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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
05 November 2020 (05.11.2020)

WIPO | PCT

(10) International Publication Number
WO 2020/221990 A1

(51) International Patent Classification:
G06K 9/00 (2006.01) *G06N 3/00* (2006.01)
G06K 9/62 (2006.01)

(21) International Application Number:
PCT/GB2020/051012

(22) International Filing Date:
24 April 2020 (24.04.2020)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
1906027.6 30 April 2019 (30.04.2019) GB

(71) Applicant: **HUAWEI TECHNOLOGIES CO., LTD.** [CN/CN]; Huawei Administration Building, Bantian, Longgang District, Shenzhen, Guangdong 518129 (CN).

(72) Inventors: **DENG, Jiankang**; c/o Facesoft Ltd, 80 Wood Lane, Shepherd's Bush, London W12 0BZ (GB). **ZAFEIRIOU, Stefanos**; c/o Facesoft Ltd, 80 Wood Lane, Shepherd's Bush, London W12 0BZ (GB).

(74) Agent: **GILL JENNINGS & EVERY LLP**; The Broadgate Tower, 20 Primrose Street, London EC2A 2ES (GB).

(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,

(54) Title: FACIAL LOCALISATION IN IMAGES

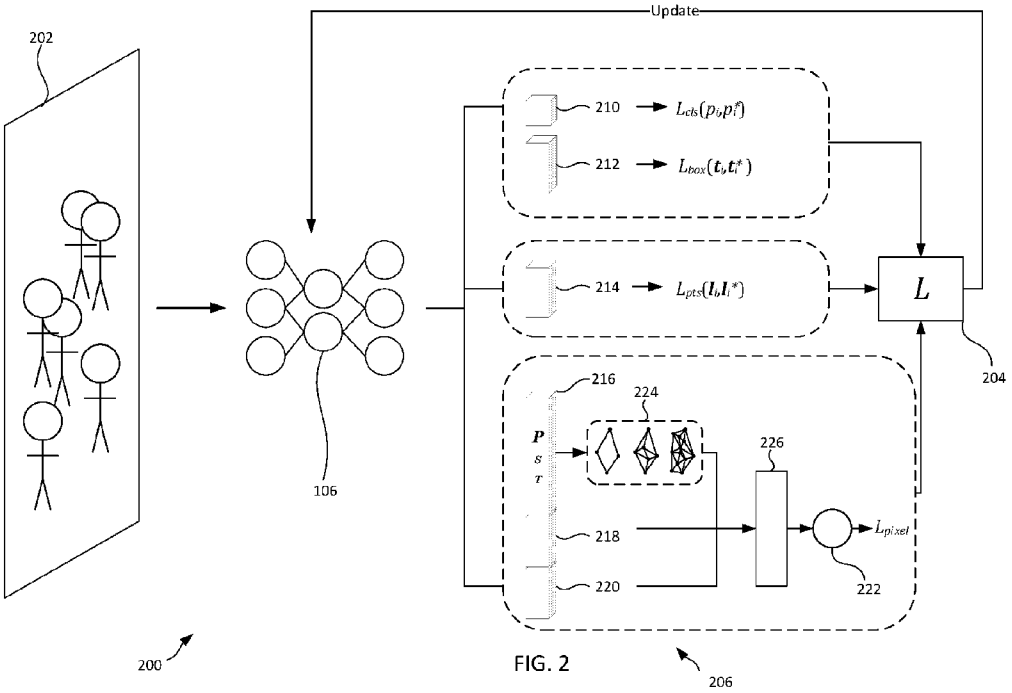

FIG. 2

(57) Abstract: This specification relates to methods of facial localisation in images. According to a first aspect of this specification, there is described a computer implemented method of training a neural network for face localisation, the method comprising: inputting, to the neural network, a training image, the training image comprising one or more faces; processing, using the neural network, the training image; outputting, from the neural network and for each of a plurality of training anchors in the training image, one or more sets of output data, each set of output data comprising a predicted facial classification, a predicted location of a corresponding face box, and one or more corresponding feature vectors; updating parameters of the neural network in dependence on an objective function. The objective function comprises, for each positive anchor in the training image: a classification loss comparing a predicted classification for the positive anchor to a known classification for the positive anchor; a box regression loss comparing a predicted locations of a face

[Continued on next page]

WO 2020/221990 A1

SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, 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))*

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

box for the positive anchor to a known location of the face box; and a feature loss comparing pixel-based properties of the one or more feature vectors for the positive anchor to known pixel-based properties of a face associated with the positive anchor.

Facial Localisation in Images

Field

This specification relates to methods of facial localisation in images. In particular, this specification relates to the use and training of a neural network for facial localisation in images.

Background

Facial localisation in images is the task of determining the presence and location of faces within an image. It may further comprise identifying the locations of bounding boxes that surround each face determined to be present in the image. Automatic face localisation is a prerequisite step of facial image analysis in many applications, such as facial attribute (e.g. expression and/or age) and/or facial identity recognition. Typically, the training of a neural network for facial localisation (a facial classification/detection neural network) in images comprises the use of only classification and box regression losses.

Facial detection differs from generic object detection in several respects. Facial detection features smaller ratio variations (for example, from 1:1 to 1:1.5) but much larger scale variations (for example, from several pixels to thousands of pixels). This can provide significant challenges when adapting generic object detectors for facial detection and localisation.

Summary

According to a first aspect of this specification, there is described a computer implemented method of training a neural network for face localisation, the method comprising: inputting, to the neural network, a training image, the training image comprising one or more faces; processing, using the neural network, the training image; outputting, from the neural network and for each of a plurality of training anchors in the training image, one or more sets of output data, each set of output data comprising a predicted facial classification, a predicted location of a corresponding face box, and one or more corresponding feature vectors; updating parameters of the neural network in dependence on an objective function. The objective function comprises, for each positive anchor in the training image: a classification loss comparing a predicted classification for the positive anchor to a known classification for the positive anchor; a box regression loss comparing a predicted locations of a face box for the positive anchor

to a known location of the face box; and a feature loss comparing pixel-based properties of the one or more feature vectors for the positive anchor to known pixel-based properties of a face associated with the positive anchor.

The one or more feature vectors may comprise predicted locations of a plurality of facial landmarks of a face, and wherein the feature loss comprises a facial landmark regression loss comparing the predicted locations of the plurality of facial landmarks to known locations of the plurality of facial landmarks.

The one or more feature vectors may comprise an encoded representation of a face within the predicted face box, and the method may further comprise: generating a three-dimensional representation of the face using a mesh decoder neural network, the mesh decoder neural network comprising one or more geometric convolutional layers; and generating a two-dimensional image of the face from the three-dimensional representation of the face using a differentiable renderer, wherein the feature loss comprises a dense regression loss comparing the generated two-dimensional image of the face to a ground truth image of the face within the predicted face box.

The feature vector may further comprise camera parameters and/or illumination parameters, and wherein the differentiable renderer uses the camera parameters and/or illumination parameters when generating the two-dimensional image of the face from the three-dimensional representation of the face.

The dense regression loss may be a pixel-wise difference between the generated two-dimensional image of the face and the ground truth image of the face within the predicted face box.

The mesh decoder may comprise a plurality of geometric convolutional layers and a plurality of upscaling layers, and wherein the plurality of upscaling layers are interlaced with the plurality of geometric convolutional layers.

The objective function may comprise, for each negative anchor in the training image, a classification loss comparing a predicted classification for the negative anchor to a known classification for the negative anchor.

The method may further comprise iterating the method using one or more further training images until a threshold condition is satisfied.

The neural network may comprise: a first plurality of convolutional layers comprising an input layer, a plurality of convolutional filters and one or more skip connections; and a second plurality of convolutional layers laterally connected to the first plurality of convolutional layers and configures to process the output of the first plurality of convolutional layers in a top-down manner. The second plurality of convolutional layers may comprise one or more deformable convolutional layers.

Each lateral connection may be configured to merge an output of a layer in the first plurality of convolutional layers with an output of a preceding layer of the second plurality of convolutional layers.

The neural network may further comprise one or more context modules, each context module configured to process the output of a layer of the second plurality of convolutional layers using one or more deformable convolutional networks.

According to a further aspect of this specification, there is described a method of face localisation, the method comprising: identifying, using a neural network, one or more faces within the input image; and identifying, using the neural network, one or more corresponding face boxes within the input image, wherein the neural network has been trained using any of the training methods described herein.

According to a further aspect of this specification, there is described a computer-implemented method of face localisation, the method comprising: inputting, to the neural network, an image, the image comprising one or more faces; processing, using the neural network, the image; outputting, from the neural network, one or more sets of output data, each set of output data comprising a predicted facial classification, a predicted location of a corresponding face box, and one or more corresponding feature vectors. The neural network comprises: a first plurality of convolutional layers comprising an input layer, a plurality of convolutional filters and one or more skip connections; and a second plurality of convolutional layers laterally connected to the first plurality of convolutional layers and configures to process the output of the first plurality of convolutional layers in a top-down manner.

Each lateral connection may be configured to merge an output of a layer in the first plurality of convolutional layers with an output of a preceding layer of the second plurality of convolutional layers. The second plurality of convolutional layers may comprise one or more deformable convolutions.

The neural network further may comprise one or more context modules, each context module configured to process the output of a layer of the second plurality of convolutional layers using one or more deformable convolutional networks.

The neural network may have been trained using any of the training methods described herein.

