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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
26 January 2017 (26.01.2017)

WIPO | PCT

(10) International Publication Number
WO 2017/015390 A1

(51) International Patent Classification:
G06K 9/00 (2006.01) *G06K 9/66* (2006.01)
G06K 9/46 (2006.01) *G06K 9/68* (2006.01)

(21) International Application Number:
PCT/US2016/043172

(22) International Filing Date:
20 July 2016 (20.07.2016)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/194,598 20 July 2015 (20.07.2015) US
62/258,788 23 November 2015 (23.11.2015) US

(71) Applicant: **UNIVERSITY OF MARYLAND, COLLEGE PARK** [US/US]; Office of Technology Commercialization, 2130 Mitchell Bldg., 7999 Regents Dr., College Park, MD 20742 (US).

(72) Inventors: **RANJAN, Rajeev**; 4805 Erie Street, College Park, MD 20740 (US). **PATEL, Vishal M.**; 13623 Early Light Lane, Silver Spring, MD 20906 (US). **CHELLAPPA, Ramalingam**; 9104 Mistwood Drive, Potomac, MD 20854-2882 (US). **CASTILLO, Carlos D.**; 11303 Estona Dr., Silver Spring, MD 20902 (US).

(74) Agents: **FLANAGAN, Peter** et al.; Squire Patton Boggs (US) LLP, 8000 Towers Crescent Dr., 14th Floor, Vienna, VA 22182-6212 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, 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).

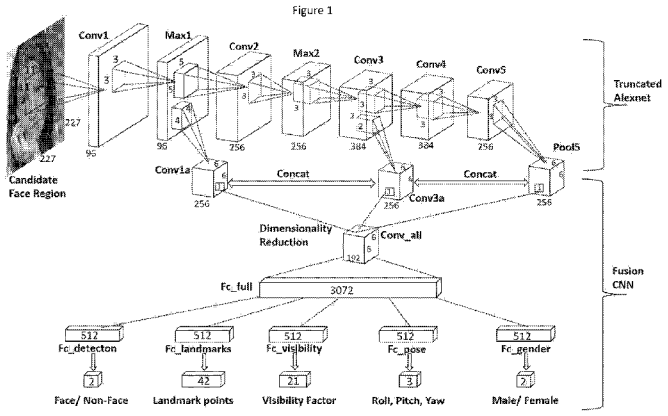
Published:

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

WO 2017/015390 A1

(54) Title: DEEP MULTI-TASK LEARNING FRAMEWORK FOR FACE DETECTION, LANDMARK LOCALIZATION, POSE ESTIMATION, AND GENDER RECOGNITION

Figure 1



(57) Abstract: Various image processing may benefit from the application deep convolutional neural networks. For example, a deep multi-task learning framework may assist face detection, for example when combined with landmark localization, pose estimation, and gender recognition. An apparatus can include a first module of at least three modules configured to generate class independent region proposals to provide a region. The apparatus can also include a second module of the at least three modules configured to classify the region as face or non-face using a multi-task analysis. The apparatus can further include a third module configured to perform post- processing on the classified region.

TITLE:

Deep Multi-task Learning Framework for Face Detection, Landmark Localization, Pose Estimation, and Gender Recognition

CROSS-REFERENCE TO RELATED APPLICATIONS:

[0001] This application is related to and claims the benefit and priority of U.S. Provisional Patent Application No. 62/194,598 filed July 20, 2015, for "A Deep Pyramid Deformable Part Model for Face Detection," the entirety of which is hereby incorporated herein by reference. This application is also related to and claims the benefit and priority of U.S. Provisional Patent Application No. 62/258,788 filed November 23, 2015, for "Hyperface: A Deep Multi-task Learning Framework for Face Detection, Landmark Localization, Pose Estimation, and Gender Recognition," the entirety of which is hereby incorporated herein by reference.

GOVERNMENT LICENSE RIGHTS:

[0002] This invention was made with government support under 201414071600012 awarded by IARPA. The government has certain rights in the invention.

BACKGROUND:

Field:

[0003] Various image processing may benefit from the application deep convolutional neural networks. For example, a deep multi-task learning framework may assist face detection, for example when combined with landmark localization, pose estimation, and gender recognition.

Description of the Related Art:

[0004] Detection and analysis of faces is a challenging issue in computer vision, and has been actively researched for its extensive applications on face verification, face tracking, person identification, and so on. Although recent

methods based on deep convolutional neural networks (CNN) have achieved results for face detection tasks it is still conventionally difficult to obtain facial landmark locations, head pose estimates and gender information from face images containing extreme pose, illumination and resolution variations. The tasks of face detection, landmark localization, pose estimation and gender classification have generally been solved as separate problems.

SUMMARY:

[0005] According to certain embodiments, an apparatus can include a first module of at least three modules. The first module can be configured to generate class independent region proposals to provide a region. The apparatus can also include a second module of the at least three modules configured to classify the region as face or non-face using a multi-task analysis. The apparatus can additionally include a third module configured to perform post-processing on the classified region.

[0006] In certain embodiments, an apparatus can include at least one processor and at least one memory including computer program instructions. The at least one memory and the computer program instructions can be configured to select a set of data for facial analysis. The at least one memory and the computer program instructions can also be configured to apply the set of data to a network comprising at least three modules. A first module of the at least three modules can be configured to generate class independent region proposals to provide a region. A second module of the at least three modules can be configured to classify the region as face or non-face using a multi-task analysis. A third module of the at least three modules can be configured to perform post-processing on the classified region.

[0007] A method, according to certain embodiments, can include selecting a set of data for facial analysis. The method can also include applying the set of data to a network comprising at least three modules. A first module of the at least three modules can be configured to generate class independent region

proposals to provide a region. A second module of the at least three modules can be configured to classify the region as face or non-face using a multi-task analysis. A third module of the at least three modules can be configured to perform post-processing on the classified region.

[0008] An apparatus, in certain embodiments, can include means for selecting a set of data for facial analysis. The apparatus can also include means for applying the set of data to a network comprising at least three modules. A first module of the at least three modules can be configured to generate class independent region proposals to provide a region. A second module of the at least three modules can be configured to classify the region as face or non-face using a multi-task analysis. A third module of the at least three modules can be configured to perform post-processing on the classified region.

