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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
12 April 2018 (12.04.2018)

WIPO | PCT

(10) International Publication Number
WO 2018/065073 A1

(51) International Patent Classification:
G06K 9/00 (2006.01) *G06K 9/62* (2006.01)
G06K 9/46 (2006.01)

(21) International Application Number:
PCT/EP2016/074088

(22) International Filing Date:
07 October 2016 (07.10.2016)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant: **TOYOTA MOTOR EUROPE** [BE/BE]; Avenue du Bourget 60, 1140 Brussels (BE).

(72) Inventors: **MEIER, Sven**; c/o TOYOTA MOTOR EUROPE, European Technical Administration Division, Avenue du Bourget 60, 1140 Brussels (BE). **KOBORI, Norimasa**; c/o TOYOTA MOTOR EUROPE, European Technical Administration Division, Avenue du Bourget 60, 1140 Brussels (BE). **KEHL, Wadim**; Tengstrasse 8, 80798 Munich (DE). **MILLETARI, Fausto**; Augustenstrasse 96a, 80798 Munich (DE).

(74) Agent: **INTÈS, Didier** et al.; Cabinet Beau De Lomenie, 158 Rue de l'Université, 75340 Paris Cedex 07 (FR).

(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, JP, KE, KG, 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, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ,

(54) Title: ELECTRONIC DEVICE, SYSTEM AND METHOD FOR RECOGNIZING AND LOCATING AN OBJECT

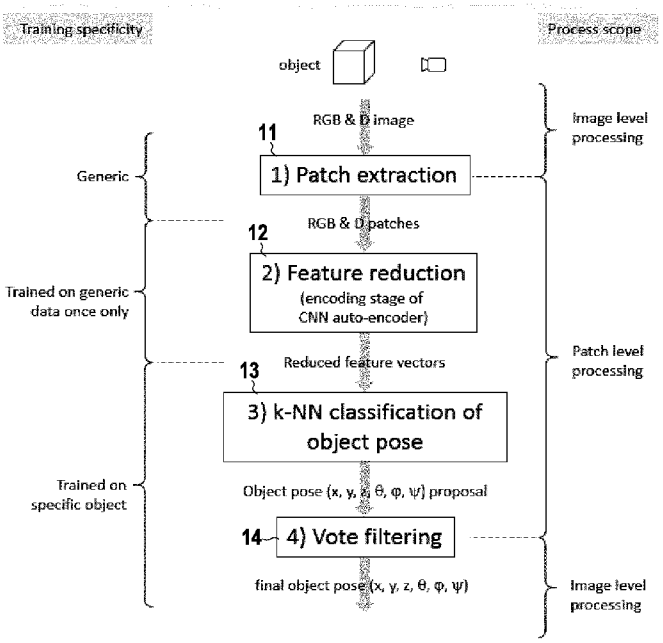


FIG.2

(57) Abstract: The invention relates to an electronic device (1) for recognizing and locating an object. The electronic device is configured to: receive 3D image data of an optical sensor (3), the sensor sensing the object (O), decompose the 3D image data into a set of patches, extract a set of features of each patch by using a pre-trained convolutional neural network (CNN) auto-encoder (12), for each patch classify the object pose by using the set of features of the patch as an input to a k-nearest neighbors algorithm (13) to match against predetermined object representations, and determine the pose of the object by combining the classified object poses of the set of patches. The invention further relates to a system and a method.

[Continued on next page]

WO 2018/065073 A1

UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

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

Electronic device, system and method for recognizing and locating an object

FIELD OF THE DISCLOSURE

[0001] The present disclosure is related to an electronic device, system and method for recognizing and locating an object, in particular for recognizing the pose of a non-stationary object.

BACKGROUND OF THE DISCLOSURE

[0002] Automation is becoming increasingly important in many fields what also implies the increased need for robotics. Meanwhile robotic systems have become common in the industrial field, their use is still rather uncommon in the environment of a domestic dwelling, e.g. to serve individual users in daily life. However, also in this field there is a high demand for robotic systems. For example, a robotic system may help an aged person to find and retrieve a specific object, e.g. a pencil.

[0003] One problem of the use of robotic systems in a domestic dwelling is that in contrast to an industrial application, many tasks cannot be standardized, i.e. predefined and tightly controlled. Thus the robotic system has to be capable to carry out individually varying tasks. Moreover the operating conditions in a domestic dwelling are more challenging, e.g. lighting, object disposition etc.

[0004] An important aspect of a robotic system is therefore its capability to find and recognize a specific object, which may be positioned in any location and in any orientation. For this purpose the robotic system may comprise an optical sensor and may be moveable, e.g. may have drivable wheels.

[0005] US2012114251 (A1) discloses a system for object recognition of a 3D object of a certain object class using a statistical shape model for recovering 3D shapes from a 2D representation of the 3D object and comparing the recovered 3D shape with known 3D to 2D representations of at least one object of the object class.

[0006] EP1603071 (A1) discloses a three-dimensional object recognizing system with a distance image generating portion for generating a distance image by using image pairs picked up by a stereoscopic camera. The discrimination of the three-dimensional object is carried out based on the output patterns of a neural network.