According to a further aspect of this specification, there is described a system comprising: one or more processors; and a memory, the memory comprising computer readable instructions that, when executed by the one or more processors, cause the system to perform one or more of the methods described herein.

According to a further aspect of this specification, there is described a computer program product comprising computer readable instructions that, when executed by a computing device, cause the computing device to perform one or more of the methods described herein.

As used herein, the terms facial localisation and/or face localisation are preferably used to connote one or more of: face detection; face alignment; pixel-wise face parsing; and/or 3D dense correspondence regression.

Brief Description of the Drawings

Embodiments will now be described by way of non-limiting examples with reference to the accompanying drawings, in which:

Figure 1 shows a schematic overview of a method of facial localisation in images;
Figure 2 shows a schematic overview of a method of training a neural network for facial localisation in images using a multi-task loss;
Figure 3 shows a flow diagram of a method of training a neural network for facial localisation in images;

Figure 4 shows an example of a neural network structure for facial localisation in images;

Figure 5 shows an example of a lateral connection in a neural network for facial localisation in images;

Figure 6 shows a further example of a neural network structure for facial localisation in images; and

Figure 7 shows an example of a computer system for performing facial localisation.

Detailed Description

In the embodiments described herein, a facial recognition neural network may be configured to output further feature vectors in addition to a face score (i.e. an indication of whether a face is present at a location or not) and a face box. For example, the facial recognition neural network may output one or more facial landmark locations. Alternatively or additionally, the facial recognition neural network may output an embedding of the region within the face box. These outputs may be used in a multi-task loss function to train the facial recognition neural network.

Use of such a multi-task loss function can result in a higher performance of the facial classification/detection neural network when compared to other facial localisation methods. For example, the facial classification/detection neural network may have a lower error rate.

In some embodiments, a lightweight backbone network is employed. This can allow the facial classification/detection neural network to be run in real-time on a single CPU core.

Figure 1 shows a schematic overview of a method 100 of facial localisation in images. An image 102 comprising one or more faces 104 is input into a neural network 106. The neural network 106 processes the image through a plurality of neural network layers, and outputs one or more sets of output facial data 108. Each set of output facial data 108 comprises a predicted facial classification for a face 106 in the image 102, a predicted location of a corresponding face box, and one or more corresponding feature vectors that encode one or more properties of a corresponding face in the image 102.

The input image 102, I, is an image comprising one or more faces. The input image 102 may comprise a set of pixel values in an array, for example a two-dimensional or three-

dimensional array. For example, in a colour image $I \in \mathbb{R}^{H \times W \times 3}$, where H is the height of the image in pixels, W is the width of the image in pixels and the image has three colour channels (e.g. RGB or CIELAB). The image 102 may, in some embodiments, be in black-and-white/greyscale.

The neural network 106 comprises a plurality of layers of nodes, each node associated with one or more parameters. The parameters of each node of the neural network may comprise one or more weights and/or biases. The nodes take as input one or more outputs of nodes in the previous layer. The one or more outputs of nodes in the previous layer are used by the node to generate an activation value using an activation function and the parameters of the neural network. One or more of the layers of the trained generator neural network 106 may be convolutional layers. One or more of the layers of the trained generator neural network 106 may be deformable convolutional layers.

In some embodiments, the neural network 106 may comprise a residual network and a feature pyramid network. The neural network 106 may comprise lateral connections between layers of the residual network and layers of the feature pyramid network. The feature pyramid may process the outputs of the residual network in a top-down manner. Examples of neural network architectures using a residual network and a feature pyramid are described below in relation to Figures 4 to 6.

The parameters of the neural network 106 may be trained using a multi-task objective function, $\mathcal{L}$, (also referred to herein as a loss function). The multi-task objective function may comprise two or more of a facial classification loss; a face box regression loss; a facial landmark regression loss; and/or a dense face regression loss. The multi-task loss function may comprise a supervised loss and/or a self-supervised loss. Examples of method of training the neural network are described below in relation to Figures 2 and 3.

The set of output facial data 108 comprises, for each potential face identified in the image 102 by the neural network, a predicted facial classification 108a for a face 106 in the image 102. The predicted classification 108a comprises a probability of an anchor in the image being a face.

The set of output facial data 108 further comprises, for each potential face identified in the image 102 by the neural network, a predicted bounding box 108b for a face 106 in the image 102. The predicted bounding box 108b may, for example, be the coordinates of the corners of a bounding box 110 of the potential face.

The set of output facial data 108 further comprises, for each potential face identified in the image 102 by the neural network, one or more feature vectors 108c encoding one or properties of the face. For example, a feature vector 108c may comprise the location of one or more facial landmarks, such as the eyes, nose, chin and/or mouth of a face. The feature vector 108c may alternatively or additionally comprise an embedding of the face contained within the bounding box 110.

Figure 2 shows a schematic overview of a method of training a neural network for facial localisation in images using a multi-task loss. The method comprises training the neural network 106 on a set of training images 102 using a multi-task loss function 204. For each of a plurality of images 202 in a training dataset, training image 202 is input into the neural network 106 and processed through a plurality of neural network layers. The neural network 106 outputs a set of predicted facial data 206 based on the input training image 202. The predicted facial data 206 is compared to known facial data for the corresponding training image 202 using a multi-task loss function 204, and the comparison is used to update parameters of the neural network 106.

The training image 202 is taken from a training set. The training set comprises a plurality of images containing faces, and a set of known facial data for each image. The known facial data comprises the locations of each face in the corresponding image, as well as the locations of the bounding boxes for each face in the corresponding image. In some embodiments, the known facial data further comprises locations of one or more facial landmarks for each face in the corresponding image.

Each training image may be associated with a plurality of training anchors. Each anchor represents a "prior" box where a face could potentially be in the image. Each anchor may be defined a reference box comprising a central location and a scale. However, it will be appreciated that the anchors may be defined in other ways. Anchors are used in facial detection to distinguish positive and negative training samples. A positive anchor is defined as an anchor for which a face is present. A negative anchor is defined as an anchor for which a face is not present. An example of the use of Anchors

is provided in "*Faster R-CNN: Towards Real-Time Object Detection with Regional Proposal Networks*" (Shaoqing Ren *et al.*, arXiv:1506.01497, the contents of which are incorporated herein by reference).

The predicted facial data 206 comprises a facial classification 210. The predicted facial classification 210 comprises a probability, p_i, of an associated anchor, i, being a face. The predicted facial classification 210 may, in some embodiments, have two components corresponding to "face present" and "face not present".

The predicted facial classification 210 for an anchor is associated with a classification loss, $\mathcal{L}_{clc}(p_i, p_i^*)$, where p_i^* is the known facial classification for anchor i (i.e. 1 for a positive anchor and 0 for a negative anchor). The classification loss compares a predicted classification for the anchor to a known classification for the anchor. The classification loss may, for example, be a softmax loss for binary cases, such as a cross entropy loss or binary hinge loss. However, the skilled person will appreciate that other loss functions may alternative be used. The classification loss may, in some embodiments, only be used for positive anchors. In other embodiments, it may be used for both positive and negative anchors.

The predicted facial data 206 comprises a predicted facial box location 212, $\boldsymbol{t}_i$. The predicted facial box location 212 comprises data identifying a predicted location of a bounding box for the corresponding face for an anchor, i. The predicted facial box location 212 may be, for example, coordinates of the corners of the predicted bounding box for the face, such as all four corners of the box, or coordinates of two diagonally opposite corners of the box. The predicted facial box location 212 may alternatively be the coordinates of the centre of the bounding box and the height and width of the box. Other examples are possible.

The predicted facial box location 212 for a positive anchor is associated with a box regression loss, $\mathcal{L}_{box}(\boldsymbol{t}_i, \boldsymbol{t}_i^*)$, where $\boldsymbol{t}_i^*$ is the known facial box location for positive anchor i. The box regression loss compares the predicted facial box location 212, $\boldsymbol{t}_i$, to the known location of the corresponding face box, $\boldsymbol{t}_i^*$. The box regression loss may, for example, be a robust loss function, such as a smooth-L1 loss function. However, the skilled person will appreciate that other loss functions may alternative be used. In some embodiments, the box regression targets may be normalised to a centre location and width and height of the box.

The predicted facial data also comprises one or more predicted feature vectors that each encode one or more pixel-based properties of the area with the facial box. The feature vectors are each associated with a feature loss, $\mathcal{L}_f$, comparing pixel-based properties of the one or more feature vectors for the positive anchor to known pixel-based properties of the face associated with the positive anchor.

An example of such a feature vector is a facial landmark data 214. The facial landmark data 214 comprises the predicted coordinates/locations of one or more facial landmarks within the predicted facial box. For example, five facial landmarks corresponding to the eyes, the corners of the mouth and the nose may be used.

The facial landmark data 214 is associated with a facial landmark regression loss, $\mathcal{L}_{pts}(\boldsymbol{l}_i, \boldsymbol{l}_i^*)$, where $\boldsymbol{l}_i^*$ are the known facial landmark locations for positive anchor i. The facial landmark regression loss compares the predicted facial landmark locations 214, $\boldsymbol{l}_i$, to the known location of the corresponding facial landmarks, $\boldsymbol{l}_i^*$. The facial landmark regression loss may, for example, be a robust loss function, such as a smooth-L1 loss function. However, the skilled person will appreciate that other loss functions may alternative be used. In some embodiments, the facial landmark location targets may be normalised to a centre location.

