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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
18 May 2017 (18.05.2017)

WIPO | PCT

(10) International Publication Number
WO 2017/079972 A1

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

(21) International Application Number:
PCT/CN2015/094572

(22) International Filing Date:
13 November 2015 (13.11.2015)

(25) Filing Language: English

(26) Publication Language: English

(72) Inventor; and
(71) Applicant : **WANG, Xiaogang** [CN/CN]; Department of Electrical Engineering, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong (CN).

(72) **Inventors**: **ZENG, Xingyu**; Department of Electronic Engineering, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong (CN). **OUYANG, Wanli**; Department of Electrical Engineering, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong (CN).

(74) **Agent**: **INSIGHT INTELLECTUAL PROPERTY LIMITED**; No. 48A Zhichun Road, Haidian District, Beijing 100098 (CN).

(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, 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, 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, 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))*

(54) **Title**: A METHOD AND A SYSTEM FOR CLASSIFYING OBJECTS IN IMAGES

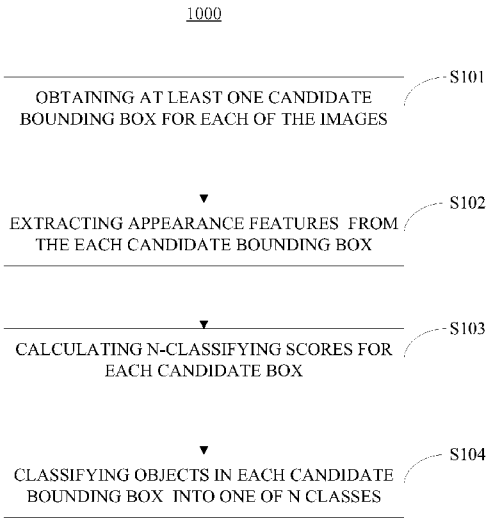

Fig. 1

(57) **Abstract**: Disclosed is a method for classifying objects in images, comprising: obtaining at least one candidate bounding box for each of the images, wherein each candidate bounding box contains one object to be classified; extracting appearance features from the each candidate bounding box; calculating n-classifying scores for each candidate box based on the extracted appearance features; and classifying objects in each candidate bounding box into one of n classes, where n is integer greater than 1.

WO 2017/079972 A1

A METHOD AND A SYSTEM FOR CLASSIFYING OBJECTS IN IMAGES

Technical Field

[0001] The disclosures relate to a method and a system for classifying objects in images.

Background

[0002] The aim of object detection is to detect instances of object of a certain class in digital images and videos. The performance of object detection systems depends heavily on image representation, of which the quality can be influenced by many kinds of variations, such as viewpoints, illuminations, poses, and occlusions. Due to such uncontrollable factors, it is non-trivial to design robust image representation that is sufficiently discriminative to represent large quantity of object classes.

[0003] Substantial efforts have been dedicated to design hand-crafted features. Typically, object detection based on hand-drafted features involves extracting multiple features on the landmarks of images with multiple scales, and concatenating them into high-dimensional feature vectors.

[0004] Deep Convolutional Neural Network (CNN) has been applied to learn features directly from raw pixels. For the object detection task, existing deep CNN learning methods pre-train the CNN by using images without bounding box ground truth, and subsequently fine-tune the deep neural net using another set of images with bounding box ground truth. Typically, the image set used for fine-tuning has lower quantity of semantic class numbers compared to the image set used for pre-training. In addition, the number of semantic classes in image set for fine-tuning equals the number of actual classes to be detect.

Summary

[0005] The following presents a simplified summary of the disclosure in order to provide a

basic understanding of some aspects of the disclosure. This summary is not an extensive overview of the disclosure. It is intended to neither identify key or critical elements of the disclosure nor delineate any scope of particular embodiments of the disclosure, or any scope of the claims. Its sole purpose is to present some concepts of the disclosure in a simplified form as a prelude to the more detailed description that is presented later.

[0006] In an aspect, disclosed is a method for classifying objects in images, comprising: obtaining at least one candidate bounding box for each of the images, wherein each candidate bounding box contains one object to be classified; extracting appearance features from the each candidate bounding box; calculating n-classifying scores for each candidate box based on the extracted appearance features; and classifying objects in each candidate bounding box into one of n classes, where n is integer greater than 1.

[0007] In another aspect of the present application, disclosed is method for classifying objects in images, comprising: means for obtaining at least one candidate bounding box for each of the images, wherein each candidate bounding box contains one object to be classified; means for extracting appearance features from the each candidate bounding box; means for calculating n-classifying scores for each candidate box based on the extracted appearance features; and means for classifying objects in each candidate bounding box into one of n classes, where n is integer greater than 1.

[0008] In another aspect of the present application, disclosed is a system for classifying objects in images, comprising: a memory that stores executable components; and a processor that executes the executable components to perform operations of the system, the executable components comprise: obtaining at least one candidate bounding box for each of the images, wherein each candidate bounding box contains one object to be classified; extracting appearance features from the each candidate bounding box; calculating n-classifying scores for each candidate box based on the extracted appearance features; and classifying objects in each candidate bounding box into one of n classes, where n is integer greater than 1.

[0009] In one embodiment of the present application, the appearance features are extracted from the each candidate bounding box by using a feature learning CNN. The CNN may be trained by steps of: training a second CNN with a first output layer having a first number of nodes, and a second output layer having integral-multiple-times (for example, 4 times) of the first number of nodes; initializing a third CNN with neuro parameters of the trained second CNN; training the third CNN with a set of fine-tuning images; adding a second number of output layers for the trained third CNN, wherein each of the second number of output layer is used for n-classes classification; initializing a fourth CNN with neuro parameters of the third CNN; replacing an output layer of the fourth CNN with a new output layer with n nodes; and fine-tuning the forth CNN with the set of fine-tuning images to obtain the feature learning CNN.

[0010] For example, the second CNN may be trained by: initializing the second CNN with neuro parameters of a first CNN pre-trained with a set of pre-training images; replacing, with two new layers, an output layer of the second CNN; and training the second CNN with the set of fine-tuning images.

