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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
02 January 2020 (02.01.2020)

WIPO | PCT

(10) International Publication Number
WO 2020/000096 A1

(51) International Patent Classification:
A61B 5/103 (2006.01) *G06T 1/40* (2006.01)
G06N 3/04 (2006.01) *G06T 7/00* (2017.01)

(21) International Application Number:
PCT/CA2019/050887

(22) International Filing Date:
27 June 2019 (27.06.2019)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/691,818 29 June 2018 (29.06.2018) US

(71) Applicant: **WRNCH INC.** [CA/CA]; 1001 rue Lenoir, Suite B113, Montreal, Québec H4C 2Z6 (CA).

(72) Inventors: **CHO, Dongwook**; 1001 rue Lenoir, Suite B113, Montreal, Québec H4C 2Z6 (CA). **ZHANG, Maggie**; 1001 rue Lenoir, B113, Montreal, Québec H4C 2Z6 (CA). **KRUSZEWSKI, Paul**; 1001 rue Lenoir, B113, Montreal, Québec H4C 2Z6 (CA).

(74) Agent: **DLA PIPER (CANADA) LLP**; Suite 6000, 1 First Canadian Place, PO Box 367, 100 King St W, Toronto, Ontario M5X 1E2 (CA).

(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: HUMAN POSE ANALYSIS SYSTEM AND METHOD

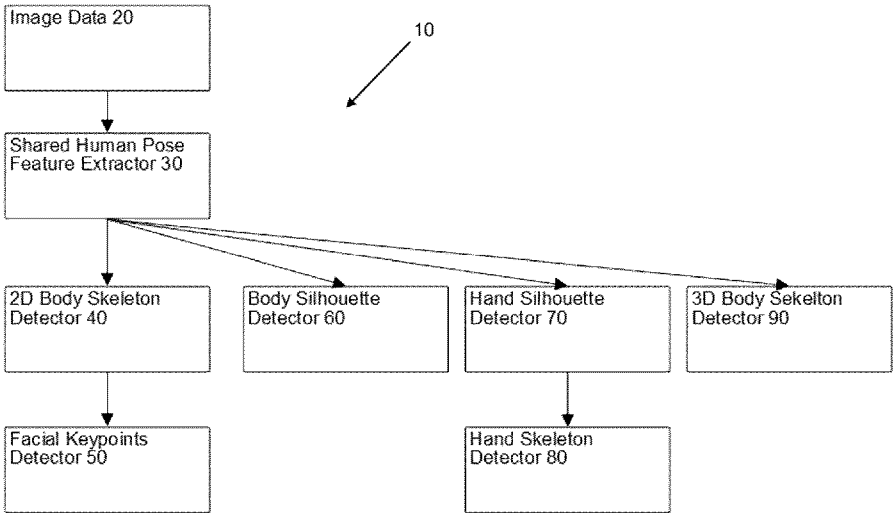

FIGURE 1

(57) Abstract: System and method for extracting human pose information from an image, comprising a feature extractor connected to a database, a convolutional neural network (CNN) with a plurality of CNN layers. Said system/method further comprising at least one of the following modules: a 2D body skeleton detector for determining 2D body skeleton information from the human-related image features; a body silhouette detector for determining body silhouette information from the human-related image features; a hand silhouette detector for determining hand silhouette detector from the human-related image features; a hand skeleton detector for determining hand skeleton from the human-related image features; a 3D body skeleton detector for determining 3D body skeleton from the human-related image features; and a facial keypoints detector for determining facial keypoints from the human-related image features.

WO 2020/000096 A1

[Continued on next page]

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))*

— *in black and white; the international application as filed contained color or greyscale and is available for download from PATENTSCOPE*

HUMAN POSE ANALYSIS SYSTEM AND METHOD

TECHNICAL FIELD

[0001] The present invention relates to the field of human pose analysis, and more particularly to human pose analysis systems and methods using lightweight convolutional neural networks (CNNs).