Alternatively or additionally, the feature vector may comprise an embedding 216, P_{ST}, of potential face within the bounding box. The embedding 216 may comprise shape and texture parameters of the face within the corresponding bounding box. In some embodiments, camera parameters 218 and/or lighting parameters 220 may also be output by the neural network 106. Camera parameters 218 may comprise one or more of: a camera location; a camera pose; and/or a focal length of the camera. Lighting (or illumination) parameters may comprise one or more of: a light source location; colour values of the light source; and/or ambient light colours. The camera parameters 218 and/or lighting parameters 220 may be referred to as rendering parameters.

The embedding 216 may be associated with a dense regression loss function, L_{pixel}. The dense regression loss does not directly compare the embedding 216 to a known embedding. Instead, the embedding is post-processed to generate a predicted facial image 222, which is then compared to the potential facial image (i.e. the image in the

bounding box) using the regression loss function. This branch of the process may therefore be considered to be self-supervised.

To generate the predicted facial image 222, the embedding may be processed using a mesh decoder neural network 224 to generate a three dimensional representation of the potential face, such a three dimensional mesh. The mesh decoder neural network 224 comprises a plurality of layers of nodes, each node associated with one or more parameters. The parameters of each node of the neural network may comprise one or more weights and/or biases. The nodes take as input one or more outputs of nodes in the previous layer. The one or more outputs of nodes in the previous layer are used by the node to generate an activation value using an activation function and the parameters of the neural network. The mesh decoder neural network 224 comprises one or more geometric convolutional layers (also referred to herein as mesh convolutional layers). A geometric convolutional layer is a type of convolutional filter layer of the neural network used in geometric deep learning that can be applied directly in the mesh domain.

The mesh decoder may further comprise one or more up-scaling layers configured to upscale the mesh (i.e. increase the number of vertices in the mesh). The up-scaling layers may be interlaced with the geometric convolutional layers. For example the geometric convolutional layers may alternate with the up-scaling layers.

The mesh decoder 224 is configured to take as input the embedding parameters 216 and to process the input using a series of neural network layers to output a shape map and a texture map of the two dimensional potential face from the training image. The output of the mesh decoder neural network may be symbolically represented as $\mathcal{D}_{P_{ST}}$. The shape and texture map comprises three dimensional shape and texture information relating to the input visual data. In some embodiments, the output is a set of parameters describing (x, y, z) coordinates of points in a mesh and corresponding (r, g, b) values for each point in the mesh. The output may, in some embodiments, be a coloured mesh within a unit sphere.

The mesh decoder may be pre-trained, and have its parameters fixed during training of the neural network 106. An example of a mesh decoder neural network and how a mesh decoder neural network may be trained is described in relation to co-pending GB

Patent Application No. 902524.6 (the contents of which are incorporated herein by reference).

A mesh may be defined in terms of an undirected connected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} \in \mathbb{R}^{n\times 6}$ is a set of n vertices containing joint shape (e.g. (x, y, z)) and texture (e.g. (r, g, b)) information and $\mathcal{E} \in \{0,1\}^{n\times n}$ is an adjacency matrix defining the connection status between vertices. However, it will be appreciated that alternative representations of a mesh may be used.

An example of a convolutional operator on a graph/mesh can be defined by formulating mesh filtering with a kernel g_θ using a recursive polynomial, such as a Chebyshev polynomial. In defining such a convolutional operator on a mesh, it is useful to define some intermediate variables in the following way. A non-normalised graph Laplacian may be defined as $L = D - \mathcal{E} \in \mathbb{R}^{n\times n}$ where $D \in \mathbb{R}^{n\times n}$ is a diagonal matrix with $D_{ii} = \sum_j \mathcal{E}_{ij}$. A normalised graph Laplacian may then be defined as $L = I_n - D^{-1/2}\mathcal{E}D^{-1/2}$, where I_n is the identity matrix.

The Laplacian can be diagonalised by the Fourier bases $U = [u_0, \dots, u_{n-1}] \in \mathbb{R}^{n\times n}$ such that $L = U\Lambda U^T$, where $\Lambda = \mathrm{diag}([\lambda_0, \dots, \lambda_{n-1}]) \in \mathbb{R}^{n\times n}$. A graph Fourier representation of a given mesh, $\mathbf{x} \in \mathbb{R}^{n\times 6}$, can then be defined as $\hat{\mathbf{x}} = U^T\mathbf{x}$ with inverse $\mathbf{x} = U\hat{\mathbf{x}}$.

Given these definitions, an example of a convolutional operator, g_θ, on a graph/mesh can be defined. The filter g_θ can be parameterised as a truncated Chebyshev polynomial expansion of order K as

$$g_\theta(\Lambda) = \sum_{k=0}^{K-1} \theta_k T_k(\tilde{\Lambda})$$

where $\boldsymbol{\theta} \in \mathbb{R}^K$ is a vector of Chebyshev coefficients and $T_k(\tilde{\Lambda}) \in \mathbb{R}^{n\times n}$ is the Chebyshev polynomial of order k evaluated at a scaled Laplacian $\tilde{\Lambda} = 2\Lambda/\lambda_{max} - I_n$. $T_k(x)$ can be recursively computed using $T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x)$, with $T_0(x) = 1$ and $T_1(x) = x$.

A spectral convolution can then be defined as

$$y_j = \sum_{i=1}^{F_{in}} g_{\theta_{i,j}}(\Lambda)\, x_i \in \mathbb{R}^n$$

where $\mathbf{x} \in \mathbb{R}^{n \times F_{in}}$ is the input and $\mathbf{y} \in \mathbb{R}^{n \times F_{out}}$. Such a filtering operation is very efficient, costing only $\mathcal{O}(K|\mathcal{E}|)$ operations.

Other types of geometric/mesh convolution (or graph convolutions) may alternatively or additionally be used. For example, other sets of orthogonal polynomials may be used in place of the Chebyshev polynomials.

The joint shape and texture map undergoes further processing using a differentiable renderer 226 to generate a 2D rendered representation 222, , $\mathcal{R}(\mathcal{D}_{P_{ST}})$, of the potential face within the bounding box. The representation may comprise a two-dimensional projection of the joint shape and texture map onto an image plane. The sets of rendering parameters 218, 220 may be used to render the representation of the two-dimensional visual data, for example to define the parameters of a camera model used in the rendering, i.e. the 2D rendered representation 222 is given by $\mathcal{R}(\mathcal{D}_{P_{ST}}, P_{cam}, P_{ill})$. The differentiable renderer 226 may, for example, be a 3D mesh renderer that projects the coloured mesh $\mathcal{D}_{P_{ST}}$ onto a 2D image plane. An example of such a differentiable renderer 226 may be found in Genova *et al.* (CVPR, 2018. 2,3), though other renderers may alternatively be used.

The dense regression loss, L_{pixel}, compares the rendered 2D face 222, $\mathcal{R}$, to the original facial image in the training image 202 (i.e. the ground truth image within the predicted facial box). A pixel-wise difference may be used to perform the comparison. An example of the dense regression loss may be given by:

$$\mathcal{L} = \frac{1}{H \times W} \sum_{i}^{W} \sum_{j}^{H} \left\| \mathcal{R}(\mathcal{D}_{P_{ST}})_{i,j} - I^*_{i,j} \right\|_n$$

where I^* is the anchor crop of the ground-truth training image and H and W are the width and height of the anchor crop respectively. $\|x\|_n$ is an L-n norm, such as an L-1 norm. Other pixel-wise loss functions may alternatively be used.

The multi-task objective function 204 comprises a combination of the classification loss, the box regression loss and the feature loss. The aim of the training is, for each training anchor i, to minimise the multi-task objective function. An example of a multi-task loss function for an anchor i is:

$$\mathcal{L} = \mathcal{L}_{clc}(p_i, p_i^*) + \lambda_1 p_i^* \mathcal{L}_{box}(\boldsymbol{t}_i, \boldsymbol{t}_i^*) + \lambda_2 p_i^* \mathcal{L}_{pts}(\boldsymbol{l}_i, \boldsymbol{l}_i^*) + \lambda_3 p_i^* \mathcal{L}_{pixel},$$

where $\{\lambda_i\}$ are constants that control the relative significance of the different loss functions in the multi-task objective function. These parameters may, for example, be in the range (0,1]. For example, λ_1=0.25, λ_2=0.1, and λ_3=0.01.

The multi-task objective function 204 is used to update the parameters of the neural network 106 during training. The parameters of the neural network 106 may be updated using an optimisation procedure that aims to optimise the multi-task loss function 208. The optimisation procedure may, for example, be a gradient descent algorithm. The optimisation procedure may use backpropagation, e.g. by back-propagating the multi-task loss function 204 to the parameters of the neural network 106. The optimisation procedure may be associated with a learning rate which varies during training. For example, the learning rate may start from an initial value, such as 10^(-3), and then increase after a first threshold number of epochs, for example to 10^(-2) after 5 epochs. The learning rate may then decay after further threshold epochs, for example reducing by a pre-defined factor at pre-defined epoch numbers. For example, the learning rate may decay by a factor of ten at 55 and 68 epochs.

Figure 3 shows a flow diagram of a method of training a neural network for facial localisation in images. The method may correspond to the method described above in relation to Figure 2.

At operation 3.1, a training image is input to the neural network, the training image comprising one or more faces. The training image may be selected from a training dataset comprising a plurality of images, each image comprising one or more faces at known locations and with known bounding boxes. One or more features of the faces, such as the locations of facial landmarks, may also be known.