[0011] In one embodiment of the present application, each of the pre-training images may have at least one ground truth bounding box, and each of the fine-training images may have at least one ground truth bounding box and at least one candidate bounding box.

[0012] In one embodiment of the present application, the first number may be determined by clustering the fine-training images according to relative locations and size relationships of each candidate bounding box and the corresponding ground truth box in each of the fine-training images.

[0013] In one embodiment of the present application, the second number may be determined by clustering the set of fine-training images according to relative locations and size relationships of the ground truth boxes in each of the fine-training images.

Brief Description of the Drawing

[0014] Exemplary non-limiting embodiments of the present invention are described below with reference to the attached drawings. The drawings are illustrative and generally not to an exact scale. The same or similar elements on different figures are referenced with the same reference numbers.

[0015] Fig. 1 illustrates method for classifying objects in images according to an embodiment of the present application.

[0016] Fig. 2 is a flowchart illustrates a method for training the feature learning CNN.

[0017] Fig. 3 illustrates an example CNN according to the present application.

[0018] Fig. 4 and 5 illustrates exemplary images presenting the difficulty to tell object category or where the ground truth bounding box is if the only the image region within the candidate window is provided.

[0019] Fig. 6 illustrates exemplary images of window-object relationship clusters obtained.

[0020] Fig. 7 illustrates a system for classifying objects in images according to an embodiment of the present application.

Detailed Description

[0021] Reference will now be made in detail to some specific embodiments of the invention including the best modes contemplated by the inventors for carrying out the invention. Examples of these specific embodiments are illustrated in the accompanying drawings. While the invention is described in conjunction with these specific embodiments, it will be understood that it is not intended to limit the invention to the described embodiments. On the contrary, it is intended to cover alternatives, modifications, and equivalents as may be included within the spirit and scope of the invention as defined by the appended claims. In the

following description, numerous specific details are set forth in order to provide a thorough understanding of the present invention. The present invention may be practiced without some or all of these specific details. In other instances, well-known process operations have not been described in detail in order not to unnecessarily obscure the present invention.

[0022] The terminology used herein is for the purpose of describing particular embodiments only and is not intended to be limiting of the invention. As used herein, the singular forms "a", "an" and "the" are intended to include the plural forms as well, unless the context clearly indicates otherwise. It will be further understood that the terms "comprises" and/or "comprising," when used in this specification, specify the presence of stated features, integers, steps, operations, elements, and/or components, but do not preclude the presence or addition of one or more other features, integers, steps, operations, elements, components, and/or groups thereof.

[0023] Fig. 1 illustrates method 1000 for classifying objects in images according to an embodiment of the present application.

[0024] As shown, at step S101, at least one candidate bounding box are obtained for each of the images, wherein each candidate bounding box contains one object to be classified. At step S102, one or more appearance features are extracted from the each candidate bounding box. At step S103, n-classifying scores are calculated for each candidate box based on the extracted appearance features. At step S104, objects in each candidate bounding box are classified into one of n classes, where n is integer greater than 1.

[0025] The extracted appearance features in step S102 are used to classify each candidate window, for example by C binary-class Support Vector Machine (SVM) detectors, each of which predicts one value for one candidate bounding box for one semantic object class. The score of each SVM detector measures the confidence on the candidate window containing a specific object class. The feature extracting step S102 can be performed with a feature learning CNN which will be described in detail below.

[0026] The feature learning CNN

[0027] Fig. 2 is a flowchart illustrates a method 2000 for training the feature learning CNN.

[0028] As shown in Fig. 2, at step S201, a first CNN is pre-trained with a set of pre-training images. At step S202, a second CNN is initialized with the neuro parameters of the pre-trained first CNN. At step S203, the output layer of the second CNN is replaced with two new layers, wherein one of the new output layer has c nodes, and the other output layer having $4c$ nodes. At step S204, the second CNN are trained with the set of fine-tuning images.

[0029] In one embodiment of the application, each of the pre-training images has at least one ground truth bounding box, and each of the fine-training images has at least one ground truth bounding box and at least one candidate bounding box. In the context of the application, for example, a candidate bounding box may contain the left/right half of a person if the candidate bounding box is at the left/right of the person. As the other examples, the candidate bounding box may contain the upper body or the legs of the person or the whole body. Compared with the ground truth bounding box, the candidate bounding box may have a larger scale with more background regions included or have a smaller scale with only part of the object included. The relationship between the candidate bounding box and the ground truth bounding box of the object, which is called the candidate-ground truth relationship in this application, provide rich information. The candidate-ground truth relationship is available with the locations and sizes of the candidate bounding box and the ground truth bounding box.

[0030] The two new output layers of the second CNN consist of a c-classes classification and a 4c regression. The label for the c-classes classification is the cluster index for each box proposal. The label for the 4c regression is the l_{loc} for each cluster. The value of c is determined by clustering the fine-training images according to relative locations and size relationships of each candidate bounding box and the corresponding ground truth box in each of the fine-training images.

[0031] For example, the overlaps between the candidate bounding box and the ground truth bounding box thereof are calculated. Those samples with overlap larger than 0.5 are assigned a positive label. Others are assigned a negative label. The loss function for training the network will take the assigned label to calculate loss for each sample. And the goal of back-propagation procedure is to minimize the loss function for all samples.

[0032] The relative location and size is formulated as

$$l_{loc}=[(x_s-x_g)/W_s,(y_s-y_g)/H_s, \log(W_s/W_g),\log(H_s/H_g)] \quad \text{equation (1)}$$

wherein, (x_s,y_s) denotes the center coordinate of the candidate bounding box; (W_s,H_s) denotes width and height of a box proposal; (x_g,y_g) denotes the center coordinate of the ground truth bounding box; and (W_g,H_g) denotes the width and height of the ground truth bounding box.

[0033] At step S205, a third CNN is initialized with neuro parameters of the trained second CNN. At step S206, the third CNN is trained with a set of fine-tuning images. At step S207, k output layers are added for the trained third CNN, wherein each of the second number of output layer is used for n-classes classification.