BRIEF DESCRIPTION OF THE DRAWINGS:

[0009] For proper understanding of the invention, reference should be made to the accompanying drawings, wherein:

[0010] Figure 1 illustrates a hypernet architecture, according to certain embodiments.

[0011] Figure 2 illustrates a modular system according to certain embodiments.

[0012] Figure 3 illustrates pseudo-code of a recursive or iterative region proposal, according to certain embodiments.

[0013] Figure 4 illustrates pseudo-code of landmarks-based non-maximum suppression (NMS), according to certain embodiments.

[0014] Figure 5a illustrates an R-CNN-based network architecture for face detection, according to certain embodiments.

[0015] Figure 5b illustrates an R-CNN-based network architecture for landmark localization, according to certain embodiments.

[0016] Figure 5c illustrates an R-CNN-based network architecture for pose

estimation, according to certain embodiments.

[0017] Figure 5d illustrates an R-CNN-based network architecture for gender recognition, according to certain embodiments.

[0018] Figure 6 illustrates a network architecture for Multitask Face.

[0019] Figure 7 illustrates a deep pyramidal deformable part model for face detection, according to certain embodiments.

[0020] Figure 8 illustrates a method according to certain embodiments.

DETAILED DESCRIPTION:

[0021] Certain embodiments provide for simultaneous face detection, landmarks localization, pose estimation and gender recognition using deep convolutional neural networks (CNN). This approach, which can be referred to as Hyperface, can fuse the intermediate layers of a deep CNN using a separate CNN and can train multi-task loss on the fused features. For example, a multi-task leaning algorithm can operate on the fused features. Certain embodiments can exploit the synergy among the tasks, which can boost their individual performances. Certain embodiments can capture both global and local information of faces and can outperform conventional approaches for each of those four tasks.

[0022] In certain embodiments, therefore, a framework can be based on CNNs and can provide simultaneous face detection, facial landmark localization, head pose estimation and gender recognition from a given image. A CNN architecture can learn common features for these tasks and build the synergy among them. Information contained in features can be hierarchically distributed throughout the network. Lower layers can respond to edges and corners, and hence can contain better localization features. These layers may be more suitable for learning landmark localization and pose estimation tasks. On the other hand, higher layers can be class-specific and suitable for learning complex tasks, which may be suitable for face detection and gender

recognition. All the intermediate layers of a deep CNN can be used in order to train different tasks under consideration. The set of intermediate layer features can be referred to as hyperfeatures.

[0023] A CNN architecture can contain multiple layers with hundreds of feature maps in each layer. The overall dimension of hyperfeatures may be too large to be efficient for learning multiple tasks. Moreover, the hyperfeatures may need to be associated in a way that they efficiently encode the features common to the multiple tasks. This can be handled using feature fusion techniques. Features fusion can transform the features to a common subspace where they can be combined linearly or non-linearly.

[0024] In certain embodiments, a separate fusion-CNN is provided to fuse the hyperfeatures. In order to learn the tasks, the deep CNN and the fusion CNN can be trained simultaneously using multiple loss functions. In this way, the features can get better understanding of faces, which can lead to improvements in the performances of the individual tasks. The deep CNN combined with the fusion-CNN can undergo learning together end-to-end. Since the network performs hyperfeatures fusion, it can be referred to as a hypernet.

[0025] The CNN architecture in certain embodiments can perform the multiple tasks of face detection, landmarks localization, pose estimation and gender recognition. Moreover, there can be two post-processing methods: recursive region proposal and landmark-based non-maximum separation, each of which can leverage the multitask information obtained from the CNN to improve the overall performance. Certain embodiments may provide improved performances on challenging unconstrained datasets for all of these four tasks.

[0026] Figure 1 illustrates a hypernet architecture, according to certain embodiments. A single CNN model, such as a hypernet, can provide simultaneous face detection, landmark localization, pose estimation and gender classification. The network architecture can be deep in both vertical and horizontal directions. Figure 1 can be understood in view of a brief overview of the system pipeline and discussion of the different components in

this pipeline in detail.

[0027] Figure 2 illustrates a modular system according to certain embodiments. Three modules can be employed in certain embodiments. The first module, module 210, can generate class independent region-proposals from a given image and can scale them to a predetermined size, such as 227 x 227 pixels. This module can be variously implemented as a software module running on hardware. The module can be located on a same computer as the other modules or on a different computer. The computer hardware running module 210, and each of the other modules, can include at least one processor and at least one memory. The at least one memory can include computer program code. The at least one memory and the computer program code can be configured to, with the at least one processor, cause the system to carry out the functions associated with the corresponding module. The software to implement the module can be any region selection mechanism, including any conventional region selection mechanism.

[0028] An example of a region selection mechanism is a sliding window mechanism that proposes overlapping windows as successive potential regions. This approach may be considered an example of an exhaustive search. Additional criteria may be included in this module, such as thresholds of difference amounts within a region. Thus may avoid, for example, suggesting a region that is a single color as a candidate region to be a face. The region sizes may vary as well. For example, a first proposal may be the entire image, to capture the case of a so-called selfie. Subsequent proposals may suggest smaller and smaller candidate regions to capture faces occupying decreasingly smaller portions of the image. The smaller and smaller regions may continue to decrease in size until some minimum number of pixels per region is reached. Other approaches, such as beginning at the smallest pixel size and proceeding to the full image, are also permitted.

[0029] In a parallel computing environment, each successive candidate region can be fed to a different parallel process for evaluation. Other approaches are

also permitted. For example, the module can be implemented in hardware running alone, without conventional software.

[0030] Second module 220 can be a CNN which takes in the resized candidate regions and classifies them as face or non-face. If a region gets classified as a face, the network can additionally provide the facial landmarks locations, estimated head pose and the gender information for the region. This module may similarly be variously constructed and implemented in software running on hardware or hardware alone. This system may provide output in the form of a report, a modified image, or any other form. The inputs for this module can include candidate regions from module 210 as well as fed back information from module 230. Module 220 can be variously implemented, for example, as illustrated in Figure 1. Because the second module 220 may also be able to generate candidate face regions, in certain embodiments the first module 210 and second module 220 can be integrated as a single module.