At operation 3.2, the input training image is processed using the neural network.

At operation 3.3, the neural network outputs, for each of a plurality of training anchors in the training image, one or more sets of output data, each set of output data comprising a predicted facial classification, a predicted location of a corresponding face box, and one or more corresponding feature vectors.

Operations 3.1 to 3.3 may be iterated over a subset of the training examples in the training data to form an ensemble of training examples. This ensemble may be used to determine the objective function in operation 3.4, for example by averaging the multi-task objective function over the ensemble. Alternatively, operation 3.4 may be performed for each training example separately.

At operation 3.4, parameters of the neural network are updated in dependence on an objective function. The objective function comprises, for each positive anchor in the training image: a classification loss comparing a predicted classification for the positive anchor to a known classification for the positive anchor; a box regression loss comparing a predicted locations of a face box for the positive anchor to a known location of the face box; and a feature loss comparing pixel-based properties of the one or more feature vectors for the positive anchor to known pixel-based properties of a face associated with the positive anchor. The update may be performed after each iteration of operations 3.1 to 3.3. Alternatively, the update may be performed after a number of iterations of operations 3.1 to 3.3. In these embodiments, the multi-task loss function may comprise one or more expectation values taken over the ensemble of training examples.

Operations 3.1 to 3.4 may be iterated over the training dataset until a threshold condition is met. The threshold condition may be a threshold number of training epochs. The threshold number of training epochs may lie in the range 50-150, such as 70-90, for example 80.

Figure 4 shows an example of a neural network structure for facial localisation in images. The neural network 106 may, in some embodiments, use this structure.

The neural network 400 comprises a backbone network 402. The backbone network 402 may comprise a plurality of convolutional layers. The backbone network 402 may further comprise a plurality of down-sampling layers that reduce the dimensionality of an input to the layer. The backbone network may be initialised as a pre-trained classification network, such as a ResNet network (e.g. ResNet-152). The structure of the backbone network 402 may, for example, be given for an input image 404 size of 640x640 colour pixels by:

layer name	152-layer	output size
Input Image Crop		640×640×3
	7×7, 64, stride 2	320×320×64
	3×3, max pool, stride 2	160×160×64
C2	[1×1, 64 3×3, 64 1×1, 256] ×3	160×160×256
C3	[1×1, 128 3×3, 128 1×1, 512] ×8	80×80×512
C4	[1×1, 256 3×3, 256 1×1, 1024] ×36	40×40×1024
C5	[1×1, 512 3×3, 512 1×1, 2048] ×3	20×20×2048
C6	3×3, 256, stride 2	10×10×256

where the convolutional building blocks are shown in square brackets and down-sampling is performed by conv3_1, conv4_1 and conv5_1 with a stride of 2. The backbone network 402 is configured to process the input image 404 through a series of convolutional layers (in this example labelled C2-C6) in a "bottom up" manner (i.e. from a low indexed layer to a high indexed layer) and generate an output for each convolutional layer. Although six layers are shown in this example, fewer or greater numbers of layers may be used. During training, the one or more layers of the pre-trained backbone network, such as the first two layers, may be fixed to improve accuracy.

During training one or more of the convolutional layers may be fixed to their initial, pre-trained value. For example, the first two layers may be fixed. This can achieve higher accuracy of the neural network once trained.

The neural network 400 further comprises a feature pyramid network 406. The feature pyramid network 406 comprises a plurality of feature pyramid levels (in this example labelled P2-P6, with P6 acting directly on/being shared with the final layer of the backbone network 402). The feature pyramid levels may comprise convolutional layers. The convolutional layers may be deformable convolutional layers. The convolutional layers may be interlaced with one or more upscaling layers. The feature pyramid 406 is configured to process the outputs of a corresponding layer in the backbone network

402 in a "top-down" manner (i.e. from a high indexed layer to a low indexed layer) in order to generate a plurality of feature maps.

An example of a feature pyramid structure is given by:

Feature Pyramid	Stride	Anchor
P_2 (160 × 160 × 256)	4	16, 20.16, 25.40
P_3 (80 × 80 × 256)	8	32, 40.32, 50.80
P_4 (40 × 40 × 256)	16	64, 80.63, 101.59
P_5 (20 × 20 × 256)	32	128, 161.26, 203.19
P_6 (10 × 10 × 256)	64	256, 322.54, 406.37

where The dimension of the output of each layer is indicated in the first column, the "anchor" column indicates the scales associated with each layer of the pyramid, and the "stride" column indicates the stride of the convolutions performed by each layer. For example, in this example the P2 layer is associated with the anchor scales 16x16, 20.16x20.16 and 25.4x25.4 and performs convolutions with stride 4. However, it will be appreciated that different scales and convolutions may alternatively be used. The lower levels of the feature pyramid may be associated with smaller scales than the higher levels and/or have a smaller stride.

The neural network 400 comprises a plurality of lateral connections 408 between the output of layers of the backbone network 402 and the layers of the feature pyramid 406. The lateral connections 408 may, in some embodiments, process the output of the associated backbone network 402 layer and then input it to the corresponding feature pyramid layer, along with the output of a higher feature pyramid layer. In other words, each feature pyramid layer after the highest layers receives as input a combination of the output of the corresponding backbone network layer and the output of the previous higher layer of the feature pyramid, e.g. layer P3 of the feature pyramid receives as input a combination of the output of layer C3 of the backbone network and the output of layer P4 of the feature pyramid. The higher layers of the feature pyramid (in this example layers P6 and P5) receive their input from the backbone network without combining it with the output of a previous layer.

Each layer of the feature pyramid outputs a set of features. The features may be further processed to generate the output data 108 and/or the multi-task loss 208. The number of features in each layer may be fixed to be the same. For example, each layer may be fixed to have 256 features.

The use of lateral connections allows the "bottom up" pathways and "top down" pathways to be combined. In general, top-down pathways may be semantically strong but spatially coarse, while the bottom-up pathways may be spatially dedicate but semantically weak. By combing the pathways using lateral connections, the performance of the neural network in facial localisation can be improved.

The backbone network 402 may, in some embodiments, be a MobileNet network, such MobileNet-0.25. This provides a more lightweight model that can run on a single CPU in substantially real time. These lightweight models may use 7x7 convolution with stride 4 on the input image to reduce the data size. Dense anchors may be tiled on P4-P5 to improve performance. In some embodiments, the deformable convolutional layers are replaced with convolutional layers to speed up the processing. During training, the one or more layers of the pre-trained backbone network, such as the first two layers, may be fixed to improve accuracy.

Figure 5 shows an example of a lateral connection in a neural network for facial localisation in images. In this example, the lateral connection is for the P3 level of the feature pyramid, though it may *mutatis mutandis* be applied to any of the lateral connections between layers.

The lateral connection 500 provides the output of a layer 502 of the backbone network 402 (in this example the C3 layer of the backbone network) to a corresponding layer of the feature pyramid network 406 (in this example the P3 layer of the feature pyramid). The lateral connection 500 may be configured to process the output of the layer 502 of the backbone network 402 to generate processed output 504. The processing may, for example be to alter the dimensionality of the output of the layer 502 of the backbone network 402 so that it can be used as input to the corresponding feature pyramid layer. In the example shown, a 1x1 convolution is used to reduce the dimensionality of the output of the layer 502 of the backbone network from 80x80x512 to 80x80x256. Each lateral connection may apply a different convolution to reduce or increase the dimensionality of the associated output layer to match the corresponding feature pyramid layer.

The output from the immediately higher layer 506 of the feature pyramid (in this example P4) undergoes upsampling to increase its dimensionality. In the example

shown, the output of the P4 layer of the pyramid network is up-scaled by a factor of two from 40x40x256 to 80x80x256. The upscaling may, for example, be performed by nearest neighbour up-sampling. The up-scaled output is then added to the processed output 504 from the lateral connection 500 in order to generate a merged map 508. Element wise addition may be used to merge the up-scaled output with the processed output 504.

The merged map 508 is used as input to the layer of the feature pyramid network 406 corresponding to the lateral connection 500 (in this example the P3 layer). The P3 layer applies a 3x3 deformable convolution to the merged map 508 in order to generate an output feature map 510. The use of a deformable convolution can reduce the aliasing effects of upsampling, as well as perform non-rigid context modelling.

Figure 6 shows a further example of a neural network structure for facial localisation in images. The neural network 600 of this example is substantially the same as that of Figure 4, but with the addition of a context module 602 for each of the feature pyramid layers. The context module 602 may increase the receptive field and enhance the rigid context modelling power of the neural network 600.

Each context module 602 takes as input the feature map 604 from a corresponding feature pyramid layer. A sequence of deformable convolutions is performed using the feature map 604. Each deformable convolution after a first deformable convolution is applied to the output 606a-c of the previous deformable convolution. In the example shown, three 3x3 deformable convolutions are applied. However, it will be appreciated that fewer or greater numbers of deformable convolutions may be applied, and that different sized deformable convolutions may be used.

The outputs of each deformable convolution 606a-c are combined to form an enhanced output 608. The enhanced output 608 may be further processed to generate the output data 108 and/or the multi-task loss 208.

Figure 7 shows a schematic example of a system/apparatus for performing any of the methods described herein. The system/apparatus shown is an example of a computing device. It will be appreciated by the skilled person that other types of computing devices/systems may alternatively be used to implement the methods described herein, such as a distributed computing system.

