



(51) International Patent Classification:
G06K 9/00 (2006.01) **G06F 17/30** (2006.01)
G06K 9/46 (2006.01)

(21) International Application Number:
PCT/US2009/005832

(22) International Filing Date:
27 October 2009 (27.10.2009)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
12/264,277 4 November 2008 (04.11.2008) US

(71) Applicant (for all designated States except US): **EASTMAN KODAK COMPANY** [US/US]; 343 State Street, Rochester, New York 14650-2201 (US).

(72) Inventors; and

(75) Inventors/Applicants (for US only): **YU, Jie** [CN/US]; 343 State Street, Rochester, New York 14650 (US).

JOSHI, Dhiraj [IN/US]; 343 State Street, Rochester, New York 14650 (US). **LUO, Jiebo** [US/US]; 343 State Street, Rochester, New York 14650 (US). **HAO, Wei** [CN/US]; 343 State Street, Rochester, New York 14650 (US).

(74) Common Representative: **EASTMAN KODAK COMPANY**; 343 State Street, Rochester, New York 14650-2201 (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, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

[Continued on next page]

(54) Title: EVENT RECOGNITION USING IMAGE AND LOCATION INFORMATION

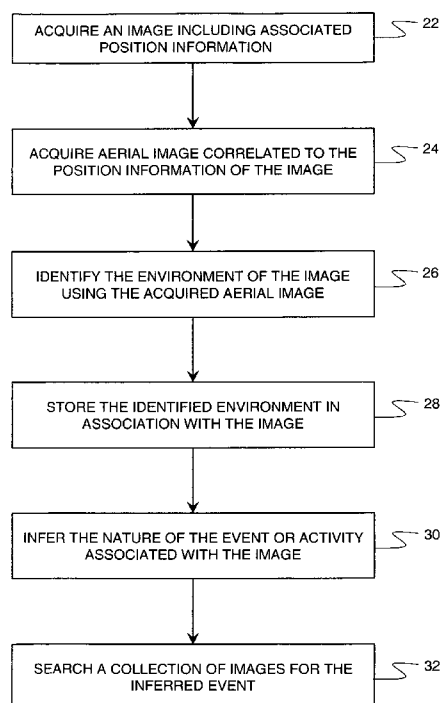


FIG. 2

(57) Abstract: A method of recognizing an event depicted in an image from the image and a location information associated with the image is disclosed. The method includes acquiring the image and its associated location information; using the location information to acquire an aerial image(s) correlated to the location information; identifying the event using both the image and the acquired aerial image(s); and storing the event in association with the image for subsequent use.



(84) **Designated States** (*unless otherwise indicated, for every kind of regional protection available*): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (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, SE, SI, SK, SM,

TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

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

EVENT RECOGNITION USING IMAGE AND LOCATION INFORMATION

FIELD OF THE INVENTION

5 The present invention relates to the recognition of the event depicted in an image by combining image content and location information associated with the location where the image was taken.

BACKGROUND OF THE INVENTION

10 Using a computer to analyze and discern the meaning of the content of digital media assets, known as semantic understanding, is an important field for enabling the creation of an enriched user experience with these digital assets. One type of semantic understanding in the digital imaging realm is the analysis that leads to identifying the type of event that the user has captured such as a birthday party, a baseball game, a concert, and many other types of events
15 where images are captured. Typically, events such as these are recognized using a probabilistic graphic model that is learned using a set of training images to permit the computation of the probability that a newly analyzed image is of a certain event type. An example of this type of model is found in the published article of L.-J. Li and L. Fei-Fei, *What, where and who? Classifying events by scene and*
20 *object recognition*, Proceedings of ICCV, 2007.

 An aerial image co-located with an image taken on the ground by personal cameras contains complementary information which can be used as additional semantic information for event recognition. The main advantage is that aerial images are free of distraction and clutter which is often an adverse factor for
25 computer vision algorithms. It is helpful to recognize the *environment* of an image (using aerial images) for event recognition. Incorporating inference from aerial images for event recognition requires fusion of information from the two modalities.

SUMMARY OF THE INVENTION

30 In accordance with the present invention, there is a method of recognizing an event depicted in the image and a location information associated with an image, comprising:

- a) acquiring the image and its associated location information;
b) using the location information to acquire an aerial image correlated to the location information;
c) identifying the event using both the image and the acquired aerial image(s); and
d) storing the event of the image in association with the image for subsequent use.

Features and advantages of the present invention include using an aerial image associated with the location information of an image to achieve a better understanding of the type of environment of an image to help predict the event or activity depicted in the image.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a pictorial of a system that can make use of the present invention;

FIG. 2 is a flow chart for practicing an embodiment of the invention;

FIG. 3 is a table showing the ontological structure of example environment types;

FIG. 4a and 4b depict examples of images and their corresponding aerial images of the environment;

FIG. 5a depicts a block diagram of an embodiment of the present invention for identifying the event of the image using both the image and the acquired aerial image;

FIG. 5b depicts a block diagram of another embodiment of the present invention for identifying the event of the image using both the image and the acquired aerial image; and

FIG. 5c depicts a block diagram of a still further embodiment of the present invention for identifying the event of the image using both the image and the acquired aerial image.

DETAILED DESCRIPTION

In FIG. 1, system 10 is shown with the elements necessary to practice the current invention including computing device 12, indexing server 14,

aerial image server 16, and communication network 20. Computing device 12 can be a personal computer for storing images where images will be understood to include both still and moving or video images. Computing device 12 communicates with a variety of devices such as digital cameras or cell phone cameras (not shown) for the purpose of storing images captured by these devices. These captured images can further include location information such as GPS data provided by the capturing device. Computing device 12 can also communicate through communication network 20 to an internet service that uses images captured without location information and allows the user to add location information to the images. In either case, images with location information are well known in the art.

Indexing server 14 is another computer processing device available on communication network 20 for the purposes of executing the algorithms in the form of computer instructions that analyze the content of images for semantic information such as event type. It will be understood that providing this functionality in system 10 as a web service via indexing server 14 is not a limitation of the invention. Computing device 12 can also be configured to execute the algorithms responsible for the analysis of images provided for indexing.

