



(51) International Patent Classification:

G06T 7/33 (2017.01) G06K 9/46 (2006.01)
G06K 9/62 (2006.01) G06F 16/532 (2019.01)

(21) International Application Number:

PCT/SG2020/050158

(22) International Filing Date:

24 March 2020 (24.03.2020)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

10201902701U 26 March 2019 (26.03.2019) SG
10201902800S 28 March 2019 (28.03.2019) SG

(71) Applicant: **AGENCY FOR SCIENCE, TECHNOLOGY AND RESEARCH** [SG/SG]; 1 Fusionopolis Way, #20-10, Connexis North Tower, Singapore 138632 (SG).

(72) Inventors: **YUAN, Miaolong**; c/o PKM, Institute for Infocomm Research, 1 Fusionopolis Way, #21-01, Connexis (South Tower), Singapore 138632 (SG). **LI, Zhengguo**; c/o PKM, Institute for Infocomm Research, 1 Fusionopolis Way, #21-01, Connexis (South Tower), Singapore 138632 (SG). **GAO, Wenchao**; c/o PKM, Institute for Infocomm Research, 1 Fusionopolis Way, #21-01, Connexis (South

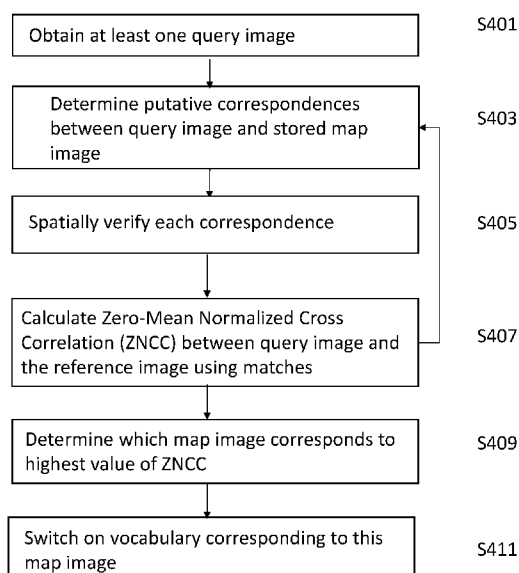
Tower), Singapore 138632 (SG). **WAN, Kong Wah**; c/o PKM, Institute for Infocomm Research, 1 Fusionopolis Way, #21-01, Connexis (South Tower), Singapore 138632 (SG). **YAU, Wei Yun**; c/o PKM, Institute for Infocomm Research, 1 Fusionopolis Way, #21-01, Connexis (South Tower), Singapore 138632 (SG).

(74) Agent: **POH, Chee Kian Daniel**; Marks & Clerk Singapore LLP, Tanjong Pagar Post Office, P O Box 636, Singapore 910816 (SG).

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

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ,

(54) Title: PLACE RECOGNITION



(57) Abstract: A method of determining which of a plurality of reference images has lighting conditions which most closely matches those of a query image, the method comprising, for each reference image: determining a set of matches between the reference image and the query image, wherein a match comprises a first feature in the query image and a second feature in the reference image, wherein the first and second features are both projections of the same point in three-dimensional space; and calculating a Zero-Normalized Cross Correlation for the determined set of matches, wherein the reference image corresponding to the set of matches having the highest value of Zero-Normalized Cross Correlation is determined to be the reference image with lighting conditions which most closely match those of the query image.

Figure 5

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, KM, ML, MR, NE, SN, TD, TG).

Published:

- *with international search report (Art. 21(3))*
- *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

PLACE RECOGNITION**Cross Reference to Related Application(s)**

The present disclosure claims the benefit of Singapore Patent Application No.

- 5 10201902701U filed on 26 Mar 2019 and Singapore Patent Application No.
10201902800S filed on 28 March 2019, are both incorporated in their entirety by
reference herein.

Technical Field

10

The present disclosure relates to visual place recognition, particularly visual place
recognition using feature matching between images for robotic navigation.

Background

- 15 Visual place recognition plays a key role in the navigation of mobile robots. It is a
challenging task, as the appearance of the environments can change drastically with
each visit due to moving objects in dynamic environments, changing weather
conditions and seasons, lighting conditions, and so on.
- 20 Appearance-based place recognition has obtained great attention in robotics over the
past decade. As one of the most successful appearance-based simultaneous

localization and mapping (SLAM) methods, Fast Appearance Based-Mapping (FAB-MAP) provides an efficient data-driven approach for mapping and loop closure detection. FAB-MAP represents images with a Bag-of-Words (BoW) obtained by quantizing the local shapes obtained using Scale-Invariant Feature Transform (SIFT) features or Speeded Up Robust Features (SURF).

The bag-of-words (BoW) technique is widely used to generate putative correspondences between images. In this technique, a vocabulary of visual words is first generated, and each image is then expressed as a histogram of frequency counts over these visual words. This enables to reduce the computation expensive of dealing with vast scenes containing thousands of features in a single image.

FAB-MAP has successfully demonstrated loop closure detection ability in trajectories of 70km and 1000 km. Similarly, Elena et al. presented a unified probabilistic framework for defining, modelling and recognizing places based on a location model in which a covisibility map is incrementally maintained over time. The bottleneck for loop closure detection is usually the extraction of the SURF or SIFT features, which is computationally expensive.

To overcome the computational burden of the SURF or SIFT detectors, more recently, computationally efficient binary features, such as Oriented FAST and Rotated BRIEF

(ORB) features, have been employed for feature matching with BoW techniques. ORB represents an image by a set P of local features. Details of ORB are given in E. Rublee, V. Rabaud, K. Konolige and G. Bradski, *ORB: An efficient alternative to SIFT or SURF*, IEEE International Conference on Computer Vision, 2564-2571, 2011.

5

Generally, in feature matching, features in the image of a location to be recognised (a “query” image) are compared with the features extracted from a reference image using similarity measures, such as Hamming distance, to determine how many matches are present. Using the results of this comparison, the current place can be recognised, and thus approximate pose of the query image can be retrieved.

10

So-called “visual maps” of particular areas may be constructed consisting of several reference images of different locations within the particular area. Feature matching may then be employed to determine which of the reference images most closely matches the query image and therefore the approximate location of the query image within the area.

15

Bag-of-Words techniques, such as those described above usually ignore the spatial relationships among the visual features during word assignments, and therefore lead to a large percentage of outliers. Further, the robustness of the BoW technique with

20

respect to lighting conditions heavily depends on the vocabulary which is usually built up off-line.

Spatial verification is typically employed to filter out any unreliable matches obtained
5 from BoW which are not geometrically consistent. Methods such as RANSAC, or Hough voting based methods are known in the art. However, such methods have disadvantages such as being computationally expensive (RANSAC) or giving rise to large numbers of outliers (Hough voting).

10 In order to address or alleviate at least one of the aforementioned problems and/or disadvantages, there is a need to provide an improved method for feature matching between images.

Summary

15 In a first aspect, a method of determining which of a plurality of reference images has lighting conditions which most closely matches those of a query image is provided, the method comprising, for each reference image: determining a set of matches between the reference image and the query image, wherein a match comprises a first feature appearing in the query image and a second feature appearing in the reference image,
20 wherein the first and second features are both projections of the same point in three-dimensional space; and calculating a Zero-Mean Normalized Cross Correlation for the

determined set of matches, wherein the reference image corresponding to the set of matches with the highest value of the Zero-Mean Normalized Cross Correlation is determined to be the reference image with lighting conditions which most closely match those of the query image.

5

Note that, in this disclosure, the term feature (as in “first feature” and “second feature”) refers to an image feature, or, equivalently, a feature point in an image. The terms feature, image feature and feature point are employed interchangeably herein.

10 In an embodiment, Zero-Normalized Cross Correlation for the set of matches is calculated only using pixels corresponding to determined matches between the reference image and the query image.

In an embodiment, the Zero-Normalized Cross Correlation for the set of matches is the
15 Zero-normalized Cross Correlation between the set of first features in the query image and the set of second features in the second image.

The term lighting conditions may refer to not only ambient lighting conditions at the time of capturing the image or images but also the exposure and other camera settings
20 employed in order to capture the images. The term “lighting conditions” when used in

conjunction with the reference images, therefore, is intended to refer broadly to any captured difference in the appearance of the same three-dimensional points in different reference images, whether or not that difference is a result of an actual difference in the ambient lighting conditions at the time that the images were
5 captured.

Some or all of the reference images may be stored in database. The query image may have been obtained in real time, or the query image may be an image stored in a database.

10

Determining a plurality of matches between the reference image and the query image may comprise: determining a plurality of match candidates between the reference image and the query image; and spatially verifying each match candidate. A match candidate is a first feature in the query image and a second feature in the reference
15 image, wherein the first and second features are both hypothesised to be projections of the same point in three-dimensional space. In an embodiment, a match is determined by spatially verifying a match candidate.

Spatially verifying a match candidate may comprise, for each of the plurality of match
20 candidates: determining an individual similarity transformation between a position of the feature of the match candidate in the first image and a position of the feature of the match candidate in the second image, and mapping the determined individual

- similarity transformations in a Hough space; partitioning the Hough space into a plurality of partitions; determining a plurality of groups, wherein a group is comprised of all of the hypothesized matches with individual similarity transforms that fall into the same partition; for each group: determining a local similarity transformation; and
- 5 verifying a match candidate by: calculating an error generated by describing the relative positions of the feature of the match candidate in the first image and the feature of the match candidate in the second image with one of the determined local similarity transformations, and determining that the error is below an error threshold.
- 10 An individual similarity transformation is the similarity transformation between the first feature of the match candidate in the first image and the second feature of the match candidate in the second image. The similarity transformation may be an affine transformation.
- 15 Advantageously, by calculating a plurality of local similarity transforms a high degree of accuracy is obtained as a large number of correct matches can be captured. Advantageously, the use of Hough space enables accurate feature matching to be achieved at high speed.
- 20 A match candidate refers to a pair of image features in the first and second images that are hypothesised to correspond to the same point in 3-D space. They are putative

correspondences between the first image and the second image. Verifying a match candidate refers to identifying whether the feature in the first image and the feature in the second do, in fact, correspond to the same point in 3-D space, i.e. that the hypothesis is correct.

5

An individual similarity transformation is a similarity transformation between the feature of a match candidate in a first image and the feature of a match candidate in a second image. An individual similarity transformation may comprise an affine transformation between the location of the feature in the first image and the location of the feature in the second image.

10

Mapping the individual similarity transformations in a Hough space, or equivalently, voting the correspondences into Hough space may comprise mapping the correspondences into the parameter space of the similarity transformation.

15 Partitioning the Hough space into a plurality of partitions may comprise dividing up the parameter space of the match similarity transformations, or equivalently, quantizing the parameter space of the match similarity transformations. Each dimension of the parameter space may correspond to a different parameter of the match similarity transformation. Every dimension of the parameter space may be divided up or only
20 some of the dimensions of the parameter space may be divided up. The dimensions may be divided up equally or differently. The partitioning of the space may be independent for each parameter.

Determining the plurality of groups of correspondences in the Hough space may comprise identifying clusters in the Hough space. A single local similarity transformation may be determined for each group or cluster of correspondences.

5

Verifying a match candidate may comprise ensuring that one of the local similarity transformations adequately represents the relationship between the point in the first image and the point in the second image. Adequately representing may mean that the match similarity transformation for that match candidate is an inlier of the local similarity transformation. Adequately representing may mean that if the local similarity transformation is applied to the position of the feature in the first, the corresponding position of the feature in the second image is obtained within a margin of error and/or vice-versa.

10

15

Determining the groups in the Hough space may or may not comprise calculating a score for each of the partitions. Determining the groups in the Hough space may or may not comprise selecting from those partitions with a score which exceeds a threshold. The score may include contributions from different levels of partitioning of the Hough space. The score may only consider the occupancy of each partition at the finest level of partitioning. The score may be employed in combination with the occupancy of each partition at the finest level of partitioning.

20

The local similarity transformation may or may not comprise an average of all of the match similarity transformations in a group. The local similarity transformation may or may not comprise the mean of all of the match similarity transformations in a group.

- 5 Calculating an error generated by describing the relative positions of the point of the match candidate in the first image and the point of the match candidate in the second image with one of the determined local similarity transformations may or may not comprise calculating two-way projection errors. Calculating an error generated by describing the relative positions of the point of the match candidate in the first image
10 and the point of the match candidate in the second image with one of the determined local similarity transformations may or may not comprise applying a Random sample consensus (RANSAC) algorithm and determining that the match similarity transform of the match candidate is an inlier of one of the determined local similarity transformations.

