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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
27 December 2018 (27.12.2018)

WIPO | PCT

(10) International Publication Number
WO 2018/232592 A1

(51) International Patent Classification:
G06T 7/12 (2017.01)

(21) International Application Number:
PCT/CN2017/089189

(22) International Filing Date:
20 June 2017 (20.06.2017)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant: MICROSOFT TECHNOLOGY LICENSING, LLC. [US/US]; One Microsoft Way, Redmond, WA Washington 98052 (US).

(72) Inventor; and
(71) Applicant *(for SC only)*: **WEI, Yichen** [CN/CN]; c/o Microsoft Asia Pacific R&D, Headquarters 14F, Building 2, No. 5, Danling Street, Haidian District, Beijing 100080 (CN).

(72) Inventors: **DAI, Jifeng**; c/o Microsoft Asia Pacific R&D, Headquarters 14F, Building 2, No. 5, Danling Street, Haidian District, Beijing 100080 (CN). **HU, Han**; c/o Microsoft Asia Pacific R&D, Headquarters 14F, Building 2, No. 5, Danling Street, Haidian District, Beijing 100080 (CN).

(74) Agent: **KING & WOOD MALLESONS**; 20th Floor, East Tower, World Financial Centre, No. 1 Dongsanhuan Zhonglu, Chaoyang District, Beijing 100020 (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, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, 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,

(54) Title: FULLY CONVOLUTIONAL INSTANCE-AWARE SEMANTIC SEGMENTATION

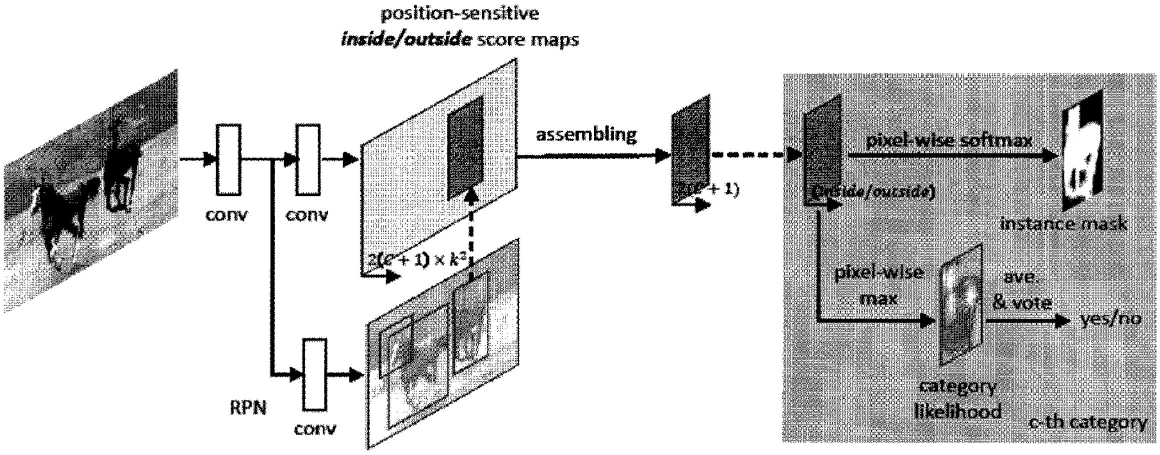

FIG. 4

(57) Abstract: Embodiments herein present a fully convolutional approach for instance-aware semantic segmentation. Embodiments herein, for a given visual image, locate predefined categories of objects therein and produce pixel masks for each object instance. One method embodiment includes receiving an image and generating a pixel-wise score map for a given region of interest within the received image for each pixel cell present therein. For each pixel cell within the region of interest, the method may detect whether the pixel cell belongs to an object to obtain a detection result and determine whether the pixel cell is inside an object instance boundary to obtain a segmentation result. The method may then fuse the results to obtain a result of inside or outside for each pixel cell and form at least one mask based on those values.

[Continued on next page]

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

Declarations under Rule 4.17:

— *of inventorship (Rule 4.17(iv))*

Published:

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

FULLY CONVOLUTIONAL INSTANCE-AWARE SEMANTIC SEGMENTATION

BACKGROUND INFORMATION

[0001] Fully convolutional networks (FCNs) have recently dominated the field of semantic image segmentation. An FCN takes an input image of arbitrary size, applies a series of convolutional layers, and produces per-pixel likelihood score maps for all semantic categories. Thanks to the simplicity, efficiency, and the local weight sharing properties of convolution, FCNs provide an accurate, fast, and end-to-end solution for semantic segmentation.

[0002] However, conventional FCNs often do not work for instance-aware semantic segmentation tasks, which often require detection and segmentation of individual object instances. This limitation is inherent in such solutions. Because convolution is translation invariant, the same image pixel receives the same responses, and thus classification scores, irrespective to its relative position in the context. However, instance-aware semantic segmentation needs to operate on region level, and the same pixel can have different semantics in different regions. This behavior is not modeled in a single FCN on the whole image.

[0003] Certain translation-variant properties are required to solve such problems. In a prevalent family of instance-aware semantic segmentation approaches, these problems are addressed by adopting different types of sub-networks in three stages: 1) an FCN is applied on the whole image to generate intermediate and shared feature maps; 2) from the shared feature maps, a pooling layer warps each region of interest (ROI) into fixed-size per-ROI feature maps; and 3) one or more fully-connected (fc) layer(s) in the last network convert the per-ROI feature maps to per-ROI masks. Note that the translation-variant property is introduced in the fc layer(s) in the last step.

[0004] But again, such solutions have several drawbacks. First, the ROI pooling step loses spatial details due to feature warping and resizing, which however, is necessary to obtain a fixed-size representation (*e.g.*, 14 x 14) for fc layers. Such distortion and fixed-size representation degrades the segmentation accuracy, especially for large objects. Second, the fc layers over-parametrize the task, without using regularization of local weight sharing. For example, the last fc layer has high dimensional 784-way output to estimate a 28 x 28 mask. Last, the per-ROI network computation in the last step is not shared among ROIs. A considerably complex sub-network in the last step is necessary to obtain good accuracy. It is therefore slow for a large number of ROIs (typically hundreds or thousands of region proposals). For example, in the MNC method (*see generally*, J. Dai, K. He, and J. Sun. Instance-aware semantic segmentation via multi-task network cascades. In CVPR, 2016), 10 layers in the ResNet-101 model are kept in the per-ROI sub-network. Such an approach has been measured to take 1.4 seconds per image, where more than 80% of the time is spent on the last per-ROI step.