[0007] However, the known systems require an extensive training which has to be done for each object, in order to become known to the system. Accordingly, in case a user wants to add a new object to the list of known objects, the system has to perform the complete training, in particular including a training of an involved neural network, what may take several hours or even days involving a high processing power.

SUMMARY OF THE DISCLOSURE

[0008] Currently, it remains desirable to provide an electronic device, system and method for recognizing and locating an object with a reduced training effort.

[0009] Therefore, according to the embodiments of the present disclosure, an electronic device for a vehicle for recognizing and locating an object is provided. The electronic device being configured to:

- receive 3D image data of an optical sensor, the sensor sensing the object,
- decompose the 3D image data into a set of patches,
- extract a set of features of each patch by using a pre-trained convolutional neural network (CNN) auto-encoder,
- for each patch classify the object pose by using the set of features of the patch as an input to a k-nearest neighbors algorithm to match against predetermined object representations, and
- determine the pose of the object by combining the classified object poses of the set of patches.

[0010] By providing such an electronic device, the training of the device can be substantially reduced when a new object is learnt by the device. The combination of feature extraction and classification implemented as CNN auto-encoder and k-NN classifier respectively has following advantages:

[0011] It is sufficient to (pre)train only once the convolutional neural network (CNN) auto-encoder (i.e. the CNN auto-encoder feature space reduction). This training can be intense and detailed and the high computational effort required is saved in all subsequent uses of this processing step.

[0012] However, the k-nearest neighbors algorithm (i.e. k-NN classifiers) is easy and quick to train, hence the computational effort to learn a new object is small.

[0013] In other words, the feature space reduction step implemented in the CNN auto-encoder reduces computational burden on the k-NN classification step.

[0014] Furthermore, the (local) patch decomposition achieves scale invariance. Accordingly, the patch data may be independent of the distance and apparent size of the object. Scale invariance is the property of object appearance descriptions to be constant as the apparent size of the object varies.

[0015] The image data may comprise a pair of a visible light image and a depth image. These data may be the input data to the device.

[0016] The visible light image may comprise the visible part of the electromagnetic spectrum, in particular decomposed into the three bands (RGB: red, green, blue) processed by the human vision system.

[0017] The set of features may be extracted by using the encoding stage of the convolutional neural network (CNN) auto-encoder. Desirably the set of features may be output by the encoding stage.

[0018] Accordingly, the used convolutional neural network (CNN) auto-encoder may comprise an encoding stage and a decoding stage. The device desirably only uses the encoding stage but not decoding stage of the convolutional neural network (CNN) auto-encoder to obtain (i.e. extract) the set of features. In other words, the set of features may be the output of the encoding stage.

[0019] The object pose may be a 6D pose comprising x, y, z location information and θ, φ, ψ rotation information.

[0020] More generally, the object pose may be a mathematical description of the location and orientation of an object in a coordinate system.

[0021] The features may comprise mathematical descriptions of the object appearance, in particular the color and/or brightness of the object.

[0022] The pose of the object may be determined by vote filtering using the classified object poses of the set of patches as votes.

[0023] Accordingly, the contribution of all patches may constitute a vote space. Applying an appropriate threshold to this vote space may lead to the most likely object pose of the object in the scene sensed by the optical sensor.

[0024] A threshold may be defined which sets a feature space distance to the vote space, i.e. which limits the votes to the most likely object poses.

[0025] The patches may be 3 dimensional portions of the 3D image data, in particular of a predetermined size and/or overlapping each other.

[0026] Accordingly, an input image may be decomposed into a set of smaller images named patches.

[0027] The convolutional neural network (CNN) auto-encoder may be pre-trained based on (i.e. by using) a generic set of patches, in particular once only.

[0028] Accordingly, the training of the convolutional neural network (CNN) auto-encoder may be generic. Once trained, the auto-encoder may be applied to previously unseen objects and images. Hence, there is no need of re-training thus greatly increasing the speed of adapting this algorithm to new objects.

[0029] The generic set of patches may be a set of patches extracted from a large and representative dataset of object data. The data may be further augmented by permuting colors and adding noise. This large dataset may be used to provide a broad training set for the convolutional auto-encoder. Once trained the encoding stage of the auto-encoder may be kept and used for the extracting descriptive features from input 3D image data.

[0030] The convolutional neural network (CNN) auto-encoder may be configured to reduce the patch to a feature vector f(x), in particular of 1024 features.

[0031] Accordingly, the activation values of the compressing feature layer F of the convolutional neural network auto-encoder may be used to reduce a patch to a set of features (i.e. to extract the set of features from the patch).

[0032] The k-nearest neighbors (k-NN) algorithm may be configured to find the most likely object pose in the feature space defined by the feature vector f(x).

[0033] The k-nearest neighbors algorithm is pre-trained based on data of the object.

[0034] Accordingly, the device may use a 3D model file of the target object as prior knowledge. This model may provide shape (3D) and appearance (color)

information of the object. The k-NN algorithm trained on the particular object properties may be used to find the best matching pose of each patch. The features of the compression feature layer are sufficiently discriminative to be usable as lookup feature space for a k-NN algorithm. Hence, although the k-NN algorithm can be rapidly trained and also the trained k-NN algorithm requires only low processing power when applied on a set of features, the device can reliably recognize and locate an object.