15

- The Hough space may four dimensional. The Hough space may be three dimensional. The Hough space may be found dimensional but one dimension may only comprise two partitions. The Hough space may be partitioned into dimensions corresponding to translation in a first, or x-direction, translation in a second, or y-direction, scale and
20 orientation. The dimension corresponding to orientation may be partitioned into two partitions or may not be partitioned at all. The Hough space may be partitioned according to the parameterization of the match similarity transformations.

In an aspect, a method of selecting a visual vocabulary for use in visual place recognition is provided, the method comprising: obtaining a first query image taken at a first exposure, wherein the first query image comprises an image obtained under the

5 lighting conditions under which visual place recognition will be performed;

determining which of a plurality of reference images has lighting conditions which most closely matches that of the first query image by determining a plurality of matches between the reference image and the query image, wherein a match comprises a first feature appearing in the query image and a second feature appearing

10 in the reference image, wherein the first and second features are both projections of the same point in three-dimensional space; and calculating a Zero-Mean Normalized Cross Correlation for the determined matches between the reference image and the query image, wherein the reference image with the highest value of the Zero-Mean Normalized Cross Correlation for the determined matches between the reference

15 image and the query image is determined to be the reference image with lighting conditions which most closely match those of the query image; and selecting a vocabulary corresponding to the reference image determined to have lighting conditions which most closely match those of the query image.

20 The visual vocabulary may comprise a series of feature images. The visual vocabulary may be suitable for use with a Bag-of-words method.

The exposure at which the query image is taken may or may not be known.

In an aspect, a method of determining which of a plurality of query images has an
5 exposure which most closely matches that of a reference image is provided, the
method comprising, for each reference image: determining a plurality of matches
between the reference image and the query image, wherein a match comprises a first
feature appearing in the query image and a second feature appearing in the reference
image, wherein the first and second features are both projections of the same point in
10 three-dimensional space; and calculating a Zero-Mean Normalized Cross Correlation
for the determined matches between the reference image and the query image,
wherein the reference image with the highest value of the Zero-Mean Normalized
Cross Correlation for the determined matches between the reference image and the
query image is determined to be the reference image with an exposure which most
15 closely matches that of the reference image.

In an aspect, a method of selecting an exposure and a visual vocabulary for use in
visual place recognition, the method comprising; obtaining a plurality of query images,
each query image being taken at a different exposure; for each query image,
20 determining a Zero-Mean Normalized Cross Correlation (ZNCC) between each of a
plurality of reference images and the query image, wherein determining a Zero-Mean

Normalized Cross Correlation between the query image and a reference image comprises: determining a plurality of matches between the reference image and the query image, wherein a match comprises a first feature appearing in the query image and a second feature appearing in the reference image, wherein the first and second
5 features are both projections of the same point in three-dimensional space; and calculating a Zero-Mean Normalized Cross Correlation for the determined matches between the reference image and the query image, and determining that the similarity value between a first query image and a first reference image gives rise to the largest Zero-Mean Normalized Cross Correlation; and selecting the exposure of the first query
10 image and the visual vocabulary corresponding to the first reference image.

In an aspect, a method of verifying a plurality of match candidates between a first image and a second image, wherein a match comprises a first feature appearing in the first image and a second feature appearing in the second image, wherein the first and
15 second features are both projections of the same point in three-dimensional space, the method comprising: for each of the plurality of match candidates, determining a match similarity transformation between a position of the first feature in the first image and a position of the second feature in the second image; grouping at least some of the plurality of match candidates into a plurality of groups according to their respective
20 match similarity transformations; for each group, determining a local similarity transformation; and verifying a match candidate by: calculating an error generated by describing the match candidate with one of the determined local similarity

transformations; and determining that the error is below an error threshold, wherein the first image and the second image are represented by a set of features using a visual vocabulary, wherein the first feature and the second feature of a match candidate are hypothesized to be the same by virtue of their corresponding to the same visual word in the visual vocabulary, and wherein the visual vocabulary is chosen by obtaining a first query image taken at a first exposure, wherein the first query image comprises an image obtained under the lighting conditions under which visual place recognition will be performed; determining which of a plurality of reference images has lighting conditions which most closely matches that of the first query image by determining a plurality of matches between the reference image and the query image, wherein a match comprises a first feature appearing in the query image and a second feature appearing in the reference image, wherein the first and second features are both projections of the same point in three-dimensional space; and calculating a Zero-Mean Normalized Cross Correlation for the determined matches between the reference image and the query image, wherein the reference image with the highest value of the Zero-Mean Normalized Cross Correlation for the determined matches between the reference image and the query image is determined to be the reference image with lighting conditions which most closely match those of the query image; and selecting a vocabulary corresponding to the reference image determined to have lighting conditions which most closely match those of the query image.

The first image and the second image may be represented by a histogram of a set of features using a visual vocabulary. The first feature and the second feature are hypothesized to be the same, and therefore to be a match candidate, by virtue of their being represented by the same feature of the visual vocabulary, or equivalently the same visual word.

In an aspect, a system for determining lighting conditions is provided, the system comprising: an input for receiving a query image representative of a lighting condition to be determined; a memory storing a plurality of reference images, the reference images being representative of different lighting conditions; and a processor configured to perform a method of determining which of a plurality of reference images has lighting conditions which most closely matches those of a query image is provided, the method comprising, for each reference image: determining a plurality of matches between the reference image and the query image, wherein a match comprises a first feature appearing in the query image and a second feature appearing in the reference image, wherein the first and second features are both projections of the same point in three-dimensional space; and calculating a Zero-Mean Normalized Cross Correlation for the determined matches between the reference image and the query image, wherein the reference image with the highest value of the Zero-Mean Normalized Cross Correlation for the determined matches between the reference image and the query image is determined to be the reference image with lighting conditions which most closely match those of the query image.

In an aspect, a system for performing visual place recognition is provided, the system comprising: an input for receiving a query image; a memory storing a plurality of reference images; and a processor configured to verify a plurality of match candidates

5 between a first image and a second image, wherein a match comprises a first feature appearing in the first image and a second feature appearing in the second image, wherein the first and second features are both projections of the same point in three-dimensional space, the method comprising: for each of the plurality of match candidates, determining a match similarity transformation between a position of the

10 first feature in the first image and a position of the second feature in the second image; grouping at least some of the plurality of match candidates into a plurality of groups according to their respective match similarity transformations; for each group, determining a local similarity transformation; and verifying a match candidate by: calculating an error generated by describing the match candidate with one of the

15 determined local similarity transformations; and determining that the error is below an error threshold, wherein the first image and the second image are represented by a visual vocabulary, such that each feature is represented by a word, and wherein the first feature and the second feature are hypothesized to be the same by virtue of their being represented by the same word, and wherein the visual vocabulary is chosen by:

20 obtaining a first query image taken at a first exposure, wherein the first query image comprises an image obtained under the lighting conditions under which visual place recognition will be performed; determining which of a plurality of reference images

has lighting conditions which most closely matches that of the first query image by determining a plurality of matches between the reference image and the query image, wherein a match comprises a first feature appearing in the query image and a second feature appearing in the reference image, wherein the first and second features are both projections of the same point in three-dimensional space; and calculating a Zero-Mean Normalized Cross Correlation for the determined matches between the reference image and the query image, wherein the reference image with the highest value of the Zero-Mean Normalized Cross Correlation for the determined matches between the reference image and the query image is determined to be the reference image with lighting conditions which most closely match those of the query image; and selecting a vocabulary corresponding to the reference image determined to have lighting conditions which most closely match those of the query image.

In an aspect, a mobile robotic device is provided, the device comprising: a camera; and a system for determining lighting conditions, the system comprising: an input for receiving a query image representative of a lighting condition to be determined; a memory storing a plurality of reference images, the reference images being representative of different lighting conditions; and a processor configured to perform a method of determining which of a plurality of reference images has lighting conditions which most closely matches those of a query image is provided, the method comprising, for each reference image: determining a plurality of matches between the reference image and the query image, wherein a match comprises a first feature

- appearing in the query image and a second feature appearing in the reference image, wherein the first and second features are both projections of the same point in three-dimensional space; and calculating a Zero-Mean Normalized Cross Correlation for the determined matches between the reference image and the query image, wherein the
- 5 reference image with the highest value of the Zero-Mean Normalized Cross Correlation for the determined matches between the reference image and the query image is determined to be the reference image with lighting conditions which most closely match those of the query image.
- 10 In an aspect, a computer readable medium configured to cause a processor to perform a method of determining which of a plurality of reference images has lighting conditions which most closely matches those of a query image is provided, the method comprising, for each reference image: determining a plurality of matches between the
- 15 reference image and the query image, wherein a match comprises a first feature appearing in the query image and a second feature appearing in the reference image, wherein the first and second features are both projections of the same point in three-dimensional space; and calculating a Zero-Mean Normalized Cross Correlation for the determined matches between the reference image and the query image, wherein the
- 20 reference image with the highest value of the Zero-Mean Normalized Cross Correlation for the determined matches between the reference image and the query image is determined to be the reference image with lighting conditions which most closely match those of the query image.

The computer readable medium may be tangible or non-tangible.

Brief Description of Figures

Methods and systems according to embodiments of the invention will now be

5 described with reference to the following Figures, in which:

Figure 1 shows a mobile robotic device comprising a system for determining lighting conditions and performing visual place recognition according to embodiments;

Figure 2 shows a flowchart of a method for generating a visual vocabulary according to an embodiment;

10 Figure 3 shows a flowchart of a method for determining lighting conditions according to an embodiment;

Figure 4 shows a flowchart for a method of feature matching according to an embodiment;

15 Figure 5 shows a flowchart of a method for determining lighting conditions according to an embodiment;

Figure 6 shows an example of combining exposure and vocabulary for use in visual place recognition;

Figure 7 shows an example of combining exposure and vocabulary for use in visual place recognition;

Figure 8 shows an example of combining exposure and vocabulary for use in visual place recognition;

Figure 9 shows a flowchart of a method of selecting exposure according to an embodiment; and

- 5 Figure 10 shows a flowchart of a method of selecting exposure according to an embodiment.

Description

- 10 For purposes of brevity and clarity, descriptions of embodiments of the present disclosure are directed to a system and method for place recognition, in accordance with the drawings. While aspects of the present disclosure will be described in conjunction with the embodiments provided herein, it will be understood that they are not intended to limit the present disclosure to these embodiments. On the contrary,
- 15 the present disclosure is intended to cover alternatives, modifications and equivalents to the embodiments described herein, which are included within the scope of the present disclosure as defined by the appended claims. Furthermore, in the following detailed description, specific details are set forth in order to provide a thorough understanding of the present disclosure. However, it will be recognized by an
- 20 individual having ordinary skill in the art, i.e. a skilled person, that the present disclosure may be practiced without specific details, and/or with multiple details

arising from combinations of aspects of particular embodiments. In a number of instances, well-known systems, methods, procedures, and components have not been described in detail so as to not unnecessarily obscure aspects of the embodiments of the present disclosure.

5

In embodiments of the present disclosure, depiction of a given element or consideration or use of a particular element number in a particular figure or a reference thereto in corresponding descriptive material can encompass the same, an equivalent, or an analogous element or element number identified in another figure or
10 descriptive material associated therewith.

References to “an embodiment / example”, “another embodiment / example”, “some embodiments / examples”, “some other embodiments / examples”, and so on, indicate that the embodiment(s) / example(s) so described may include a particular
15 feature, structure, characteristic, property, element, or limitation, but that not every embodiment / example necessarily includes that particular feature, structure, characteristic, property, element or limitation. Furthermore, repeated use of the phrase “in an embodiment / example” or “in another embodiment / example” does not necessarily refer to the same embodiment / example.

20

The terms “comprising”, “including”, “having”, and the like do not exclude the presence of other features / elements / steps than those listed in an embodiment.

Recitation of certain features / elements / steps in mutually different embodiments does not indicate that a combination of these features / elements / steps cannot be used in an embodiment.

5 As used herein, the terms “a” and “an” are defined as one or more than one. The use of “/” in a figure or associated text is understood to mean “and/or” unless otherwise indicated. The term “set” is defined as a non-empty finite organization of elements that mathematically exhibits a cardinality of at least one (e.g. a set as defined herein can correspond to a unit, singlet, or single-element set, or a multiple-element set), in
10 accordance with known mathematical definitions. The recitation of a particular numerical value or value range herein is understood to include or be a recitation of an approximate numerical value or value range. The recitation of a particular numerical value or value range herein is understood to include or be a recitation of an approximate numerical value or value range.

15

In this disclosure, the term match is employed to mean that a feature appearing in a first image and a feature appearing in the second image correspond to the same point in a 3D scene. For example, if a projection of a corner of a building appears in both the first image and the second image then the projections of that corner in the images are
20 a match for each other.

In this disclosure, the term “vocabulary” is employed to mean a set of images of feature descriptors extracted from historical images. In the bag-of-words technique, an image is represented as a histogram of these feature descriptors.