[0034] At step S208, a fourth CNN with neuro parameters of the pre-trained third CNN is initialized. At step S209, an output layer of the fourth CNN is replaced with a new output layer with n nodes. At step S210, the forth CNN is fine-tuned with the set of fine-tuning images to obtain the feature learning CNN.

[0035] In one embodiment of the application, the value of k is determined by clustering the set of fine-training images according to relative locations and size relationships of the ground truth boxes in each of the fine-training images, which can be formulated as:

$$l_{loc}=[(x_{g1}-x_{g2})/W_{g1},(y_{g1}-y_{g2})/H_{g1}, \log(W_{g1}/W_{g2}),\log(H_{g1}/H_{g2})] \quad \text{equation (2)}$$

wherein, (x_{g1},y_{g1}) denotes the center coordinate of the first ground truth box; (W_{g1},H_{g1}) denotes width and height of the first ground truth box; (x_{g2},y_{g2}) denotes the center coordinate of the second ground truth box; and (W_{g2},H_{g2}) denotes the width and height of the second ground truth box.

[0036] Fig 3 shows an example CNN according to the present application. As shown, the CNN structure consists of a data layers, several polling layers, convolutional layers and fully connected layers.

[0037] The data layer receives images $\{x_{i,j}\}_{i=1,j=1}^{k,d}$ and its labels $\{y_{i,j}\}_{i=1,j=1}^{k,n}$, where $x_{i,j}$ is the *j*-th bit value of the *d*-dimension feature vector of the *i*-th input image region, $y_{i,j}$ is the *j*-th bit value of the *n* dimension label vector of the *i*-th input image region.

[0038] The convolution layer receives the output from the data layer and performs convolution, padding, sampling, and non-linear transformation operations.

[0039] The fully connected layer takes the fully connected layer takes the output from the convolution layer and performs inner-product operation. It aims to project the input to final prediction. This prediction can be the probability for each class, and can also be the relative location.

[0040] The trained CNN can be back-propagated with any known art for neural network training. The training process can try to minimize the loss function and can be divided into many updating steps in which the algorithm calculates the gradient with respect to all the neural network parameters, including the convolution filters, deformational layer weights, fully connected weights, and bias. Updates of the parameters can be performed using the production of one prefixed learning rate and the corresponding gradients.

[0041] In one embodiment of the application, the candidate windows for extracting appearance features are processed as following.

[0042] Firstly, for the candidate window b_s=(x, y, W, H) with size (W, H) and center (x, y), crop images I (λ, b_s) with sizes (λW, λH), $\lambda \in \Lambda$ and center (x, y). The cropped images and the candidate window have the same center location (x, y). λ is the scale of contextual region. Choice of the scale set Λ will be discussed in detailed later.

[0043] Secondly, the cropped image are rotated by degrees $r \in R$ to obtain I (r, λ, b_s), for example, the value of R can be R= {0°, 45°, 90°}.

[0044] Thirdly, the cropped images I (r, λ, b_s) with different sizes and rotations are warped into the same size and treated as the input of the feature learning CNN for extracting their features, i.e., $f_{r,\lambda} = f(r, \lambda, I(r, \lambda, b_s))$ where $f(r, \lambda, *)$ denotes the CNN for extracting features from $I(r, \lambda, b_s)$, $f_{r,\lambda}$ denotes the vector of features extracted for rotation r and scale λ. In one embodiment of the application, for the candidate window b_s, there are six cropped images I (r, λ, b_s) with (r, λ) being (0°, 0.8), (0 , 1.2), (45°, 1.2), (90°, 1.2), (0°, 1.8), and (0°, 2.7). To be understand, the approaches of the application is not constrained to network structure.

[0045] Finally, the extracted features are then concatenated into $F = concat_{(r,\lambda)}\{f_{r,\lambda}\}$, where $concat_{(r,\lambda)}$ is the operation for concatenating features into a vector.

[0046] When the location and size of the candidate bounding box is different from that of the ground truth bounding box, the candidate bounding box only have partial visual content of the object. The limited view results in difficulty for CNN to figure out the visual difference between object classes. For example, it is hard to tell whether it is an ipod or a monitor if one can only see the screen, but it becomes much easier if the whole object and its contextual region is provided, as shown in Fig. 4. As another problem, occlusion frequently happens. Thus the ground truth bounding boxes of an object class may contain different amount of object parts and thus different bounding box sizes. Without a region larger than the ground truth as input, it is confusing for the network to decide the bounding box size. As shown in

Fig. 5, the ground truth box for a standing un-occluded person should cover more parts of human body than the one with legs occluded. When the image cropped from candidate bounding box only covers the upper body of this person, it is difficult to predict whether the person's legs are occluded or not. CNN should output a smaller box if occluded, but a larger box otherwise. CNN can handle this difficulty when the input contains a larger region than the ground truth.

[0047] To handle the problems above, multiple scales of contextual regions are used as the input for CNN. In one embodiment of the application, 4 scales for cropping images, 0.8, 1.2, 1.8, 2.7, which are linear in log scale are used. 1.2 are set as default value. The cropped image with scale 2.7 has been proved to be sufficient in covering most of the ground-truth regions when the overlap between the window and the object is greater than 0.5. Even if the overlap between the candidate bounding box and ground truth bounding box is 0.37, the cropped image with scale 2.7 is able to cover more than 50% of the ground truth region. 1.8 is obtained by linear interpolation between 1.2 and 2.7 in log scale. 0.8 is chosen because some candidate bounding boxes can be larger than the ground truth bounding box, as shown by the first image in Fig.6. A cropped image with a smaller scale can help these windows to fit the actual scale of the object.

[0048] In the case that the method for classifying objects in images as disclosed above is implemented with software, the system 7000 for classifying objects in images may include a general purpose computer, a computer cluster, a mainstream computer, a computing device dedicated for providing online contents, or a computer network comprising a group of computers operating in a centralized or distributed fashion. As shown in Fig. 7, the system 7000 may include one or more processors (processors 702, 704, 706 etc.), a memory 712, a storage device 716, a communication interface 714, and a bus to facilitate information exchange among various components of system 7000. Processors 702-706 may include a central processing unit ("CPU"), a graphic processing unit ("GPU"), or other suitable information processing devices. Depending on the type of hardware being used, processors

702-706 can include one or more printed circuit boards, and/or one or more microprocessor chips. Processors 702-706 can execute sequences of computer program instructions to perform various methods or run the modules that implement the function for classifying objects in images.