Aerial image server 16 communicates with other computing devices via communication network 20 and upon request, aerial image server 16 provides an aerial image correlated to the provided location information. Aerial images stored on aerial image server 16 are captured by a variety of manned aircraft including airplanes or unmanned aircraft including satellites. Aerial images can be captured from cameras that look directly down or orthogonally to a specific altitude or they can be captured obliquely to preserve third dimensional information of the captured image. Pictometry International Corp. is one company that provides obliquely captured aerial images correlated to location information. Well known internet services provide satellite aerial images correlated to provided location information such as physical address, GPS data, or other forms of location information such as the name of a business that has a known location. Aerial images can have metadata associated with them, which

can be stored in association with the aerial images on the aerial image server.

Metadata associated with the aerial images can include information such as names of businesses located within the physical area represented by the image, historical landmarks, parks, and much more. These metadata can be provided with an aerial
5 image upon request. Aerial images are also provided in a variety of resolutions representing capture or simulated capture from a specific altitude. High resolution aerial images show more detail and are typically more expensive to acquire. High resolution aerial images are larger files requiring more bandwidth and hence time to transmit.

10 In FIG. 2, a process diagram is illustrated showing the sequence of steps necessary to practice the invention. In step 22, an image is acquired that includes associated location information such as GPS data. The location information is preferably associated with the image in the form of metadata, but can be merely supplied in association with the image without deviating from the
15 scope of the invention. The image can be provided by computing device 12 from its internal storage or from any storage device or system accessible by computing device 12 such as a local network storage device or an online image storage site. Using the location information provided in step 22, computing device 12 provides the location information to aerial image server 16 in step 24 to acquire an aerial
20 image correlated to the location information associated with the image.

Using the acquired aerial image of step 24, computing device 12 identifies the environment of the image in step 26 using classifiers in the following manner. Recently, many people have shown the efficacy of representing an image as an unordered set of image patches or “bag of visual
25 words” (F.-F. Li and P. Perona, A Bayesian hierarchical model for learning natural scene categories, Proceedings of CVPR, 2005; S. Lazebnik, C. Schmid, and J. Ponce, Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, Proceedings of CVPR, 2006). Suitable descriptions (e.g., so called SIFT descriptors) are computed for each of training images, which are
30 further clustered into bins to construct a “visual vocabulary” composed of “visual words.” The intention is to cluster the SIFT descriptors into “visual words” and then represent an image in terms of their occurrence frequencies in it. The well-

known k-means algorithm is used with cosine distance measure for clustering these descriptors.

While this representation throws away the information about the spatial arrangement of these patches, the performances of systems using this type of representation on classification or recognition tasks are impressive. In particular, an image is partitioned by a fixed grid and represented as an unordered set of image patches. Suitable descriptions are computed for such image patches and clustered into bins to form a “visual vocabulary.” In the present invention, the same methodology is adopted and it is extended to consider both color and texture features for characterizing each image grid.

To extract color features, an image grid is further partitioned into 2×2 equal size sub-grids. Then for each subgrid, one can extract the mean R, G, and B values to form a $4 \times 3 = 12$ feature vector which characterizes the color information of 4 sub-grids. To extract texture features, one can apply a 2×2 array of histograms with 8 orientation bins in each sub-grid. Thus a $4 \times 8 = 32$ -dimensional SIFT descriptor is applied to characterize the structure within each image grid, similar in spirit to Lazebnik et al. In a preferred embodiment of the present invention, if an image is larger than 200,000 pixels, it is first resized to 200,000 pixels. The image grid size is then set to 16×16 with overlapping sampling interval 8×8 . Typically, one image generates 117 such grids.

After extracting all the raw image features from image grids, the present invention builds separate color and texture vocabularies by clustering all the image grids in the training dataset through k-means clustering. In the present invention, both vocabularies are set to size 500. By accumulating all the grids in the training set of images, one obtains two normalized histograms for an event, h_c and h_t , corresponding to the word distribution of color and texture vocabularies, respectively. Concatenating h_c and h_t , the result is a normalized word histogram of size 1000. Each bin in the histogram indicates the occurrence frequency of the corresponding word.

For each input aerial image that was obtained based on the location coordinates, both color and texture bags of words are computed as described above. The images corresponding to a particular type (or class) of environment

form the training set for that class. The training sets of all the classes form the total training set. A classifier is trained using this training set.

In a preferred embodiment of the present invention, a classifier known as SAMME (by J. Zhu, S. Rosset, H. Zou, and T. Hastie. Multi-class Adaboost. Technique Report, 2005) is selected as the classifier for solving the multi-class problem. Compared with another algorithm called AdaBoost.MH, which needs to perform K one-against-all classifications, SAMME performs K-class classification directly. It only requires weak classifiers better than random guess (e.g. correct probability larger than $1/K$), rather than better than $1/2$ as the standard two-class AdaBoost requires.

For each input aerial image that needs to be classified, both color and texture bags of words are computed as described above. The trained classifier is applied to the bags of words of the input image to produce a class of the environment for the input image.

It will be understood that some locations can have multiple environments. For example, Rotary Park on North Main Street in Honeoye Falls, New York includes parking environments, picnic environments, fishing environments, and soccer environments. Broadly by way of example, but not limitation, identified environments can include baseball fields, football fields, tennis courts, golf courses, residential areas, commercial properties, schools, shopping malls, playgrounds, parks, beaches, forests, or theme parks. As shown in table 35 of FIG. 3, environments can have pre-established simplified environment ontologies associated with them. In both computer science and information science, an ontology is a data model that represents a set of concepts within a domain and the relationships between those concepts. It is used to reason about the objects within that domain. In the example of FIG. 3, a very simplified ontology is shown for two example environments. Relationships are not shown between the concepts listed in FIG. 3, but if established can assist in the use of the ontology of the environment for inferring the nature of an activity or event.