[0035] The disclosure further relates to a system for recognizing and locating an object, the system comprising:

- an electronic device, in particular as described above, and
- an optical sensor configured to sense the object, the sensor being in particular a 3D camera or a stereo camera.

[0036] Accordingly, the system may be configured to autonomously recognize and locate an object. For example it may be realized as a moveable robotic system, e.g. with means for retrieving the object.

[0037] The disclosure further relates to a method of recognizing and locating an object. The method comprises the steps of:

- receiving 3D image data of an optical sensor, the sensor sensing the object,
- decomposing the 3D image data into a set of patches,
- extracting a set of features of each patch by using a pre-trained convolutional neural network (CNN) auto-encoder;
- for each patch classifying the object pose by using the set of features of the patch as an input to a k-nearest neighbors algorithm to match against predetermined object representations, and
- determining the pose of the object by combining the classified object poses of the set of patches.

[0038] The method may comprise further method steps which correspond to the functions of the electronic device as described above. The further desirable method steps are described in the following.

[0039] The image data may comprise a pair of a visible light image and a depth image.

[0040] The visible light image may comprise the visible part of the electromagnetic spectrum, in particular decomposed into the three bands (RGB) processed by the human vision system.

[0041] The set of features may be extracted by using the encoding stage of the convolutional neural network (CNN) auto-encoder. Desirably the set of features may be output by the encoding stage.

[0042] The object pose may be a 6D pose comprising x, y, z location information and θ, φ, ψ rotation information.

[0043] The features may comprise mathematical descriptions of the object appearance, in particular the color and/or brightness of the object.

[0044] The pose of the object may be determined by vote filtering using the classified object poses of the set of patches as votes.

[0045] The patches may be 3 dimensional portions of the 3D image data, in particular of a predetermined size and/or overlapping each other.

[0046] The convolutional neural network (CNN) auto-encoder may be pre-trained based on a generic set of patches, in particular once only.

[0047] The convolutional neural network (CNN) auto-encoder may reduce the patch to a feature vector f(x), in particular of 1024 features.

[0048] The k-nearest neighbors algorithm may find the most likely object pose in the feature space defined by the feature vector f(x).

[0049] The k-nearest neighbors algorithm may be pre-trained based on data of the object.

[0050] It is intended that combinations of the above-described elements and those within the specification may be made, except where otherwise contradictory.

[0051] It is to be understood that both the foregoing general description and the following detailed description are exemplary and explanatory only and are not restrictive of the disclosure, as claimed.

[0052] The accompanying drawings, which are incorporated in and constitute a part of this specification, illustrate embodiments of the disclosure and together with the description, and serve to explain the principles thereof.

BRIEF DESCRIPTION OF THE DRAWINGS

[0053] Fig. 1 shows a block diagram of a system with an electronic device according to embodiments of the present disclosure;

[0054] Fig. 2 shows a schematic flow chart illustrating an exemplary method of recognizing and locating an object according to embodiments of the present disclosure;

[0055] Fig. 3 shows a flow chart illustrating the exemplary method of Fig. 2 in context of a specific example;

[0056] Fig. 4 shows an example of a convolutional auto-encoder according to embodiments of the present disclosure;

[0057] Fig. 5 shows a depiction of the auto-encoder of fig. 4;

[0058] Fig. 6 shows an exemplary scene where an object is recognized and located according to embodiments of the present disclosure.

DESCRIPTION OF THE EMBODIMENTS

[0059] Reference will now be made in detail to exemplary embodiments of the disclosure, examples of which are illustrated in the accompanying drawings. Wherever possible, the same reference numbers will be used throughout the drawings to refer to the same or like parts.

[0060] Fig. 1 shows a block diagram of a system 30 with an electronic device 1 according to embodiments of the present disclosure. The system may comprise a robotic system 10, which may have various functions. For example it may be moveable, e.g. has drivable wheels, and it may have means for retrieving an object, e.g. at least one gripper.

[0061] The electronic device 1 carries out a computer vision algorithm for detecting the presence and location of objects in a scene. Robotic systems require this information to be able to find, locate and manipulate objects. The input to the electronic device 1 is a pair of visible light (RGB) and depth images (D). The output of the electronic device 1 is the 6D pose (x, y, z location and θ, φ, ψ rotation around x, y, z) of the target object.

[0062] The electronic device 1 is connected to or comprises a data storage 2. Said data storage may be used to store a target object in the form of a 3D model file providing shape (3D) and appearance (color) information of the scene. Furthermore the k-NN algorithm and/or the CNN auto-encoder are stored in the data storage 2. Accordingly, the electronic device 1 comprises the k-NN algorithm and the CNN auto-encoder and is configured to perform the k-NN algorithm and the CNN auto-encoder.

[0063] The electronic device 1 may additionally carry out further functions in the system 30. For example, the electronic device may also act as the general purpose ECU (electronic control unit) of the system. The electronic device 1 may comprise an electronic circuit, a processor (shared, dedicated, or group), a

combinational logic circuit, a memory that executes one or more software programs, and/or other suitable components that provide the described functionality. In other words, device 1 may be a computer device.