The apparatus (or system) 700 comprises one or more processors 702. The one or more processors control operation of other components of the system/apparatus 700. The one or more processors 702 may, for example, comprise a general purpose processor. The one or more processors 702 may be a single core device or a multiple core device. The one or more processors 702 may comprise a central processing unit (CPU) or a graphical processing unit (GPU). Alternatively, the one or more processors 702 may comprise specialised processing hardware, for instance a RISC processor or programmable hardware with embedded firmware. Multiple processors may be included.

The system/apparatus comprises a working or volatile memory 704. The one or more processors may access the volatile memory 704 in order to process data and may control the storage of data in memory. The volatile memory 704 may comprise RAM of any type, for example Static RAM (SRAM), Dynamic RAM (DRAM), or it may comprise Flash memory, such as an SD-Card.

The system/apparatus comprises a non-volatile memory 706. The non-volatile memory 706 stores a set of operation instructions 708 for controlling the operation of the processors 702 in the form of computer readable instructions. The non-volatile memory 706 may be a memory of any kind such as a Read Only Memory (ROM), a Flash memory or a magnetic drive memory.

The one or more processors 702 are configured to execute operating instructions 408 to cause the system/apparatus to perform any of the methods described herein. The operating instructions 708 may comprise code (i.e. drivers) relating to the hardware components of the system/apparatus 700, as well as code relating to the basic operation of the system/apparatus 700. Generally speaking, the one or more processors 702 execute one or more instructions of the operating instructions 708, which are stored permanently or semi-permanently in the non-volatile memory 706, using the volatile memory 704 to temporarily store data generated during execution of said operating instructions 708.

Implementations of the methods described herein may be realised as in digital electronic circuitry, integrated circuitry, specially designed ASICs (application specific integrated circuits), computer hardware, firmware, software, and/or combinations

thereof. These may include computer program products (such as software stored on e.g. magnetic discs, optical disks, memory, Programmable Logic Devices) comprising computer readable instructions that, when executed by a computer, such as that described in relation to Figure 7, cause the computer to perform one or more of the methods described herein.

Any system feature as described herein may also be provided as a method feature, and vice versa. As used herein, means plus function features may be expressed alternatively in terms of their corresponding structure. In particular, method aspects may be applied to system aspects, and vice versa.

Furthermore, any, some and/or all features in one aspect can be applied to any, some and/or all features in any other aspect, in any appropriate combination. It should also be appreciated that particular combinations of the various features described and defined in any aspects of the invention can be implemented and/or supplied and/or used independently.

Although several embodiments have been shown and described, it would be appreciated by those skilled in the art that changes may be made in these embodiments without departing from the principles of this disclosure, the scope of which is defined in the claims.

Claims

1. A computer implemented method of training a neural network for face localisation, the method comprising:

inputting, to the neural network, a training image, the training image comprising one or more faces;

processing, using the neural network, the training image;

outputting, from the neural network and for each of a plurality of training anchors in the training image, one or more sets of output data, each set of output data comprising a predicted facial classification, a predicted location of a corresponding face box, and one or more corresponding feature vectors;

updating parameters of the neural network in dependence on an objective function, wherein the objective function comprises, for each positive anchor in the training image:

a classification loss comparing a predicted classification for the positive anchor to a known classification for the positive anchor;

a box regression loss comparing a predicted locations of a face box for the positive anchor to a known location of the face box; and

a feature loss comparing pixel-based properties of the one or more feature vectors for the positive anchor to known pixel-based properties of a face associated with the positive anchor.

2. The method of claim 1, wherein the one or more feature vectors comprises predicted locations of a plurality of facial landmarks of a face, and wherein the feature loss comprises a facial landmark regression loss comparing the predicted locations of the plurality of facial landmarks to known locations of the plurality of facial landmarks.

3. The method of claim 1 or 2, wherein the one or more feature vectors comprises an encoded representation of a face within the predicted face box, and wherein the method comprises:

generating a three-dimensional representation of the face using a mesh decoder neural network, the mesh decoder neural network comprising one or more geometric convolutional layers; and

generating a two-dimensional image of the face from the three-dimensional representation of the face using a differentiable renderer,

wherein the feature loss comprises a dense regression loss comparing the generated two-dimensional image of the face to a ground truth image of the face within the predicted face box.

4. The method of claim 3, wherein the feature vector further comprises camera parameters and/or illumination parameters, and wherein the differentiable renderer uses the camera parameters and/or illumination parameters when generating the two-dimensional image of the face from the three-dimensional representation of the face.

5. The method of any of claims 3 or 4, wherein the dense regression loss is a pixel-wise difference between the generated two-dimensional image of the face and the ground truth image of the face within the predicted face box.

6. The method of any of claims 3 to 5, wherein the mesh decoder comprises a plurality of geometric convolutional layers and a plurality of upscaling layers, and wherein the plurality of upscaling layers are interlaced with the plurality of geometric convolutional layers.

7. The method of any preceding claim, wherein the objective function comprises, for each negative anchor in the training image, a classification loss comparing a predicted classification for the negative anchor to a known classification for the negative anchor.

8. The method of any preceding claim, further comprising iterating the method using one or more further training images until a threshold condition is satisfied.

9. The method of any preceding claim, wherein the neural network comprises:

a first plurality of convolutional layers comprising an input layer, a plurality of convolutional filters and one or more skip connections; and

a second plurality of convolutional layers laterally connected to the first plurality of convolutional layers and configures to process the output of the first plurality of convolutional layers in a top-down manner.

10. The method of claim 9, wherein each lateral connection is configured to merge an output of a layer in the first plurality of convolutional layers with an output of a preceding layer of the second plurality of convolutional layers.

11. The method of claims 9 or 10, wherein the second plurality of convolutional layers comprises one or more deformable convolutions.

12. The method of any of claims 9 to 11, wherein the neural network further comprises one or more context modules, each context module configured to process the output of a layer of the second plurality of convolutional layers using one or more deformable convolutional networks.

13. A computer implemented method of face localisation, the method comprising:

identifying, using a neural network, one or more faces within the input image; and

identifying, using the neural network, one or more corresponding face boxes within the input image,

wherein the neural network has been trained using the method of any preceding claim.

14. A computer implemented method of face localisation, the method comprising:

inputting, to the neural network, an image, the image comprising one or more faces;

processing, using the neural network, the image;

outputting, from the neural network, one or more sets of output data, each set of output data comprising a predicted facial classification, a predicted location of a corresponding face box, and one or more corresponding feature vectors,

wherein the neural network comprises:

a first plurality of convolutional layers comprising an input layer, a plurality of convolutional filters and one or more skip connections; and

a second plurality of convolutional layers laterally connected to the first plurality of convolutional layers and configures to process the output of the first plurality of convolutional layers in a top-down manner.

15. The method of claim 14, wherein each lateral connection is configured to merge an output of a layer in the first plurality of convolutional layers with an output of a preceding layer of the second plurality of convolutional layers.

16. The method of claims 14 or 15, wherein the second plurality of convolutional layers comprises one or more deformable convolutions.

17. The method of any of claims 14 to 16, wherein the neural network further comprises one or more context modules, each context module configured to process the output of a layer of the second plurality of convolutional layers using one or more deformable convolutional networks.

18. The method of any of claims 14-17, wherein the neural network has been trained using the method of any of claims 1-8

19. A system comprising:

one or more processors; and

a memory, the memory comprising computer readable instructions that, when executed by the one or more processors, cause the system to perform the method of any preceding claim.

20. A computer program product comprising computer readable instructions that, when executed by a computing device, cause the computing device to perform the method of any of claims 1-18.