The identified environment of the image is then stored in association with the image in step 28. Preferably the identified environment is stored as metadata included with the image such as EXIF data, but the invention is

not so limited. The identified environment can be stored as data in a database in association with the image. In step 30, inferences regarding the nature of the event or activity associated with the image are made from the stored environment of the image. Indexing server 14 or computing device 12 can infer from the stored environment of the image, for example a baseball field, that the nature of the event was a baseball game. This inference can lead to a valid classification of the event as a baseball game. If indexing server 14 can analyze the image as shown in FIG. 4a and detect the presence of a baseball glove and a baseball, then these objects also being found in the ontology of baseball field environment give more conclusive evidence that the event is properly classified as a baseball game. If however, as shown in FIG. 4b, objects such as guitar, people, and drum set are determined by analyzing the image and the environment is still a baseball field then the classification of the event as a baseball game is likely invalid. The proper inference, a concert, is made by looking at both the stored environment of the image and the image. If metadata have been provided with the aerial image, further information can be considered in combination with any of the other data when making the inference as to the nature of the event or activity associated with the image. Metadata provided with the acquired aerial image can be stored additively to the ontology of stored environment if desired for subsequent use.

Having inferred the nature of the activity or event allows computing device 12 to search a collection of images for the inferred event in step 32. The event type, the stored environment, the elements of the stored environment ontology or any metadata associated with the image or the aerial image may be used alone or in combination to frame the search of the collection of images. The collection of images to be searched may belong to the user of computing device 12 and resident thereon or it could be a public collection of images available for searching via communication network 20 with a tool such as Google™ Image search implemented by a web browser running on computing device 12. Another popular collection of images available for searching online with a web browser is maintained by FlickrR™ at <http://www.flickr.com> where text tags are used to search for images.

In FIG. 4a, image 40 is shown along with acquired aerial image 42 provided by aerial image server 16 in response to the provision of location information associated with image 40. List 44 shows the objects determined from the indexing analysis of image 40 while list 46 shows the metadata associated with the acquired aerial image 42. The inference regarding the potential nature of the event being a baseball game is fairly conclusive and the terms in the list can be used to search for images in the collection that have similar content. Likewise in FIG. 4b, image 48 captured with the same location information at a different time and date is shown along with acquired aerial image 42. List 50 shows the objects determined from the indexing analysis of image 48. Comparing lists 50 and 46, the inference regarding the potential nature of the event as a baseball game is invalid. The conclusive inference of a concert is reached only upon consideration of both the stored environment of the image and the image.

In order to infer the event accurately, information from the images and the corresponding co-located aerial images can be fused at three levels: feature level, score level, and decision level. Feature level fusion requires concatenation of features from both image and the corresponding aerial image to form a unified feature vector. Score level fusion uses the output scores from multiple classifiers across all of the classes and feeds them to a fusion or meta classifier. Decision level fusion trains a fusion classifier that takes the prediction labels of both the image and the co-located aerial image classifiers.

FIG. 5a depicts the feature level fusion as one embodiment of the invention. Image 52 is fed to a feature extractor 56 (using SIFT and color feature extraction approach as described previously) to obtain visual information (feature1). The corresponding co-located aerial image 54 is fed to another feature extractor 58 for visual feature (feature2) related to aerial images. Both image feature (feature1) and the co-located aerial image feature (feature2) are combined in the feature level fusion module 60 by simple concatenation. Using the fused feature, a trained classifier 62 will infer the event of the image and make prediction of the event category 64.

FIG. 5b depicts the score level fusion as a second embodiment of the invention. Image 52 is fed to a feature extractor 56 (using SIFT and color

feature extraction approach as described previously) to obtain visual information. Classifier 66 uses only the image visual feature to infer the event of the image and output confidence related scores (score1) for each event category. The corresponding co-located aerial image 54 is fed to another feature extractor 58 for visual feature related to aerial images. Classifier 68 uses only the aerial image's visual feature to infer the event of the image and output another set of confidence related scores (score2). The scores of both image classifier 66 and corresponding aerial image classifier 68 are concatenated to form a fused score vector and fed to score level fusion module 70. Using the fused score, a trained classifier 72 will infer the event of the image and make prediction of the event category 64.

FIG. 5c depicts the decision level fusion as a third embodiment of the invention. Image 52 is fed to a feature extractor 56 (using SIFT and color feature extraction approach as described previously) to obtain visual information. Classifier 66 uses only the image visual feature to infer the event depicted in the image and make prediction 74 of the most relevant event category based on the image (decision1). The corresponding co-located aerial image 54 is fed to another feature extractor 58 for visual feature related to aerial images. Classifier 68 uses only the aerial image's visual feature to infer the event shown in the image and make prediction 76 of the most relevant event category based on aerial image (decision2). Both image prediction 74 and corresponding aerial image prediction 76 are combined and fed to a decision level fusion module 78. Using the fused decision, a trained classifier 80 will infer the event of the image and make prediction of the event category 64.

PARTS LIST

- 10 system
- 12 computing device
- 14 indexing server
- 16 aerial image server
- 20 communication network
- 22 step
- 24 step
- 26 step
- 28 step
- 30 step
- 32 step
- 35 table
- 40 image
- 42 acquired aerial image
- 44 list
- 46 list
- 48 image
- 50 list
- 52 image acquiring step
- 54 co-located aerial image acquiring step
- 56 image feature extraction step
- 58 co-located aerial image feature extraction step
- 60 feature level fusion step
- 62 classifier trained on features of images
- 64 classifier prediction based on image and co-located aerial image
- 66 classifier trained on features of images
- 68 classifier trained on features of co-located aerial images
- 70 score level fusion step
- 72 score level classifier
- 74 classifier prediction based on image
- 76 classifier prediction based on co-located aerial image

Parts List - continued

78 decision level fusion

80 classifier trained on prediction output of image and co-located aerial image
classifiers

CLAIMS:

1. A method of recognizing an event depicted in an image from the image and a location information associated with the image, comprising:
- 5 a) acquiring the image and its associated location information;
b) using the location information to acquire an aerial image(s) correlated to the location information;
c) identifying the event using both the image and the acquired aerial image(s); and
10 d) storing the event in association with the image for subsequent use.
2. The method of Claim 1, wherein step c) further comprises:
- 15 i) extracting features from the image and the aerial image(s);
ii) combining the features; and
iii) classifying the event depicted in the image using the combined features.
3. The method of Claim 1, wherein step c) further comprises:
- 20 i) producing a first score from the image using a first pre-defined image classifier;
ii) producing a second score from the aerial image(s) using a second pre-defined image classifier; and
iii) combining the first and the second scores to classify the event
25 depicted in the image.
4. The method of Claim 1, wherein step c) further comprises:
- i) producing a first decision from the image using a first pre-defined image classifier;
30 ii) producing a second decision from the aerial image(s) using a second pre-defined image classifier; and

iii) combining the first and the second decisions to classify the event depicted in the image.