5 In this disclosure the term “pose” is employed to mean the 3-D position of a device.

As noted above, visual place recognition using feature matching comprises matching features between a query image and a reference image of a known location in order to determine the location of the query image. The reference image may form part of a
10 series of reference images forming a visual map. In an embodiment, feature matching is performed as a two-step process comprising employing a visual vocabulary to represent each image and determine putative correspondences between them followed by spatial verification of the putative correspondences.

15 Figure 1 shows a computer system 100 configured to perform feature matching between two images according to an embodiment. The system 100 may further be configured to perform feature matching in order to determining lighting conditions according to an embodiment. The computer system 100 comprises a processor 101, such as a CPU which performs the feature matching according to an embodiment. The
20 processor may be further configured to determine lighting conditions. The system may comprise an input 105 by which it may receive one or more of the images to be to be matched. The input 105 may comprise a camera, a hard drive, a disk reader, an

ethernet connection or any other means for receiving an image or data relating to an image. The input 105 may instead or further be configured to receive instructions which cause the processor 101 to perform image matching according to an embodiment. Although only one input 105 is shown in Figure 1, the person skilled in the art will appreciate that the system 100 may have multiple inputs, one or more of which may be configured to received data in the form of images and one or more of which may be configured to receive instructions for the processor 101. In an embodiment, the computer system 100 may also comprise a memory 103. According to embodiments, the memory 103 may be configured to store images. The memory 103 may be further configured to store geographical information about the locations captured in those images. The memory 103 may additionally, or alternatively be configured to store instructions configured to cause the processor to perform a method of feature matching or localization according to an embodiment and/or a visual vocabulary database. Such a database will be discussed below. The memory 103 may be configured to store a visual vocabulary. Visual vocabularies are discussed further below. The computer system 100 may also comprise an output 107 by which data relating to the feature matching or lighting condition determination performed by the processor 101 is output. Examples of output 107 may include a screen, an external hard drive, or a disk writer, or any other means suitable for outputting data. In an embodiment, the output 107 may include instructions for controlling the movement of a mobile device. In an embodiment, the system 100 forms part of a mobile device 10,

such as a mobile robotic device. In this embodiment, the output 107 may comprise the control module for the movement of the mobile device.

As described above, in many techniques of feature matching and visual place
5 recognition, the first step of feature matching between two images is to determine putative correspondences using a visual vocabulary technique such as the Bag-of-Words method. In Bag-of-words, a visual vocabulary comprising a series of feature images is used to describe each image. The images are then represented as histograms
10 of the features in the visual vocabulary. The robustness of the BoW technique with respect to lighting conditions heavily depends on the vocabulary which is usually built up off-line.

A flowchart showing a method of generating a visual vocabulary according to an embodiment is shown in Figure 2

15

In step S2201, an image or a series of images from which the vocabulary will be derived is obtained.

In step S2203, a feature detector such as an ORB detector is applied to image or series
20 of images in order to extract image features. Features consist of keypoints and associated descriptors.

Next, in step S2205, the feature descriptions are clustered in order to generate the visual vocabulary. In an embodiment, this is done using K-means clustering. The final visual vocabulary therefore comprises clusters of feature descriptions.

- 5 When a vocabulary constructed in this way is applied to another, new image, the new image is represented as a histogram of the occurrences of each feature in the vocabulary. The putative correspondences required as a first step in feature matching are derived from the resulting representation of each of the images, with each putative correspondence comprising one feature from the vocabulary which has been
- 10 identified in both images.

As features of an image will vary with lighting conditions, the vocabularies generated from images, or sets of images obtained under different lighting conditions will also differ according to the lighting conditions.

15

- One option for building up a vocabulary which can deal with different lighting conditions is to build up a general vocabulary using features from the images of predefined positions under a range of different lighting conditions. However, it is very challenging to determine the sizes of clusters so as to capture all the variation in the
- 20 descriptors due to different lighting conditions. If the clusters are too large, distinct features under the similar lighting conditions could be grouped together. On the other

hand, if the clusters are too small, the same feature can be classified into various clusters due to different lighting conditions.

In an embodiment, a group of small vocabularies is built up, each corresponding to a different lighting condition. In this embodiment, a suitable vocabulary is selected depending on the lighting conditions at the time in question. In this approach, each vocabulary is built up using features from the images of predefined positions under each typical lighting condition.

The small vocabulary approach is advantageous because, besides the accuracy of the representation, the memory size for storing the corresponding visual map is an important index. When the former approach is employed, all of the images under the different lighting conditions must be loaded in the memory so as to capture all the variation in the descriptions due to different lighting conditions. However, in the latter approach according to an embodiment, many fewer images are required to be loaded into the memory as only images corresponding to the current lighting condition are required.

In approaches according to the above described embodiment employing lighting-specific vocabularies, in order to select the vocabulary with the most suitable lighting conditions it is necessary to determine the lighting conditions of an image upon which feature matching will be performed. If the feature matching and place recognition is to

be done in real time, this will comprise selecting the vocabulary corresponding to the present lighting conditions.

In an embodiment, a plurality of vocabularies, each corresponding to a different lighting condition is stored in the memory 103 of the system 100. In an embodiment, each vocabulary is generated from a series of training images captured under a single lighting condition, as described above in relation to Figure 2. Under the same lighting conditions, a visual map is also built up with reference images which are also stored in the memory 103 of the system 100. In an embodiment, reference images may be selected from the training images. However, in other embodiments they may not.

As the vocabularies are generated from images captured under different lighting conditions, in order to ensure that the best vocabulary is employed (or “switched on”) for visual place recognition (as well as ensuring that the best set of reference images is employed as a visual map), therefore, it is necessary to determine which vocabulary was trained using images captured under lighting conditions that most closely correspond to those of the present time.

In an embodiment, one reference image from each of the visual maps is compared to a query image in order to determine the best match. A method of performing this comparison according to an embodiment is shown in Figure 3.

In Step **S301**, at least one query image is obtained. In an embodiment, in order to select the best vocabulary, a few positions are selected around the current position of the robot. One image (a few images) is (are) captured at each position.

- 5 In an embodiment, the robot orientation for capturing the one or more query images is kept as close as possible that of the corresponding reference images. This ensures that a high number of feature matches are obtained between the images. In an embodiment, the approximate robot orientation is determined using wheel odometry or Adaptive Monte Carlo Localization (AMCL) in order to ensure that it is as close as
- 10 possible to that of the reference images. This is discussed in detail further below.

- In step **S303**, a set of putative correspondences is determined between one of the reference images and the query image. In an embodiment, these are themselves determined using a bag of words method. In an embodiment, this is based on an
- 15 offline trained vocabulary stored in the memory 103.

- In an embodiment, a temporary vocabulary is required to code features of two image sequences. In one embodiment, a general vocabulary, comprising features that were obtained from images over a range of lighting conditions is employed. In another
- 20 embodiment, one of the stored light-condition-specific vocabularies is selected. In an embodiment, single-pass sequential clustering is employed to select a suitable vocabulary.

In this step, each of the local features is converted to a visual word u . In an embodiment, given two images P and Q , a pair of features $p \in P$ and $q \in Q$ are considered as a putative correspondence $c = (p, q)$ if they share the same visual word, i.e., $u(p) = u(q)$.

In step **S305**, each of the correspondences determined in step S303 is spatially verified using a spatial verification technique. In embodiments, spatial verification techniques including, but not limited to RANSAC and vote-and-verify may be employed. In general, these spatial verification techniques comprise determining a similarity transformation between the features in the one image and the features in the other image and verifying the correspondences using the similarity transformation.

In one embodiment, a spatial verification technique according to the flow chart of Figure 4 is employed. In this embodiment, local similarity transforms are employed in order to verify the matches. Techniques according to this embodiment are both computationally efficient and ensure a high level of accuracy and will now be described in detail.

In step **S203**, a similarity transform is determined for each correspondence $c = (p, q)$.

Given a local image feature $p \in P$, such as an ORB feature, we assume that its position (x_p, y_p) , scale s_p , orientation θ_p and its visual word $u(p)$ are given. The local shape and position of each feature can be described by a 3×3 matrix given by:

$$F(p) = \begin{bmatrix} M(p) & t(p) \\ 0^T & 1 \end{bmatrix}, \quad (1)$$

5 where $M(p) = s_p R(p)$ and $R(p) = \begin{bmatrix} \cos(\theta_p) & -\sin(\theta_p) \\ \sin(\theta_p) & \cos(\theta_p) \end{bmatrix}$ is a 2×2 rotation matrix and

$t(p) = (x_p, y_p)$ is the position in the x, y axis. Intuitively, $F(p)$ defines a similarity transformation from the current image framework to a normalized image patch, centered at the origin (0, 0) with scale $s = 1$ and orientation $\theta = 0$.

10 A similarity transformation (i.e., a relative transformation) from p to q is given by:

$$F(c) = F(p)F(q)^{-1} \begin{bmatrix} M(c) & t(c) \\ 0^T & 1 \end{bmatrix} \quad (2)$$

Where $M(c) = s(c)R(c)$ and $t(c) = t(p) - M(c)t(q)$ and $s(c) = s_q / s_p$,

$R(c) = R(q)R(p)^{-1}$ denote relative scaling and rotation from p to q, respectively.

15 In step **S205**, the similarity transforms, $F(c)$ are clustered. In an embodiment, this is done by mapping the similarity transform for each correspondence in a Hough space and determining which of the correspondences fall into the same bins (thereby forming a cluster) in the Hough space. This process according to an embodiment will now be explained in detail.

$F(c)$ can be written as a 4D transformation vector by

$$F(c) = (t_x(c), t_y(c), s(c), \theta(c)) \quad (3)$$

5 where $\theta(c) = \theta_q - \theta_p$ and $(t_x(c), t_y(c)) = t(c)$.

In one embodiment, each correspondence $c = (p, q)$ is mapped as a point in a 4-D Hough transformation space, with each dimension corresponding to one of the parameters of $F(c)$ shown in Equation (3). In order to enable efficient implementation, the parameter space for each of the four parameters (i.e. translation $(x(c), y(c))$, scale $s(c)$ and orientation $\theta(c)$) is independently quantized into n_x , n_y , n_s and n_θ bins, respectively. The transformation space L may also be partitioned at different resolution levels $l_0, l_1, \dots, l_{L-1}$, with quantization of the parameter space at each resolution level.

15

In order to cluster the correspondences, each of the four parameters is normalized for each correspondence as:

$$s(c) = (\log(s(c) + \log(s_{\max})) / 2 \log(s_{\max})) \quad (4)$$

$$x(c) = (x(c) + w * s(c)) / 2w * s(c) \quad (5)$$

$$20 \quad y(c) = (y(c) + h * s(c)) / 2h * s(c) \quad (6)$$

$$\theta(c) = (\theta(c) + 2\pi) / 4\pi \quad (7)$$

where w and h are width and height of the image, $s_{\max}$ is the maximal scale of the local feature shape. Once the normalized parameters are calculated, their values are then mapped (“voted”) into the Hough space.

- 5 In another embodiment, rotation is largely ignored in the Hough transformation space as will now be explained in detail.

A 3D point $[X \ Y \ Z]^T$ in the world frame, has positions $[X_p \ Y_p \ Z_p]^T$ and $[\tilde{X}_q \ \tilde{Y}_q \ \tilde{Z}_q]^T$ in two different local camera frames.

10

The relationship between the two positions is given as

$$\begin{bmatrix} X_p \\ Y_p \\ Z_p \end{bmatrix} = \begin{bmatrix} R_{11} & R_{12} & R_{13} \\ R_{21} & R_{22} & R_{23} \\ R_{31} & R_{32} & R_{33} \end{bmatrix} \begin{bmatrix} \tilde{X}_q \\ \tilde{Y}_q \\ \tilde{Z}_q \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix} \quad (8)$$

where the rotation matrix and translation vector depend on the movement of a robot

- 15 and they are independent of features.