[0005] Recently, a fully convolutional approach has been brought forth for instance mask proposal generation. This approach extends the translation invariant score maps in conventional FCNs to *position-sensitive* score maps, which are somewhat translation-variant. However, this approach is only used for mask proposal generation and presents several drawbacks. It is blind to semantic categories and requires a downstream network for detection. The object segmentation and detection sub-tasks are separated and the solution is not end-to-end. It operates on square, fixed-size sliding windows (224 x 224 pixels) and adopts a time-consuming image pyramid scanning to find instances at different scales.

SUMMARY

[0006] The various embodiments herein present systems, methods, and software of an end-to-end, fully convolutional approach for instance-aware semantic segmentation. Instance-aware semantic segmentation is a fundamental technique for many vision applications, such as photograph management and advanced driver assistance systems (ADAS), among other applications. Embodiments herein, for a given visual image, locate predefined categories of objects therein and produce pixel masks for each object instance. While the background section above sets forth some solutions to obtain such outputs, the present embodiments do so with greater efficiency (e.g., speed) and accuracy.

[0007] One such embodiment, in the form of a method, includes receiving an image and generating a pixel-wise score map for a given region of interest (ROI) within the received image, a score generated within the ROI for each pixel cell present therein. The method then, for each pixel cell within the region of interest, may detect whether the pixel cell belongs to an object to obtain a detection result and determine whether the pixel cell is inside an object instance boundary to obtain a segmentation result. The method may then fuse the detecting and segmentation results of each pixel cell to obtain a result of inside or outside for each respective pixel cell and form at least one mask based on the inside and outside values of at least one of the pixel cells.

[0008] Another embodiment is a system that includes an input, at least one processor, and a memory device that stores instructions executable on the at least one processor to perform data processing activities. The data processing activities may include generating a pixel-wise score map for a given ROI within an image received via the input, a score generated within the ROI for each pixel cell present therein. The data processing activities also include processing pixel cells of the ROI to detect whether the pixel cell belongs to an object to obtain a detection and determining whether the pixel cell is inside an object instance boundary to obtain a segmentation result. The data processing activities may then fuse the detecting and segmentation results of each pixel cell to obtain a result of inside or outside for each respective pixel cell and form at least one mask based on the inside and outside values of at least one of the pixel cells.

Brief Description of the Drawings

[0009] FIG. 1 illustrates an example of a fully convolutional instance-aware semantic segmentation method, according to an example embodiment.

[0010] FIG. 2 illustrates an instance fully convolutional network according to some instance segment embodiments.

[0011] FIG. 3 illustrates segmentation and classification results of different regions of interest for a category “person”, according to an example embodiment.

[0012] FIG. 4 illustrates an architecture of a fully convolutional instance-aware semantic segmentation network, according to an example embodiment.

[0013] FIG. 5 is a block flow diagram of a method, according to an example embodiment.

[0014] FIG. 6 is a block diagram of a computing device, according to an example embodiment.

DETAILED DESCRIPTION

[0015] The various embodiments herein present systems, methods, and software of an end-to-end, fully convolutional approach for instance-aware semantic segmentation. Instance-aware semantic segmentation is a fundamental technique for many vision applications, such as photograph management and advanced driver assistance systems (ADAS), among other applications. Embodiments herein, for a given visual image, locate predefined categories of objects therein and produce pixel masks for each object instance. While the background section above sets forth some solutions to obtain such outputs, the present embodiments do so with greater efficiency (e.g., speed) and accuracy.

[0016] Referred to as fully convolutional instance-aware semantic segmentation, or FCIS, embodiments herein resolve challenges not addressed in prior solutions, such as discussed above in the Background section, while exploiting the merits of FCNs for end-to-end instance-aware semantic segmentation. In some such embodiments, the underlying convolutional representation and the score maps are fully shared for object segmentation and detection sub-tasks, via a novel, joint formulation with no extra parameters. The FCN network structure is also highly integrated and efficient. The per-ROI computation is simple, fast, and does not involve any image warping or resizing operations. The approach is briefly illustrated in FIG. 1. Some of these embodiments operate on box proposals instead of sliding windows.

[0017] Extensive experiments have verified that the approach of the embodiments herein is state-of-the-art in both accuracy and efficiency. For example, the background section above refers to a prior solution that took 1.4 seconds to process each image. On the same test, embodiments as set forth herein achieved 0.24 second per image speeds. This is an 83-percent improvement. At the same time, accuracy of the various embodiments here in was compared against prior solutions. The accuracy of these embodiments exceeded all other solutions measured, outperforming the next closest with a twelve-percent accuracy improvement.

[0018] These and other embodiments are described herein with reference to the figures.

[0019] In the following detailed description, reference is made to the accompanying drawings that form a part hereof, and in which is shown by way of illustration specific embodiments in which the inventive subject matter may be practiced. These embodiments are described in sufficient detail to enable those skilled in the art to practice them, and it is to be understood that other embodiments may be utilized and that structural, logical, and electrical changes may be made without departing from the scope of the inventive subject matter. Such embodiments of the inventive subject matter may be referred to, individually and/or collectively, herein by the term "invention" merely for convenience and without intending to voluntarily limit the scope of this application to any single invention or inventive concept if more than one is in fact disclosed.

[0020] The following description is, therefore, not to be taken in a limited sense, and the scope of the inventive subject matter is defined by the appended claims.

[0021] The functions or algorithms described herein are implemented in hardware, software or a combination of software and hardware in one embodiment. The software comprises computer executable instructions stored on computer readable media such as memory or other type of storage devices. Further, described functions may correspond to modules, which may be software, hardware, firmware, or any combination thereof. Multiple functions are performed in one or more modules as desired, and the embodiments described are merely examples. The software is executed on a digital signal processor, ASIC, microprocessor, or other type of processor operating on a system, such as a personal computer, server, a router, or other device capable of processing data including network interconnection devices.

[0022] Some embodiments implement the functions in two or more specific interconnected hardware modules or devices with related control and data signals communicated between and through the modules, or as portions of an application-specific integrated circuit. Thus, the exemplary process flow is applicable to software, firmware, and hardware implementations.