5 5. The method of Claim 1, wherein step b) includes acquiring the aerial image(s) at a plurality of resolutions.

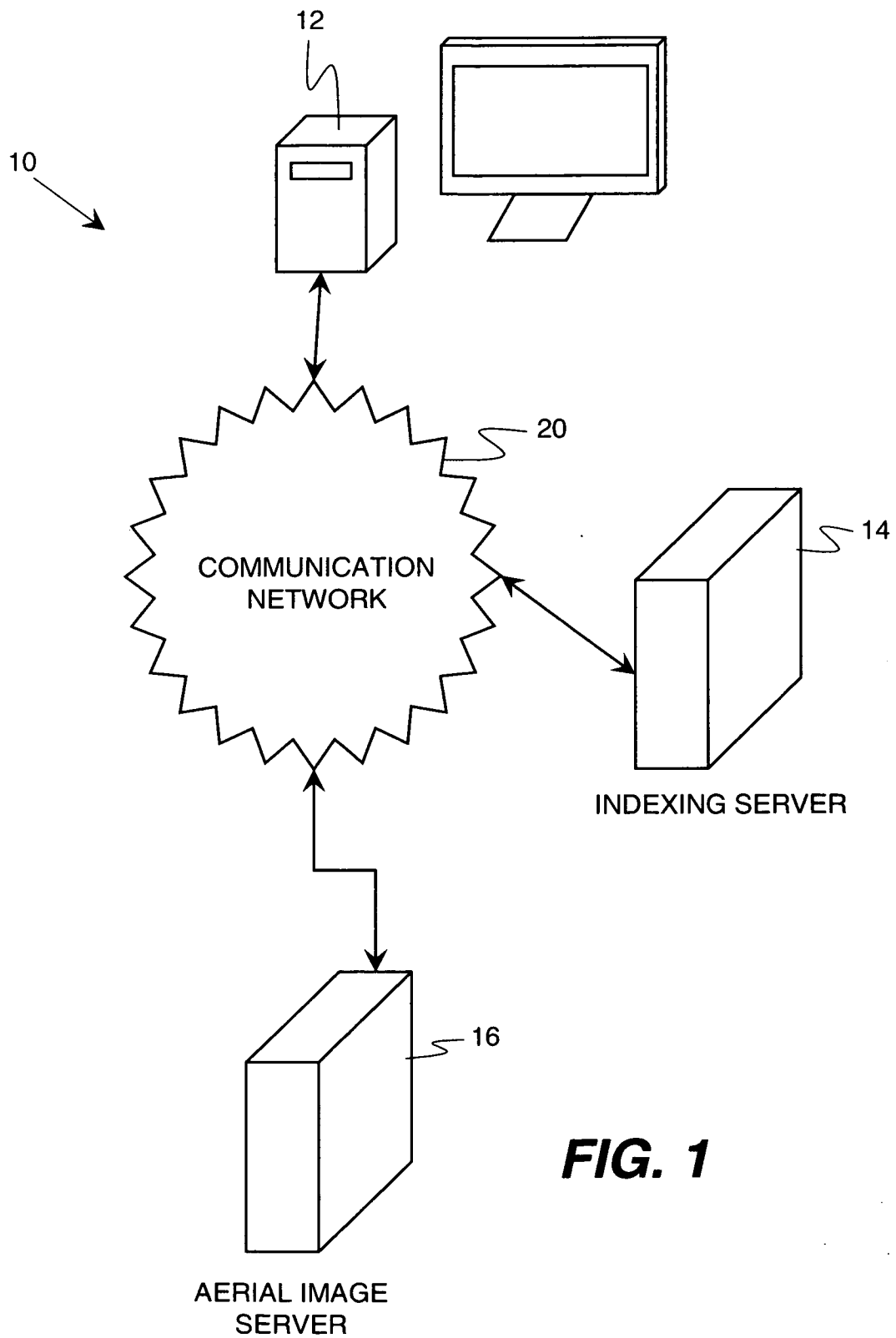
6. The method of Claim 1, wherein the event includes baseball games, football games, tennis games, golfing, residential activities, commercial activities, school activities, shopping, playground activities, park activities, beach
10 activities, hiking, or theme park visits.

7. The method of Claim 1, further comprising:
e) applying steps a) to d) a plurality of times to a plurality of
images; and
15 f) searching the plurality of images based on the stored events for images of a particular event.