Let (f_x, f_y) and (c_x, c_y) be the focal lengths and principle point of the first camera in pixels, $(\tilde{f}_x, \tilde{f}_y)$ and $(\tilde{c}_x, \tilde{c}_y)$ be the focal lengths and principle point of the second

camera in pixels. Let (x_p, y_p) and $(\tilde{x}_q, \tilde{y}_q)$ be the positions of the features in the images, respectively. From the pin-hole camera model it can be shown that

$$\begin{bmatrix} x_p \\ y_p \end{bmatrix} = \begin{bmatrix} f_x \frac{X_p}{Z_p} + c_x \\ f_y \frac{Y_p}{Z_p} + c_y \end{bmatrix} \quad (9)$$

$$\begin{bmatrix} \tilde{x}_q \\ \tilde{y}_q \end{bmatrix} = \begin{bmatrix} \tilde{f}_x \frac{\tilde{X}_q}{\tilde{Z}_q} + \tilde{c}_x \\ \tilde{f}_y \frac{\tilde{Y}_q}{\tilde{Z}_q} + \tilde{c}_y \end{bmatrix} \quad (10)$$

From Eq. (8), it can be shown that

$$X_p = R_{11}\tilde{X}_q + R_{12}\tilde{Y}_q + R_{13}\tilde{Z}_q + \tilde{c}_x \quad (11)$$

$$Y_p = R_{21}\tilde{X}_q + R_{22}\tilde{Y}_q + R_{23}\tilde{Z}_q + \tilde{c}_y \quad (12)$$

Subsequently, it follows that

$$\begin{bmatrix} x_p \\ y_p \end{bmatrix} = \hat{\sigma}_c \hat{R} \begin{bmatrix} \tilde{x}_q \\ \tilde{y}_q \end{bmatrix} + \hat{t}_c \quad (13)$$

where the constant $\hat{\sigma}_c$, matrix $\hat{R}$ and vector $\hat{t}_c$ are computed as

$$\hat{\sigma}_c = \frac{\tilde{Z}_q}{Z_p} \quad (14)$$

$$\hat{R} = \begin{bmatrix} \frac{f_x R_{11}}{\tilde{f}_x} & \frac{f_x R_{12}}{\tilde{f}_y} \\ \frac{f_y R_{21}}{\tilde{f}_x} & \frac{f_y R_{22}}{\tilde{f}_y} \end{bmatrix} \quad (15)$$

$$\hat{t}_c = \begin{bmatrix} c_x \\ c_y \end{bmatrix} + \hat{\sigma}_c \left(\begin{bmatrix} f_x R_{13} \\ f_y R_{23} \end{bmatrix} - \hat{R} \begin{bmatrix} \tilde{c}_x \\ \tilde{c}_y \end{bmatrix} \right) + \frac{1}{Z_p} \begin{bmatrix} f_x t_x \\ f_y t_y \end{bmatrix} \quad (16)$$

It is worth noting that the matrix $\hat{R}$ is not a rotational matrix. Let the singular value decomposition of $\hat{R}$ be denoted as

$$\hat{R} = U \begin{bmatrix} \gamma_1 & 0 \\ 0 & \gamma_2 \end{bmatrix} V \quad (17)$$

The scale factor $s(c)$, rational matrix $R(c)$, and translational vector $t(c)$ are then given as

$$s(c) = \frac{\gamma_1 + \gamma_2}{2} \hat{\sigma}_c \quad (18)$$

$$R(c) = UV \quad (19)$$

$$t(c) = \begin{bmatrix} c_x \\ c_y \end{bmatrix} + \hat{\sigma}_c \begin{bmatrix} f_x R_{13} \\ f_y R_{23} \end{bmatrix} - s(c) R(c) \begin{bmatrix} \tilde{c}_x \\ \tilde{c}_y \end{bmatrix} + \frac{1}{Z_p} \left(\begin{bmatrix} f_x t_x \\ f_y t_y \end{bmatrix} - \left(\hat{R} - \frac{\gamma_1 + \gamma_2}{2} R(c) \right) \begin{bmatrix} \tilde{f}_x \tilde{X}_q \\ \tilde{f}_y \tilde{Y}_q \end{bmatrix} \right) \quad (20)$$

From Eqn. (19) it can be seen that the value of $\theta(c)$ is independent of the features. This implies that one bin is enough for the rotational angles $\theta(c)$'s if their values are accurate. In other words, the quantization on $\theta(c)$ can be ignored from a theoretic point of view. However, in practice, noise will arise in the computed values of $\theta(c)$'s, thus, in an embodiment, two bins are adopted for rotation in order to reduce the effect of noise in the computed values of θ_q and θ_p .

In this embodiment therefore, quantization is only performed on translation $(t_x(c), t_y(c))$, and scale $s(c)$ for each similarity transformation of each correspondence $c = (p, q)$, independently. As in the embodiment described above, the three parameters are quantized, and normalized before voting. The transformation space is partitioned at L different resolution levels $l_0, l_1, \dots, l_{L-1}$. At each level, the three parameters are quantized into n_x , n_y , and n_s bins.

Once the quantization and the voting are complete according to either of the embodiments described above, the score of each bin in Hough space is then calculated. It is expected that the most groups of correspondences at the finest level, i.e., l_0 , will be the most geometrically consistent and thus considered as inliers. In an embodiment, only the score of each bin at the finest level ($l=0$) is taken into account.

The score of the bins at level l_0 are calculated using the following equation:

$$score(b_{0i}) = \sum_{l=0}^{L-1} 2^{-\alpha l} \|b_{li}\| \quad (21)$$

where b_{li} denotes the count of the i^{th} bin at level l and $2^{-\alpha l}$ denotes the contributions to the overall score of its corresponding bin at level l .

5

In an embodiment, $\alpha=1$.

In an embodiment, a cluster is identified where the count of a given bin in Hough space exceeds a threshold p .

10

In another embodiment, a given number of bins are identified with the highest scores. Clusters are then identified from those identified bins with a count higher than a threshold p

15 In step **S207**, a local similarity transform is determined for each identified cluster.

In an embodiment, the local similarity transform for an individual cluster is the mean of all of the similarity transforms in that cluster, i.e. the mean of all of the transforms that were voted into the same bin.

20 In Step **S209**, each local similarity transform determined in step S207, is employed to verify the putative correspondences determined in step S303.

In an embodiment, verification is done using a two-way re-projection error threshold δ_{ep} and determining which correspondences are inliers for each individual local similarity transform. In an embodiment, if a correspondence is found to be an inlier for one of the local similarity transforms, it is regarded as verified and the features in the two images are considered to be a match. Thus, minimizing the reprojection error in this way is used for estimating the error from point correspondences between two images. The reprojection error is a geometric error corresponding to the distance between a projected point and the corresponding point on the same image. It is used to quantify how closely an estimate of a feature recreates the point's true projection.

10

In an embodiment, the verification is performed using Random sample consensus (RANSAC), an iterative method to estimate parameters of a mathematical model from a set of observed data that contains outliers, when outliers are to be accorded no influence on the values of the estimates. Therefore, it also can be interpreted as an outlier detection method.

15

Thus, in the above described embodiment, Hough voting is employed to determine local similarity transforms, i.e. a plurality of transforms, each of which describes the geometric relationship of a different group of correspondences. Calculating a plurality of local similarity transforms as opposed to a single, global similarity transformation is advantageous because Hough voting is sensitive to mismatches due to uniform quantization or to feature detection errors.

20

Note that although the spatial verification technique described above in relation to Figure 4 is advantageous as it provides an accurate method of spatially verifying correspondences in a computationally efficient way, any other suitable method of spatially verifying the correspondences between images may be employed according to embodiments.

In an embodiment, Steps S303 and S305 are performed for every reference image under consideration.

10

In step **S307**, the reference image with the highest number of spatially verified correspondences is determined. In an embodiment, this reference image is taken to be the closest match for the lighting condition in the query image.

15 In step **S309**, the vocabulary corresponding to the reference image with the highest number of correspondences is “switched on”. In practice that means that for a given period of time, this vocabulary will be employed to determine putative correspondences in visual place recognition, such as the method of Figure 2.

20 Figure 5 shows a method of determining the best lighting conditions according to another embodiment. In this embodiment, steps **S401**, **S403** and **S405** proceed as

steps S301, S303 and S305 in the embodiment of Figure 3 above. The number of spatially verified correspondences is N.

In step **S407**, the Zero-Mean Normalized Cross-Correlation between the query image
5 and the reference image are determined according to an embodiment. In an
embodiment, this is done by calculating the zero-mean normalized cross correlation
between the two images using each of the verified matches. In practice this means
determining the cross correlation for the set of matches, i.e. the cross correlation
between the set of determined features in the query image and set of determined
10 features in the reference image.

This process will be described in detail below according to an embodiment.

For simplicity, let Z_1 be the on-line captured image and Z_2 be the corresponding off-
15 line captured reference image with $Z_1(p)$ the grayscale value of the pixel at point p in
the on-line image and $Z_2(q)$ the grayscale value of the pixel at point q in the off-line
image. As described above, spatially verification techniques generally comprise
determining a similarity transformation between the features in the one image and the
features in the other image and verifying the correspondences using the similarity
20 transformation.

Once the matching process is complete, therefore, for each matched correspondence $c=(p,q)$, we have a similarity transformation describing the relationship between the position of the matched feature in the reference image and the corresponding matched feature in the query image. A similarity transformation from one image to the other may generally be parameterized by scale, rotation and translation in x and y , i.e.

$$s_p, R(p) \text{ and } t(p).$$

Using these parameters, in an embodiment, the grayscale value of an intermediate pixel $\hat{Z}_1(q)$ is defined as

$$\hat{Z}_1(q) = Z_1 \left(\frac{R^T(p)(q - t(p))}{s_p} \right) \quad (22)$$

Clearly, $\hat{Z}_1(q)$ is aligned with the feature $Z_2(q)$, when $Z_2(q)$ is a match with feature $Z_1(p)$.

The Zero-Mean Normalized Cross Correlation (ZNCC) between the images Z_1 and Z_2 is then defined as

$$\Psi(Z_1, Z_2) = \frac{\sum_q (\hat{Z}_1(q) - \bar{\hat{Z}}_1)(Z_2(q) - \bar{Z}_2)}{\sqrt{\sum_q (\hat{Z}_1(q) - \bar{\hat{Z}}_1)^2} \sqrt{\sum_q (Z_2(q) - \bar{Z}_2)^2}} \quad (23)$$

where $\bar{\hat{Z}}_1$ and $\bar{Z}_2$ are computed as

$$\bar{\hat{Z}}_1 = \frac{1}{N} \sum_q \hat{Z}_1(q), \bar{Z}_2 = \frac{1}{N} \sum_q Z_2(q) \quad (24)$$

Thus, in this embodiment, only the matched points are considered in the calculation of the ZNCC.

5 It can be easily verified that

$$\Psi(Z_1, Z_2) = 1 \Leftrightarrow Z_1(p) = a(Z_2(q) - \bar{Z}_2) + \bar{Z}_1 \quad (25)$$

where $a > 0$ is a constant. It follows that

$$Z_1(p_1) - Z_1(p_2) = a(Z_2(q_1) - Z_2(q_2)) \quad (26)$$

10 Subsequently, it can be derived that

$$Z_1(p_1) \geq Z_1(p_2) \Leftrightarrow Z_2(q_1) \geq Z_2(q_2) \quad (27)$$

$$Z_1(p_1) \leq Z_1(p_2) \Leftrightarrow Z_2(q_1) \leq Z_2(q_2) \quad (28)$$

Therefore, the orders for two pairs of pixels $(Z_1(p_1), Z_1(p_2))$ and $(Z_2(q_1), Z_2(q_2))$ have
 15 more chance to be preserved if the value of $\Psi(Z_1, Z_2)$ is larger. Since an ORB feature is
 based on the order among different pixels in an image, the corresponding ORB
 features in the two images Z_1 and Z_2 have more chance to be the same if the value of
 $\Psi(Z_1, Z_2)$ is larger. Using the value of $\Psi(Z_1, Z_2)$, therefore, the best matched visual
 vocabulary can be selected.

20

In step **S409**, the reference image with the highest value of $\Psi(Z_1, Z_2)$ is determined.

In an embodiment, where more than one reference image has the same value of $\Psi(Z_1, Z_2)$, the values of $\hat{\bar{Z}}_1$ and $\bar{\bar{Z}}_2$ are further applied to detect the similarity between the images Z_1 and Z_2 as follows:

$$\Phi(Z_1, Z_2) = \left| 1 - \frac{\hat{\bar{Z}}_1}{\bar{\bar{Z}}_2} \right| \quad (29)$$

- 5 The vocabulary with the smallest $\Phi(Z_1, Z_2)$ is selected. Note that the performance of the proposed matching method is usually improved if the number of matched pairs is increased.

Note that while any method of correspondence matching can be employed in the methods of Figures 4 and 5 the use of local similarity transformations according to
10 embodiments (as described, for example in association with Figure 2 above) for feature matching is advantageous for use in the methods of Figures 4 and 5 because such methods of matching correspondences ensure a large number of matches is obtained with computational efficiency.

15

Feature matching depends not only on the environmental light conditions, but also on the capturing conditions such as aperture, exposure time, ISO, etc. In an embodiment, the camera exposure is fine-tuned such that the number of matches and/or the value of $\Psi(Z_1, Z_2)$ is further enlarged.

20

Figure 6 shows a schematic representation of a process of determining the best exposure as well as lighting conditions according to an embodiment.