[0023] In FCNs, a classifier is trained to predict each pixel's likelihood score of "*the pixel belongs to some object category*". Such predictions are translation invariant and unaware of individual object instances. For example, the same pixel can be foreground with regard to one object but background with regard to another, or adjacent, object. A single score map per-category is insufficient to distinguish these two cases.

[0024] To introduce translation-variant properties, a fully convolutional solution is first applied in some embodiments. Some such embodiments utilize k^2 position-sensitive score maps that correspond to k x k evenly partitioned cells of objects. This is illustrated in Figure 2 ($k = 3$). Each score map has the same spatial extent of the original image (in a lower resolution, *e.g.*, 16x smaller). Each score represents the likelihood of "*the pixel belongs to some object instance at a relative position*". For example, the first map is for "at top left position" in Figure 2.

[0025] During training and inference, for a fixed-size square sliding window (224 x 224 pixels), its pixel-wise foreground likelihood map may be produced by assembling its k x k cells from the corresponding score maps. In this way, a pixel can have different scores in different instances as long as the pixel is at different relative positions in the instances. This approach is works well for the object mask proposal task. However, this approach may also be limited by the task. Only a fixed-size square sliding window is used in some such embodiments. The FCN is then applied on multi-scale images to find object instances of different sizes. The approach is therefore blind to the object categories. Only a separate "objectness" classification sub-network is used to categorize the window as object or background.

For the instance-aware semantic segmentation task, a separate downstream network may be used to further classify the mask proposals into object categories.

[0026] For the instance-aware semantic segmentation task, other approaches, such as simultaneous detection and segmentation ("SDS"), Hypercolumn, convolution feature masking, instance aware semantic segmentation via multi-task network cascades ("MNC"), and MultiPathNet, share a similar structure. For example, two sub-networks may be used in some embodiments for object segmentation and detection sub-tasks, *separately and sequentially*. However, the design choices in such efforts, *e.g.*, the two-network structure, parameters and execution order, are kind of arbitrary. These design choices appear to have been made in some instances for convenience rather than for fundamental considerations. It appears that the separated sub-network design has not fully exploited a tight correlation between the two tasks.

[0027] Various embodiments herein enhance position-sensitive score mapping to perform object segmentation and detection sub-tasks *jointly and simultaneously*. The same set of score maps may then be shared for the two sub-tasks, as well as the underlying convolutional representation. In some embodiments, this approach brings no extra parameters and eliminates non-essential design choices to better exploit the strong correlation between the two sub-tasks.

[0028] The approach of some such embodiments is illustrated in FIG. 1 and FIG. 3. For example, given a region-of-interest (ROI), pixel-wise score maps may be produced by an assembling operation within the ROI. In some embodiments, for each pixel in a ROI, there are two tasks: 1) detection: whether the pixel belongs to an object bounding box at a relative position (detection+) or not (detection-); and 2) segmentation: whether the pixel is inside an object instance's boundary (segmentation+) or not (segmentation-). A simple solution of some embodiments is to train two classifiers, separately.

[0029] In some such embodiments, the two answers may then be fused into two scores: inside and outside. There are three possible result cases in such embodiments. These three cases are: 1) high inside score and low outside score: detection+, segmentation+; 2) low inside score and high outside score: detection+, segmentation-; 3) both scores are low: detection-, segmentation-. The two scores answer the two questions jointly via softmax and max operations/functions. For detection, some embodiments use the max operation to differentiate cases 1)-2) (detection+) from case 3) (detection-). The detection score of the whole ROI may then be obtained via average pooling over all pixels' likelihoods (followed by a softmax operator across all the categories). For segmentation, some embodiments may use softmax to differentiate cases 1) (segmentation+) from 2) (segmentation-), at each pixel. The foreground mask, considered in probabilities, of the ROI is the union of the per-pixel segmentation scores for each category. Similarly, the two sets of scores are from two 1 x 1 convolution layer. The inside/outside classifiers may be trained jointly in some embodiments as they receive the back-propagated gradients from both segmentation and detection losses.

[0030] Such embodiments have many desirable properties. All the per-ROI components as in FIG. 1 do not have free parameters. The score maps in some embodiments are produced by a single FCN, without involving any feature warping, resizing or fc layers. All the features and score maps in some such embodiments respect the aspect ratio of the original image. The local weight sharing property of FCNs is preserved and serves as a regularization mechanism. All per-ROI computation is simple (k^2 cell division, score map copying, softmax, max, average pooling) and fast, giving rise to a negligible per-ROI computation cost.

[0031] FIG. 4 illustrates an architecture of an end-to-end solution of some embodiments. While any convolutional network architecture can be used, some embodiments may be deployed that adopt the ResNet model. In some such embodiments, a last fully-connected layer for 1000 way-classification may discarded, retaining instead only the previous convolutional layers. The resulting feature maps in some embodiments have 2048 channels. Additionally, a 1 x 1 convolutional layer may be added in some such embodiments to reduce the dimension to 1024.

[0032] In the original ResNet, the effective feature stride (i.e., the decrease in feature map resolution) at the top of the network is 32. This may be too coarse for some embodiments of instance-aware semantic segmentation. To reduce the feature stride and maintain the field of view, the "hole algorithm" is applied. The stride in the first block of conv5 convolutional layers may be decreased in such embodiments from 2 to 1. The effective feature stride may thus be reduced to 16. To maintain the field of view, the "hole algorithm" may be applied on all the convolutional layers of conv5 by setting the dilation as 2.

[0033] Some embodiments may also use a region proposal network (RPN) to generate ROIs. For fair comparison with the MNC method, the RPN may be added on top of the conv4 layers in the same way. Note that RPN is also fully convolutional.

[0034] From the conv5 feature maps, $2k^2$ (C +1) score maps are produced (C object categories, one background category, two sets of k^2 score maps per category, $k = 7$ by default in some embodiments) using a 1 x 1 convolutional layer. Over the score maps, each ROI may be projected into a 16x smaller region. The segmentation probability maps and classification scores over all the categories may be computed as described above and elsewhere herein.

[0035] Bounding box (bbox) regression may then be used in some embodiments to refine the initial input ROIs. A sibling 1 x 1 convolutional layer with $4k^2$ channels may be added on the conv5 feature maps to estimate the bounding box shift in location and size.