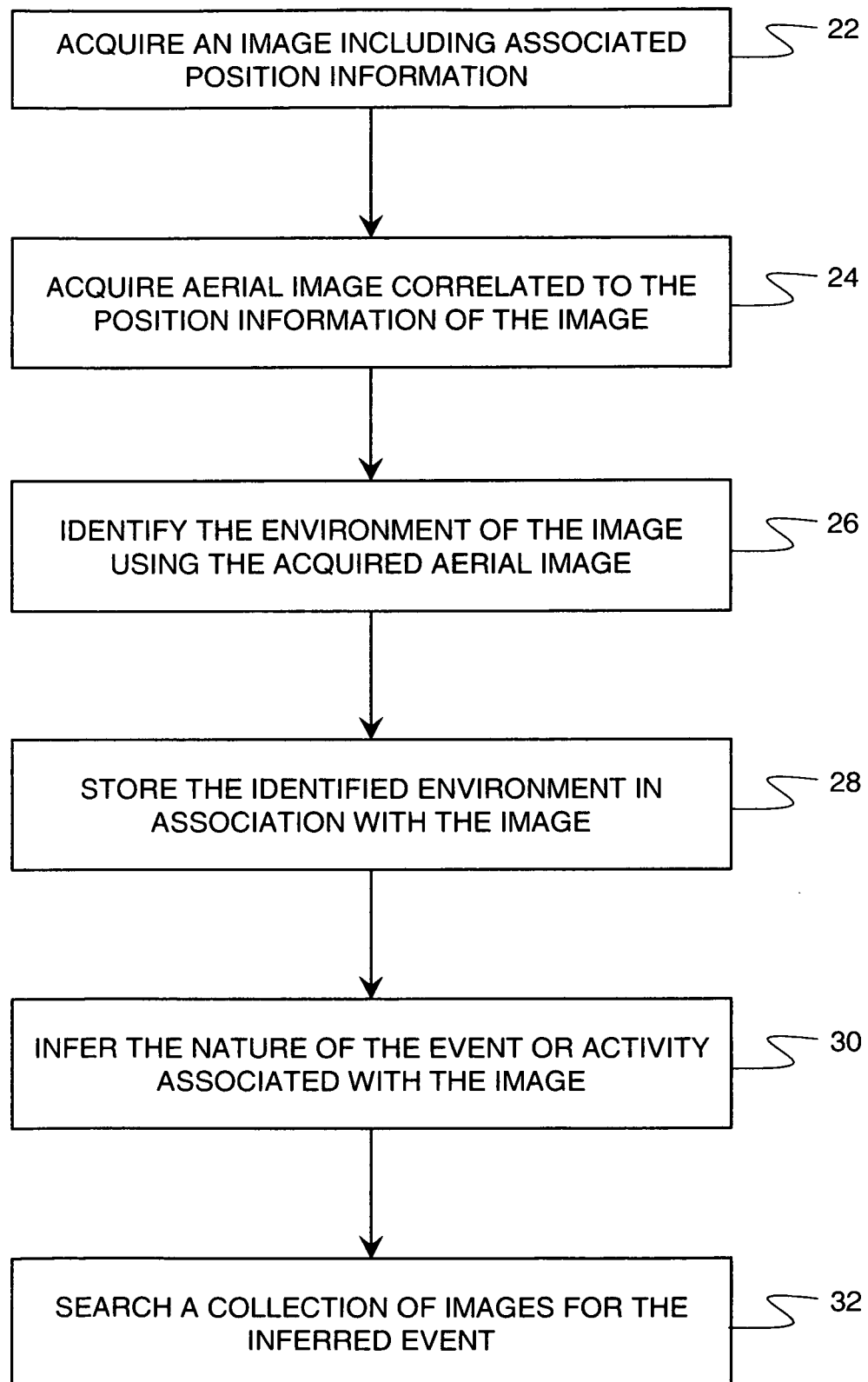
8. The method of Claim 1, further comprising acquiring metadata associated with the acquired aerial image(s) and storing such metadata
20 for subsequent use in association with the stored event.

9. The method of Claim 8, further comprising:
g) applying steps a) to d) a plurality of times to a plurality of
images; and
25 h) searching the plurality of images based on the stored events for images of a particular event.

1/8

**FIG. 1**

2/8

**FIG. 2**

3/8

ENVIRONMENT	ENVIRONMENT ONTOLOGY	35
BASEBALL FIELD	BASEBALL	
	BAT	
	GLOVE	
	UNIFORM	
	BASEBALL HAT	
	PEOPLE	
TENNIS COURT	TENNIS RACKET	
	TENNIS BALL	
	NET	
	PEOPLE	