[0064] Device 1 may be external to the (movable) robotic system 10 which is configured to find and retrieve the object. In other words, the computational resources on board of the robotic system 10 may be limited, e.g. they may only transmit the 3D data to the external (and e.g. stationary) electronic device 1, e.g. over wifi. The result determined by the device 1 may be sent back to the robot.

[0065] The electronic device 1 is further connected to an optical sensor, in particular a 3D digital camera 3, e.g. a stereo camera or e.g. a Microsoft Kinect ® camera. The electronic device 1 and the digital camera may be comprised by a robotic system 10. The digital camera 3 is configured such that it can record a 3 dimensional scene, and in particular output digital data providing shape (3D) and appearance (color) information of the scene.

[0066] The output of the digital camera 3 is transmitted to the electronic device 1. Desirably, the output is transmitted instantaneously, i.e. in real time or in quasi real time. Hence, a searched object can also be recognized and located by the electronic device in real time or in quasi real time.

[0067] The system 30 may comprise additionally a server 20. The server 20 is used to train and eventually update the k-NN algorithm and/or the CNN auto-encoder. For this purpose, the electronic device 1 may be connectable to the server. For example the electronic device 1 may be connected to the server 20 via a wireless connection. Alternatively or additionally the electronic device 1 may be connectable to the server 20 via a fixed connection, e.g. via a cable. It is also possible that a data transfer between the electronic device 1 and server 20 is achieved by using a portable data storage, e.g. an USB stick.

[0068] Fig. 2 shows a schematic flow chart illustrating an exemplary method and system of recognizing and locating an object according to embodiments of the present disclosure. In particular the method is an object detection and pose estimation procedure as follows.

[0069] In step 11 a local patch extraction is carried out. For this purpose the image (i.e. the 3D image data) is decomposed into small patches, e.g. each of same physical size. The mean distance from the camera is subtracted from the patch data. Depth and color image values are normalized to range from -1 to 1.

[0070] In more detail, the image is decomposed into patches of identical physical size of 5x5x5cm. The size information is provided by the depth channel of the input data. Using the patch representation renders the overall system more robust to occlusions and lighting.

[0071] In order to normalize data for the next processing step the depth patch is clipped to the 5x5x5cm cube and then normalized to range from -1 to 1. The same applies to each of the color channels. This yields a normalized input dataset to the following processing step.

[0072] Local patches may be extracted from different locations of the object in the image. At each location the RGB image patch and matching depth image patch are extracted for further processing. In other words a searched object may consist of several patches in the image.

[0073] In step 12 a feature extraction is carried out for each patch. A convolutional neural network auto encoder is trained on a generic set of patches to produce a concise representation of the patches in a set of features, i.e. a feature vector of e.g. 1024 values. This step is generic and once trained, can be applied to previously unseen object types and images.

[0074] In step 13 a classification of the pose for each set of features is carried out. The data from local patches reduced through the convolutional neural network auto-encoder are used as input to a k-NN classifier to match against known object representations. This stage (i.e. of step 13) is trained and tuned to every specific object.

[0075] In step 14 a vote filtering is carried out. Each patch contributes votes and after filtering votes the most likely 6D pose remains. The contribution of all patches constitutes a vote space and applying an appropriate threshold to this vote space leads to the most likely object pose in the scene.

[0076] The combined use of all 4 steps leads to a reliable object detection ability, wherein computational costs for training time and at run-time can be reduced.

[0077] The electronic device may be configured to recognize several searched objects in a scene. For example the above described procedure can be run for each searched object using the same extracted set of features (cf. step 12) but different model files (i.e. object representations) when classifying the pose (cf. step 13).

[0078] Fig. 3 shows a flow chart illustrating the exemplary method of Fig. 2 in context of a specific example.

[0079] First, a patch decomposition algorithm 11 is carried out (cf. left side of fig. 3). An exemplary scene is shown which is sampled to extract scale invariant patches from the RGB and D images. The patch comprises a pair of a color image (left) and a depth image (right). Subsequently, a convolutional neural network auto-encoder 12 is carried out which reduces the image patch 11 to a feature vector f(x) (cf. center of fig. 3). A k-NN classifier algorithm 13 (i.e. a codebook) is carried out which finds the most likely object pose in this feature space. At the end, a vote filtering algorithm is carried out (cf. right side of fig. 3). Each patch casts a vote for a likely object pose and the location in the scene with maximum votes indicates the most likely object pose.

[0080] Fig. 4 shows an example of a convolutional auto-encoder according to embodiments of the present disclosure. A previously trained convolutional neural network auto-encoder is used to reduce the normalized RGB-D input patch to a vector of e.g. 1024 features. This uses the activation values of the compressing feature layer F of the convolutional neural network auto-encoder.

[0081] In fig. 4 x is the input RGB-D patch, y the output (reconstructed) RGB-D patch, F the compression feature layer. The decoding stage (lower branch in fig. 4) is not used. Advantageously, only the processing leading from x to F is required to extract a set of features of each patch. Anyway, the set of features of the compression feature layer is sufficiently discriminative to be usable as lookup feature space for a k-NN classifier.