[0036] With regard to inference, for an input image in some embodiments, 300 ROIs with highest scores may be generated from RPN. The ROIs may pass through the bbox regression branch and give rise to another 300 ROIs. For each ROI, classification scores and foreground mask (in probability) may be obtained for all categories. FIG. 3 illustrates such an example. Non-maximum suppression (NMS) with an intersection-over-union (IoU) threshold 0.3 may used in some embodiments to filter out highly overlapping ROIs. The remaining ROIs may be classified as the categories with highest classification scores. Their foreground masks may be obtained by mask voting in some embodiments as follows. For an ROI under consideration, find all the ROIs from the 600 ROIs, or other number depending on the particular embodiment, with IoU scores higher than 0.5. Foreground masks of a category may be averaged on a per-pixel basis and weighted by their classification scores. The averaged mask may then be binarized as the output in some embodiments.

[0037] With regard to training, an ROI may be positive in some embodiments when its box IoU with respect to the nearest ground truth object is larger than 0.5, otherwise it is negative. Each ROI has three loss terms in equal weights: a softmax detection loss over $C + 1$ categories, a softmax segmentation loss *over the foreground mask of the ground-truth category only*, and a bbox regression loss. This may sum per-pixel loses over the ROI and normalize the such by the ROI's size in some embodiments. The latter two loss terms may be effective only on the positive ROIs.

[0038] During training, the model may be initialized from a pre-trained model on ImageNet classification in some embodiments. Layers absent in the pre-trained model may be randomly initialized. The training images may be resized to have a shorter side, such as of 600 pixels. Some embodiments use Stochastic gradient descent ("SGD") optimization.

[0039] Generally, as the per-ROI computation is negligible, the training benefits from inspecting more ROIs at small training cost. Some embodiments may apply online hard example mining (OHEM) during training. In some embodiments, in each mini batch, forward propagation may be performed on 300 proposed ROIs on one image. Among them, 128 ROIs with the highest losses may be selected to back-propagate their error gradients.

[0040] For the RPN proposals, 9 anchors (3 scales x 3 aspect ratios) may be used by default. Three additional anchors at a finer scale may be used in some embodiments for experiments, such as on the well-known COCO dataset. To enable feature sharing between FCIS and RPN in such sharing embodiments, joint training is performed.

[0041] FIG. 5 is a block flow diagram of a method 500, according to an example embodiment. The method 500 is an example method that may be performed on a computing device, such as a personal computer or server when processing photographs or a computing device of an advanced driver assistance system that controls operation of an automobile.

[0042] The method 500 includes receiving 502 an image and generating 504 a pixel-wise score map for a given ROI within the received 502 image, a score generated within the ROI for each pixel cell present therein. The method 500 then continues, for each pixel cell within the region of interest, by detecting 506 whether the pixel cell belongs to an object to obtain a detection result, such as detection+ or detection-. Also with regard to each pixel, the method 500 includes determining 508 whether the pixel cell is inside an object instance boundary to obtain a segmentation result, such as segmentation+ or segmentation-. The method 500 may then fuse 510 the detecting and segmentation results of each pixel cell to obtain a result of inside or outside for each respective pixel cell. In some such embodiments, a result pair of (detection+, segmentation +) and (detection+, segmentation-) indicate inside the and (detection-, segmentation+) or (detection-, segmentation-) indicate outside the object. The method 500 may then form 512 at least one mask based on the inside and outside values of at least one of the pixel cells, although some embodiments may take into account values of all the pixel cells.

[0043] In some embodiments of the method 500, generating 504 a pixel-wise score map includes determining a probability that the respective pixel cell belongs to some object at the relative position. In some such embodiments, the score map is position-sensitive score map within the received 502 image. In these and some other embodiments of the method 500, each pixel cell of the pixel-wise score map is of a k^2 size, where k is a number of pixels. In some such embodiments, each pixel cell may be formed by evenly partitioning a region of interest into k^2 parts, where k is a number of pixels. Further, the method 500 may be performed iteratively against the received 502 image with varied values of k to identify object locations at a plurality of resolutions.

[0044] In some embodiments of the method 500, forming 512 the at least one mask includes forming a foreground mask and a background mask. In such embodiments, the foreground mask may include pixel cells with an inside value and the background mask may include pixel cells with an outside value.

[0045] The method 500 may be performed iteratively in some embodiments against a sequence of images received from an imaging device. In some such embodiments, the method 500 is performed by, and the imaging device is that of, an advanced driver assistance system.

[0046] FIG. 6 is a block diagram of a computing device, according to an example embodiment. In one embodiment, multiple such computer systems are utilized in a distributed network to implement multiple components in a transaction based environment. An object-oriented, service-oriented, or other architecture may be used to implement such functions and communicate between the multiple systems and components. One example computing device in the form of a computer 610, may include a processing unit 602 (which may be or include a graphics processing unit), memory 604, removable storage 612, and non-removable storage 614. Memory 604 may include volatile memory 606 and non-volatile memory 608. Computer 610 may include – or have access to a computing environment that includes – a variety of computer-readable media, such as volatile memory 606 and non-volatile memory 608, removable storage 612 and non-removable storage 614. Computer storage includes random access memory (RAM), read only memory (ROM), erasable programmable read-only memory (EPROM) & electrically erasable programmable read-only memory (EEPROM), flash memory or other memory technologies, compact disc read-only memory (CD ROM), Digital Versatile Disks (DVD) or other optical disk storage, magnetic cassettes, magnetic tape, magnetic disk storage or other magnetic storage devices, or any other medium capable of storing computer-readable instructions. Computer 610 may include or have access to a computing environment that includes input 616, output 618, and a communication connection 620. The computer may operate in a networked environment using a communication connection to connect to one or more remote computers, such as database servers. The remote computer may include a personal computer (PC), server, router, network PC, a peer device or other common network node, or the like. The communication connection may include a Local Area Network (LAN), a Wide Area Network (WAN) or other networks.

[0047] Computer-readable instructions stored on a computer-readable medium are executable by the processing unit 602 of the computer 610. A hard drive, CD-ROM, and RAM are some examples of articles including a non-transitory computer-readable medium. For example, a computer program 625 capable of providing a generic technique to perform access control check for data access and/or for doing an operation on one of the servers in a component object model (COM) based system according to the teachings of the present invention may be included on a CD-ROM and loaded from the CD-ROM to a hard drive. The computer-readable instructions allow computer 610 to provide generic access controls in a COM based computer network system having multiple users and servers.

[0048] It will be readily understood to those skilled in the art that various other changes in the details, material, and arrangements of the parts and method stages which have been described and illustrated in order to explain the nature of the inventive subject matter may be made without departing from the principles and scope of the inventive subject matter as expressed in the subjoined claims.