[0031] Third module 230 can provide post-processing. The post-processing can include recursive region-proposals and landmarks-based *k*-NMS. The post processing can boost the face detection score and improve the performance of individual tasks. This module may similarly be variously constructed and implemented in software running on hardware or hardware alone. More details of the recursive region-proposals and landmarks-based *k*-NMS are discussed below.

[0032] Referring to Figure 1, the network shown can be constructed by beginning with an Alexnet network for image classification. The network can include five convolutional layers along with three fully connected layers. The network can be initialized with predetermined weights, such as those used in the Imagenet Challenge with the Caffe implementation.

[0033] All the fully connected layers can be removed as they encode image-classification task specific information, which may not be desirable for face related tasks.

[0034] As mentioned above, the features in CNN can be distributed hierarchically in the network. While the lower layer features may be informative for localization tasks such as landmarks and pose estimation, the higher layer features may be suitable for more complex tasks such as detection or classification. Learning multiple correlated tasks simultaneously can build a synergy and improve the performance of individual tasks. Certain embodiments can simultaneously learn face detection, landmarks, pose and gender, for example using a fusion of the features from the intermediate layers of the network (hyperfeatures), and can learn multiple tasks on top of that fusion. Since the adjacent layers may be highly correlated, not all the intermediate layers may be considered for fusion.

[0035] Certain embodiments can fuse the max1, conv3 and pool5 layers of Alexnet, using a separate network. A naive way for fusion is directly concatenating the features. Since the feature maps for these layers have different dimensions 27 x 27 x 96, 13 x 13 x 384, 6 x 6 x 256, respectively, they cannot be concatenated directly. Therefore, add conv1a and conv3a convolutional layers can be added to pool1, conv3 layers to obtain consistent feature maps of dimensions 6 x 6 x 256 at the output. The output of these layers can then be concatenated along with pool5 to form a 6 x 6 x 768 dimensional feature map. The dimension is still quite high to train a multi-task framework. Hence, a 1 x 1 kernel convolution layer (conv_all) can be added to reduce the dimensions to 6 x 6 x 192. A fully connected layer (fc_full) can be to conv_all, which can output a 3072 dimensional feature vector.

[0036] The network can then be split into five separate branches corresponding to the different tasks. Fully connected layers fc_detection, fc_landmarks, fc_visibility, fc_pose, and fc_gender, each of dimension 512, can be added to fc_full. Finally, a fully connected layer can be added to each of the branches to predict the individual task labels.

[0037] After every convolution or fully connected layer, a rectified layer unit (RelU) non-linearity can be deployed. Pooling operation in the fusion network

can be omitted, as such pooling may provide local invariance which may not be desired for the face landmark localization task. Task-specific loss functions can then be used to learn the weights of the network.

[0038] The network can then be trained on a data set that includes faces in real-world images with pre-annotated attributes. For testing purposes, a subset of the data can be retained to against which to test the training.

[0039] The fully connected layers fc_detection, fc_landmarks, fc_visibility, fc_pose, and fc_gender can be variously implemented to provide face detection, landmark localization, visibility, pose, and gender, respectively. For example, the detection may be implemented by any face detection mechanism such as a deep pyramidal detection mechanism as described below.

[0040] Landmark localization can be implemented using a 21 point markup for face landmark location, or any other suitable mechanism. Because the faces have full pose variations, some of the landmark points may be invisible.

[0041] A visibility factor can be determined to test the presence of a predicted landmark. The pose estimation can provide head pose estimates of roll, pitch, and yaw. Furthermore, the gender recognition can compute a perceived sex or gender of the face.

[0042] The landmark locations obtained along with the detections can permit various post processing steps that can benefit all the tasks. At least two post-processing steps are possible: recursive region proposals and landmark-based *k*-NMS to improve the recall performance.

[0043] For recursive region proposals, a fast version of Selective Search can extract around 2000 regions from an image. Some faces with poor illumination or small size may fail to get captured by any candidate region with a high overlap. The network may fail to detect that face due to low score. In these situations, a candidate box that precisely captures the face may be beneficial. Hence, a new candidate bounding box from the predicted landmark points can be provided, for example using a FaceRectCalculator. The new region, being

more localized, can yield a higher detection score and the corresponding tasks output, thus increasing the recall. This procedure can be repeated iteratively or recursively. Figure 3 illustrates pseudo-code of a recursive or iterative region proposal, according to certain embodiments.

[0044] Regarding landmarks-based *k*-NMS, the traditional way of nonmaximum separation involves selecting the top scoring region and discarding all the other regions with overlap more than a certain threshold. This method can fail in the following two scenarios. First, if a region corresponding to the same detected face has less overlap with the highest scoring region, it can be detected as a separate face. Second, the highest scoring region might not always be localized well for the face, which can create some discrepancy if two faces are close together. To overcome these issues, NMS can be performed on a new region whose bounding box is defined by the boundary co-ordinates as $[\min_i x_i, \min_i y_i, \max_i x_i, \max_i y_i]$. In this way, the candidate regions would get close to each other, thus decreasing the ambiguity of the overlap and improving the face localization.

[0045] Landmarks-based k-NMS can be applied to keep the top *k* boxes, based on the detection scores. The detected face can correspond to the region with maximum score. The landmark points, pose estimates and gender classification scores can be decided by the median of the top *k* boxes obtained. Hence, the predictions do not rely only on one face region, but can consider the votes from top *k* regions for generating the final output. For example, the value of *k* may be 5. Figure 4 illustrates pseudo-code of landmarks-based NMS, according to certain embodiments.

[0046] Figure 5a illustrates an R-CNN-based network architecture for face detection, Figure 5b illustrates an R-CNN-based network architecture for landmark localization, Figure 5c illustrates an R-CNN-based network architecture for pose estimation, and Figure 5d illustrates an R-CNN-based network architecture for gender recognition, according to certain embodiments. The numbers on the left denote the kernel size and the numbers

on the right denote the cardinality of feature maps for a given layer. These are just examples of networks that can be used.