[0082] Fig. 5 shows a depiction of the auto-encoder of fig. 4. In fig. 4 the compressing feature layer is labelled F, the input image patch pairs are labelled x and reconstructed image patch pairs y. Advantageously, the auto-encoder used for feature extraction is trained only once on generic data. It can be applied to a broad variety of previously unseen objects without the need of re-training, thus greatly increasing the speed of adapting this algorithm to new object types.

[0083] A convolutional auto-encoder, i.e. an auto-encoder with a number of convolutional layers, is trained on a set of patches extracted from a large and representative dataset of object data. The data is further augmented by permuting colors and adding noise. This large dataset is used to provide a broad training set for the convolutional auto-encoder.

[0084] Once trained the encoding stage of the auto-encoder is kept and used for the extracting descriptive features from input images. For this the activation signals of the compression layer F is used. Practical tests have shown that using a compression layer with 1024 nodes performs well but this invention is not limited to this number.

[0085] Fig. 6 shows an exemplary scene where an object is recognized and located according to embodiments of the present disclosure. In the example the searched target object O is located substantially in the center of the scene. Votes casting is shown for varying feature space distance thresholds (left and right). The object locations corresponding to the vote centers are indicated in the images. It can be seen how votes accumulate near the target object O. Accordingly, in the present example the higher threshold applied in the right image is more suitable.

[0086] Throughout the description, including the claims, the term "comprising a" should be understood as being synonymous with "comprising at least one" unless otherwise stated. In addition, any range set forth in the description, including the claims should be understood as including its end value(s) unless otherwise stated. Specific values for described elements should be understood to be within accepted manufacturing or industry tolerances known to one of skill in the art, and any use of the terms "substantially" and/or "approximately" and/or "generally" should be understood to mean falling within such accepted tolerances.

[0087] Although the present disclosure herein has been described with reference to particular embodiments, it is to be understood that these embodiments are merely illustrative of the principles and applications of the present disclosure.

[0088] It is intended that the specification and examples be considered as exemplary only, with a true scope of the disclosure being indicated by the following claims.

CLAIMS

1. An electronic device (1) for recognizing and locating an object, the electronic device being configured to:
receive 3D image data of an optical sensor (3), the sensor sensing the object (O),
decompose the 3D image data into a set of patches,
extract a set of features of each patch by using a pre-trained convolutional neural network (CNN) auto-encoder (12),
for each patch classify the object pose by using the set of features of the patch as an input to a k-nearest neighbors algorithm (13) to match against predetermined object representations, and
determine the pose of the object (O) by combining the classified object poses of the set of patches.

2. The electronic device (1) according to claim 1, wherein
the image data comprises a pair of a visible light image and a depth image.

3. The electronic device (1) according to claim 2, wherein
the visible light image comprises the visible part of the electromagnetic spectrum, in particular decomposed into the three bands (RGB) processed by the human vision system.

4. The electronic device (1) according to any one of the preceding claims, wherein the set of features is extracted by using the encoding stage of the convolutional neural network (CNN) auto-encoder, the set of features being in particular output by the encoding stage.

5. The electronic device (1) according to any one of the preceding claims, wherein
the object pose is a 6D pose comprising x, y, z location information and θ, φ, ψ rotation information.

6. The electronic device (1) according to any one of the preceding claims, wherein the features comprise mathematical descriptions of the object appearance, in particular the color and/or brightness of the object.

7. The electronic device (1) according to any one of the preceding claims, wherein the pose of the object is determined by vote filtering using the classified object poses of the set of patches as votes.

8. The electronic device (1) according to any one of the preceding claims, wherein the patches are 3 dimensional portions of the 3D image data, in particular of a predetermined size and/or overlapping each other.

9. The electronic device (1) according to any one of the preceding claims, wherein the convolutional neural network (CNN) auto-encoder is pre-trained based on a generic set of patches, in particular once only.

10. The electronic device (1) according to any one of the preceding claims, wherein the convolutional neural network (CNN) auto-encoder is configured to reduce the patch to a feature vector f(x), in particular of 1024 features.

11. The electronic device (1) according to the preceding claim, wherein the k-nearest neighbors algorithm is configured to find the most likely object pose in the feature space defined by the feature vector f(x).

12. The electronic device (1) according to any one of the preceding claims, wherein the k-nearest neighbors algorithm is pre-trained based on data of the object.

13. A system (30) for recognizing and locating an object, the system comprising:
an electronic (1) device according to any one of the preceding claims, and
an optical sensor configured to sense the object, the sensor being in particular a 3D camera or a stereo camera.

14. A method of recognizing and locating an object,
the method comprising the steps of:
receiving 3D image data of an optical sensor, the sensor sensing the object (O),
decomposing the 3D image data into a set of patches,
extracting a set of features of each patch by using a pre-trained convolutional neural network (CNN) auto-encoder (12),
for each patch classifying the object pose by using the set of features of the patch as an input to a k-nearest neighbors algorithm (13) to match against predetermined object representations, and
determining the pose of the object (O) by combining the classified object poses of the set of patches.

15. The method according to claim 14, wherein
the set of features is extracted by using the encoding stage of the convolutional neural network (CNN) auto-encoder, the set of features being in particular output by the encoding stage.