BACKGROUND

[0002] Early approaches for human pose analysis use visible markers attached on a person's body to be recognized by a camera or use images captured by depth sensor to understand shape of person or localize body parts. There have been attempts to analyze commonly available color images using classical computer vision techniques such as image feature detection approaches or structural analysis. These methods were not robust enough to handle a variety of natural images.

[0003] More recently robust methods to localize human body joints and construct human skeletons in 2D image space were proposed. These methods are implemented based on deep neural network models that are trained with large scale image database.

[0004] Multiple aspects of analysis can be made for a person in an image such as body skeleton in image, body shape, 3-dimensional body skeleton, detailed poses of each body part such as hands. Most of existing methods focus on analysing a single aspect of a person. Some methods localize a person and segment the body silhouette in image. Other methods localize only a person's hands and their joints. A unified analysis of a person's image makes possible a better understanding of human pose.

[0005] Also, most of robust methods require heavy computations for real-time analysis, which prohibits the implementation in inexpensive devices such as consumer electronics or mobile devices.

[0006] Therefore, there is a need for an improved method and system for human pose analysis.

SUMMARY

[0007] According to a first broad aspect, there is provided a system for extracting human pose information from an image, comprising: a feature extractor for extracting human-related image features from the image, the feature extractor being connectable to a database comprising a dataset of reference images and provided with a first convolutional neural network (CNN) architecture including a first plurality of CNN layers, each convolutional layer applies convolutional operation to its input data using trained kernel weights; and at least one of the following modules: a 2D body skeleton detector for determining 2D body skeleton information from the human-related image features; a body silhouette detector for determining body silhouette information from the human-related image features; a hand silhouette detector for determining hand silhouette detector from the human-related image features; a hand skeleton detector for determining hand skeleton from the human-related image features; a 3D body skeleton detector for determining 3D body skeleton from the human-related image features; and a facial keypoints detector for determining facial keypoints from the human-related image features, wherein each one of the 2D body skeleton detector, the body silhouette detector, the hand silhouette detector, the hand skeleton detector, the 3D body skeleton detector and the facial keypoints detector is provided with a second convolutional neural network (CNN) architecture including a second plurality of CNN layers.

[0008] In one embodiment of the system, the feature extractor comprises: a low-level feature extractor for extracting low-level features from the image; and an intermediate feature extractor for extracting intermediate features, the low-level features and the intermediate features forming together the human-related image features.

[0009] In one embodiment of the system, at least one of the first and second architecture comprises a deep CNN architecture.

[0010] In one embodiment of the system, one of the first and second CNN layers comprise lightweight layers.

[0011] According to another broad aspect, there is provided a method for extracting human pose information from an image, comprising: receiving an image; extracting human-related image features from the image using a feature extractor, the feature extractor being connectable to a database comprising a dataset of reference images and provided with a first convolutional neural network (CNN) architecture including a first plurality of CNN layers, each convolutional layer applies convolutional operation to its input data using trained kernel weights; and determining the human pose information using at least one of the following modules: a 2D body skeleton detector for determining 2D body skeleton information from the human-related image features; a body silhouette detector for determining body silhouette information from the human-related image features; a hand silhouette detector for determining hand silhouette detector from the human-related image features; a hand skeleton detector for determining hand skeleton from the human-related image features; a 3D body skeleton detector for determining 3D body skeleton from the human-related image features; and a facial keypoints detector for determining facial keypoints from the human-related image features, wherein each one of the 2D body skeleton detector, the body silhouette detector, the hand silhouette detector, the hand skeleton detector, the 3D body skeleton detector and the facial keypoints detector is provided with a second convolutional neural network (CNN) architecture including a second plurality of CNN layers.

[0012] In one embodiment of the method, the feature extractor comprises: a low-level feature extractor for extracting low-level features from the image; and an intermediate feature extractor for extracting intermediate features, the low-level features and the intermediate features forming together the human-related image features.

[0013] In one embodiment of the method, at least one of the first and second architecture comprises a deep CNN architecture.

[0014] In one embodiment of the method, one of the first and second CNN layers comprise lightweight layers.