In the embodiment of Fig. 6, a number (in this case 3) of differently exposed images
5 701 are captured. Further several vocabularies 703, corresponding to different light conditions LC1-LC5 are stored in the memory 103. In an embodiment the best match between vocabulary and exposure is determined by comparing all combinations of exposure and light condition, as represented by the lines in Figure 6. In other words, the final best matching of captured lighting conditions is a combined configuration of
10 one vocabulary and a particular exposure time setting. Two typical examples of possible scenarios for vocabulary and exposure matching are illustrated in Fig. 7 and Fig. 8.

In the examples of Figure 7 and 8, the lines joining the three exposures 701 and the
15 vocabularies 703 indicate the best match (i.e. the one giving rise to the highest number of feature matches, and/or highest value of ZNCC) between each exposure and a vocabulary

In the example of Figure 7, the best vocabulary match for all three exposures 701 is V2.
20 In the example of Figure 8, in contrast, a different vocabulary provides the best match for each of the three employed exposures.

In either example, the vocabulary and exposure which together provide the largest overall number of matches are selected for use in Visual Place Recognition, in accordance with an embodiment.

- 5 A flowchart for a method of determining the exposure according to an embodiment is shown in Figure 9.

In Step **S1001**, a plurality of query images is obtained. In an embodiment, at least one image is obtained at a plurality of exposures.

10

As in the embodiments discussed above in relation to Figures 3 and 4, in an embodiment, the robot orientation for capturing the query images is kept as close as possible to that of the corresponding reference images. In an embodiment, the approximate robot orientation is determined using wheel odometry or Adaptive

- 15 Monte Carlo Localization in order to ensure that it is as close as possible to that of the reference images. This is discussed in detail further below.

In step **S1003**, a set of putative correspondences is determined between one of the reference images and one of the query images. In an embodiment, these are

- 20 determined as described above in step S303 of Figure 3.

In step **S1005**, each of the correspondences determined in step S1003 is spatially verified using a spatial verification technique. In an embodiment, the spatial verification technique employed is one of determining local similarity transformations between correspondences as described above in relation to Figure 2.

5

In an embodiment, steps S1003 and S1005 are performed for every reference image in combination with every exposure, such that all of the combinations of possible vocabulary and exposure are explored (see Figure 6 for an example with 5 lighting conditions and 3 exposures).

10

In step **S1007**, the combination of reference image and query image giving rise to the highest number of spatially verified correspondences is determined. In an embodiment, this reference image is taken to be the closest match for the lighting condition in the query image.

15

In step **S1009**, the vocabulary corresponding to the above combination is “switched on”. In practice that means that for a given period of time, this vocabulary will be employed to determine putative correspondences in visual place recognition. Likewise the camera settings are switched to those of the exposure corresponding to the

20 combination determined in S1007.

Figure 10 shows a method of determining the best lighting conditions and exposure according to another embodiment. In this embodiment, steps **S1101**, **S1103** and **S1105** proceed as steps S1001, S1003 and S1005 in the embodiment of Figure 10 above.

- 5 In step **S1007**, the Zero-Mean Normalized Cross-Correlation between the query image and the reference image under consideration are determined according to an embodiment. This proceeds in exactly the way described above in relation to step S407 of Figure 4.
- 10 In step **S1009**, the reference image and exposure combination with the highest value of $\Psi(Z_1, Z_2)$ is determined.

- In an embodiment, where more than one reference image and exposure combination has the same value of $\Psi(Z_1, Z_2)$, if the ZNCCs of more than one vocabulary are the
- 15 same, the values of $\hat{\bar{Z}}_1$ and $\bar{\bar{Z}}_2$ are further applied to detect the similarity between the images Z_1 and Z_2 as follows:

$$\Phi(Z_1, Z_2) = \left| 1 - \frac{\hat{\bar{Z}}_1}{\bar{\bar{Z}}_2} \right| \quad (30)$$

The vocabulary and exposure combination with the smallest $\Phi(Z_1, Z_2)$ is selected.

In step **S1111**, the vocabulary corresponding to the above combination is “switched on”. Likewise, the camera settings are switched to those of the exposure corresponding to the combination determined in S1109.

- 5 Note that while the discussion of Figures 6 to 10 above assumes that stored vocabularies and reference images corresponding to different lighting conditions are available, there may only a single vocabulary stored in the memory 103. Nevertheless, the skilled person will appreciate that the method described above is equally applicable for determining the best exposure for visual place recognition when only a
- 10 single vocabulary is available.

The skilled person will, of course, appreciate that the reference images themselves will not only be affected by ambient lighting conditions but also the exposure and other camera settings employed in order to capture the images. The term “lighting

15 conditions” when used in conjunction with the reference images, therefore, is intended to refer broadly to any captured difference in the appearance of same features in different reference images, whether or not that difference is a result of an actual difference in the ambient lighting conditions at the time that the images were captured.

Likewise, the term “exposure” when used in conjunction with a query image is intended to refer broadly to any variable that will affect the appearance of features in the captured image.

5 Table 1 shows results obtained by capturing image sequences under fourteen different lighting conditions. In this example, a dense visual map was utilized with the objective of reducing possible effects from other components. Fifteen vocabularies were built up, fourteen of them (V1-V14) using images at only one lighting condition and one (ATV) built up via the images at all the lighting conditions. Fourteen images (1-14) were
10 randomly selected from each of the fourteen different light condition image sequences as the query images. Table 1 shows results obtained for the difference between the ground-truth pose and the detected pose via the VPR according to an embodiment. Clearly, the matching difference was smallest when the lighting conditions of the query image and the vocabulary were the same (shown in bold). Even though the matching
15 difference using the all time vocabulary (ATV) was also small, all of the images at all of the different lighting conditions were required for this vocabulary, therefore the memory requirement was large.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
V1	0.12	3.38	2.09	2.22	2.1	2.28	2.37	2.89	3.53	3.84	1.98	2.5	2.88	4.65
V2	4.03	0.14	3.03	3.21	3.32	3.2	5.55	5.44	7.01	7.24	2.85	2.55	2.77	2.65
V3	1.9	2.44	0.34	1.38	1.45	1.56	2.55	2.62	4.02	5	1.47	1.59	2.24	4.02
V4	1.87	2.75	1.46	0.32	1.47	1.76	2.01	2.56	4.32	4.69	1.53	1.63	2.42	4.74
V5	1.71	2.22	1.56	1.57	0.27	1.37	2.48	2.62	3.91	4.09	1.82	1.69	2.45	4.90
V6	1.56	1.85	1.39	1.69	1.52	0.37	2.25	2.82	3.73	4.18	1.28	1.42	1.89	4.49
V7	1.69	5.32	2.57	2.64	3.06	2.30	0.27	1.77	2.92	3.81	2.82	2.55	4.15	5.99
V8	1.87	4.27	2.29	2.38	2.50	2.38	1.28	0.25	2.59	3.49	2.34	2.09	2.65	5.26
V9	3.11	4.79	3.97	4.03	4.11	3.59	2.68	2.60	0.31	1.48	3.99	4.10	4.88	5.32
V10	3.29	4.83	4.39	3.90	4.55	3.89	2.84	3.05	1.39	0.26	4.32	4.40	4.56	5.32
V11	1.61	1.76	1.89	1.61	1.50	1.68	2.72	2.30	3.92	4.62	0.24	1.12	1.29	4.0
V12	2.15	2.62	1.78	1.95	2.37	1.95	3.03	2.53	3.88	4.61	1.59	0.23	1.56	3.53
V13	2.55	2.19	2.19	2.53	2.62	2.25	4.14	3.48	4.38	5.33	1.57	1.40	0.33	2.52
V14	5.47	2.18	4.61	4.95	5.20	4.95	5.92	5.90	6.16	6.38	3.95	3.49	2.54	0.35
ATV	0.32	0.58	0.64	0.57	0.51	0.54	0.57	0.74	0.66	0.80	0.54	0.45	0.76	1.11

Table 1: Difference between ground-truth pose and detected pose via the VPR for different combination of query images and vocabulary

5

A new framework of switched vocabulary has been described in accordance with embodiments above. According to embodiments, a simplified Hough quantization approach, an on-line zero-mean normalized cross correlation (ZNCC) based matching method to select the best vocabulary from a set of vocabularies built up off-line, a

10 differently exposed images-based approach to determine the most suitable exposures, and a structure from motion-based method to correct the pose after the visual place recognition is provided.

Although the above description of embodiments is directed to robotic vision for

15 navigation and localization, methods according to embodiments may also be employed in other applications, including but not limited to self-driving car navigation, 3D reconstruction, image stitching, etc .

In the foregoing detailed description, embodiments of the present disclosure are described with reference to the provided figures. The description of the various embodiments herein is not intended to call out or be limited only to specific or
5 particular representations of the present disclosure, but merely to illustrate non-limiting examples of the present disclosure. The present disclosure serves to address at least one of the mentioned problems and issues associated with the prior art. Although only some embodiments of the present disclosure are disclosed herein, it will be apparent to a person having ordinary skill in the art in view of this disclosure that a
10 variety of changes and/or modifications can be made to the disclosed embodiments without departing from the scope of the present disclosure. Therefore, the scope of the disclosure as well as the scope of the following claims is not limited to embodiments described herein.

Claims

1. A method of determining which of a plurality of reference images has lighting conditions which most closely match those of a query image, the method comprising, for each reference image:
 - 5 determining a set of matches between the reference image and the query image, wherein a match comprises a first feature in the query image and a second feature in the reference image, wherein the first and second features are both projections of the same point in three-dimensional space; and
 - 10 calculating a Zero-Normalized Cross Correlation for the determined set of matches,
 - wherein the reference image corresponding to the set of matches having the highest value of Zero- Normalized Cross Correlation is determined to be the reference image with lighting conditions which most closely match those of
 - 15 the query image.
2. The method of any of claims 1 or 2, wherein determining the set of matches between the reference image and the query image comprises:
 - determining a plurality of match candidates between the reference
 - 20 image and the query image; and
 - spatially verifying each match candidate.

3. The method of claim 1 or claim 2, wherein spatially verifying each match candidate comprises determining an individual similarity transform between a position of the first feature in the first image and a position of the second feature in the second image.

5

4. The method of either of claims 1 or 2, wherein spatially verifying a match candidate comprises, for each of the plurality of match candidates:

determining an individual similarity transformation between a position of the first feature of the match candidate in the first image and a

10

position of the second feature of the match candidate in the second image, and

mapping the determined individual similarity transformations into a Hough space;

partitioning the Hough space into a plurality of partitions;

15

determining a plurality of groups,

wherein a group is comprised of all of the match candidates with

individual similarity transforms that fall into the same partition;

for each group: determining a local similarity transformation; and

verifying a match candidate by:

20

calculating an error generated by describing the relative positions of the first feature of the match candidate in the first image and the second

feature of the match candidate in the second image with one of the determined local similarity transformations, and determining that the error is below an error threshold.

- 5 5. The method of claim 4, further comprising: determining a number of individual similarity transformations falling within a partition, and wherein a group is comprised of all the match candidates with individual similarity transforms that fall into the same partition, wherein the number of individual similarity transforms falling within the partition exceeds a
- 10 threshold.
6. The method of claim 3, further comprising:
- determining a score for each of the plurality of partitions, wherein the score of a partition depends on a number of individual similarity
- 15 transformations falling within a partition;
- determining a given number of partitions with the highest scores;
- determining a number of individual similarity transformations falling within each partition of the given number of partitions with the highest scores, and wherein
- 20 a group is comprised of all of the match candidates with individual similarity transforms that fall into the same partition of one of the given number of partitions with the highest scores, wherein the number of individual similarity transforms falling within the partition exceeds a threshold.

7. The method of any one of claims 4 to 6, wherein determining a local similarity transformation for a group comprises calculating a mean of all of the individual similarity transformations of the match candidates in that group.
8. The method of any one of claims 4 to 7, wherein calculating an error generated by describing the relative positions of the point of the match candidate in the first image and the point of the match candidate in the second image with one of the determined local similarity transformations comprises calculating a two-way projection error.
9. The method of any one of claims 4 to 7, wherein determining that the error is below an error threshold comprises applying a Random sample consensus (RANSAC) algorithm and determining that the individual similarity transform of the match candidate is an inlier of one of the determined local similarity transformations.
10. The method of any one of claims 4 to 9, wherein the Hough space is 4 dimensional, with each dimension corresponding to one of: translation in a first direction, translation in a second direction, scale, and orientation, and

wherein the dimension corresponding to orientation is partitioned into no more than two partitions.