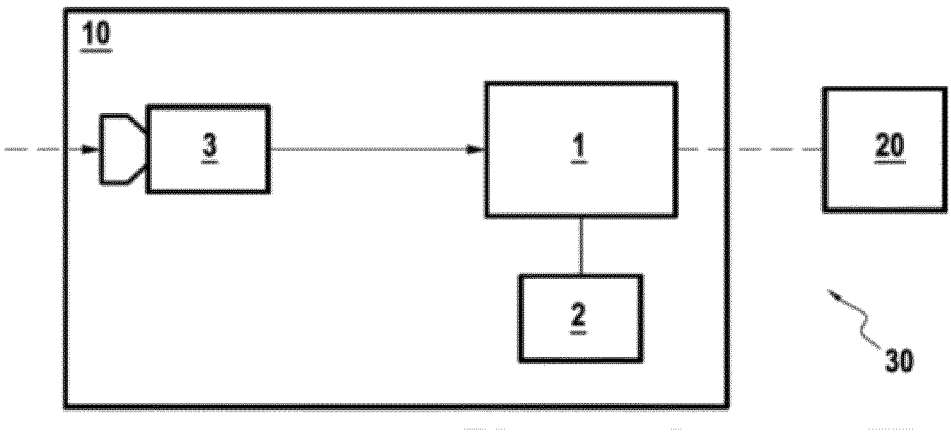

10
3
1
20
2
30

FIG.1

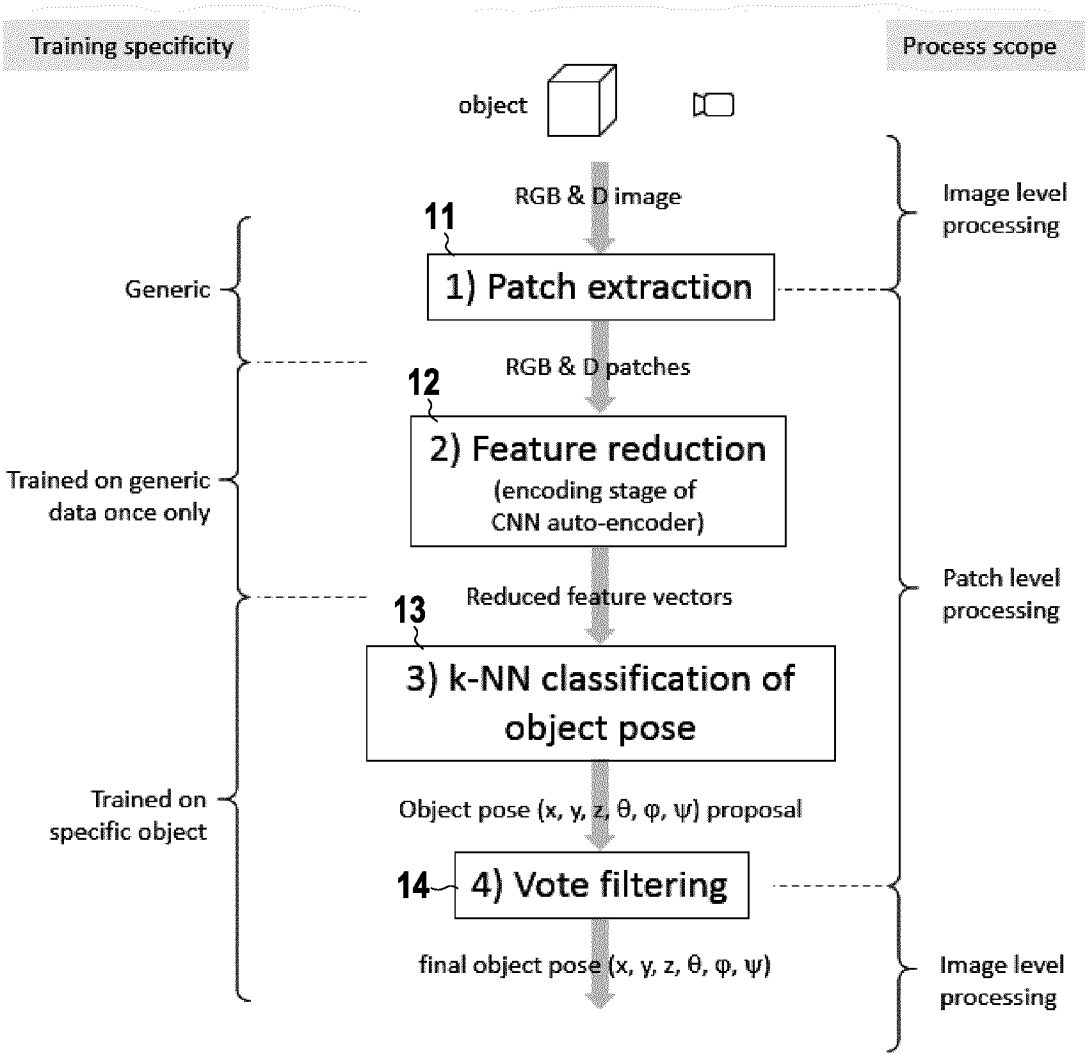

Training specificity
Process scope
object
RGB & D image
11
1) Patch extraction
Image level processing
Generic
RGB & D patches
12
2) Feature reduction
(encoding stage of CNN auto-encoder)
Trained on generic data once only
Reduced feature vectors
13
3) k-NN classification of object pose
Patch level processing
Trained on specific object
Object pose (x, y, z, θ, φ, ψ) proposal
14
4) Vote filtering
final object pose (x, y, z, θ, φ, ψ)
Image level processing