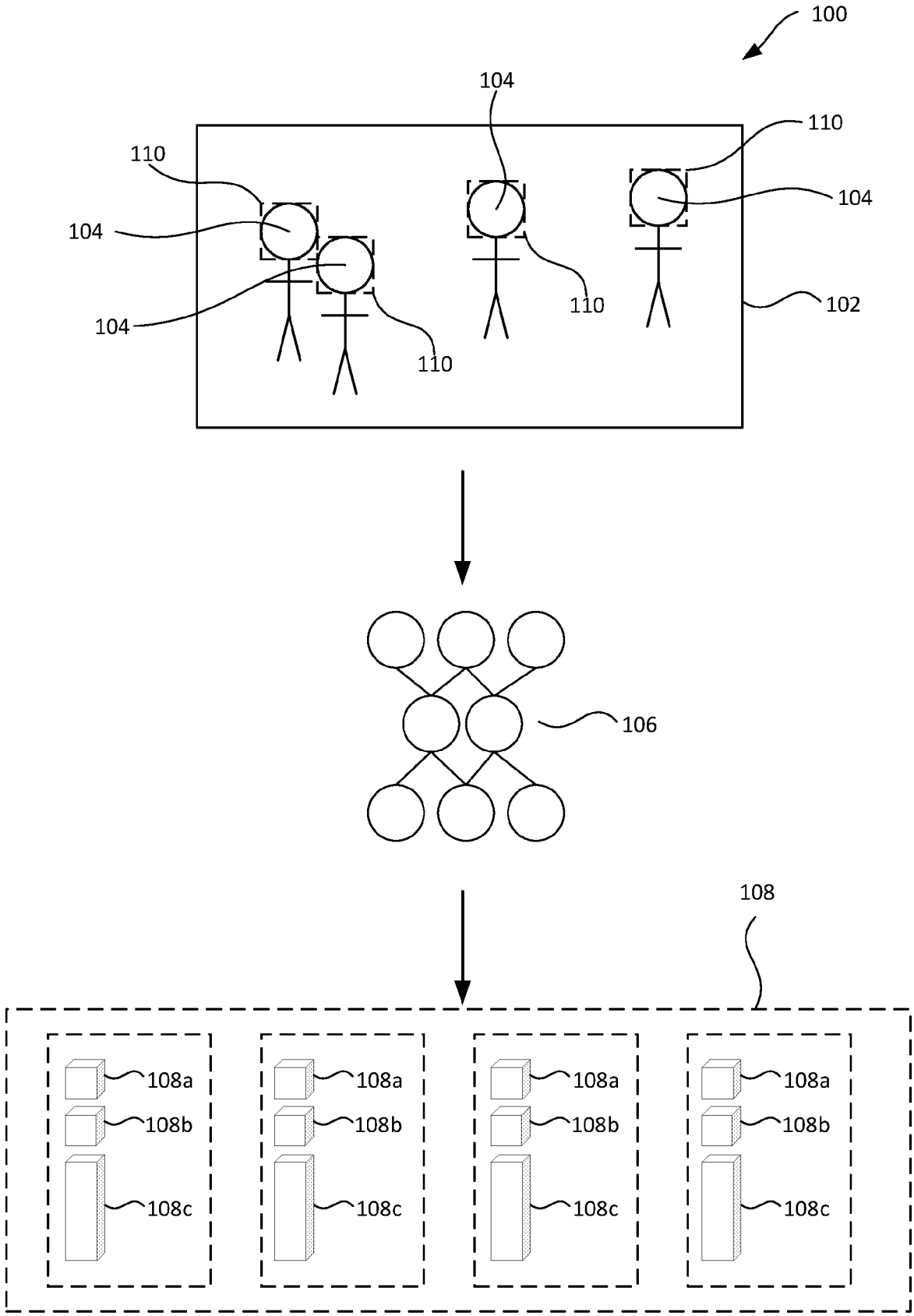
100
104
110
110
104
104
110
110
102
104
106
108
108a
108b
108c
108a
108b
108c
108a
108b
108c
108a
108b
108c

FIG. 1

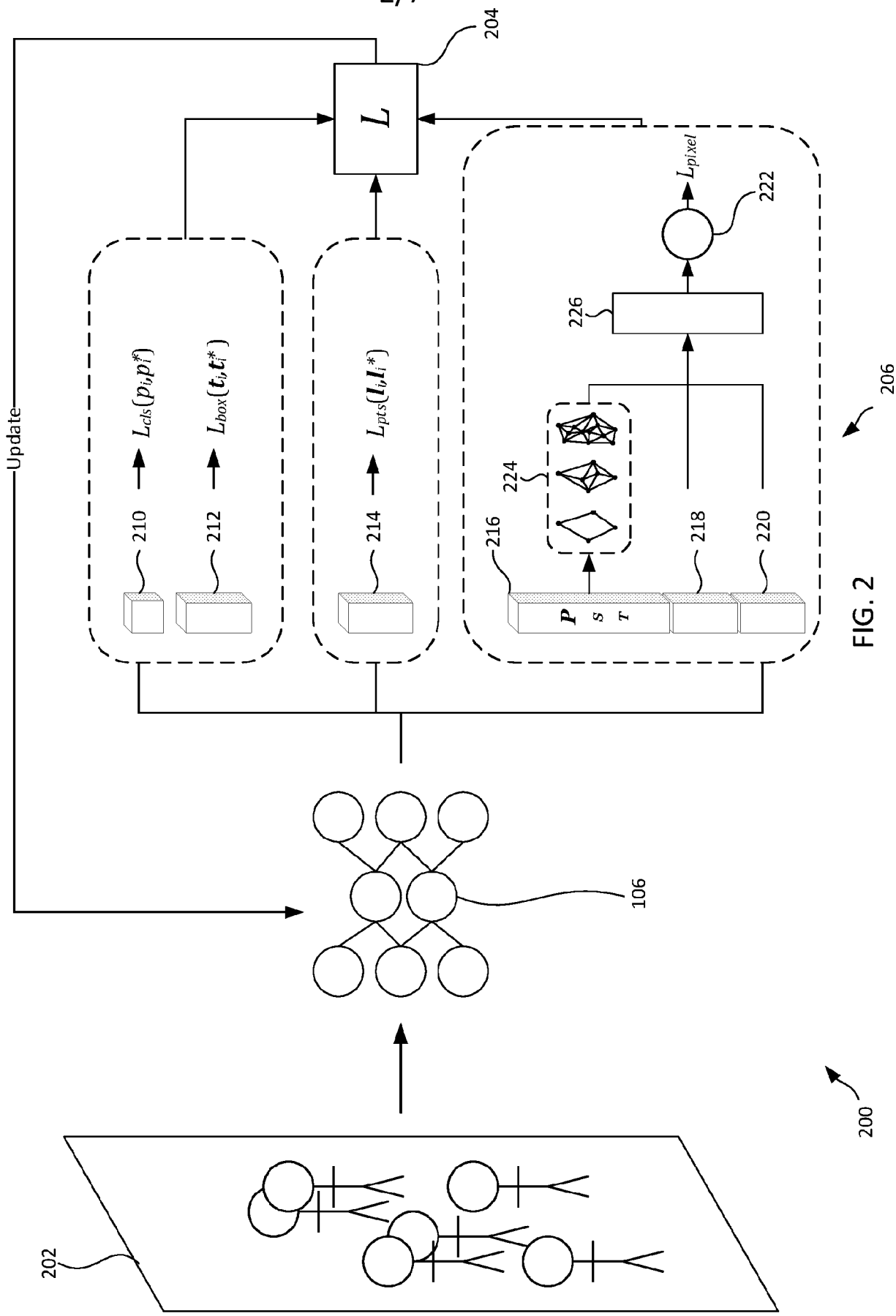
204
L
Update
$L_{cls}(p_i, p_i^*)$
$L_{box}(t_i, t_i^*)$
$L_{pts}(l_i, l_i^*)$
L_{pixel}
210
212
214
216
218
220
222
224
226
P
S
T
206
106
202
200

FIG. 2

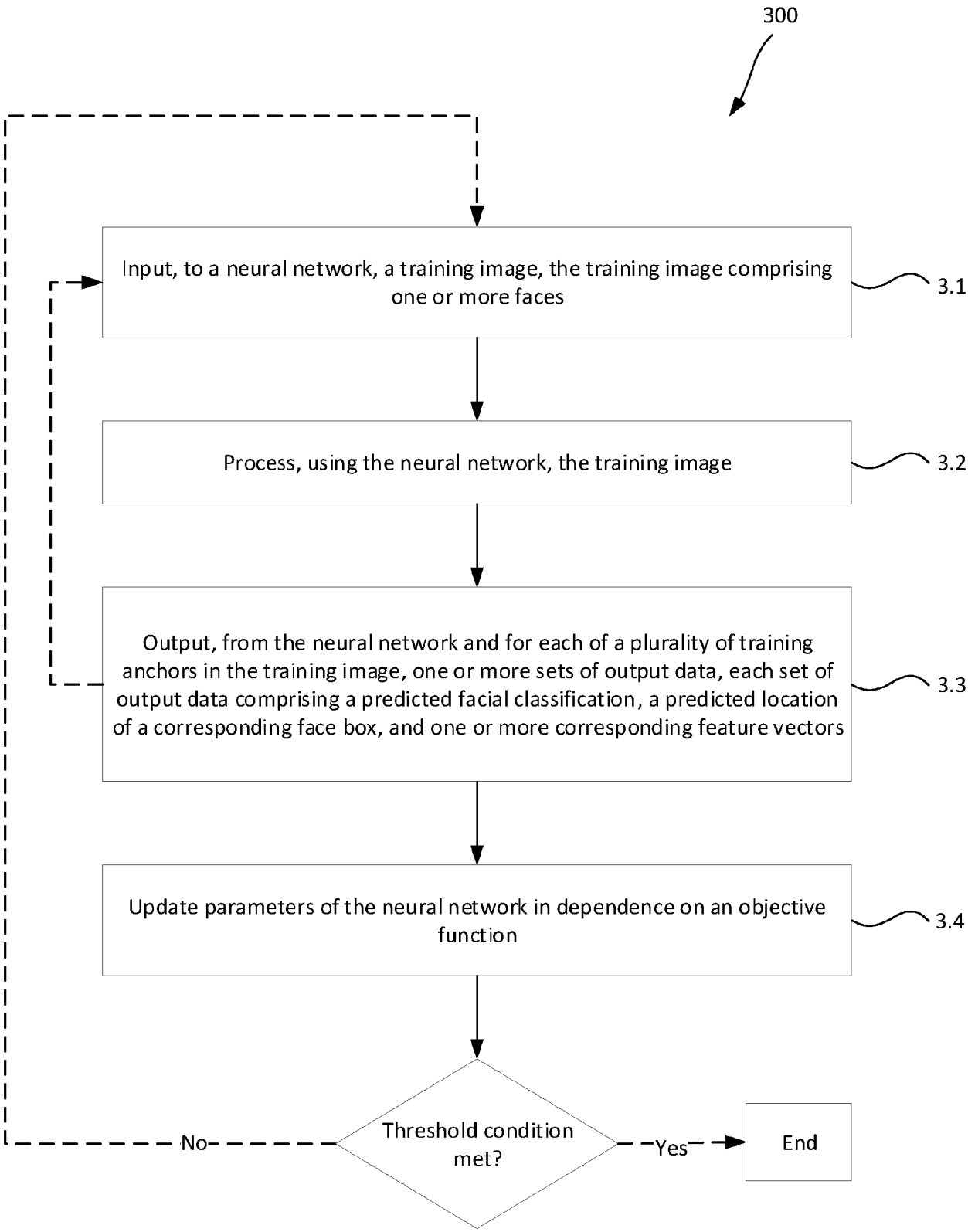
300
Input, to a neural network, a training image, the training image comprising one or more faces
3.1
Process, using the neural network, the training image
3.2
Output, from the neural network and for each of a plurality of training anchors in the training image, one or more sets of output data, each set of output data comprising a predicted facial classification, a predicted location of a corresponding face box, and one or more corresponding feature vectors
3.3
Update parameters of the neural network in dependence on an objective function
3.4
Threshold condition met?
No
Yes
End