BRIEF DESCRIPTION OF THE DRAWINGS

[0015] Further features and advantages of the present invention will become apparent from the following detailed description, taken in combination with the appended drawings, in which:

[0016] Figure 1 is a block diagram illustrating a system for extracting human pose information from an image, the system comprising a feature extractor, a 2D body skeleton detector, a body silhouette detector, a hand silhouette detector and 3D body skeleton detector, a facial keypoint detector and a hand skeleton detector, in accordance with an embodiment;

[0017] Figure 2 is a block diagram illustrating the feature extractor of Figure 1, in accordance with an embodiment;

[0018] Figure 3 is a block diagram illustrating the 2D body skeleton detector of Figure 1, in accordance with an embodiment;

[0019] Figure 4 is a block diagram illustrating the body silhouette detector of Figure 1, in accordance with an embodiment;

[0020] Figure 5 is a block diagram illustrating the hand silhouette detector of Figure 1, in accordance with an embodiment;

[0021] Figure 6 is a block diagram illustrating the hand skeleton detector of Figure 1, in accordance with an embodiment;

[0022] Figure 7 is a block diagram illustrating the 3D body skeleton detector of Figure 1, in accordance with an embodiment;

[0023] Figure 8 is a block diagram illustrating the facial keypoint detector of Figure 1, in accordance with an embodiment; and

[0024] Figure 9 is a block diagram of a processing module adapted to execute at least some of the steps of the extraction of human pose information, in accordance with an embodiment.

[0025] It will be noted that throughout the appended drawings, like features are identified by like reference numerals.

DETAILED DESCRIPTION

[0026] Figure 1 illustrates one embodiment of a system 10 for extracting human pose information from an image. The system 10 is configured for receiving images, localize humans within the received images and automatically infer human pose information from each image.

[0027] In one embodiment, the human pose information comprises geometric information of human skeletons and body part shapes. Human skeletons may be expressed by bone joint locations and/or bone orientations with lengths and body part shapes can be expressed as silhouettes and/or surface meshes with locations. For example, human pose information may include information such as 2D and/or 3D body skeletons with human joints, body shapes or silhouettes, and/or skeletons and silhouettes of body part like hand, etc.

[0028] The system 10 is configured to first extract from the images human-related image features that are learned by an image dataset and determine the human pose information from the extracted human-related image features.

[0029] In one embodiment, the human-related image features comprise primitive information related to human bodies and human body parts obtained from an image, such as points, edges, lines, contours, intensities, gradients, contrasts of small to large objects in an image, relations of those objects, etc.

[0030] In one embodiment, the data set comprises a set of reference images with and without human beings and ground-truth labels related to human body geometry. Labels may include 2D body joint locations (x,y) and visibilities (such as not available, visible,

existing in image but occluded) in image; 2D hand joint locations and visibilities in image; 2D face keypoints locations and visibilities in image, a silhouette of a human body, a silhouette of a hand, 3D body joint locations, etc. It should be understood that not all reference images contained in the data set have all the labels associated thereto.

[0031] In one embodiment, the data set of reference images comprises at least tens of thousands of images for training and may be qualified as being large scale.

[0032] The system 10 uses convolutional neural network (CNN) architecture for robust estimation of the pose information. CNNs are composed of convolutional neural network layers, hereinafter referred to as convolutional layers. Each convolutional layer receives input data or processed data from previous convolutional layer(s) and sends its output data to the following layer(s) after applying a convolutional operation to its input data. In one embodiment, the output of a convolutional layer is in the form of a tensor or a multi-dimensional array.

[0033] Each convolutional layer applies convolutional operation to its input data using trained kernel weights. The training of the weights of convolutional layers is performed by backpropagation technique using the dataset of reference images. In one embodiment, each convolutional layer is configured for applying a nonlinear activation function such as a rectified linear unit (ReLU) to the input data to allow for more robust decision of CNNs. It should be understood that functions other than ReLU functions may be used by the convolutional layers.