FIG.2

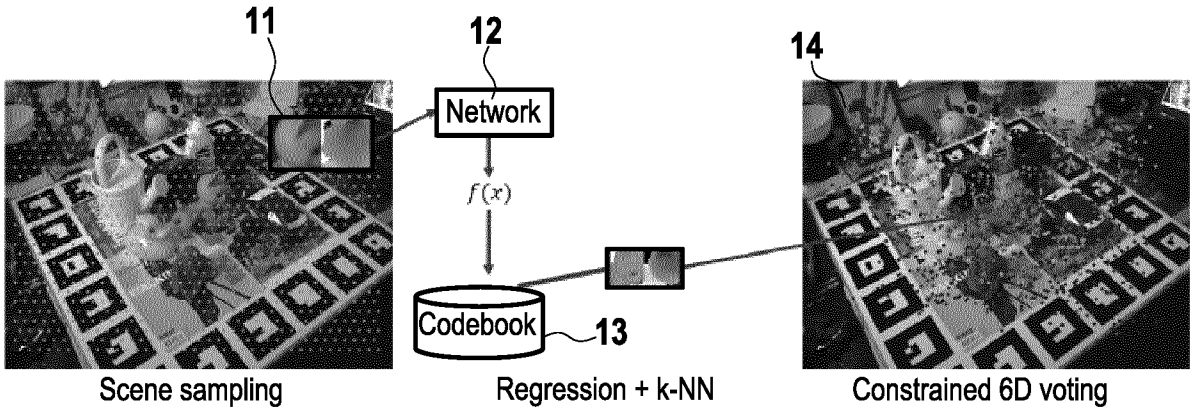

11
12
14
Network
$f(x)$
Codebook
13
Scene sampling
Regression + k-NN
Constrained 6D voting

FIG.3

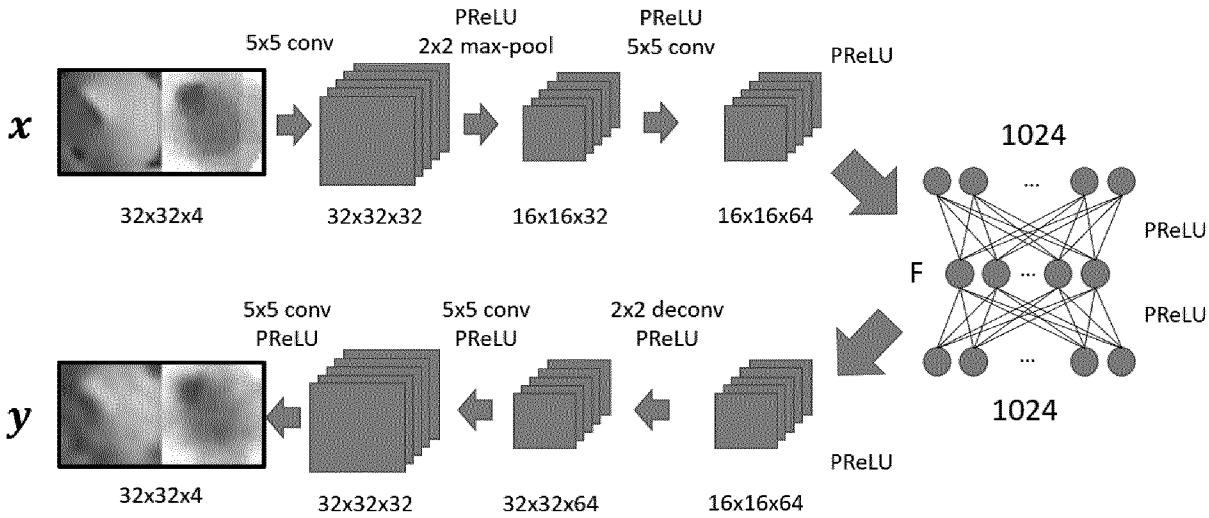

x
5x5 conv
PReLU
2x2 max-pool
PReLU
5x5 conv
PReLU
32x32x4
32x32x32
16x16x32
16x16x64
1024
F
PReLU
PReLU
1024
y
5x5 conv
PReLU
5x5 conv
PReLU
2x2 deconv
PReLU
PReLU
32x32x4
32x32x32
32x32x64
16x16x64