WHAT IS CLAIMED IS:

1. A method comprising:

receiving an image;

generating a pixel-wise score map for a given region of interest (ROI) within the received image, a score generated within the ROI for each pixel cell present therein;

for each pixel cell within the region of interest:

detecting whether the pixel cell belongs to an object to obtain a detection result of detection+ or detection-; and

determining whether the pixel cell is inside an object instance boundary to obtain a segmentation result of segmentation+ or segmentation-;

fusing the detecting and segmentation results of each pixel cell to obtain a result of inside or outside for each respective pixel cell, a result pair of (detection+, segmentation +) and (detection+, segmentation-) indicating inside and (detection-, segmentation+) or (detection-, segmentation-) indicating outside; and

forming at least one mask based on the inside and outside values of at least one of the pixel cells.

2. The method of claim 1, wherein generating a pixel-wise score map includes determining a probability that the respective pixel cell belongs to some object at the relative position.

3. The method of claim 2, wherein the score map is position-sensitive score map within the received image.

4. The method of claim 3, wherein each pixel cell is formed by evenly partitioning a region of interest into k^2 parts.

5. The method of claim 4, wherein where k is a number of pixels.

6. The method of claim 1 wherein, wherein determining whether the pixel cell is inside an object instance boundary includes applying a softmax operation.

7. The method of claim 1, wherein forming the at least one mask includes forming a foreground mask and a background mask, the foreground mask including pixel cells with an inside value and the background mask including pixel cells with an outside value.

8. The method of claim 1, wherein the method is performed iteratively against a sequence of images received from an imaging device.

9. The method of claim 8, wherein the method is performed by, and the imaging device is that of, an advanced driver assistance system.

10. The method of claim 1, wherein the method is performed with regard to a plurality regions of interest within the received image.

11. The method of claim 10, wherein the plurality of regions of interest are generated by a region proposal network.

12. The method of claim 10, further comprising:

performing a bounding box regression to refine the plurality of regions of interest.

13. A system comprising:

an input;

at least one processor; and

a memory device storing instructions executable on the at least one processor to perform data processing activities comprising:

generating a pixel-wise score map for a given region of interest (ROI) within an image received via the input, a score generated within the ROI for each pixel cell present therein;

for each pixel cell within the region of interest:

detecting whether the pixel cell belongs to an object to obtain a detection result; and

determining whether the pixel cell is inside an object instance boundary to obtain a segmentation result;

fusing the detecting and segmentation results of each pixel cell to obtain a result of inside or outside for each respective pixel cell; and

forming at least one mask based on the inside and outside values of at least one of the pixel cells.

14. The system of claim 13, wherein the input is an imaging device.

15. The system of claim 14, wherein the system is an advanced driver assistance system and the data processing activities are performed iteratively against a stream of images received from the imaging device.

16. The system of claim 13, wherein the at least one processor includes at least one graphics processing unit (GPU).

17. The system of claim 13, wherein:

the score map is position-sensitive score map within the received image; and

each pixel cell is formed by evenly partitioning a region of interest into k^2 parts, where k is a number of pixels.

18. The system of claim 13, wherein forming the at least one mask includes forming a foreground mask and a background mask, the foreground mask including pixel cells with an inside value and the background mask including pixel cells with an outside value.

19. The system of claim 13, wherein the data processing activities are performed with regard to each of a plurality regions of interest within each of a plurality of received images.

20. The system of claim 19, wherein the plurality of regions of interest are generated by a region proposal network and are refined by performing a bounding box regression on the plurality of regions of interest.

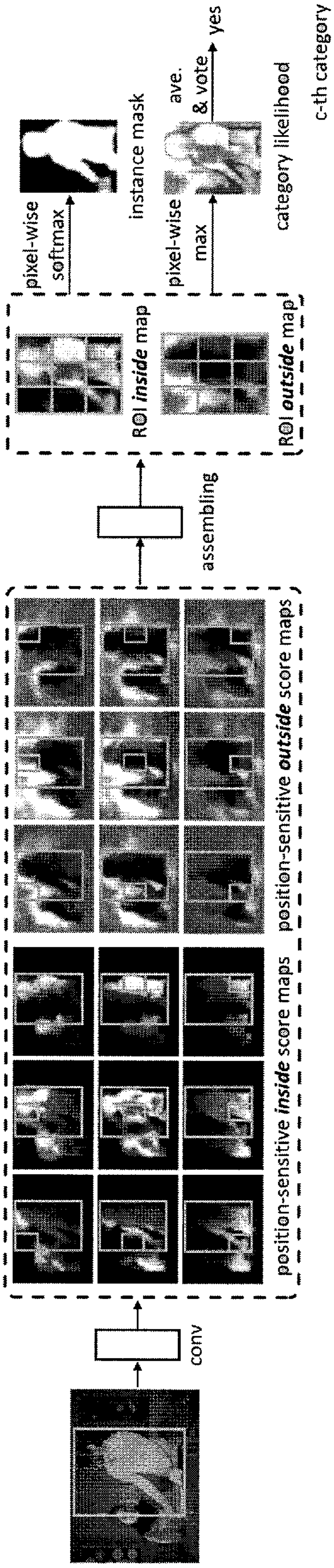
conv
position-sensitive *inside* score maps
position-sensitive *outside* score maps
assembling
ROI *inside* map
ROI *outside* map
pixel-wise softmax
instance mask
pixel-wise max
category likelihood
ave. & vote
yes
c-th category