[0047] Figure 6 illustrates a network architecture for Multitask Face. The numbers on the left denote the kernel size and the numbers on the right denote the cardinality of feature maps for a given layer. Similar to HyperFace, this model can be used to simultaneously detect face, localize landmarks, estimate pose and predict its gender. The main difference between Multitask Face and HyperFace is that HyperFace can fuse the intermediate layers of the network whereas Multitask Face can combine the tasks using the common fully connected layer at the end of the network as shown in Figure 6. Since it provides the landmarks and face score, Multitask Face can leverage iterative region proposals and landmark-based NMS post-processing algorithms during evaluation.

[0048] Figure 7 illustrates a deep pyramidal deformable part model for face detection, according to certain embodiments. As shown in Figure 7, an image pyramid can be built from a color input image with level 1 being the lowest size. Each pyramid level can be forward propagated through a deep pyramid CNN that ends at max variant of convolutional layer 5 (max5). The result can be a pyramid of max5 feature maps, each at 1/16th the spatial resolution of its corresponding image pyramid level. Each max5 level feature can be normalized using z-score to form a norm5 feature pyramid. Each norm5 feature level can get convoluted with every root-filter of a C-component DPM to generate a pyramid of DPM score. The detector can output a bounding box for face location in the image after non-maximum suppression and bounding box regression. The inclusion of a normalization layer can help reduce bias in face sizes.

[0049] Figure 8 illustrates a method according to certain embodiments. As shown in Figure 8, a method can include, at 810, selecting a set of data for facial analysis. The method can also include, at 820, applying the set of data to a network comprising at least three modules. A first module of the at least

three modules can be configured to generate class independent region proposals to provide a region. A second module of the at least three modules can be configured to classify the region as face or non-face using a multi-task analysis. A third module of the at least three modules can be configured to perform post-processing on the classified region.

[0050] The method of Figure 8 can be variously implemented in one or more processor and one more memory as with the other modules and processes described above. Processors may be embodied by any computational or data processing device, such as a central processing unit (CPU), digital signal processor (DSP), application specific integrated circuit (ASIC), programmable logic devices (PLDs), field programmable gate arrays (FPGAs), digitally enhanced circuits, or comparable device or a combination thereof. The processors may be implemented as a single controller, or a plurality of controllers or processors. Additionally, the processors may be implemented as a pool of processors in a local configuration, in a cloud configuration, or in a combination thereof.

[0051] For firmware or software, the implementation may include modules or units of at least one chip set (e.g., procedures, functions, and so on). Memories may independently be any suitable storage device, such as a non-transitory computer-readable medium. A hard disk drive (HDD), random access memory (RAM), flash memory, or other suitable memory may be used. The memories may be combined on a single integrated circuit as the processor, or may be separate therefrom. Furthermore, the computer program instructions may be stored in the memory and which may be processed by the processors can be any suitable form of computer program code, for example, a compiled or interpreted computer program written in any suitable programming language. The memory or data storage entity is typically internal but may also be external or a combination thereof, such as in the case when additional memory capacity is obtained from a service provider. The memory may be fixed or removable.

[0052] The memory and the computer program instructions may be configured, with the processor for the particular device, to cause a hardware apparatus such as network element 310 and/or UE 320, to perform any of the processes described above (see, for example, Figure 1). Therefore, in certain embodiments, a non-transitory computer-readable medium may be encoded with computer instructions or one or more computer program (such as added or updated software routine, applet or macro) that, when executed in hardware, may perform a process such as one of the processes described herein. Computer programs may be coded by a programming language, which may be a high-level programming language, such as objective-C, C, C++, C#, Java, etc., or a low-level programming language, such as a machine language, or assembler. Alternatively, certain embodiments of the invention may be performed entirely in hardware.

[0053] One having ordinary skill in the art will readily understand that the invention as discussed above may be practiced with steps in a different order, and/or with hardware elements in configurations which are different than those which are disclosed. Therefore, although the invention has been described based upon these preferred embodiments, it would be apparent to those of skill in the art that certain modifications, variations, and alternative constructions would be apparent, while remaining within the spirit and scope of the invention.

WE CLAIM:

1. An apparatus, comprising:

a first module of at least three modules, wherein the first module configured to generate class independent region proposals to provide a region;

a second module of the at least three modules is configured to classify the region as face or non-face using a multi-task analysis; and

a third module of the at least three modules is configured to perform post-processing on the classified region.

2. The apparatus of claim 1, wherein the second module comprises a five convolutional layers with three fully connected layers.

3. The apparatus of claim 2, wherein the second module further comprises a network configured to fuse the three fully connected layers.

4. The apparatus of claim 3, wherein the second module further comprises separate networks for face detection, landmark detection, visibility determination, pose estimation, and gender determination.

5. The apparatus of claim 1, wherein the third module comprises at least one of an iterative region proposal or landmark-based non-maximum suppression.

6. An apparatus, comprising:

at least one processor; and

at least one memory including computer program instructions,

wherein the at least one memory and the computer program instructions are configured to

select a set of data for facial analysis; and

apply the set of data to a network comprising at least three modules,

wherein a first module of the at least three modules is configured to generate class independent region proposals to provide a region,

wherein a second module of the at least three modules is configured to classify the region as face or non-face using a multi-task analysis,

wherein a third module of the at least three modules is configured to perform post-processing on the classified region.

7. The apparatus of claim 6, wherein the second module comprises a five convolutional layers with three fully connected layers.

8. The apparatus of claim 7, wherein the second module further comprises a network configured to fuse the three fully connected layers.

9. The apparatus of claim 8, wherein the second module further comprises separate networks for face detection, landmark detection, visibility determination, pose estimation, and gender determination.

10. The apparatus of claim 6, wherein the third module comprises at least one of an iterative region proposal or landmark-based non-maximum suppression.

11. A method, comprising:

selecting a set of data for facial analysis; and

applying the set of data to a network comprising at least three modules,

wherein a first module of the at least three modules is configured to generate class independent region proposals to provide a region,

wherein a second module of the at least three modules is configured to classify the region as face or non-face using a multi-task analysis,

wherein a third module of the at least three modules is configured to perform post-processing on the classified region.

