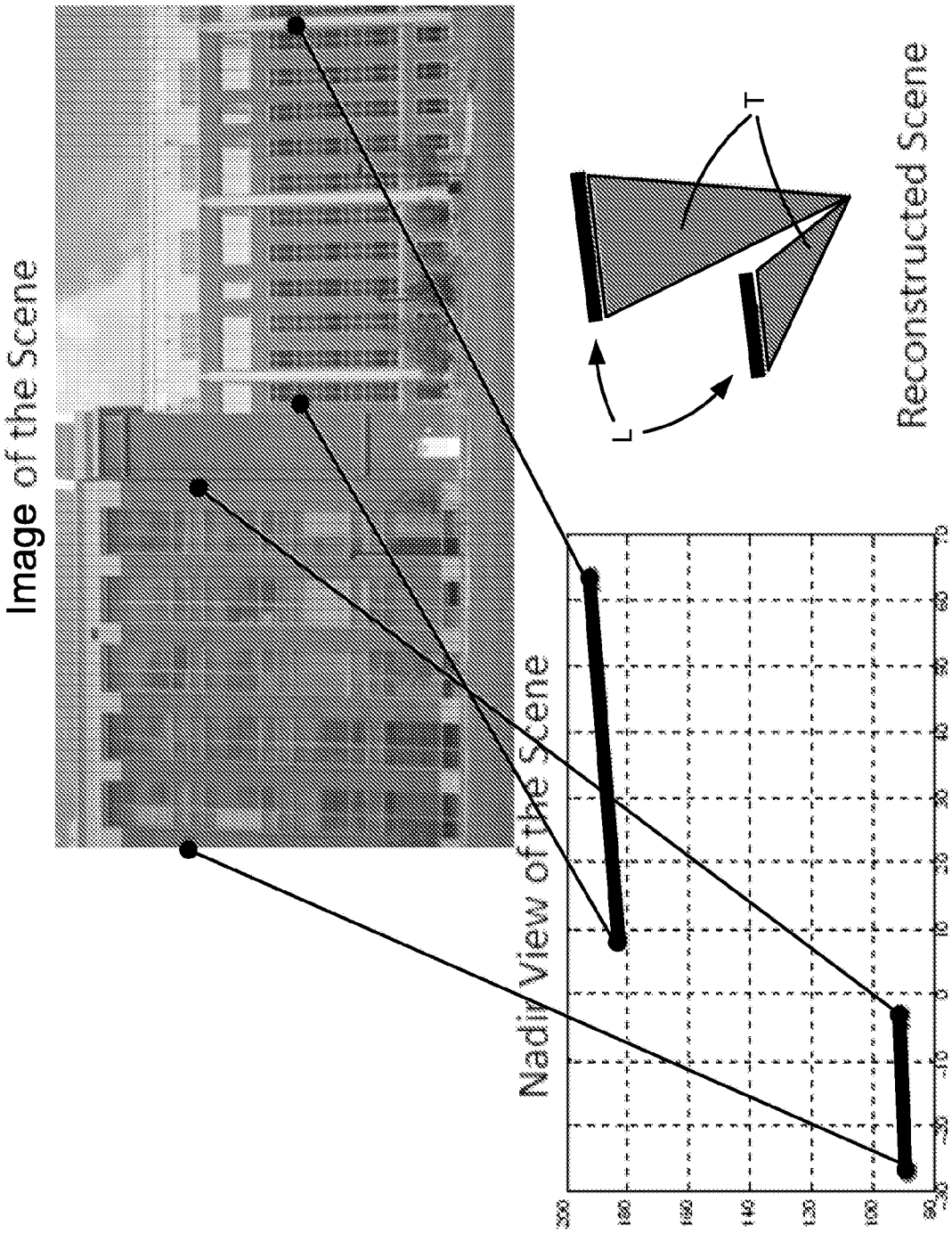
Image of the Scene
Nadir View of the Scene
Reconstructed Scene
T
L