[0034] In one embodiment, the system 10 uses deep CNNs. In one embodiment, the CNNs comprise at least three convolutional layers. In comparison to a shallow architecture with a small number of layers, a deep CNN architecture preserves more neurons or weights and possibly accommodates a variety of input data and analyzes them robustly without being influenced by noise or clutter.

[0035] In the same or another embodiment, the CNN architecture comprises lightweight convolutional layers. In this case, each convolutional layer is made "computationally light" by reducing the number of kernels and/or their size and/or by

applying down-sampling. In this case, the architecture may be adequate for real-time human pose analysis performed on a low-end device.

[0036] In one embodiment, the following approach for the CNN architecture is followed.

[0037] The convolutional layers that do not significantly influence the accuracy of estimated results may be eliminated. For example, pooling layers that also perform down-sampling may be removed and a neighboring convolution layer located before a pooling layer may perform down-sampling during its convolutional operation.

[0038] A minimal input image resolution may be chosen by considering common person size in an image. For example, a person of 80x80 pixels in an image can be robustly analyzed without losing much human-related image features. A lower resolution image may present a lack of details, but it may be sufficient for a good approximation of body pose. In one embodiment, the resolution of the image is 48x48. In another embodiment, the resolution of the image is 96x96. In a further embodiment, the resolution of the image is 256x144. In still another embodiment, the resolution of the image is 400x320.

[0039] Receptive fields of a person may be considered in order to decide the number of convolutional layers and their kernel sizes by limiting the maximum resolution to be analyzed. For instance, a region with 84x84 pixels can be covered by two convolution layers with 11x11 kernels after down-sampling an input image by 4. Ten 3x3 convolution layers can cover the same region with more layers yet less computational cost.

[0040] The output depth size defined in each convolutional layer may be reduced as long as the resulting accuracy is higher than a minimum target accuracy chosen by the user. The computational cost is proportional to the kernel size in each dimension (kernel width, height, and depth) as well as output depth size. Size of weights may be decided by the multiplied sum of kernel width, height, and depth in addition to the number of biases.

[0041] A CNN model is a collection of weights and biases learned by a machine given a dataset and designed architecture. The CNN model may be chosen empirically to provide the highest accuracy.

[0042] Referring back to Figure 1, the system comprises a feature extractor 30 in communication with a database 20. The system also comprises a 2D body skeleton detector 40, a body silhouette detector 60, a hand silhouette detector 70 and 3D body skeleton detector 90, which are all in communication with the feature extractor 30. The system 10 further comprises a facial keypoint detector 50 in communication with the 2D body skeleton detector 40 and a hand skeleton detector 80 in communication with the hand silhouette detector 70.

[0043] The database 20 comprises the data set of reference images stored therein. In one embodiment, the database 20 is stored in a memory that is comprised in the system 10. In another embodiment, the database 20 is stored in a memory that is not included in the system 10.

[0044] As illustrated in Figure 2, the feature extractor 30 comprises a low-level feature extractor 110 and at least one intermediate feature extractor 120.

[0045] With reference to Figure 2, the feature extractor 30 is composed of a low-level feature extractor 110 and one or more intermediate feature extractor 120. The low-level feature extractor 110 is configured for receiving an image and extracting from the image low-level features that represent elemental characteristics of local regions in an image such as intensities, edges, gradients, curvatures, points, object shapes and/or the like. The intermediate feature extractor 120 is (are) configured for receiving the low-level features and determining intermediate features which correspond to high-level features obtained by correlating the low-level features extracted by the low-level feature extractor 110 and are related to human pose information such as shapes and/or relations of body parts. The low-level features and the intermediate features form together the human-related image features outputted by the feature extractor 30.

[0046] As illustrated in Figure 2, the low-level extractor 110 comprises repeated blocks in different image scales and each block comprises a series of convolutional layers with ReLU activations.

[0047] The low-level feature extractor 110 preserves generic image features such as edges, contours, blobs, their orientations, or some other observations learned from large scale image dataset.