12. The method of claim 11, wherein the second module comprises a five convolutional layers with three fully connected layers.

13. The method of claim 12, wherein the second module further comprises a network configured to fuse the three fully connected layers.

14. The method of claim 13, wherein the second module further comprises separate networks for face detection, landmark detection, visibility determination, pose estimation, and gender determination.

15. The apparatus of claim 1, wherein the third module comprises at least one of an iterative region proposal or landmark-based non-maximum suppression.

16. An apparatus, comprising:

means for selecting a set of data for facial analysis; and

means for applying the set of data to a network comprising at least three modules,

wherein a first module of the at least three modules is configured to generate class independent region proposals to provide a region,

wherein a second module of the at least three modules is configured to classify the region as face or non-face using a multi-task analysis,

wherein a third module of the at least three modules is configured to perform post-processing on the classified region.

17. The apparatus of claim 16, wherein the second module comprises a five convolutional layers with three fully connected layers.

18. The apparatus of claim 17, wherein the second module further

comprises a network configured to fuse the three fully connected layers.

19. The apparatus of claim 18, wherein the second module further comprises separate networks for face detection, landmark detection, visibility determination, pose estimation, and gender determination.

20. The apparatus of claim 16, wherein the third module comprises at least one of an iterative region proposal or landmark-based non-maximum suppression.

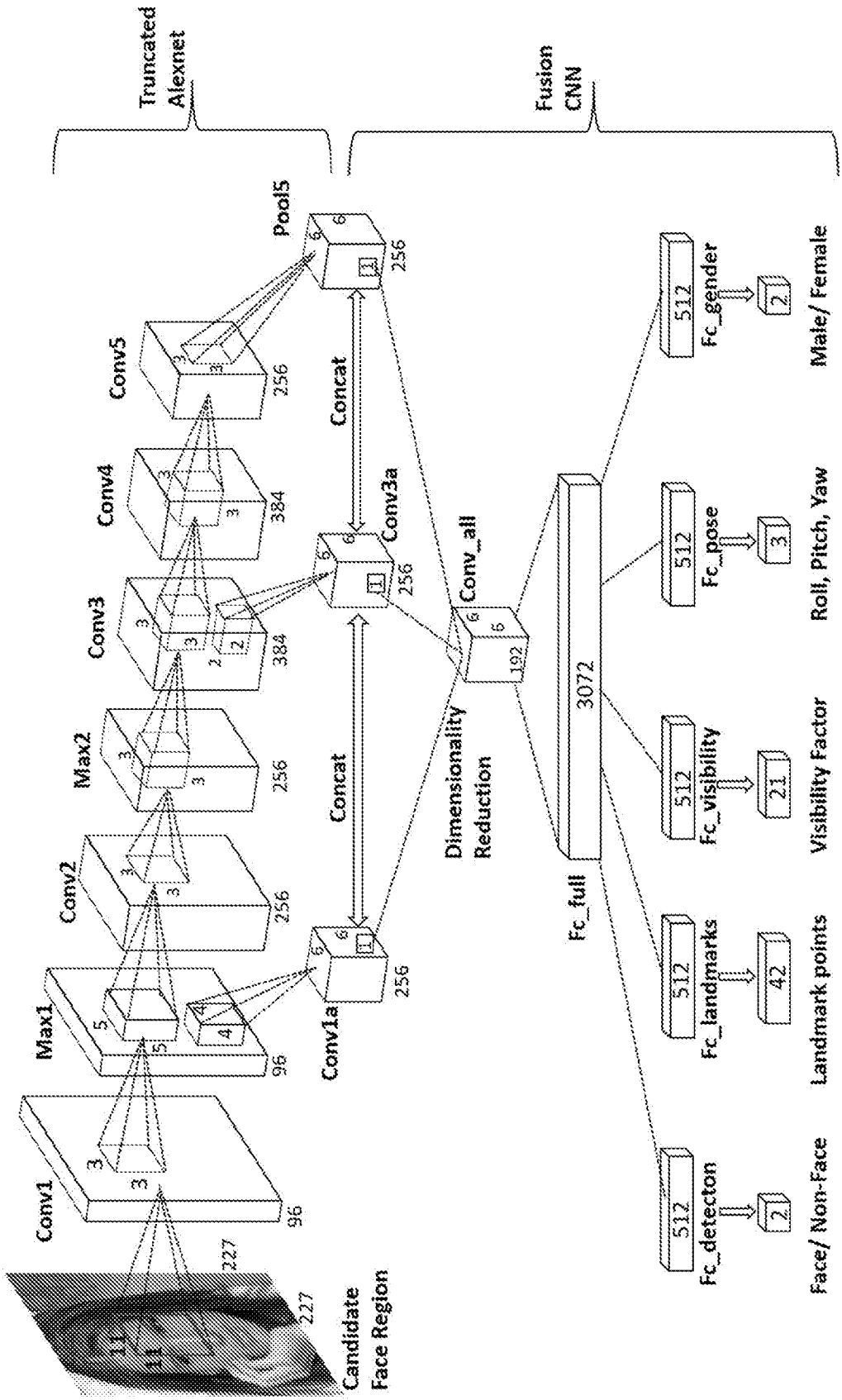

Truncated Alexnet
Fusion CNN
Candidate Face Region
227
Conv1
96
Max1
Conv2
256
Max2
Conv3
384
Conv4
384
Conv5
256
Pool5
Concat
Conv1a
Conv3a
Concat
Dimensionality Reduction
Conv_all
192
Fc_full
3072
Fc_detecton
Face/ Non-Face
Fc_landmarks
Landmark points
Fc_visibility
Visibility Factor
Fc_pose
Roll, Pitch, Yaw
Fc_gender
Male/ Female