FIG. 3

4/8

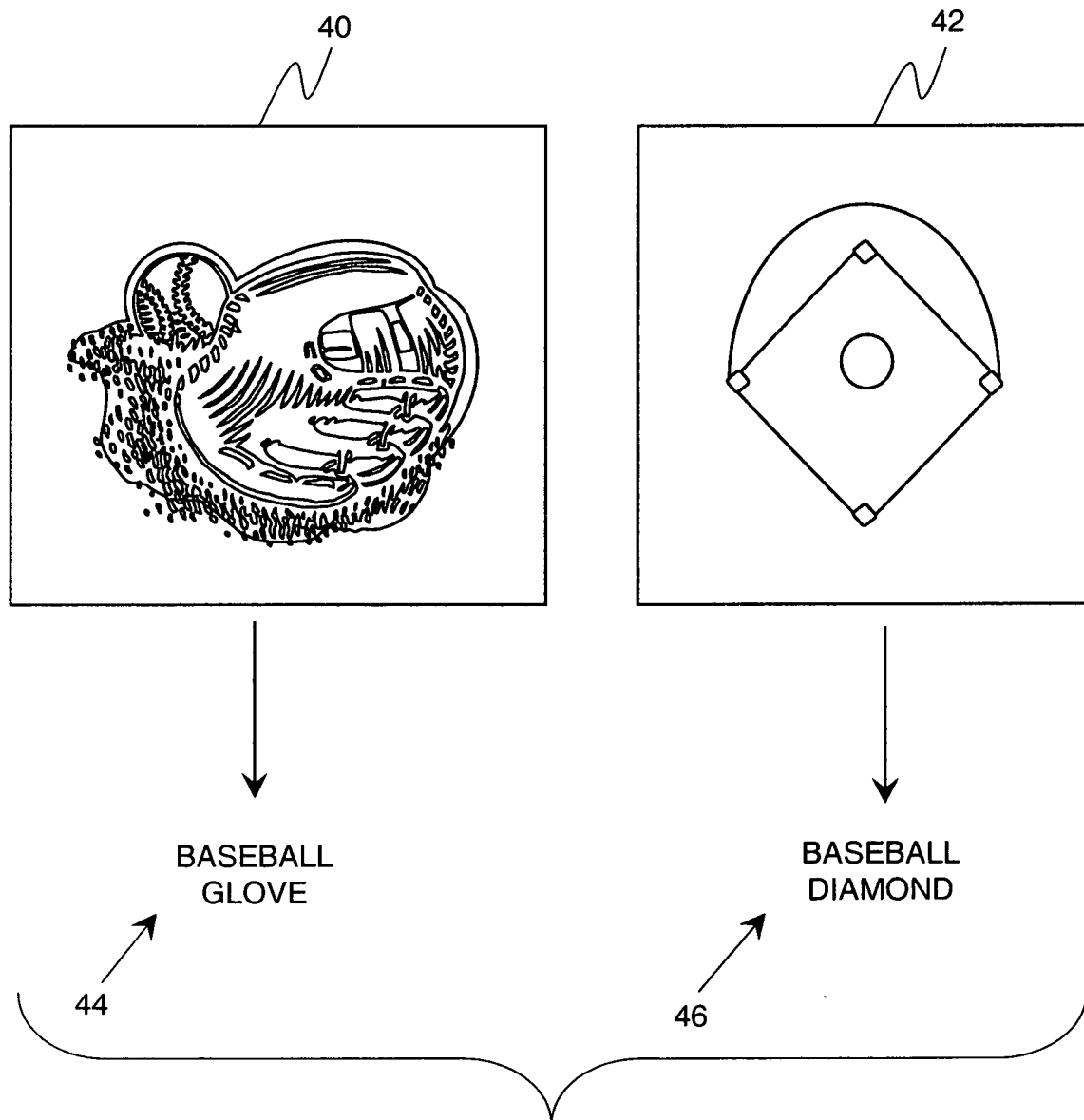


FIG. 4a

5/8

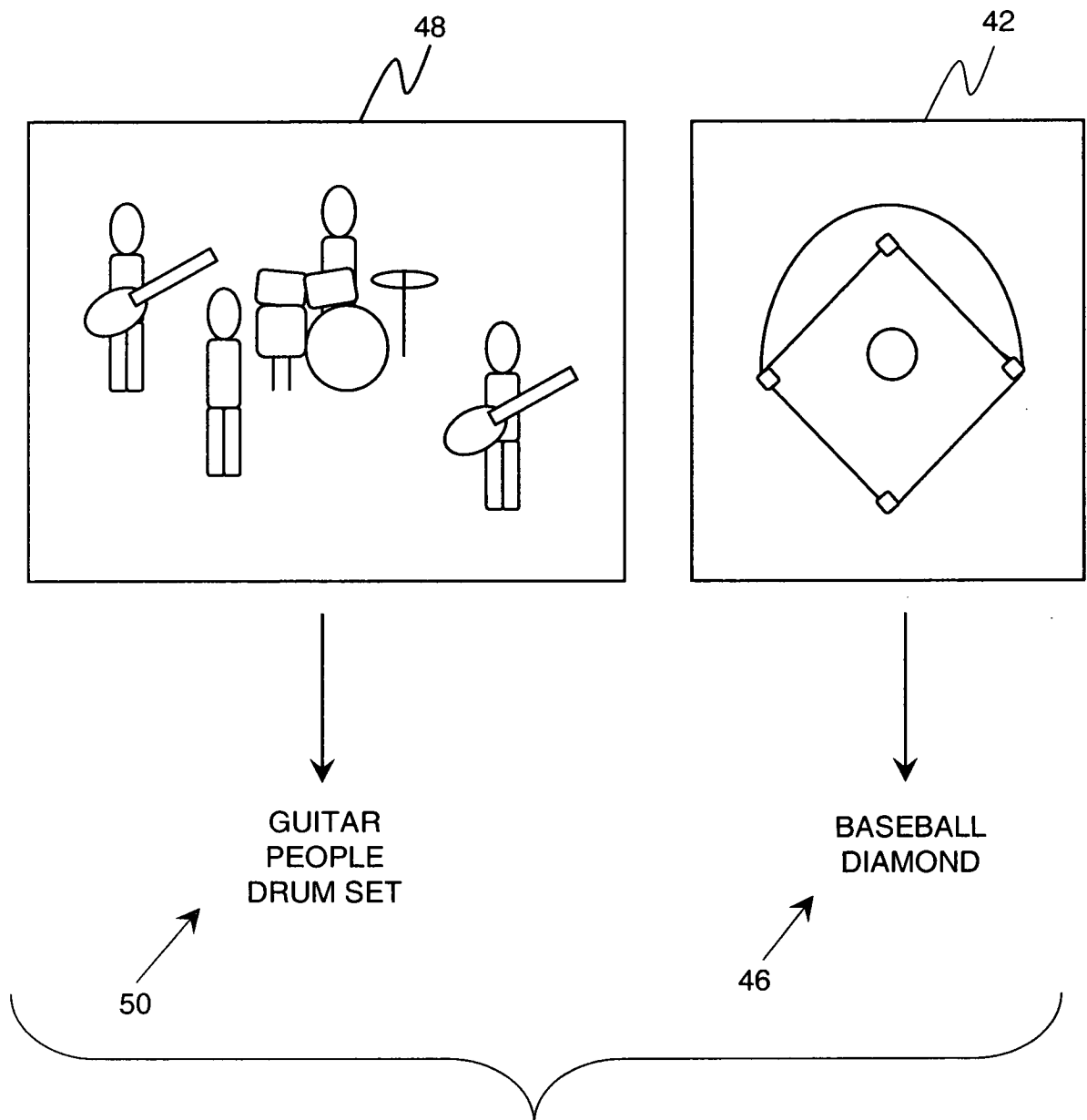