FIG. 3

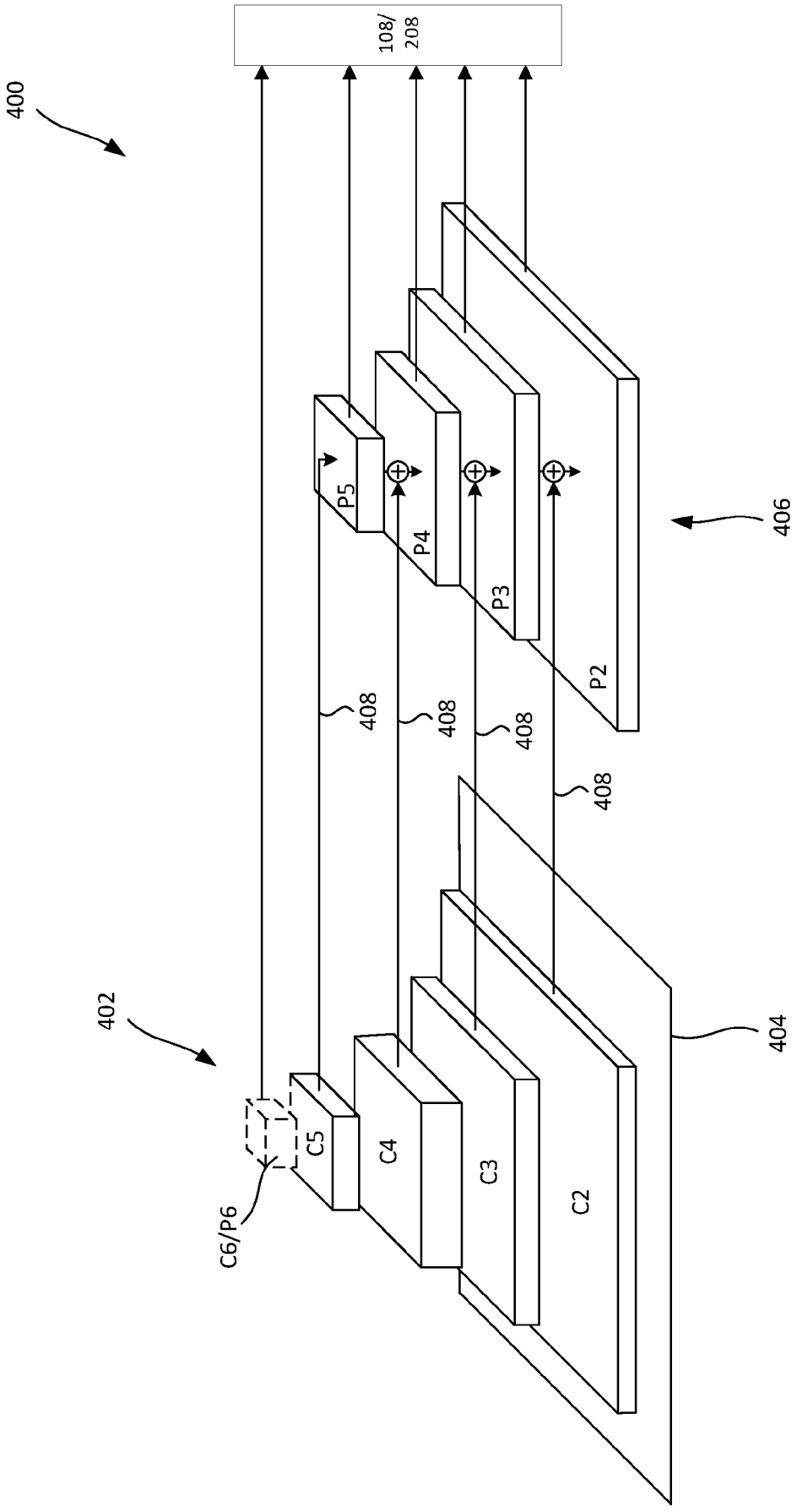
400
108/
208
406
P5
P4
P3
P2
408
408
408
408
402
404
C5
C4
C3
C2
C6/P6

FIG. 4

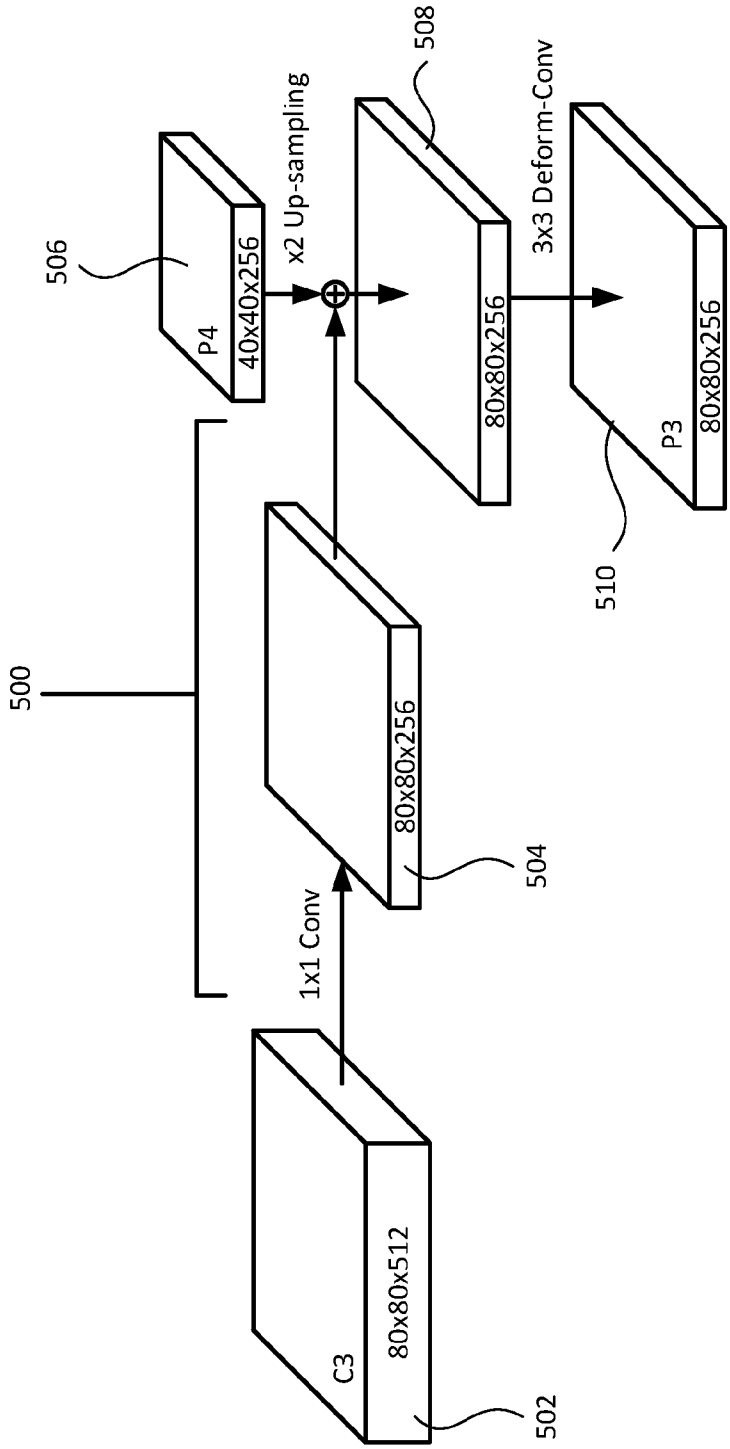
500
502
C3
80x80x512
504
80x80x256
1x1 Conv
506
P4
40x40x256
x2 Up-sampling
508
80x80x256
3x3 Deform-Conv
510
P3
80x80x256