FIG. 1

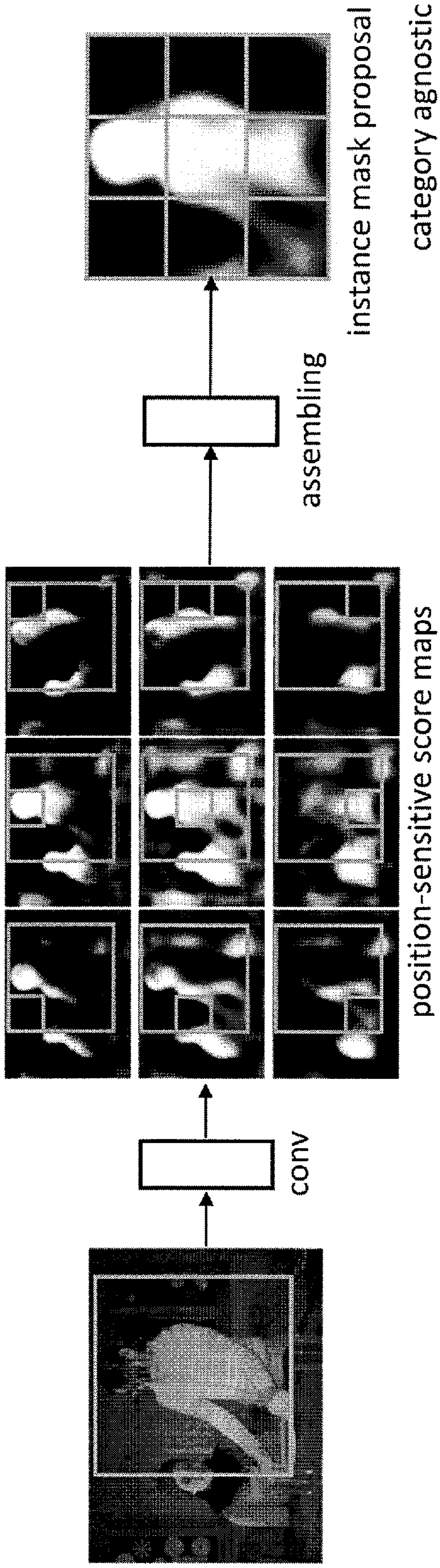
conv
position-sensitive score maps
assembling
instance mask proposal
category agnostic

FIG. 2

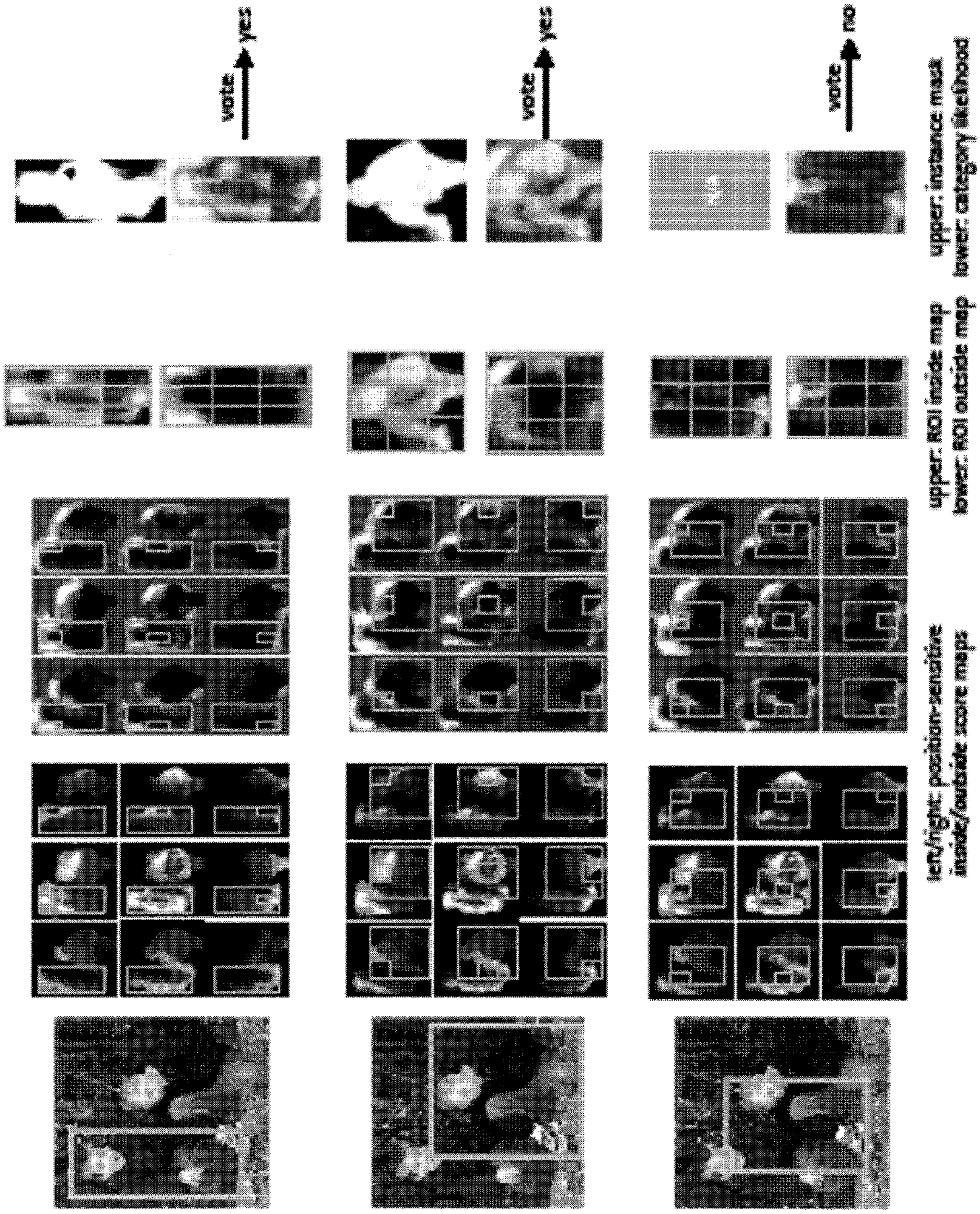
vote
yes
vote
yes
vote
no
upper: instance mask
lower: category likelihood
upper: ROI inside map
lower: ROI outside map
left/right: position-sensitive
inside/outside score maps