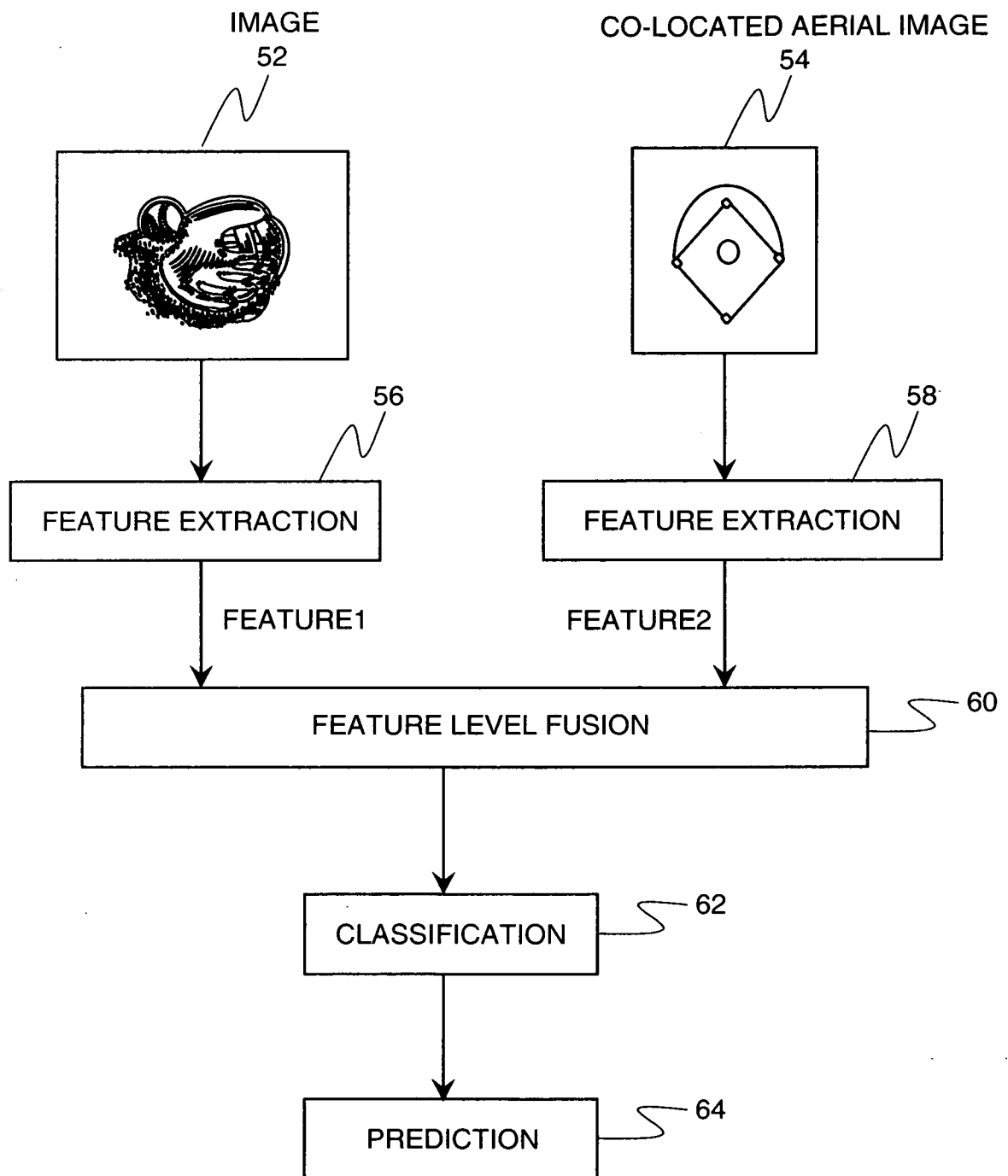
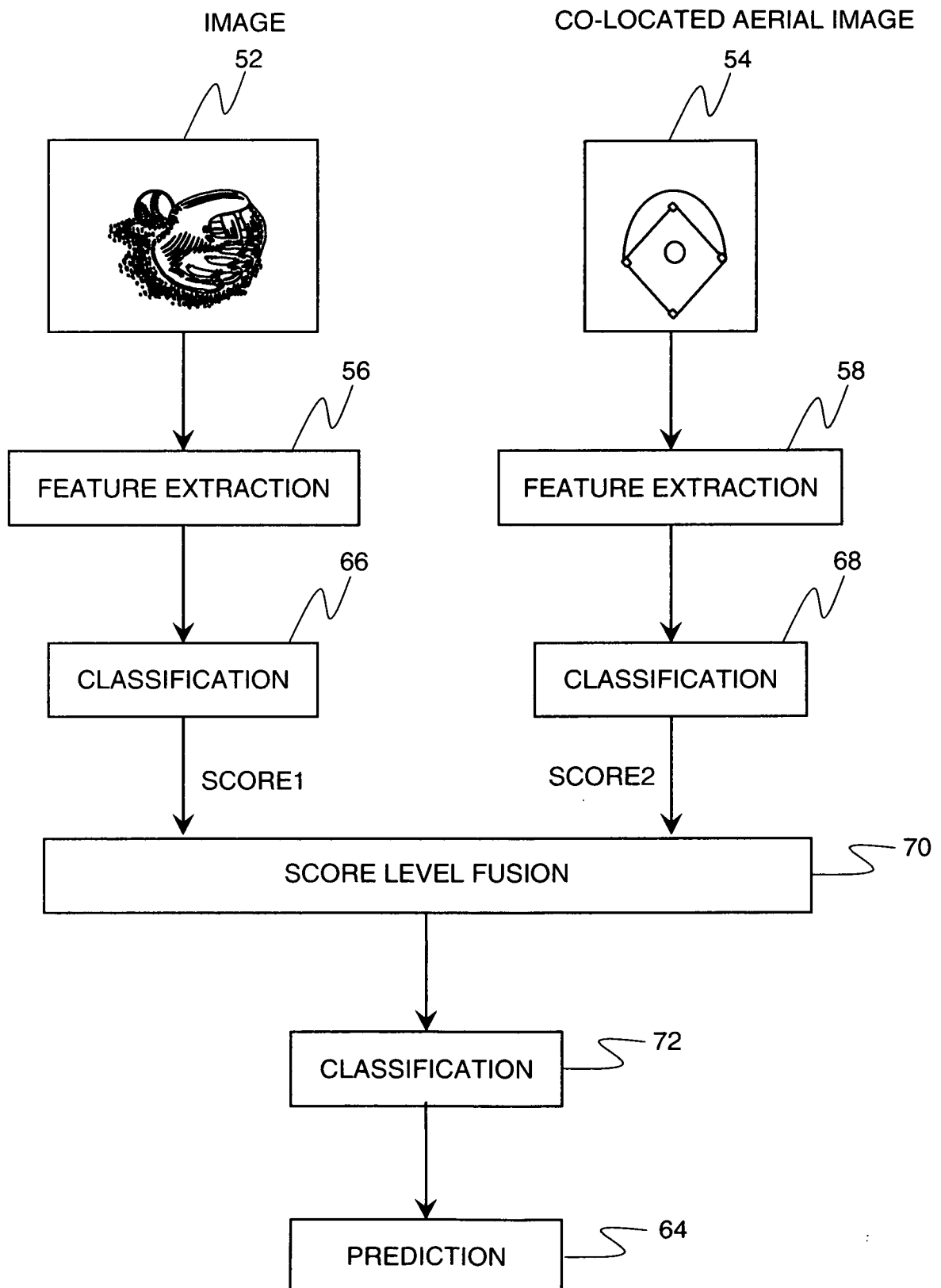