[0049] Memory 712 can include, among other things, a random access memory ("RAM") and a read-only memory ("ROM"). Computer program instructions can be stored, accessed, and read from memory 712 for execution by one or more of processors 702-706. For example, memory 712 may store one or more software applications. Further, memory 712 may store an entire software application or only a part of a software application that is executable by one or more of processors 702-706 to carry out the functions as disclosed above for classifying objects in images. It is noted that although only one block is shown in Fig. 7, memory 712 may include multiple physical devices installed on a central computing device or on different computing devices.

[0050] Although the preferred examples of the present invention have been described, those skilled in the art can make variations or modifications to these examples upon knowing the basic inventive concept. The appended claims are intended to be considered as comprising the preferred examples and all the variations or modifications fell into the scope of the present invention.

[0051] Obviously, those skilled in the art can make variations or modifications to the present invention without departing the spirit and scope of the present invention. As such, if these variations or modifications belong to the scope of the claims and equivalent technique, they may also fall into the scope of the present invention.

What is claimed is:

1. A method for classifying objects in images, comprising:

obtaining at least one candidate bounding box for each of the images, wherein each candidate bounding box contains one object to be classified;

extracting appearance features from the each candidate bounding box;

calculating n-classifying scores for each candidate box based on the extracted appearance features; and

classifying objects in each candidate bounding box into one of n classes, where n is integer greater than 1.

2. The method of claim 1, wherein the extracting is performed with a feature learning CNN trained by steps of:

training a second CNN with a first output layer having a first number of nodes, and a second output layer having integral-multiple-times of the first number of nodes;

initializing a third CNN with neuro parameters of the trained second CNN;

training the third CNN with a set of fine-tuning images;

adding a second number of output layers for the trained third CNN, wherein each of the second number of output layer is used for n-classes classification;

initializing a fourth CNN with neuro parameters of the third CNN;

replacing an output layer of the fourth CNN with a new output layer with n nodes; and

fine-tuning the forth CNN with the set of fine-tuning images to obtain the feature learning CNN.

3. The method of claim 2, wherein the second CNN is trained by:

initializing the second CNN with neuro parameters of a first CNN pre-trained with a set of pre-training images;

replacing, with two new neuro layers, an output layer of the second CNN; and

training the second CNN with the set of fine-tuning images.

4. The method of claim 3, wherein each of the pre-training images has at least one ground truth bounding box, and each of the fine-training images has at least one ground truth bounding box and at least one candidate bounding box.

5. The method of claim 4, wherein the first number is determined by clustering the fine-training images according to relative locations and size relationships of each candidate bounding box and the corresponding ground truth box in each of the fine-training images.

6. The method of claim 4, wherein the second number is determined by clustering the set of fine-training images according to relative locations and size relationships of the ground truth boxes in each of the fine-training images.

7. The method of claim 4, further comprising:

calculating a loss function based on overlaps between the ground truth bounding boxes and the corresponding candidate bounding boxes in each of the fine-training images; and

back-propagating at least one of the first CNN, the second CNN and the third CNN until the loss function is minimized so as to solidify neuro parameters for the first, second and third CNNs.

8. The method of claim 1, wherein, the method further comprises before the extracting:

cropping each of the candidate bounding boxes; and

rotating the cropped boxes.

9. The method of claim 1, wherein, the method further comprises before the extracting:

scaling each of the candidate bounding boxes with a factor before extracting features.

10. A system for classifying objects in images, comprising:

means for obtaining at least one candidate bounding box for each of the images, wherein each candidate bounding box contains one object to be classified;

means for extracting appearance features from the each candidate bounding box;

means for calculating n-classifying scores for each candidate box based on the extracted appearance features; and

means for classifying objects in each candidate bounding box into one of n classes, where n is integer greater than 1.

11. The system of claim 10, wherein the means for extracting comprises a feature learning CNN trained by steps of:

training a second CNN with a first output layer having a first number of nodes, and a second output layer having 4 times of the first number of nodes;

initializing a third CNN with neuro parameters of the trained second CNN;

training the third CNN with a set of fine-tuning images;

adding a second number of output layers for the trained third CNN, wherein each of the second number of output layer is used for n-classes classification;

initializing a fourth CNN with neuro parameters of the third CNN;

replacing an output layer of the fourth CNN with a new output layer with n nodes; and

fine-tuning the forth CNN with the set of fine-tuning images to obtain the feature learning CNN.

12. The system of claim 11, wherein the second CNN is trained by:

initializing the second CNN with neuro parameters of a first CNN pre-trained with a set of pre-training images;

replacing, with two new layers, an output layer of the second CNN； and

training the second CNN with the set of fine-tuning images.

13. The system of claim 12, wherein each of the pre-training images has at least one ground truth bounding box, and each of the fine-training images has at least one ground truth bounding box and at least one candidate bounding box.

14. The system of claim 13, wherein the first number is determined by clustering the fine-training images according to relative locations and size relationships of each candidate bounding box and the corresponding ground truth box in each of the fine-training images.

15. The system of claim 13, wherein the second number is determined by clustering the set of fine-training images according to relative locations and size relationships of the ground truth boxes in each of the fine-training images.

16. The system of claim 12, further comprising:

means for calculating a loss function based on overlaps between the ground truth bounding boxes and the corresponding candidate bounding boxes in each of the fine-training images; and

means for back-propagating at least one of the first CNN, the second CNN and the third CNN until the loss function is minimized so as to solidify neuro parameters for the first, second and third CNNs.

17. The system of claim 10,further comprising:

means for cropping each of the candidate bounding boxes before the extracting; and

means for rotating the cropped boxes.

18. The system of claim 10, further comprising:

means for, before the extracting, scaling each of the candidate bounding boxes with a factor before extracting features.