FIG. 3

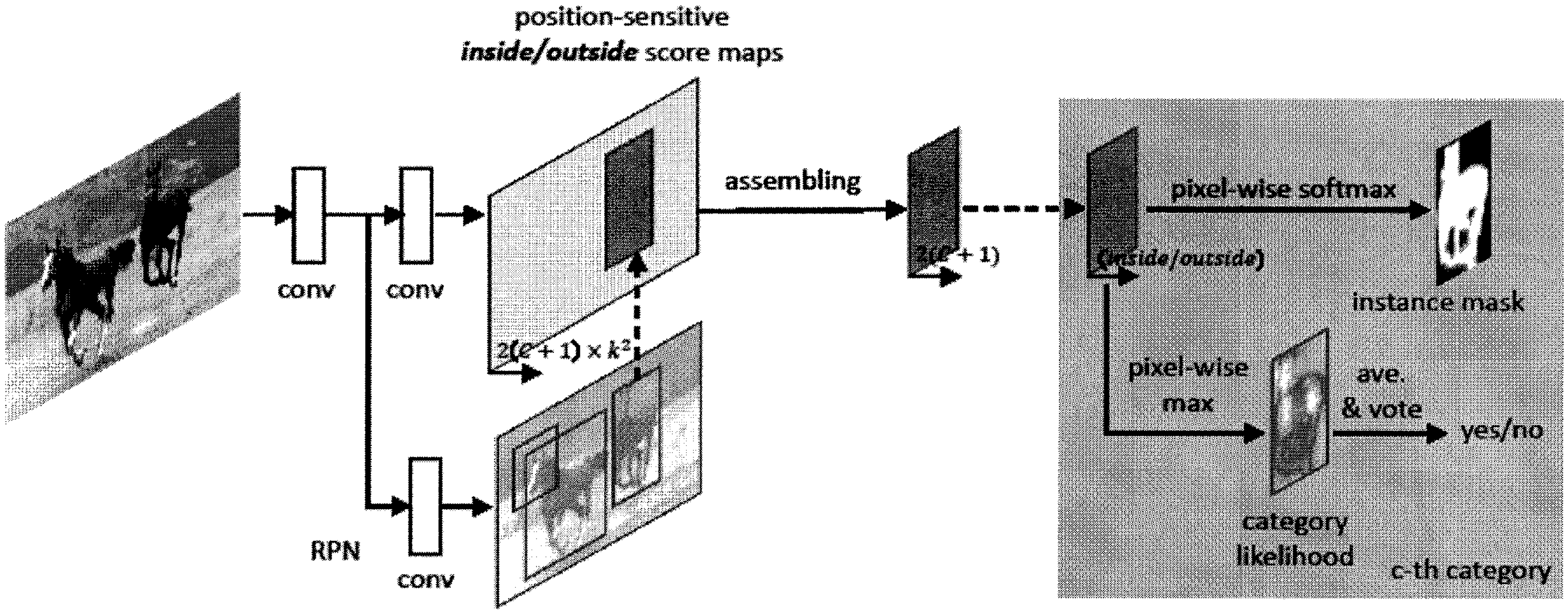
position-sensitive
inside/outside score maps
conv
conv
RPN
conv
2(C + 1) × k²
assembling
2(C + 1)
(inside/outside)
pixel-wise softmax
instance mask
pixel-wise
max
category
likelihood
ave.
& vote
yes/no
c-th category

FIG. 4

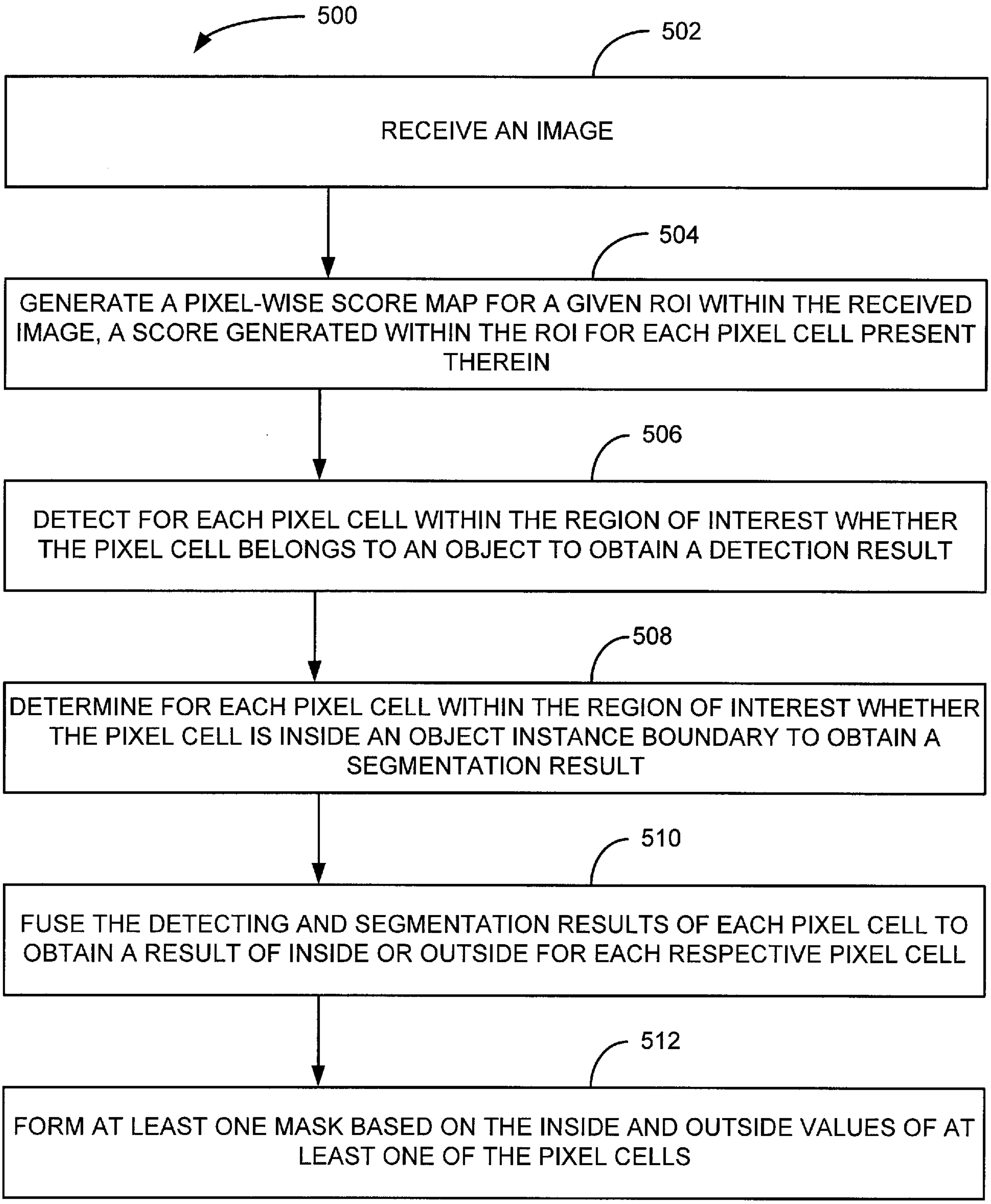
500
502
RECEIVE AN IMAGE
504
GENERATE A PIXEL-WISE SCORE MAP FOR A GIVEN ROI WITHIN THE RECEIVED IMAGE, A SCORE GENERATED WITHIN THE ROI FOR EACH PIXEL CELL PRESENT THEREIN
506
DETECT FOR EACH PIXEL CELL WITHIN THE REGION OF INTEREST WHETHER THE PIXEL CELL BELONGS TO AN OBJECT TO OBTAIN A DETECTION RESULT
508
DETERMINE FOR EACH PIXEL CELL WITHIN THE REGION OF INTEREST WHETHER THE PIXEL CELL IS INSIDE AN OBJECT INSTANCE BOUNDARY TO OBTAIN A SEGMENTATION RESULT
510
FUSE THE DETECTING AND SEGMENTATION RESULTS OF EACH PIXEL CELL TO OBTAIN A RESULT OF INSIDE OR OUTSIDE FOR EACH RESPECTIVE PIXEL CELL
512
FORM AT LEAST ONE MASK BASED ON THE INSIDE AND OUTSIDE VALUES OF AT LEAST ONE OF THE PIXEL CELLS