11. The method of any one of claims 4 to 9, wherein the Hough space is 3
5 dimensional, with each dimension corresponding to one of: translation in a first direction, translation in a second direction and scale.
12. A method of selecting a visual vocabulary for use in visual place recognition,
10 the method comprising:
- obtaining a first query image taken at a first exposure, wherein the first query image comprises an image obtained under the lighting conditions under which visual place recognition will be performed;
- 15 determining which of a plurality of reference images has lighting conditions which most closely matches that of the first query image according to any of the preceding claims; and
- selecting a vocabulary corresponding to the reference image determined to have lighting conditions which most closely match those of the query image.
- 20
13. A method of determining which of a plurality of query images has an exposure which most closely matches that of a reference image,

the method comprising, for each query image:

determining a set of matches between the reference image and the query image, wherein a match comprises a first feature appearing in the query image and a second feature appearing in the reference image, wherein the first and second features are both projections of the same point in three-dimensional space; and calculating a Zero-Normalized Cross Correlation for the determined set of matches,

wherein the query image with the highest value of the Zero-Mean

Normalized Cross Correlation for the set of determined matches is determined to be the query image with an exposure which most closely matches that of the reference image.

14. A method of selecting an exposure for obtaining images for use with a visual vocabulary in visual place recognition, the method comprising

obtaining a plurality of query images, each query image being taken at a different exposure;

determining which of the plurality of query images has an exposure which most closely matches that of a reference image taken from a visual vocabulary according to the method of claim 13;

selecting the exposure corresponding to the query image determined to have an exposure which most closely matches that of the reference image.

- 5 15. A method of selecting an exposure and a visual vocabulary for use in visual place recognition, the method comprising;
- obtaining a plurality of query images, each query image being taken at a different exposure;
- for each query image, determining a Zero-Mean Normalized Cross
- 10 Correlation between each of a plurality of reference images and the query image, wherein determining a Zero-Mean Normalized Cross Correlation between the query image and a reference image comprises:
- determining a set of matches between the reference image and the query image, wherein a match comprises a first feature appearing in the
- 15 query image and a second feature appearing in the reference image, wherein the first and second features are both projections of the same point in three-dimensional space; and
- calculating a Zero-Mean Normalized Cross Correlation for the set of matches, and
- 20 determining which combination of a query image and reference image gives rise to the largest value of Zero-Mean Normalized Cross Correlation; and

selecting the exposure of the query image and the visual vocabulary of the reference image corresponding to that combination.

16. A method of verifying a plurality of match candidates between a first image
5 and a second image, wherein a match candidate comprises a first feature appearing in the first image and a second feature appearing in the second image, wherein the first and second features are hypothesised to be projections of the same point in three-dimensional space, the method comprising:
- 10 for each of the plurality of match candidates, determining an individual similarity transformation between a position of the first feature in the first image and a position of the second feature in the second image;
grouping at least some of the plurality of match candidates into a plurality of groups according to their respective individual similarity transformations;
15 for each group, determining a local similarity transformation; and
verifying a match candidate by: calculating an error generated by describing the match candidate with one of the determined local similarity transformations; and
determining that the error is below an error threshold, wherein the first
20 image and the second image are represented by a histogram of a set of features using a visual vocabulary, and wherein the first feature and the second feature of a match candidate are hypothesized to be the same by

virtue of their corresponding to the same feature of the visual vocabulary,
and wherein the visual vocabulary is chosen using the method of claim 12.

17. A system for determining lighting conditions, the system comprising:

5 An input for receiving a query image representative of a lighting condition to
be determined;

A memory storing a plurality of reference images, the reference images
being representative of different lighting conditions; and

A processor configured to perform the method of any of claims 1 to 11.

10

18. A system for performing visual place recognition, the system comprising:

an input for receiving a query image;

a memory storing a plurality of reference images; and

a processor configured to perform the method of claim 16.

15

19. A mobile robotic device comprising:

a camera; and

the system of claim 17.