FIG. 4

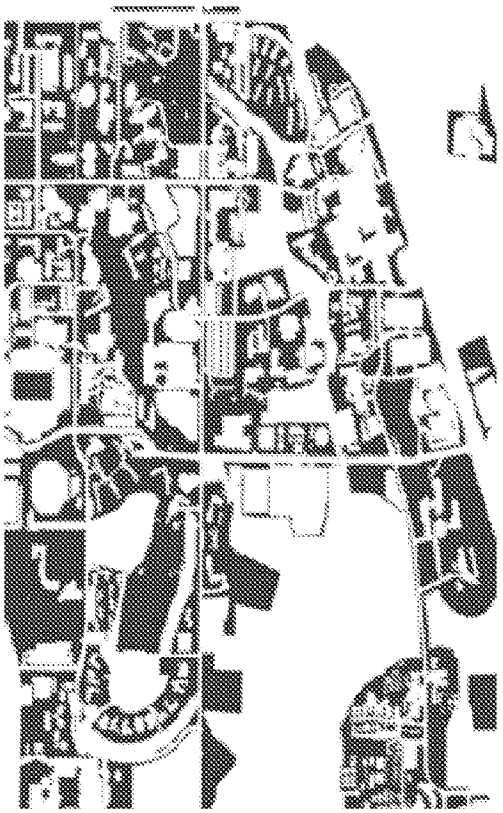

FIG. 6B

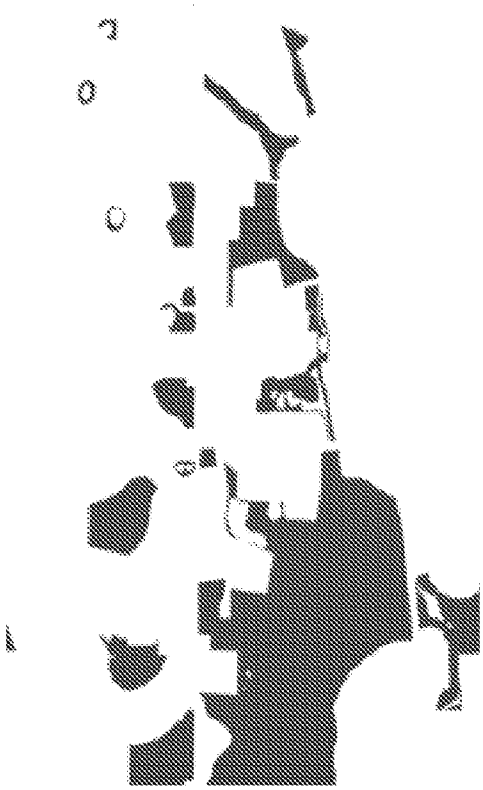

FIG. 6D

FIG. 6A

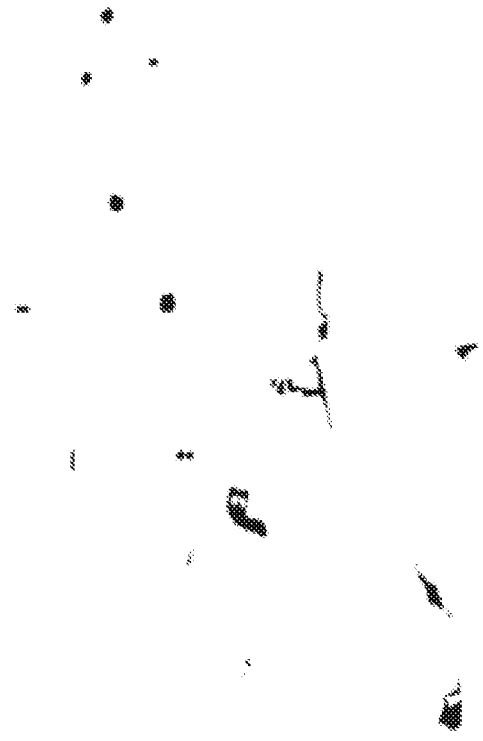

FIG. 6C

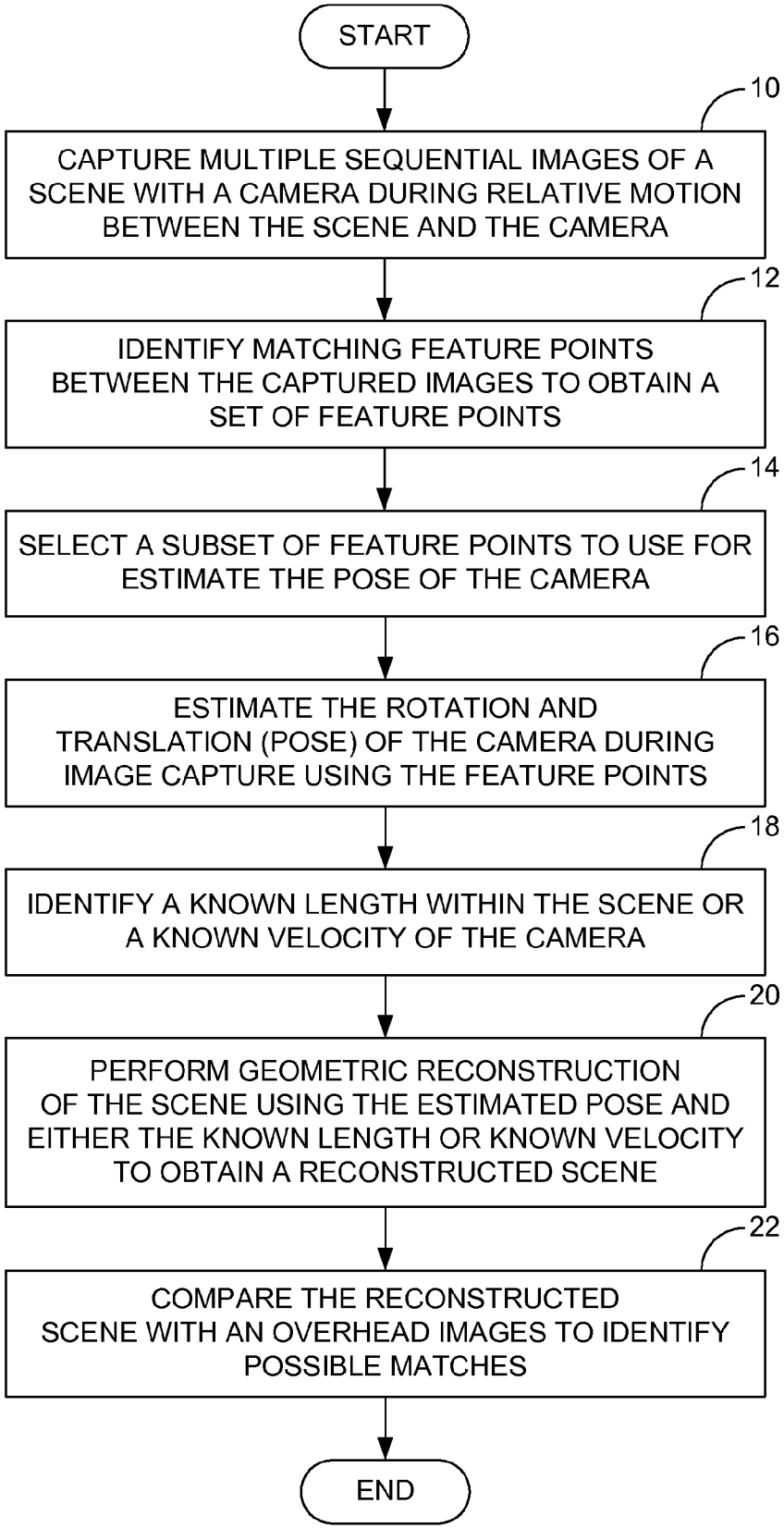
START
10
CAPTURE MULTIPLE SEQUENTIAL IMAGES OF A SCENE WITH A CAMERA DURING RELATIVE MOTION BETWEEN THE SCENE AND THE CAMERA
12
IDENTIFY MATCHING FEATURE POINTS BETWEEN THE CAPTURED IMAGES TO OBTAIN A SET OF FEATURE POINTS
14
SELECT A SUBSET OF FEATURE POINTS TO USE FOR ESTIMATE THE POSE OF THE CAMERA
16
ESTIMATE THE ROTATION AND TRANSLATION (POSE) OF THE CAMERA DURING IMAGE CAPTURE USING THE FEATURE POINTS
18
IDENTIFY A KNOWN LENGTH WITHIN THE SCENE OR A KNOWN VELOCITY OF THE CAMERA
20
PERFORM GEOMETRIC RECONSTRUCTION OF THE SCENE USING THE ESTIMATED POSE AND EITHER THE KNOWN LENGTH OR KNOWN VELOCITY TO OBTAIN A RECONSTRUCTED SCENE
22
COMPARE THE RECONSTRUCTED SCENE WITH AN OVERHEAD IMAGES TO IDENTIFY POSSIBLE MATCHES
END