FIG. 5

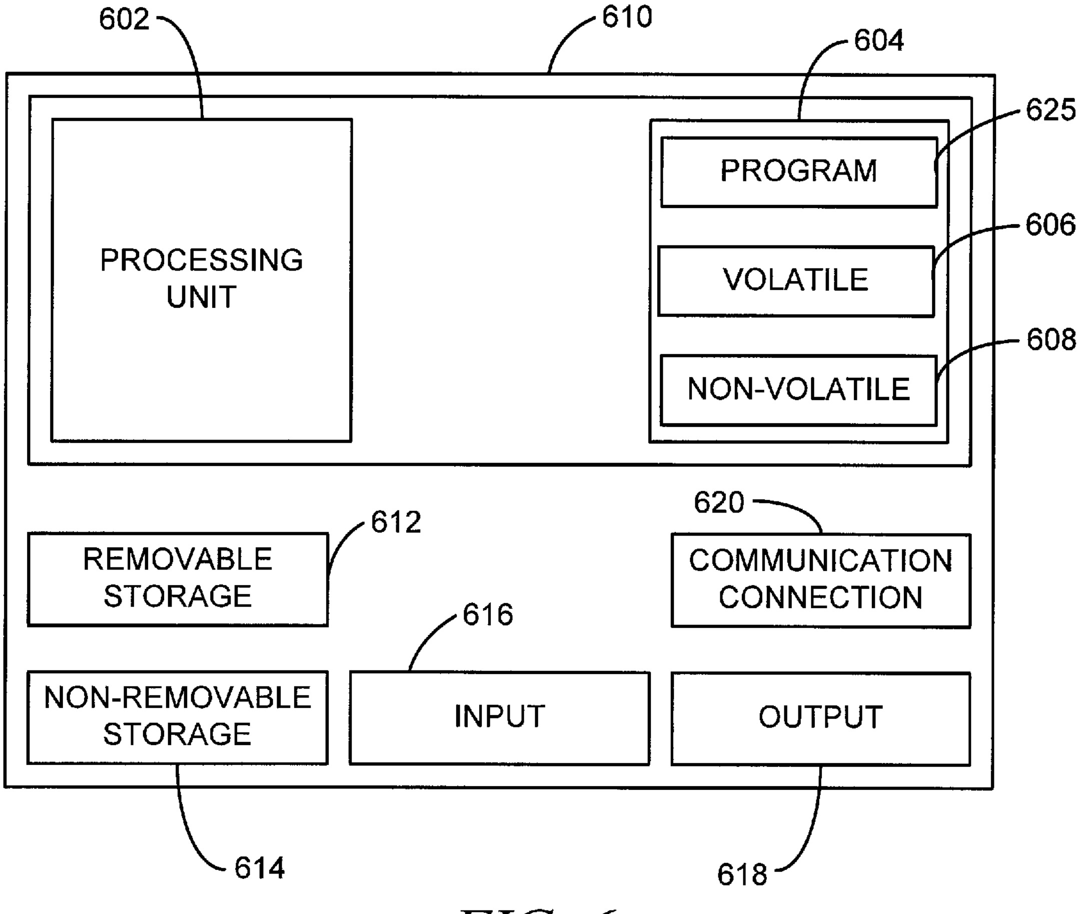
602
610
604
625
PROGRAM
606
VOLATILE
608
NON-VOLATILE
PROCESSING UNIT
620
612
REMOVABLE STORAGE
COMMUNICATION CONNECTION
616
NON-REMOVABLE STORAGE
INPUT
OUTPUT
614
618

FIG. 6

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2017/089189

A. CLASSIFICATION OF SUBJECT MATTER

G06T 7/12(2017.01)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)

G06T; H01L

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，DWPI, SIPOABS：image, detect segment, pixel, map, mask, convolutional,FCN

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	CN 105574513 A (BEIJING MEGVII TECHNOLOGY CO LTD ET AL.) 11 May 2016 (2016-05-11) claims 1-10，paragraphs 32-119，figures 3-11	1-3，13-14
Y	CN 106709568 A (UNIV BEIJING TECHNOLOGY) 24 May 2017 (2017-05-24) paragraphs 11-19	1-3，13-14
Y	CN 106296728 A (UNIV KUNMING SCI & TECHNOLOGY) 04 January 2017 (2017-01-04) paragraphs 31-49，figures 1-4	1-3，13-14
A	WO 2009078957 A1 (FLASHFOTO INC) 25 June 2009 (2009-06-25) the whole document	1-20
A	US 9058664 B2 (LIAO RUI ET AL.) 16 June 2015 (2015-06-16) the whole document	1-20

□ 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 **09 March 2018**	Date of mailing of the international search report **22 March 2018**
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**	Authorized officer **ZHONG,Yi**
Facsimile No. **(86-10)62019451**	Telephone No. **(86-10)62411332**

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2017/089189

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 105574513 A	11 May 2016	CN 105574513 B	24 November 2017
CN 106709568 A	24 May 2017	None	
CN 106296728 A	04 January 2017	None	
WO 2009078957 A1	25 June 2009	US 2014219560 A1	07 August 2014
		US 2010278426 A1	04 November 2010
		US 9042650 B2	26 May 2015
		US 8682029 B2	25 March 2014
US 9058664 B2	16 June 2015	US 2013057569 A1	07 March 2013