FIG.4

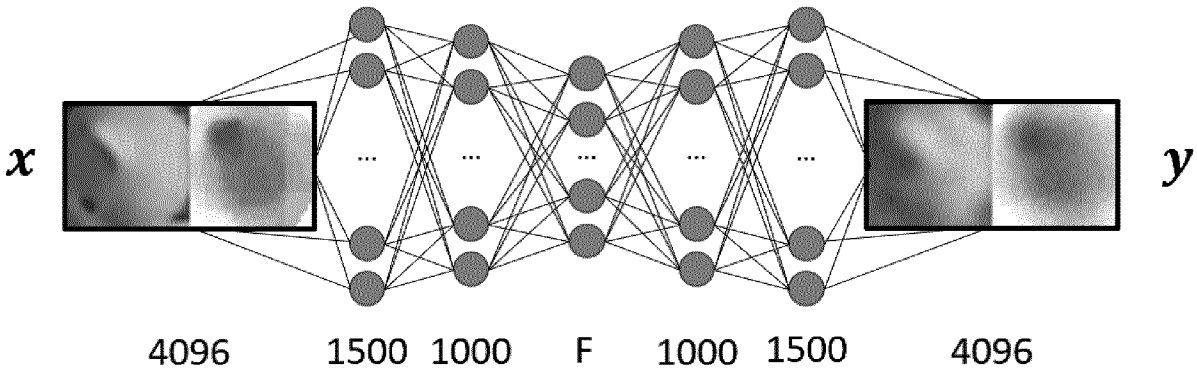


FIG.5

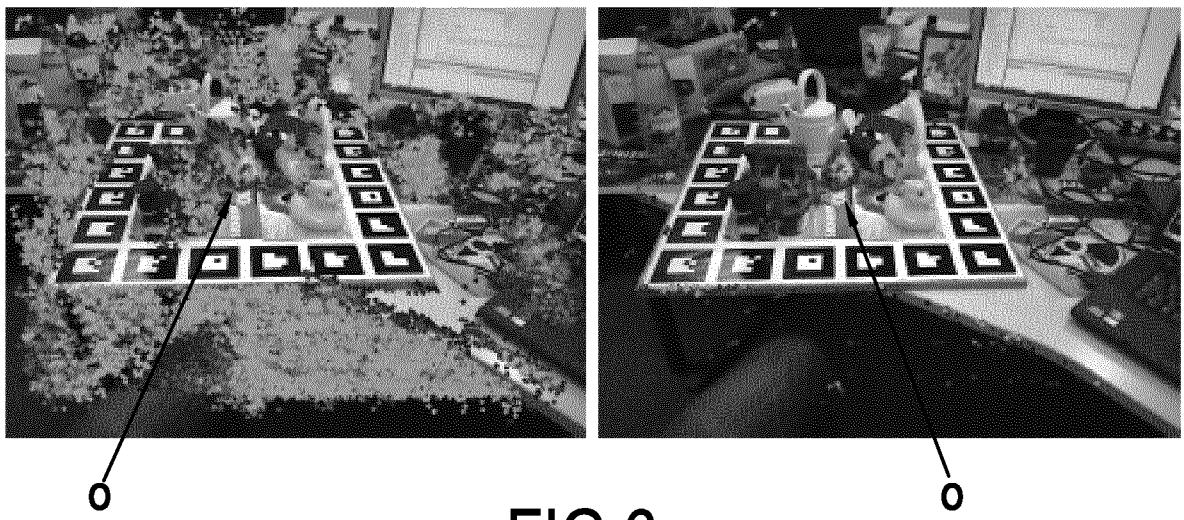


FIG.6

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2016/074088

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06K9/00 G06K9/46 G06K9/62
ADD.

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
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)
EPO-Internal, INSPEC, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	KEHL WADIM ET AL: "Deep Learning of Local RGB-D Patches for 3D Object Detection and 6D Pose Estimation", 17 September 2016 (2016-09-17), NETWORK AND PARALLEL COMPUTING; [LECTURE NOTES IN COMPUTER SCIENCE; LECT.NOTES COMPUTER], SPRINGER INTERNATIONAL PUBLISHING, CHAM, PAGE(S) 205 - 220, XP047355337, ISSN: 0302-9743 ISBN: 978-3-642-38529-2 [retrieved on 2016-09-17] Section 1 Sections 3.1-3.3 Sections 4.1-4.2 figures 1-6 ----- -/--	1-15

[X] Further documents are listed in the continuation of Box C. [ ] See patent family annex.

* Special categories of cited documents :

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

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

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

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

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

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

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

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

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
26 May 2017	02/06/2017

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Moreno, Marta

1

Form PCT/ISA/210 (second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2016/074088

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	MANUEL BLUM ET AL: "A learned feature descriptor for object recognition in RGB-D data", IEEE INTERNATIONAL CONFERENCE ON ROBOTICS AND AUTOMATION (ICRA), 14 May 2012 (2012-05-14), pages 1298-1303, XP032450919, IEEE, Piscataway, NJ, USA DOI: 10.1109/ICRA.2012.6225188 ISBN: 978-1-4673-1403-9 Section I. Section III. -----	1-3,5,6, 8,10,11, 13,14
A	MAX SCHWARZ ET AL: "RGB-D object recognition and pose estimation based on pre-trained convolutional neural network features", 2015 IEEE INTERNATIONAL CONFERENCE ON ROBOTICS AND AUTOMATION (ICRA), 1 May 2015 (2015-05-01), pages 1329-1335, XP55376263, DOI: 10.1109/ICRA.2015.7139363 ISBN: 978-1-4799-6923-4 Section I. Section III., first paragraph Section IV. -----	1-3,5,6, 9,10,13, 14

1