FIG. 5

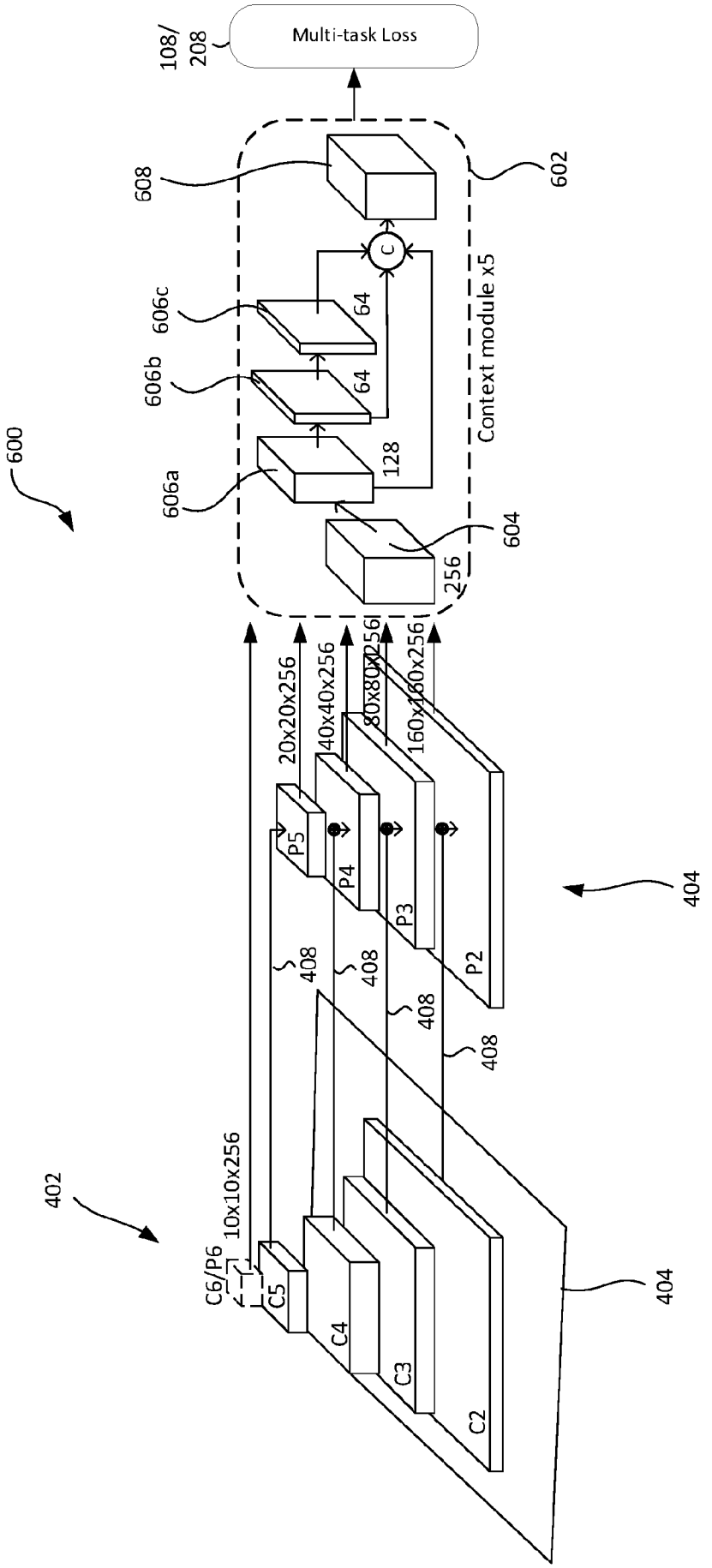
Multi-task Loss
108/
208
608
602
606c
606b
606a
64
64
128
C
Context module x5
600
604
256
20x20x256
40x40x256
80x80x256
160x160x256
P5
P4
P3
P2
404
408
408
408
408
402
10x10x256
C6/P6
C5
C4
C3
C2
404

FIG. 6

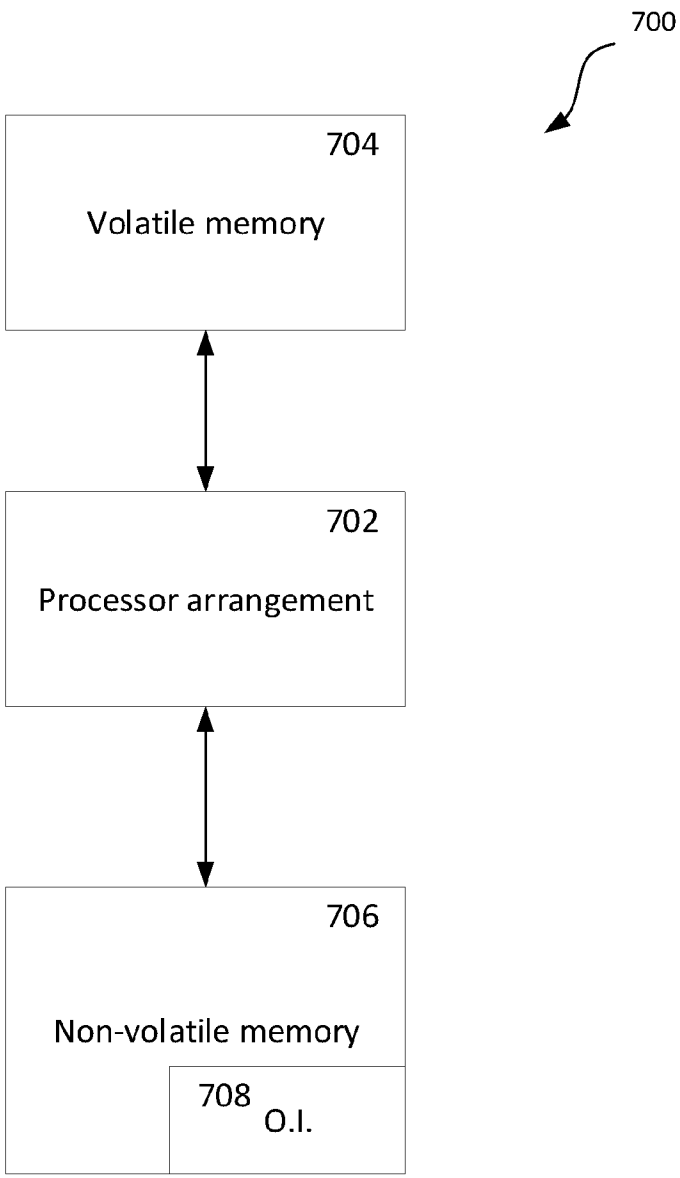
700
704
Volatile memory
702
Processor arrangement
706
Non-volatile memory
708
O.I.

FIG. 7

INTERNATIONAL SEARCH REPORT

International application No
PCT/GB2020/051012

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06K9/00 G06K9/62 G06N3/00
ADD.

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06K G06N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A,P	JIANKANG DENG ET AL: "RetinaFace: Single-stage Dense Face Localisation in the Wild", ARXIV.ORG, CORNELL UNIVERSITY LIBRARY, 201 OLIN LIBRARY CORNELL UNIVERSITY ITHACA, NY 14853, 2 May 2019 (2019-05-02), XP081271467, the whole document ----- Hao Wang ET AL: "Face R-CNN", Arxiv, 4 June 2017 (2017-06-04), XP055713126, Retrieved from the Internet: URL:https://arxiv.org/pdf/1706.01061.pdf%20http://arxiv.org/abs/1706.01061.pdf [retrieved on 2020-07-09] ----- -/--	1-8,19, 20
X		1,7,8, 19,20
Y	abstract; figure 1 Section 3.2	2-6

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

* Special categories of cited documents :

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

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

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

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

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

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

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

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

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
9 July 2020	08/09/2020

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

1

INTERNATIONAL SEARCH REPORT

International application No
PCT/GB2020/051012

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	ZHANG KAIPENG ET AL: "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks", IEEE SIGNAL PROCESSING LETTERS, IEEE SERVICE CENTER, PISCATAWAY, NJ, US, vol. 23, no. 10, 1 October 2016 (2016-10-01), pages 1499-1503, XP011622636, ISSN: 1070-9908, DOI: 10.1109/LSP.2016.2603342 [retrieved on 2016-09-09] Section II.C 3) and 4).	2
A	FORRESTER COLE ET AL: "Synthesizing Normalized Faces from Facial Identity Features", IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR 2017), 17 October 2017 (2017-10-17), pages 3386-3395, XP055541837, DOI: 10.1109/CVPR.2017.361 ISBN: 978-1-5386-0457-1 Section 3.3; figure 4	1
A	LAI YING-HSIU ET AL: "Emotion-Preserving Representation Learning via Generative Adversarial Network for Multi-View Facial Expression Recognition", 2018 13TH IEEE INTERNATIONAL CONFERENCE ON AUTOMATIC FACE & GESTURE RECOGNITION (FG 2018), IEEE, 15 May 2018 (2018-05-15), pages 263-270, XP033354708, DOI: 10.1109/FG.2018.00046 Section III.B Equation (4)	1
Y	YUXIANG ZHOU ET AL: "Dense 3D Face Decoding over 2500FPS: Joint Texture & Shape Convolutional Mesh Decoders", ARXIV.ORG, CORNELL UNIVERSITY LIBRARY, 201 OLIN LIBRARY CORNELL UNIVERSITY ITHACA, NY 14853, 6 April 2019 (2019-04-06), XP081165962, Section 3	3-6

INTERNATIONAL SEARCH REPORT

International application No.
PCT/GB2020/051012

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

see additional sheet

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying an additional fees, this Authority did not invite payment of additional fees.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

1-8(completely); 19, 20(partially)

Remark on Protest

☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.

☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

☐ No protest accompanied the payment of additional search fees.

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

This International Searching Authority found multiple (groups of) inventions in this international application, as follows:

1. claims: 1-8(completely); 19, 20(partially)

   Obtain dense 3d positioning and correspondence of each facial pixel

   ---

2. claims: 9-18(completely); 19, 20(partially)

   Perform hierarchical face localisation with feature pyramids

   ---