Figure 1

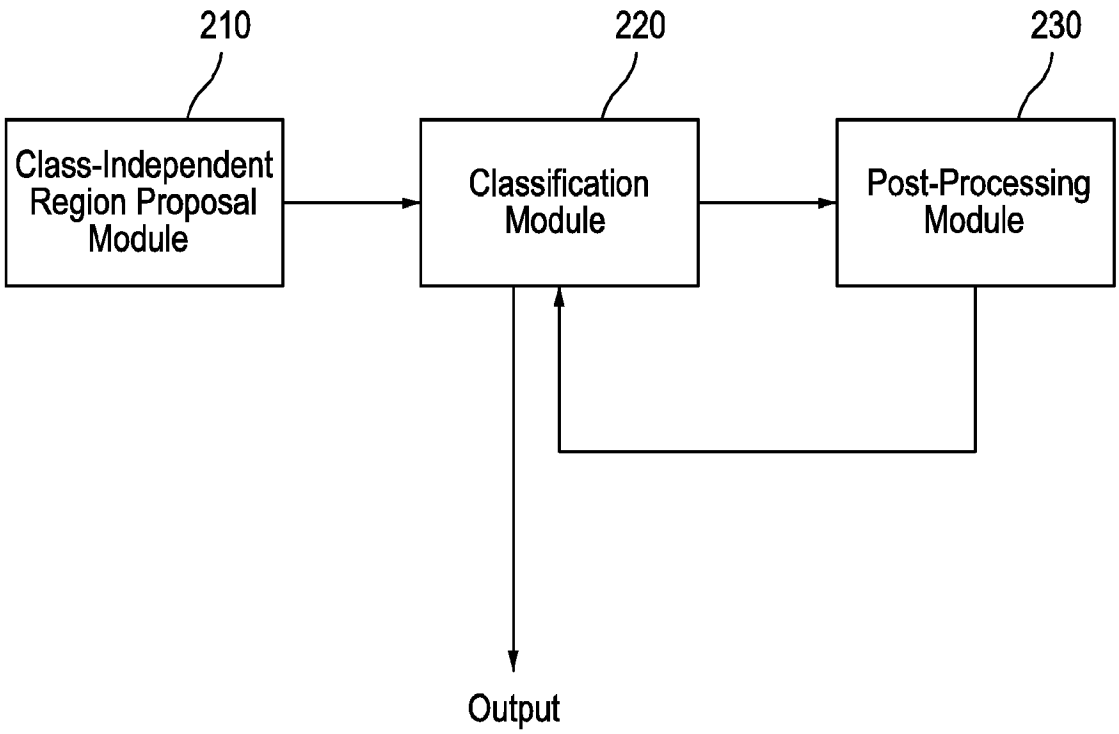

210
220
230
Class-Independent Region Proposal Module
Classification Module
Post-Processing Module
Output

Figure 2

Algorithm 1 Iterative Region Proposals

```
boxes ← selective_search(image)
scores ← get_hyperface_scores(boxes)
detected_boxes ← boxes(scores ≥ threshold)
new_boxes ← detected_boxes
for stage = 1 to T do
    fids ← get_hyperface_fiducials(new_boxes)
    new_boxes ← FaceRectCalculator(fids)
    deteced_boxes ← [deteced_boxes|new_boxes]
end
final_scores ← get_hyperface_scores(detected_boxes)
```

Figure 3

Algorithm 2 Landmarks-based NMS

1: Get **detected_boxes** from Algorithm 1
2: **fids** $\leftarrow$ $get\_hyperface\_fiducials$(**detected_boxes**)
3: **precise_boxes** $\leftarrow$ $[min_x, min_y, max_x, max_y]$(**fids**)
4: **faces** $\leftarrow$ nms(**precise_boxes**, $overlap$)
5: **for each** `face in faces` **do**
6: **top-k_boxes** $\leftarrow$ Get top-k scoring boxes
7: **final_fids** $\leftarrow$ $median(fids$(**top-k_boxes**$))$
8: **final_pose** $\leftarrow$ $median(pose$(**top-k_boxes**$))$
9: **final_gender** $\leftarrow$ $median(gender$(**top-k_boxes**$))$
10: **final_visibility** $\leftarrow$ $median(visibility$(**top-k_boxes**$))$
11: **final_bounding_box** $\leftarrow$ $FaceRectCalculator$(**final_fids**)
12: **end**