FIG. 4b

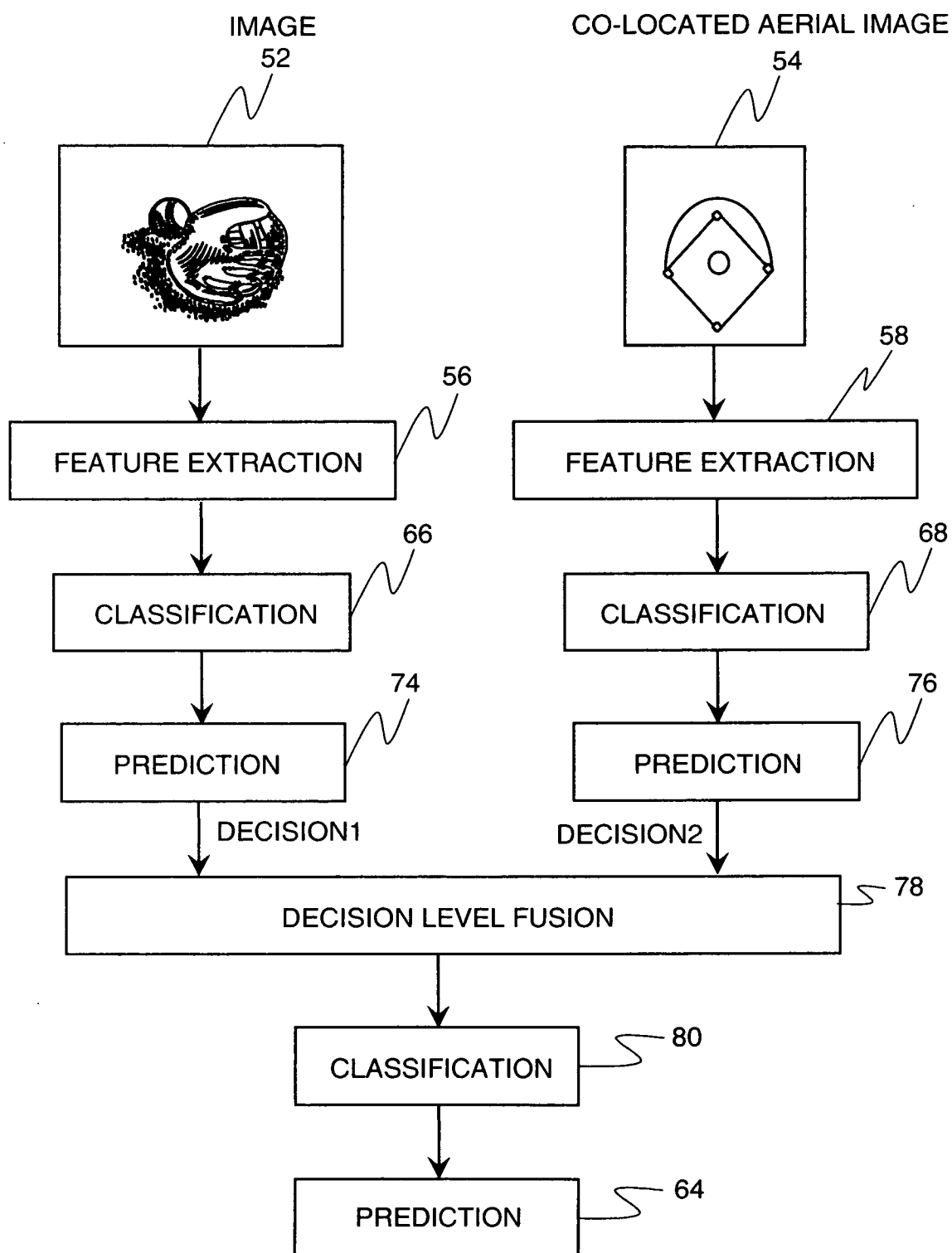
6/8

**FIG. 5a**

7/8

**FIG. 5b**

8/8

**FIG. 5c**

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2009/005832

A. CLASSIFICATION OF SUBJECT MATTER

INV. G06K9/00 G06K9/46 G06F17/30

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06K G06F

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 practical, search terms used)

EPO-Internal

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	J. LUO ET AL.: "Event recognition: Viewing the World with a Third Eye" PROCEEDING OF THE 16TH ACM INTERNATIONAL CONFERENCE ON MULTIMEDIA, [Online] 26 October 2008 (2008-10-26), pages 1071-1080, XP002563907 Retrieved from the Internet: URL: http://portal.acm.org/citation.cfm?id=1459574 [retrieved on 2010-01-18] the whole document, in particular Section 3.4. "Event Recognition using Both Views" ----- -/--	1-9

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

☐ See patent family annex.

* Special categories of cited documents :

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

"E" earlier document 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

19 January 2010

Date of mailing of the international search report

29/01/2010

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040,
Fax: (+31-70) 340-3016

Authorized officer

Grigorescu, Cosmin

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2009/005832

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	B. CHANDRASEKARAN ET AL.: "What are ontologies, and why do we need them?" IEEE INTELLIGENT SYSTEMS, vol. 14, no. 1, January 1999 (1999-01), pages 20-26, XP002563908 p. 23, Section "Task dependence of ontologies"	1-9
A	Toni Navarette: "Semantic integration of thematic geographic information in a multimedia context" June 2006 (2006-06), Ph.D. Thesis, Univ. Pompeu Fabra , XP002563909 , pages 201-251 the whole document	1-9