19. A system for classifying objects in images, comprising:

a memory that stores executable components; and

a processor that executes the executable components to:

obtain at least one candidate bounding box for each of the images, wherein each candidate bounding box contains one object to be classified;

extract appearance features from the each candidate bounding box;

calculate n-classifying scores for each candidate box based on the extracted appearance features; and

classify objects in each candidate bounding box into one of n classes, where n is integer greater than 1.

20. The system of claim 19, wherein the means for extracting comprises a feature learning CNN trained by steps of:

training a second CNN with a first output layer having a first number of nodes, and a second output layer having 4 times of the first number of nodes;

initializing a third CNN with neuro parameters of the trained second CNN;

training the third CNN with a set of fine-tuning images;

adding a second number of output layers for the trained third CNN, wherein each of the second number of output layer is used for n-classes classification;

initializing a fourth CNN with neuro parameters of the third CNN;

replacing an output layer of the fourth CNN with a new output layer with n nodes; and

fine-tuning the forth CNN with the set of fine-tuning images to obtain the feature learning CNN.

21. The system of claim 20, wherein the second CNN is trained by:

initializing the second CNN with neuro parameters of a first CNN pre-trained with a set of pre-training images;

replacing, with two new layers, an output layer of the second CNN； and

training the second CNN with the set of fine-tuning images.

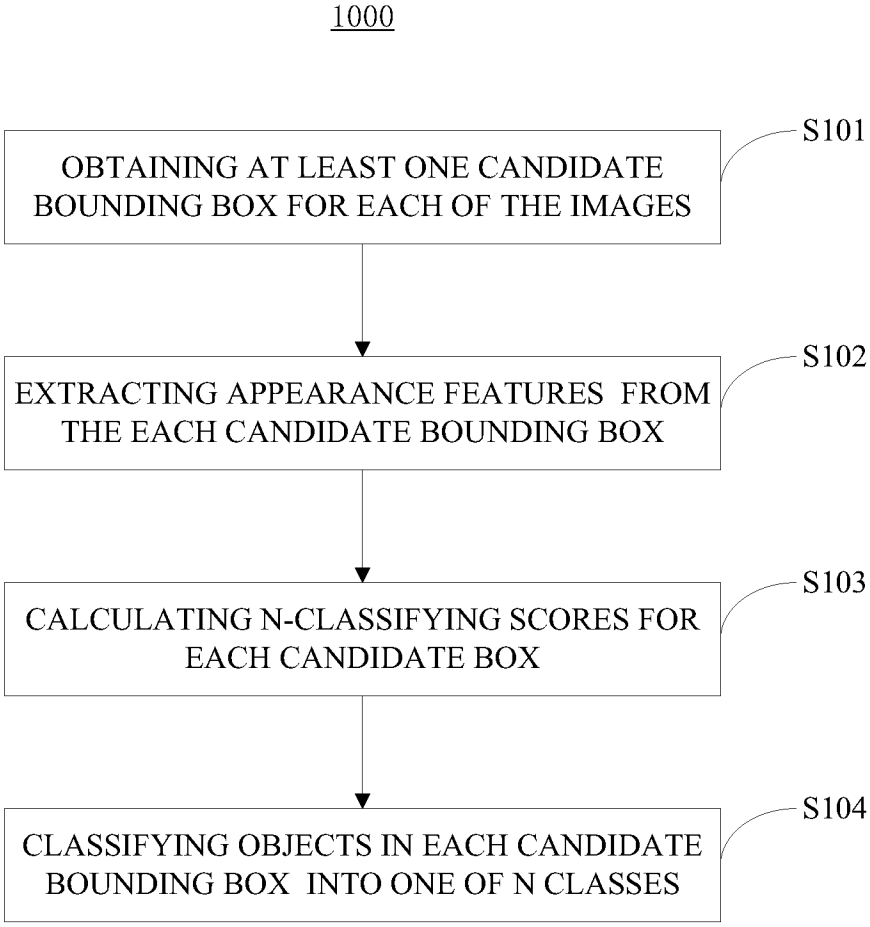
1000
OBTAINING AT LEAST ONE CANDIDATE BOUNDING BOX FOR EACH OF THE IMAGES
S101
EXTRACTING APPEARANCE FEATURES FROM THE EACH CANDIDATE BOUNDING BOX
S102
CALCULATING N-CLASSIFYING SCORES FOR EACH CANDIDATE BOX
S103
CLASSIFYING OBJECTS IN EACH CANDIDATE BOUNDING BOX INTO ONE OF N CLASSES
S104