20

20. A computer readable medium configured to cause a processor to perform
the method of any one of claims 1 to 11.

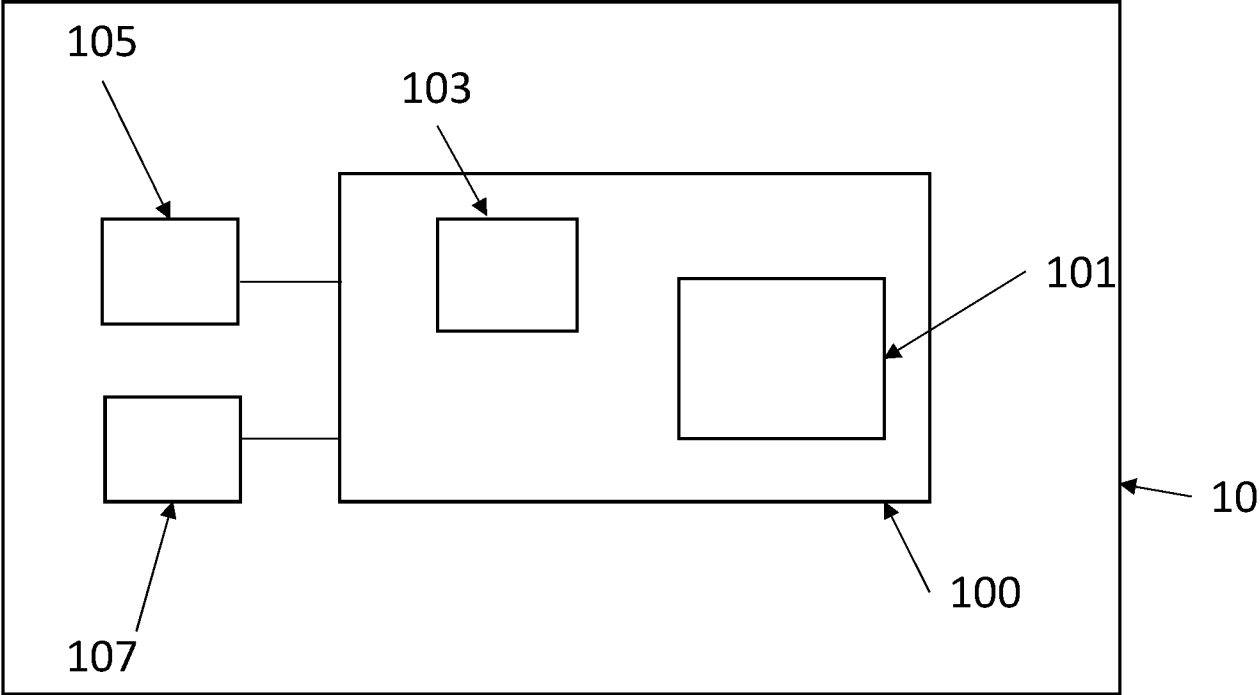


Figure 1

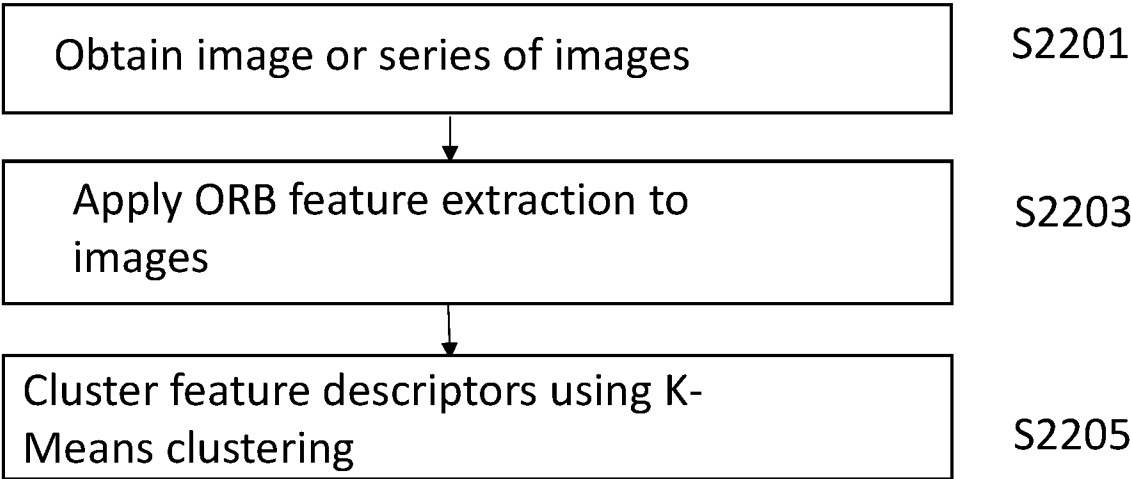


Figure 2

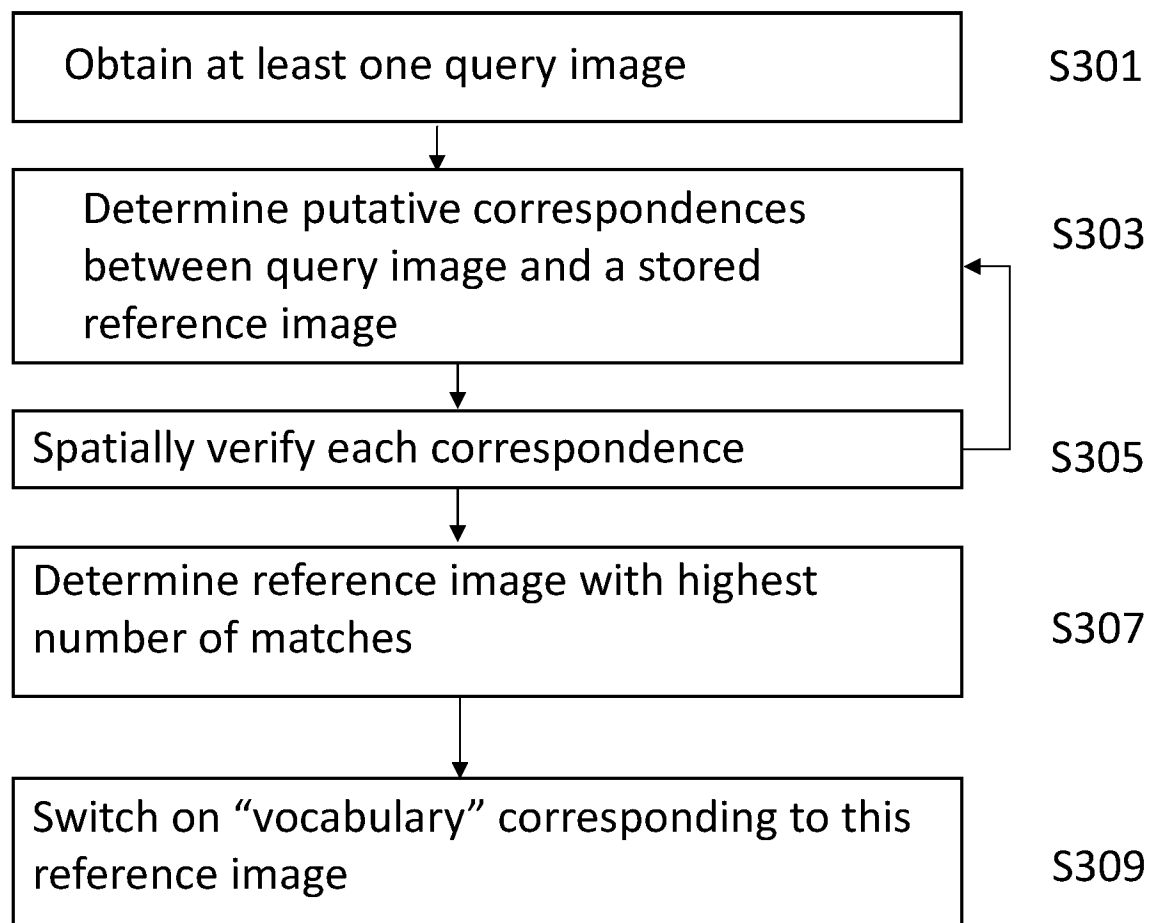


Figure 3

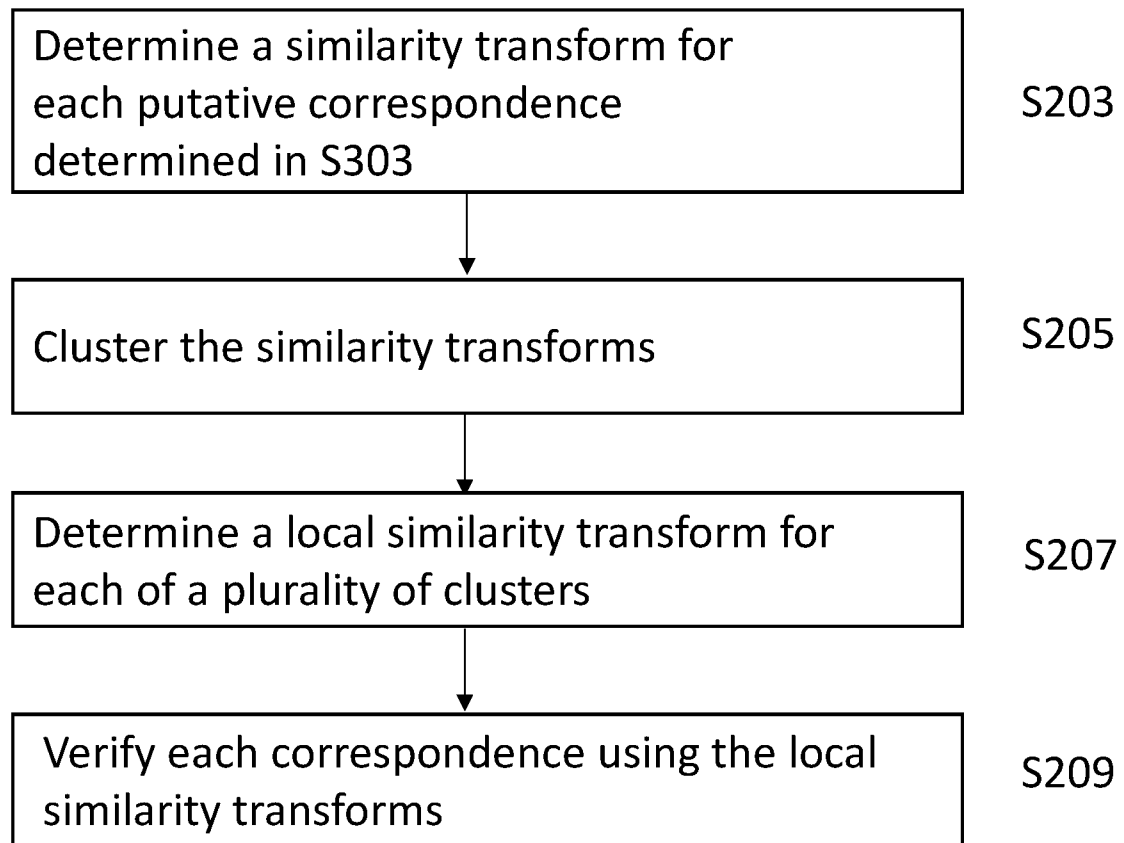


Figure 4

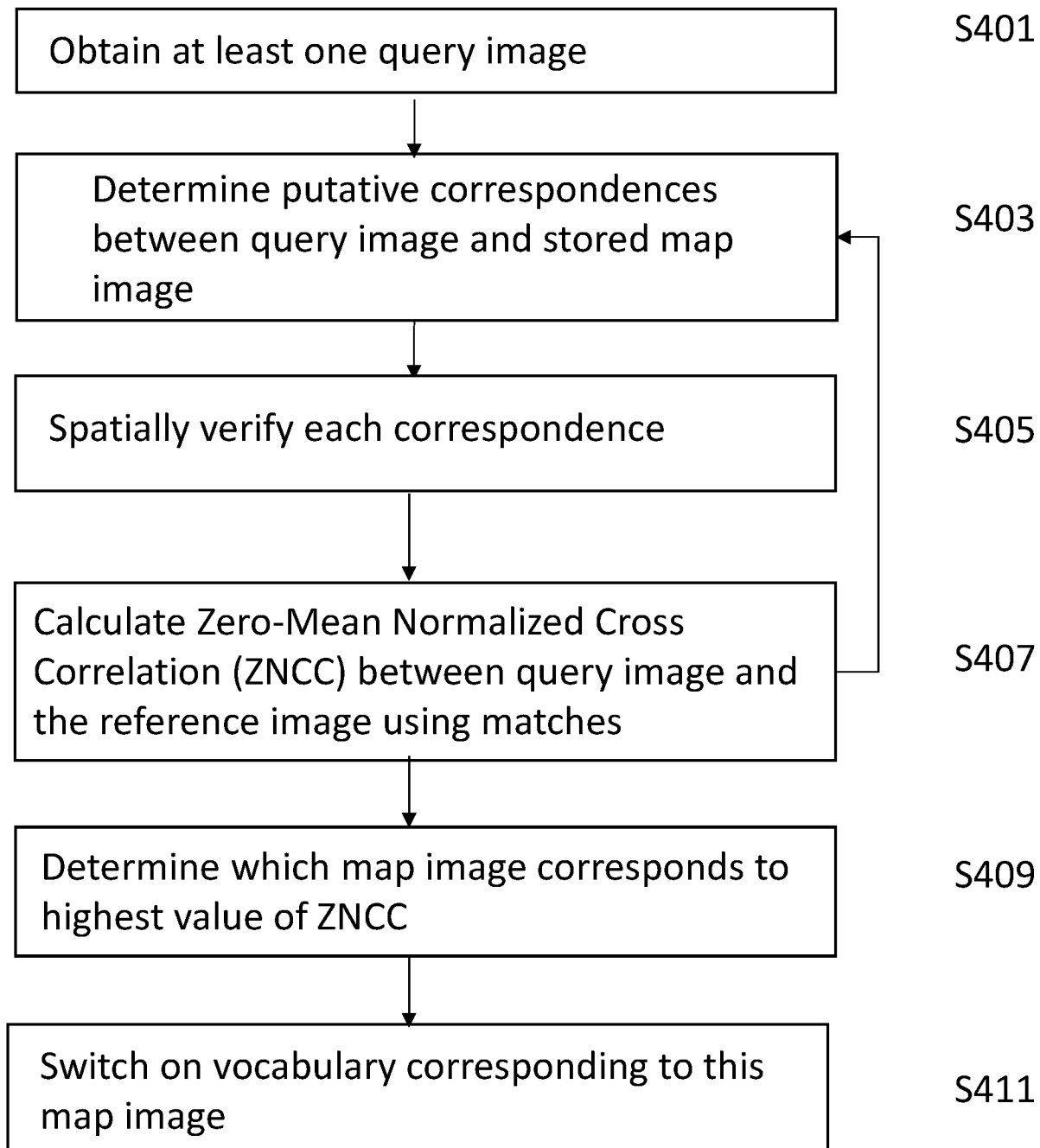


Figure 5

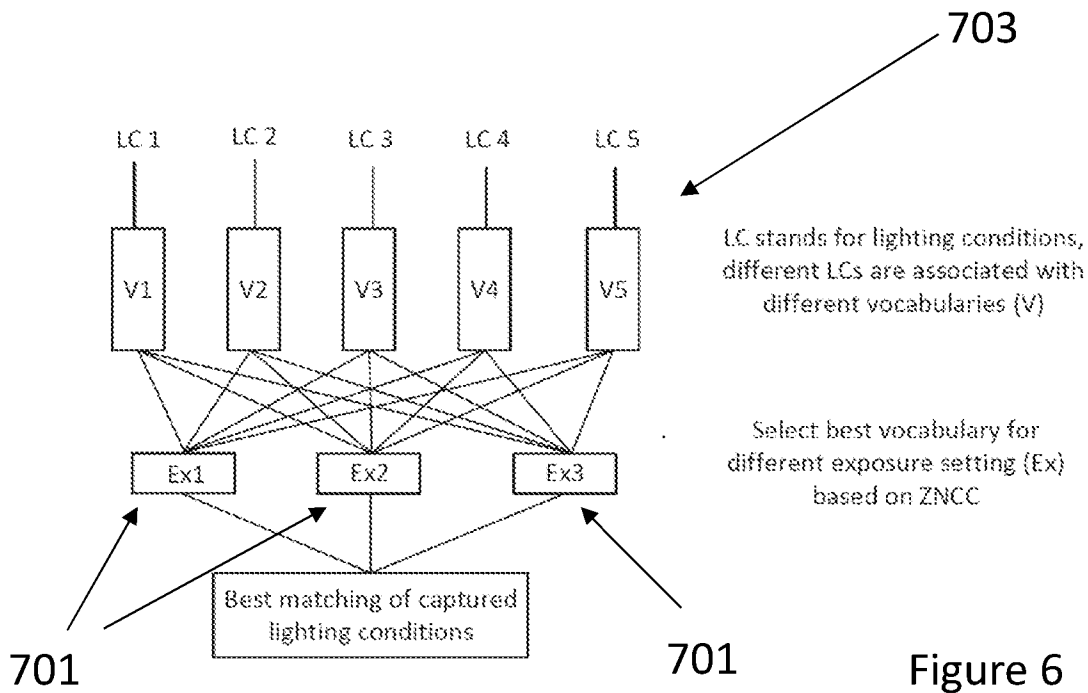


Figure 6

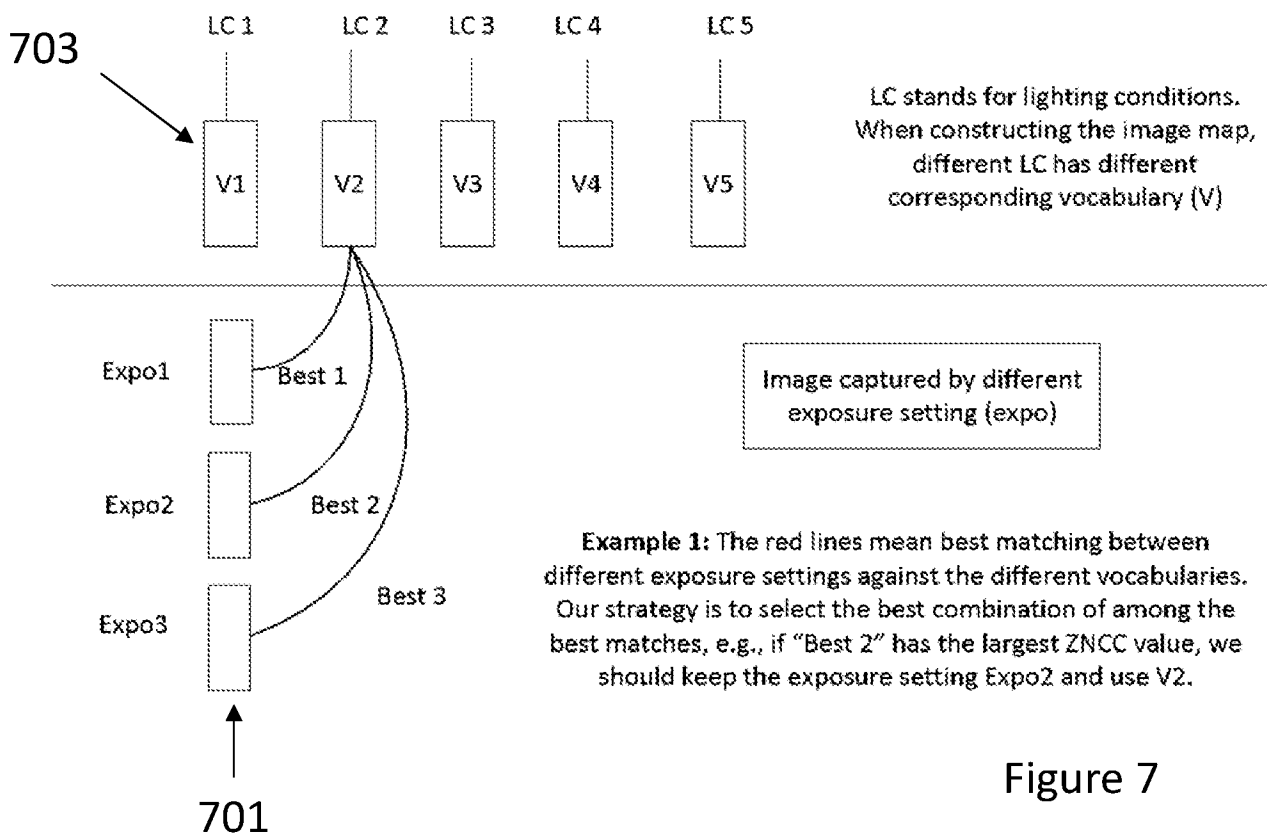


Figure 7

6/8

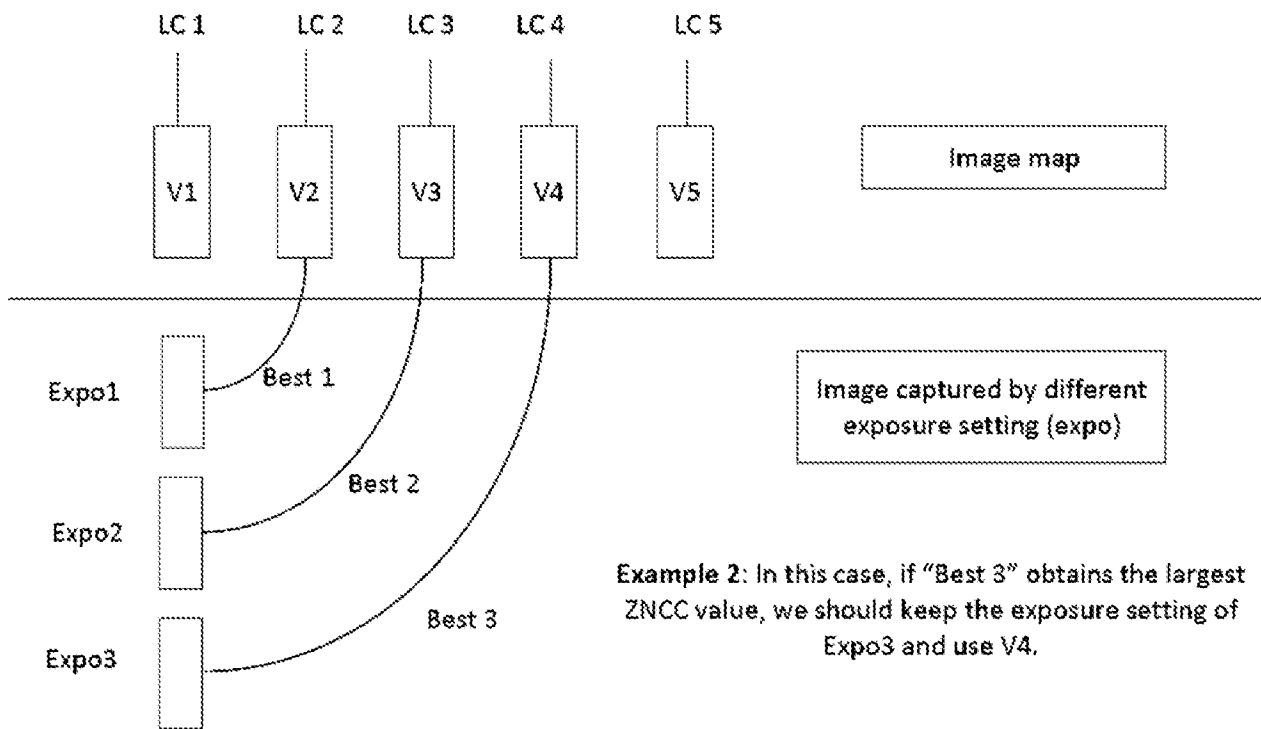


Figure 8

7/8

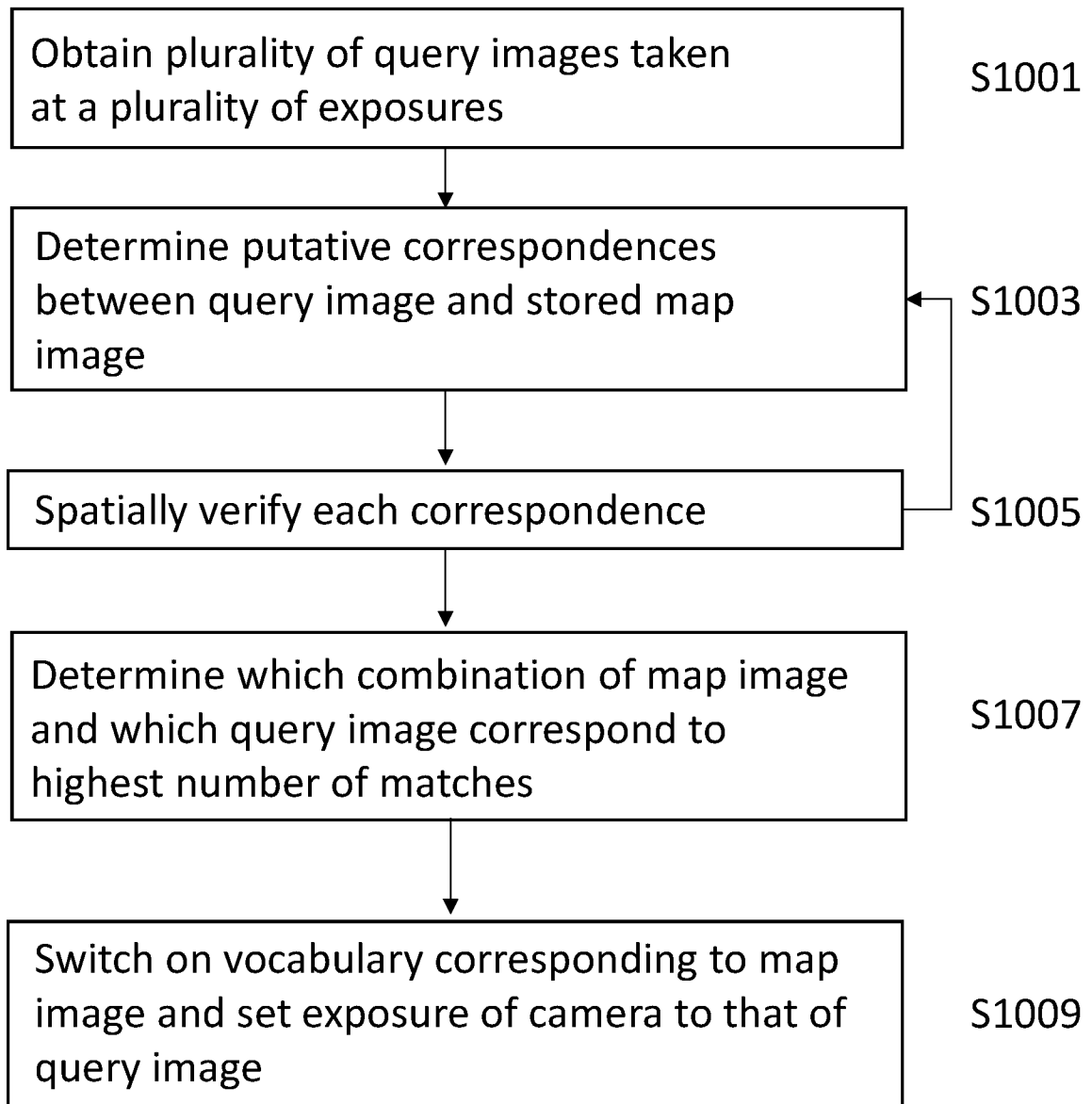


Figure 9

8/8

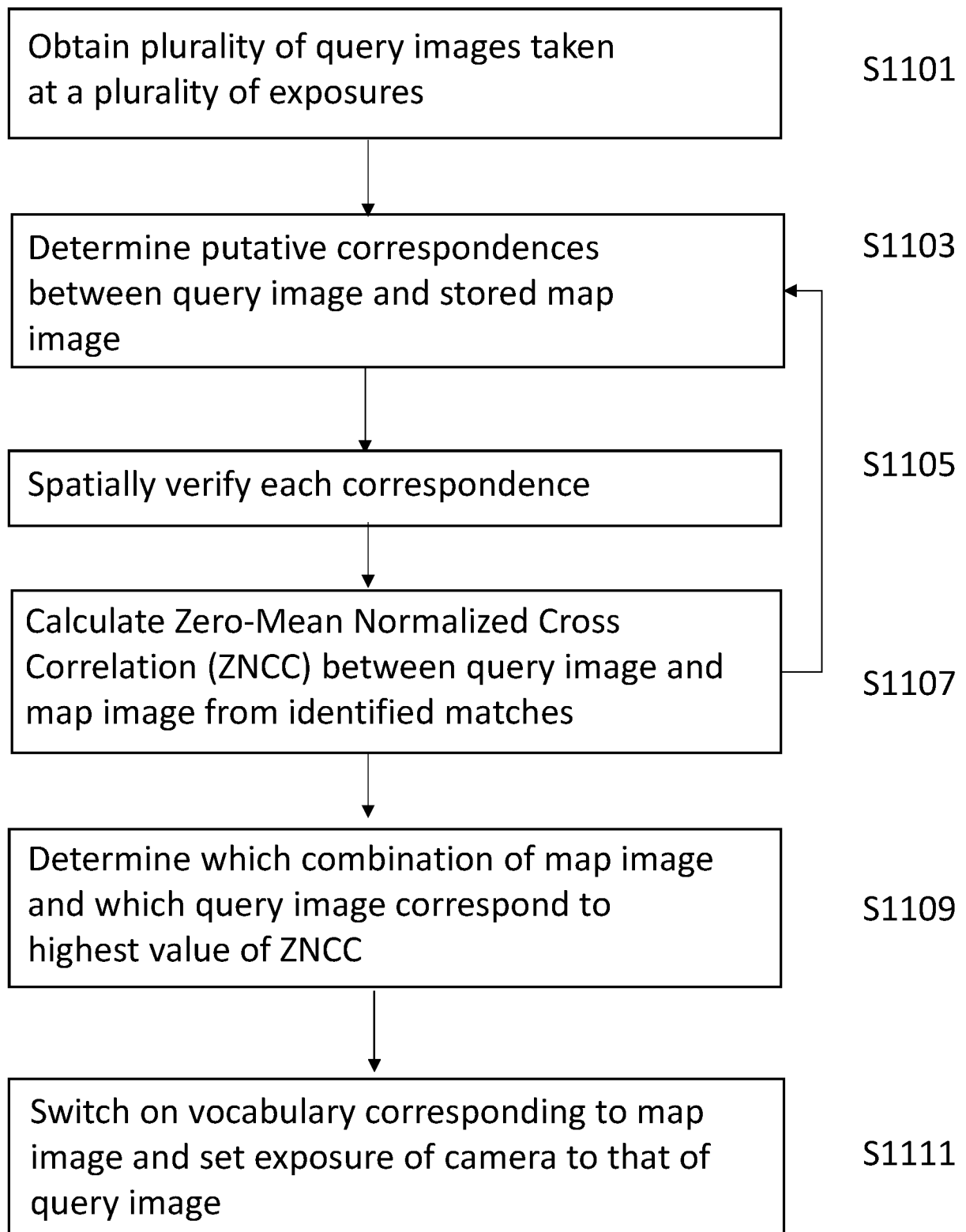


Figure 10

INTERNATIONAL SEARCH REPORT

International application No.

PCT/SG2020/050158

A. CLASSIFICATION OF SUBJECT MATTER

G06T 7/33 (2017.01) G06K 9/62 (2006.01) G06K 9/46 (2006.01) G06F 16/532 (2019.01)

According to International Patent Classification (IPC)

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06T; G06K

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

Database: FamPat

Search terms: feature, point, correspond, match, compare, recognize, map, identify, Zero-Mean Normalized Cross Correlation, ZNCC, visual, vocabulary, dictionary, word, 局部, 比较, 匹配, 对比, 映射, 相对, 对应, 零均值归一化, 类别, 名称, 种类, 标识, 物种 and other related terms

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X Y	US 2012/0201469 A1 (LIVET N. ET AL.) 9 August 2012 Paragraphs [0008], [0009], [0070]-[0077], [0116]; Figures 3a, 3b	1-11, 13, 17, 19, 20 12, 14-16, 18
Y	US 2018/0089832 A1 (LIU H. ET AL.) 29 March 2018 Paragraphs [0031], [0032] [0041]-[0045]; Figures 1, 2, 5, 6	12, 14-16, 18
A	YUAN M. ET AL., Outlier Detection using Hierarchical Spatial Verification for Visual Place Recognition. <i>2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV)</i> , 21 November 2018, pages 1-6 [Retrieved on 2020-08-24] <DOI: 10.1109/ICARCV.2018.8581070>	