Figure 4

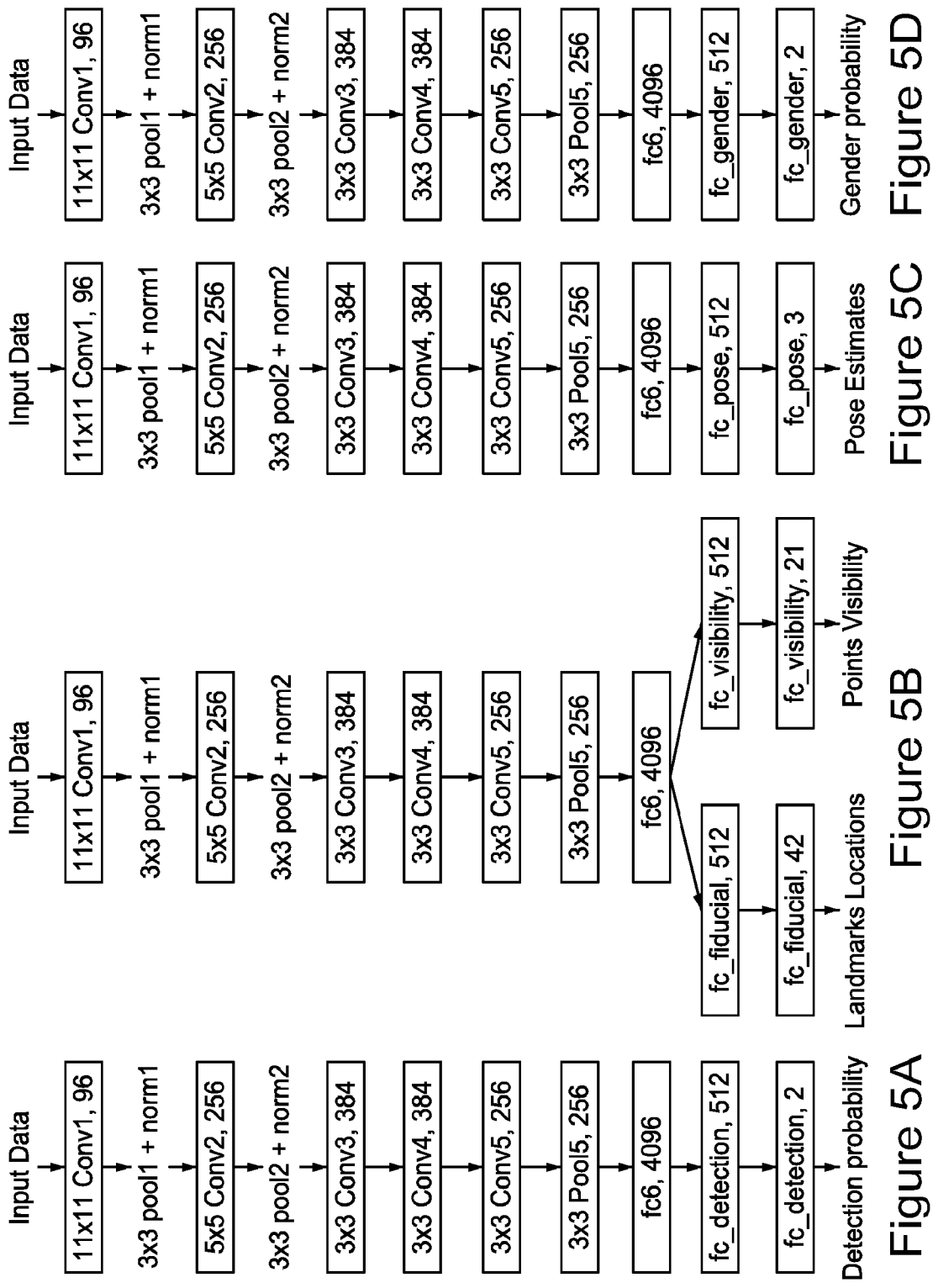


Figure 5A Figure 5B Figure 5C Figure 5D

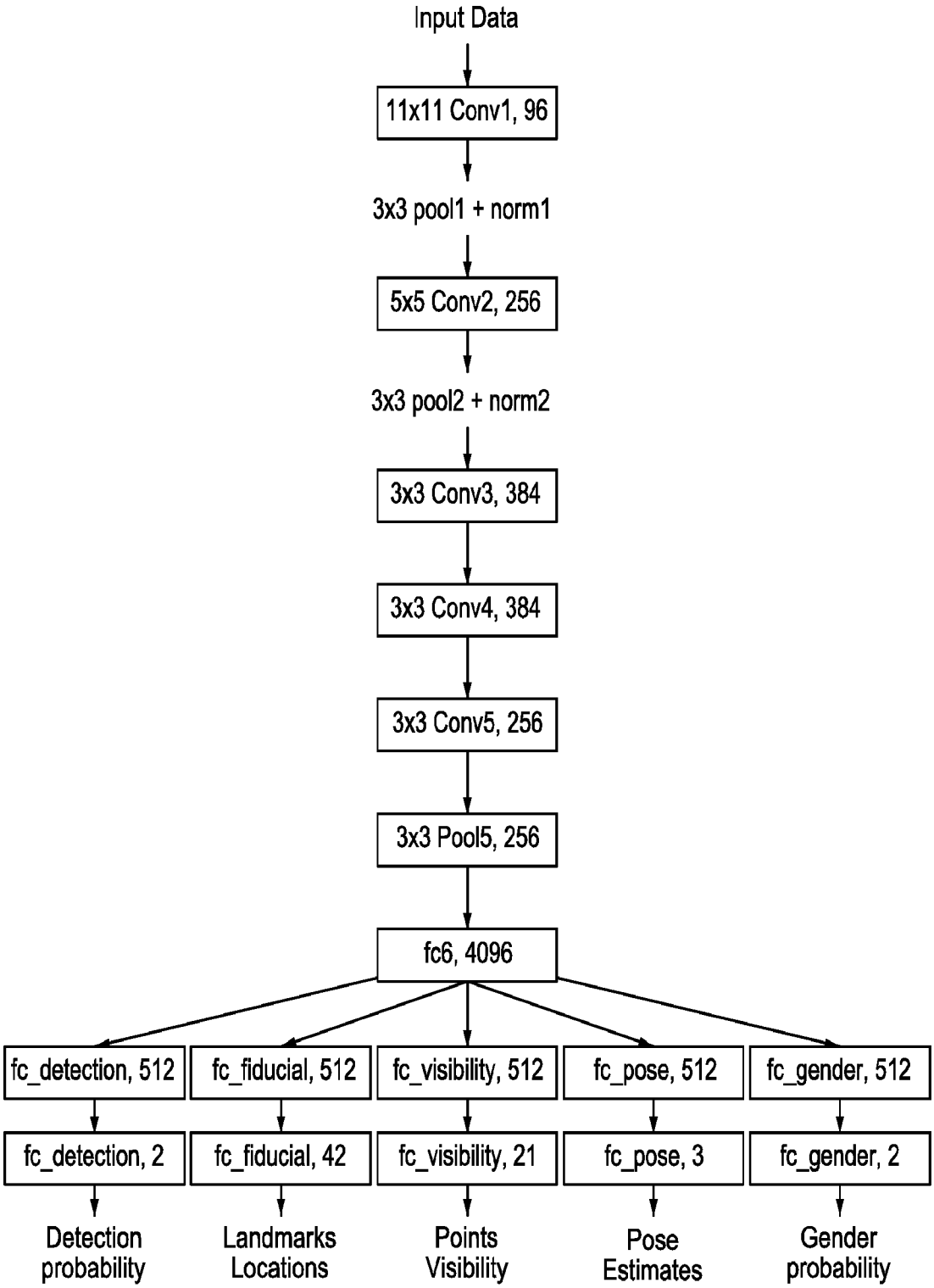

Input Data
11x11 Conv1, 96
3x3 pool1 + norm1
5x5 Conv2, 256
3x3 pool2 + norm2
3x3 Conv3, 384
3x3 Conv4, 384
3x3 Conv5, 256
3x3 Pool5, 256
fc6, 4096
fc_detection, 512
fc_fiducial, 512
fc_visibility, 512
fc_pose, 512
fc_gender, 512
fc_detection, 2
fc_fiducial, 42
fc_visibility, 21
fc_pose, 3
fc_gender, 2
Detection probability
Landmarks Locations
Points Visibility
Pose Estimates
Gender probability