Fig. 1

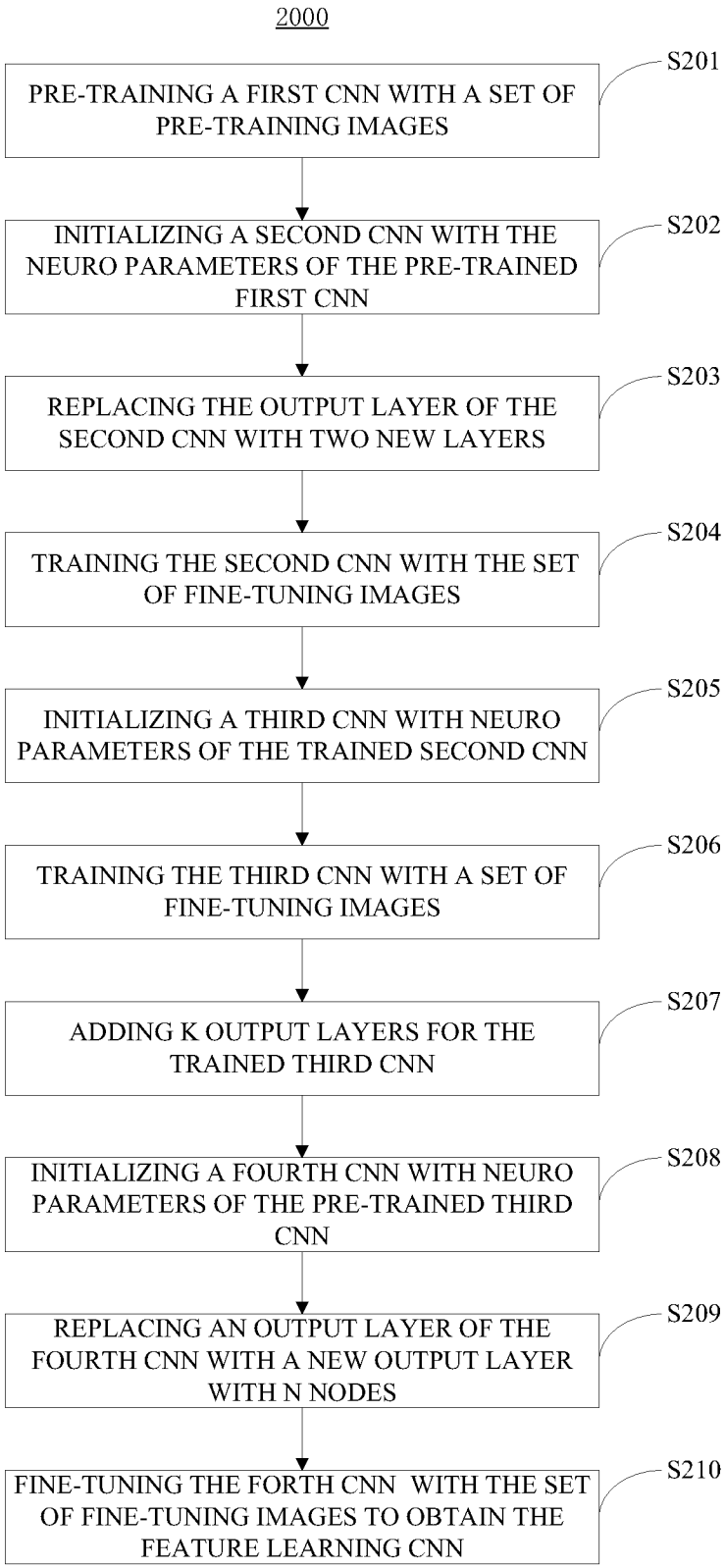
2000
PRE-TRAINING A FIRST CNN WITH A SET OF PRE-TRAINING IMAGES
S201
INITIALIZING A SECOND CNN WITH THE NEURO PARAMETERS OF THE PRE-TRAINED FIRST CNN
S202
REPLACING THE OUTPUT LAYER OF THE SECOND CNN WITH TWO NEW LAYERS
S203
TRAINING THE SECOND CNN WITH THE SET OF FINE-TUNING IMAGES
S204
INITIALIZING A THIRD CNN WITH NEURO PARAMETERS OF THE TRAINED SECOND CNN
S205
TRAINING THE THIRD CNN WITH A SET OF FINE-TUNING IMAGES
S206
ADDING K OUTPUT LAYERS FOR THE TRAINED THIRD CNN
S207
INITIALIZING A FOURTH CNN WITH NEURO PARAMETERS OF THE PRE-TRAINED THIRD CNN
S208
REPLACING AN OUTPUT LAYER OF THE FOURTH CNN WITH A NEW OUTPUT LAYER WITH N NODES
S209
FINE-TUNING THE FORTH CNN WITH THE SET OF FINE-TUNING IMAGES TO OBTAIN THE FEATURE LEARNING CNN
S210

Fig. 2

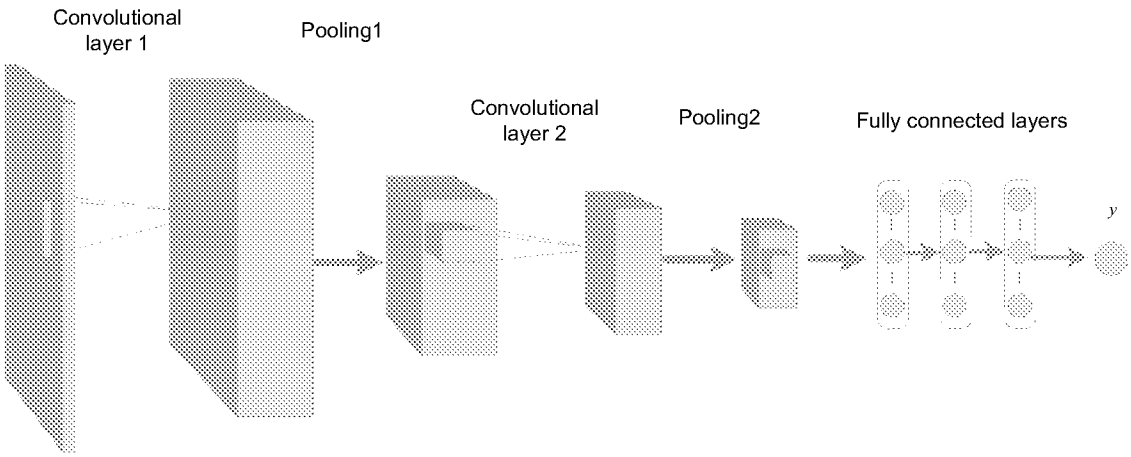
Convolutional layer 1
Pooling1
Convolutional layer 2
Pooling2
Fully connected layers
y