FIG. 7

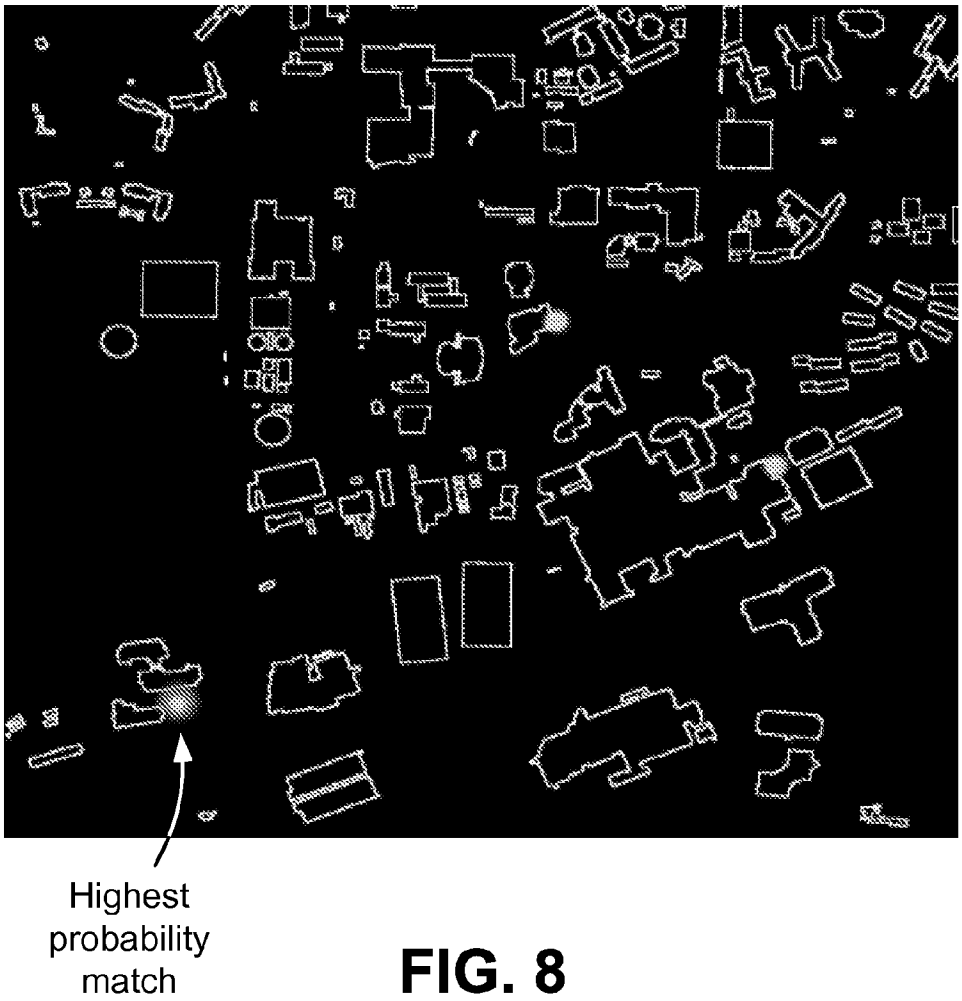
Highest
probability
match

FIG. 8

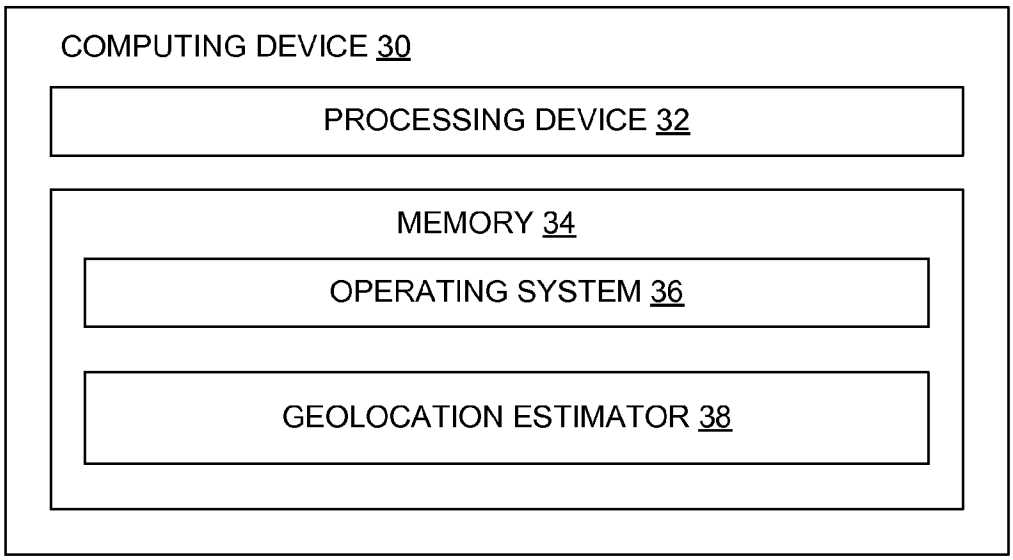
COMPUTING DEVICE 30
PROCESSING DEVICE 32
MEMORY 34
OPERATING SYSTEM 36
GEOLOCATION ESTIMATOR 38

FIG. 9