Figure 6

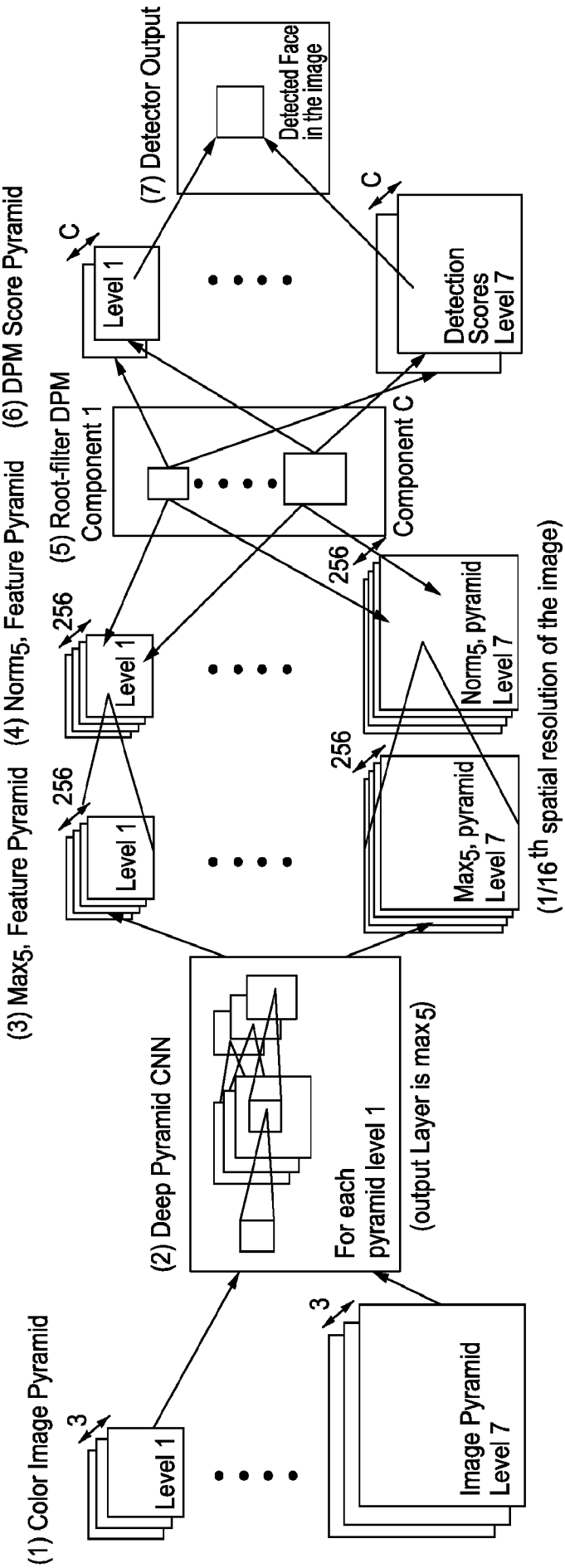

(1) Color Image Pyramid
3
Level 1
Image Pyramid Level 7
3
(2) Deep Pyramid CNN
For each pyramid level 1
(output Layer is max5)
(3) Max5, Feature Pyramid
256
Level 1
256
Level 1
(4) Norm5, Feature Pyramid
Max5, pyramid Level 7
256
Norm5, pyramid Level 7
256
(5) Root-filter DPM
Component 1
Component C
(6) DPM Score Pyramid
C
Level 1
C
Detection Scores Level 7
(7) Detector Output
Detected Face in the image
(1/16th spatial resolution of the image)

Figure 7

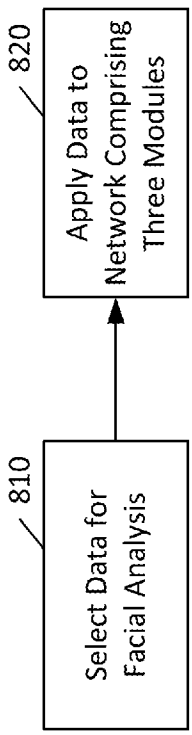

810
Select Data for Facial Analysis
820
Apply Data to Network Comprising Three Modules

Figure 8

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2016/043172

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - G06K 9/00; G06K 9/46; G06K 9/66; G06K 9/68 (2016.01)
CPC - G06K 9/00; G06K 9/46; G06K 9/66; G06K 9/68 (2016.08)
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)
IPC - G06K 9/00; G06K 9/46; G06K 9/66; G06K 9/68
CPC - G06K 9/00; G06K 9/46; G06K 9/66; G06K 9/68

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
USPC - 382/118; 382/159; 382/181; 382/224 (keyword delimited)

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
Orbit, Google Patents, Google Scholar, Google
Search terms used: facial recognition, pattern, detection, modules, images

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X --- Y	US 2005/0180627 A1 (YANG et al) 18 August 2005 (18.08.2005) entire document	1, 6, 11, 16 --- 2-5, 7-10, 12-15, 17-20
Y	US 2006/0110040 A1 (SIMARD et al) 25 May 2006 (25.05.2006) entire document	2-4, 7-9, 12-14, 17-19
Y	WO 2014/068567 A1 (WILF ITZHAK et al) 08 May 2014 (08.05.2014) entire document	4, 9, 14, 19
Y	US 2012/0237109 A1 (RAJPOOT NASIR MAHMOOD et al) 20 September 2012 (20.09.2012) entire document	5, 10, 15, 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	Date of mailing of the international search report
09 September 2016	22 SEP 2016
Name and mailing address of the ISA/ Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, VA 22313-1450 Facsimile No. 571-273-8300	Authorized officer Blaine R. Copenheaver PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774