A	CN 104794703 A (INST ADVANCED TECHNOLOGY UNIV SCIENCE & TECHNOLOGY CHINA) 22 July 2015 Whole document of the machine translation	

☒ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

*Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"D" document cited by the applicant in the international application

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

24/08/2020

(day/month/year)

Date of mailing of the international search report

26/08/2020

(day/month/year)

Name and mailing address of the ISA/SG

Intellectual Property Office of Singapore
1 Paya Lebar Link, #11-03
PLQ 1, Paya Lebar Quarter
Singapore 408533

Email: pct@ipos.gov.sg

Authorized officer

Nae Win Aung (Mr)

IPOS Customer Service Tel. No.: (+65) 6339 8616

INTERNATIONAL SEARCH REPORT

International application No.

PCT/SG2020/050158**C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT**

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 108986168 A (SHENZHEN GANDONG INTELLIGENT TECH CO LTD, UNIV PEKING SHENZHEN GRADUATE SCHOOL) 11 December 2018 Whole document of the machine translation	

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.

PCT/SG2020/050158

Note: This Annex lists known patent family members relating to the patent documents cited in this International Search Report. This Authority is in no way liable for these particulars which are merely given for the purpose of information.

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2012/0201469 A1	09/08/2012	FR 2951565 A1 JP 2013508844 A WO 2011/048302 A1 KR 20120087952 A CN 102640185 A HK 1171851 A1 EP 2491532 A1	22/04/2011 07/03/2013 28/04/2011 07/08/2012 15/08/2012 29/04/2016 29/08/2012
US 2018/0089832 A1	29/03/2018	WO 2018/063608 A1	05/04/2018
CN 104794703 A	22/07/2015	NONE	
CN 108986168 A	11/12/2018	NONE	