Fig. 3

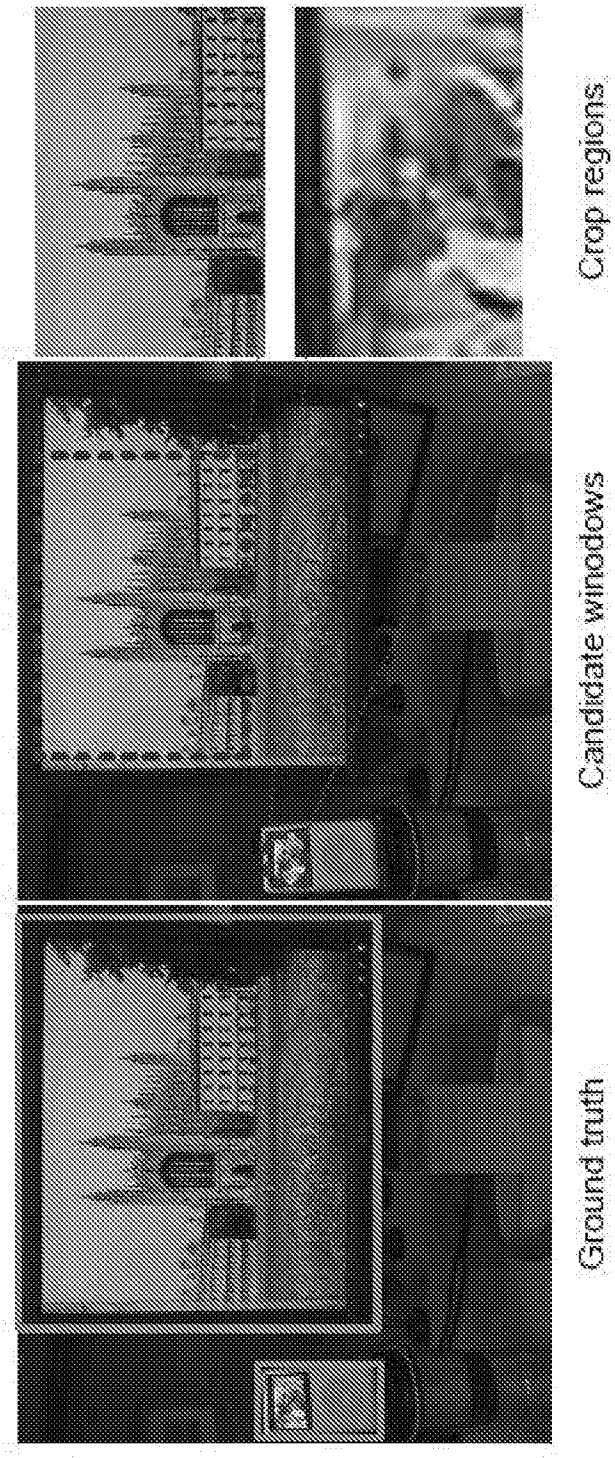

Fig. 4

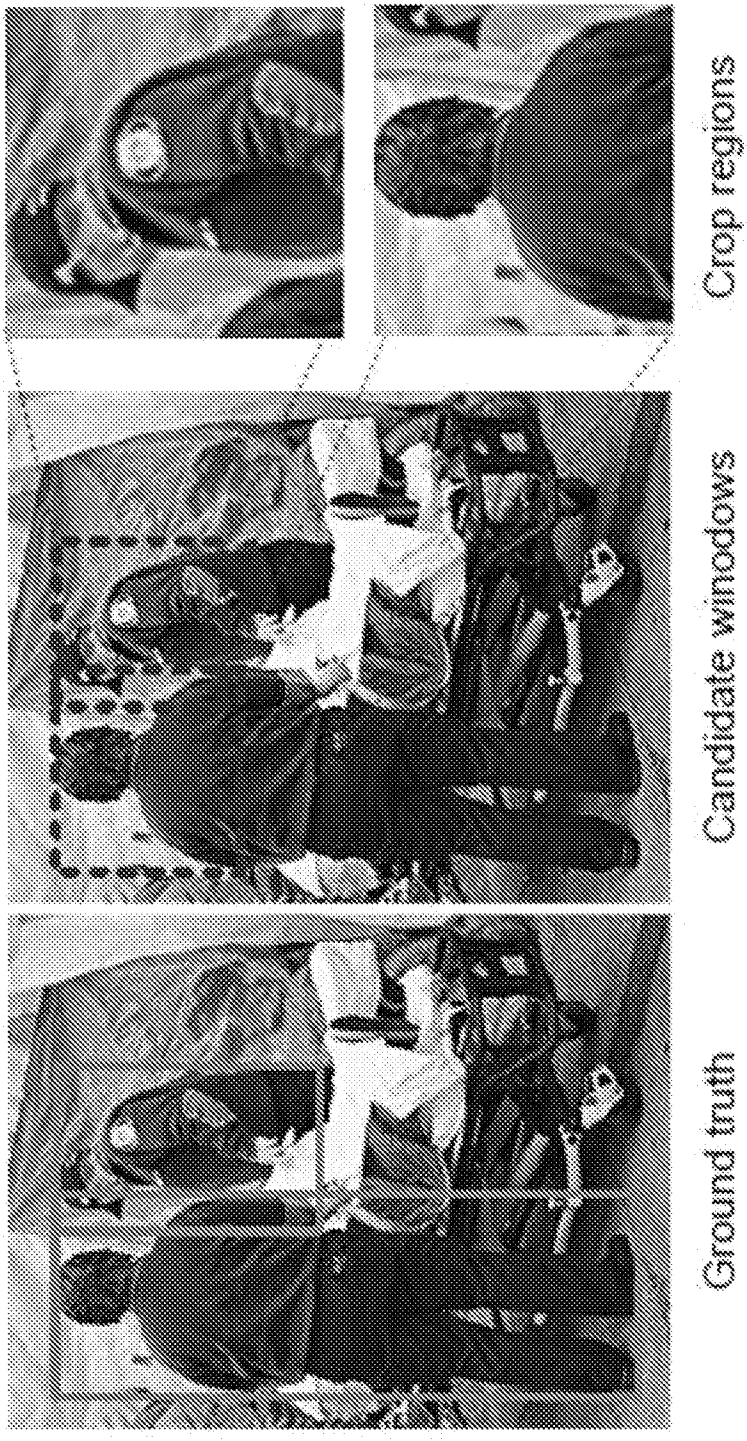

Fig. 5

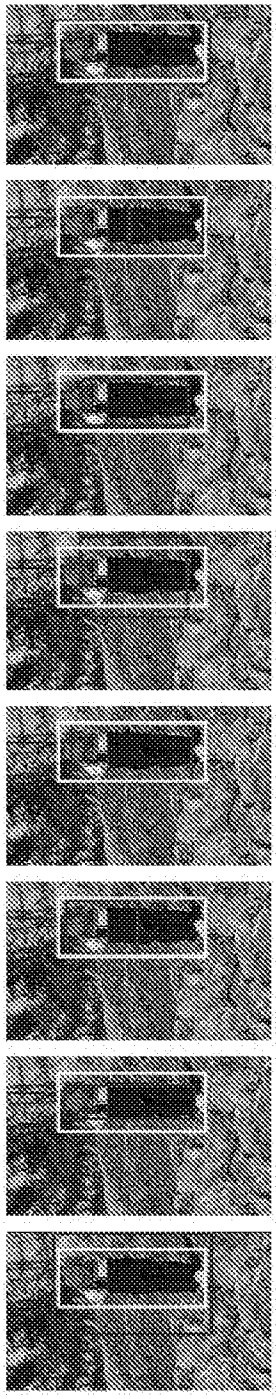

Fig. 6

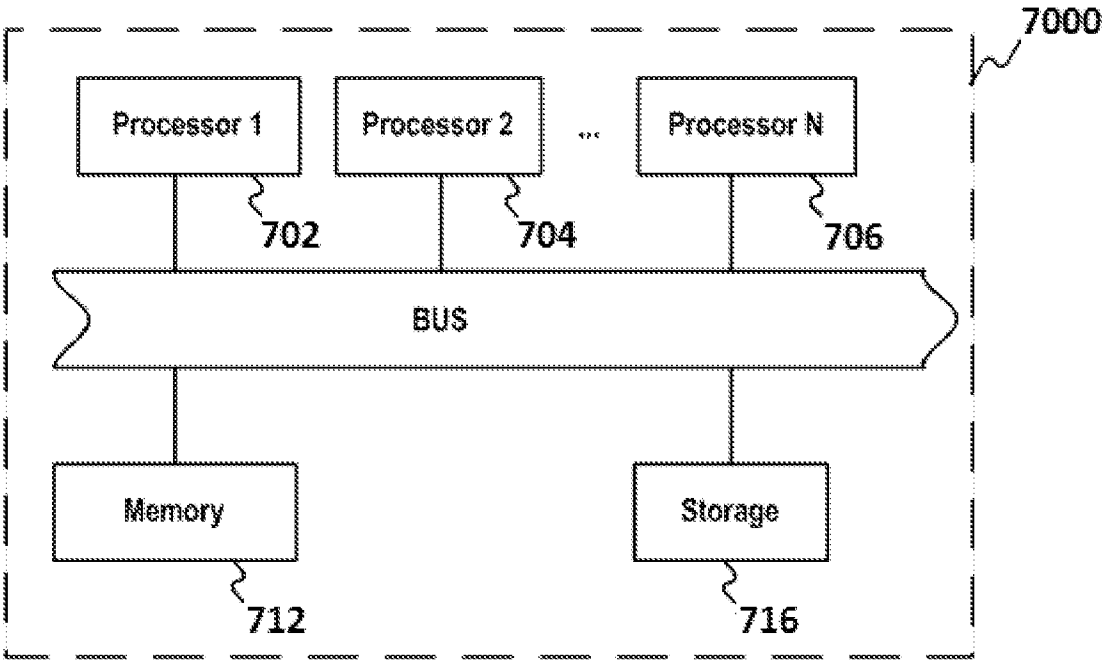
7000
Processor 1
Processor 2
...
Processor N
702
704
706
BUS
Memory
Storage
712
716

Fig. 7

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2015/094572

A. CLASSIFICATION OF SUBJECT MATTER

G06K 9/62(2006.01)i

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; H04N

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)

CNABS, CNTXT, CNKI, SIPOABS, VEN: image, picture, candidate, bounding box, score, extract, feature, object, identify, CNN, convolutional neural network, learn, train, classify, category, sort, appearance, face, SVM, support vector machine

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 2014126773 A1 (TOSHIBA TEC KK) 08 May 2014 (2014-05-08) description, paragraphs [0090] and [0091]; claims 1-6	1, 8-10, 17-19
Y	CN 103093214 A (ZHEJIANG ICARE VISION TECHNOLOGY CO LTD) 08 May 2013 (2013-05-08) description, paragraphs [0019]-[0028]; abstract	1, 8-10, 17-19
Y	CN 101701916 A (UNIV CHINA AGRICULTURAL) 05 May 2010 (2010-05-05) description, paragraphs [0037] and [0043]; claims 1-5	1, 8-10, 17-19
Y	US 2014241623 A1 (NEC LAB AMERICA INC) 28 August 2014 (2014-08-28) claims 1-20; abstract	1, 8-10, 17-19
Y	Erjin Zhou et al. "Extensive Facial Landmark Localization with Coarse-to-fine Convolutional Network Cascade" *2013 IEEE International Conference on Computer Vision Workshops*, 31 December 2013 (2013-12-31), abstract, figure2, page 387-388	1, 8-10, 17-19

☑ 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 July 2016	**19 August 2016**
Name and mailing address of the ISA/CN **STATE INTELLECTUAL PROPERTY OFFICE OF THE P.R.CHINA** **6, Xitucheng Rd., Jimen Bridge, Haidian District, Beijing 100088, China** Facsimile No. **(86-10)62019451**	Authorized officer **LIU,Dan** Telephone No. **(86-10)62411509**

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 103824054 A (BEIJING KUAGNSHI TECHNOLOGY CO LTD) 28 May 2014 (2014-05-28) the whole document	1-21
A	CN 103824049 A (BEIJING KUAGNSHI TECHNOLOGY CO LTD) 28 May 2014 (2014-05-28) the whole document	1-21
Y	De Cheng et al. "Cascade Object Detection with Complementary Features and Algorithms" *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)*, 09 February 2015 (2015-02-09), page 33 paragraph 1, page 35 paragraph 2, page 36 part C	1, 8-10, 17-19

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2015/094572

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
US	2014126773	A1	08 May 2014	JP	5707375	B2	30 April 2015
				JP	2014092975	A	19 May 2014
				CN	103810468	A	21 May 2014
CN	103093214	A	08 May 2013	CN	103093214	B	19 August 2015
CN	101701916	A	05 May 2010	CN	101701916	B	18 May 2011
US	2014241623	A1	28 August 2014	US	9020248	B2	28 April 2015
CN	103824054	A	28 May 2014		None		
CN	103824049	A	28 May 2014		None		