[0048] A proven CNN architecture such as Inception, VGG, and ResNet can be considered for backbone networks. Lightweight backbone networks can be designed for reduced computational cost while preserving the minimum human pose-related features as mentioned above.

[0049] The intermediate feature extractor 120 is configured for intermediate supervision when a CNN model is trained. Intermediate supervision allows for the training of a CNN model by adding loss layers in the middle layers (or output layers of intermediate feature extractors) in addition to the last output layer. In neural networks, a loss layer compares difference between the output layer and ground-truth data and propagate backward to train weights and biases in each layer.

[0050] The number of convolutional layers present in the intermediate feature extractors 120 and their parameters for each intermediate stage are tailored by the size of input image and target objects, i.e. humans, as described above. Each intermediate stage is trained using the dataset of reference images. For example, a stack of 2D joint heat map in which the human joints in an image are marked in the same location may be generated using 2D joint locations. The exact joint location on the heat map has high response value while the location has low or no response value if the distance from the joint location is farther. The ground-truth heat maps that are generated from the dataset using the annotated 2D joint locations are compared to the estimated heat maps that are inferred from the training model during the model training. The model is trained by adjusting weight and bias values by repeating forward and backward propagation process throughout the connected layers in the neural networks.

[0051] In one embodiment, by training multiple stages of the convolutional layers of the intermediate feature extractors 120, the features related to human poses are refined through deeper network layers and therefore more robust results may be obtained. In addition, the model training becomes more efficient.

[0052] The output of each layer in the low-level feature extractor 110 and the intermediate feature extractors 120 form the human-related image features which can be presented as human-related image feature tensors. Depending on the purpose, a subset of the human-related image feature tensors can be used for detailed human pose analysis.

[0053] Figure 3 illustrates one embodiment of a 2D body skeleton detector 40 which comprises a 2D body joint estimation network 210 and a post-processing module 220.

[0054] The 2D body skeleton detector 40 receives as input a subset of the human-related image features generated by the feature extractor 30 and generates 2D joint heat maps as output. The subset of the human-related image features comprises a combination of output feature tensors of different convolution layers of the feature extractor 30 that preserve distinctive features related to human joints and shapes.

[0055] In one embodiment, it may be difficult to measure the quality of each output feature tensor. In this case, the convolution layers that are close to the end of the low-level feature extractor 110 and/or the intermediate feature extractor 120 can be considered since they are normally refined throughout the convolution layers. For example, the output feature tensors of the last convolution layers in the low-level feature extractor 110 and N-th intermediate feature extractor 110 can be chosen to provide data to the 2D body skeleton detector 40. Once the input feature subset is processed, the 2D body skeleton detector 40 infers the estimated heat maps, which are used to decide the candidates of joint locations that are local maxima in the heat maps and a heat map response value is over a manually defined threshold. When a plurality of persons is present in an image, joint clustering is performed to separate the persons and construct skeletons during the post-processing step.

[0056] Figure 4 illustrates one embodiment of the body silhouette detector 60 which comprises a body silhouette segmentation module 310 comprising convolutional layers and a post-processing module 320.

[0057] The body silhouette detector 60 is configured for segmenting all the human bodies in an image and generating a mask image for human bodies. The convolutional layers of the body silhouette segmentation 310 receive the human-related image feature tensors from the feature extractor 30 and construct a body mask image with human body silhouettes. Masks are used to segment different objects in an image by applying bitwise masking to each pixel. A body mask image is a binary image where a mask value is 1 if a pixel belongs to a human and non-human pixel is 0. Since the scale of the human-related image feature tensors is reduced normally by factor of 2 to 16 compared to an input image width and height, upscaling can be performed during the convolutional operations to increase the body mask image resolution and preserve more details.

[0058] The post-processing module 320 takes the inferred mask image from the body silhouette segmentation module 310 and resizes the mask image to the same resolution as the source input image. The body mask image can then be used for identifying the location and shape of a person in an image.

[0059] Figure 5 illustrates one embodiment for the hand silhouette detector 70 which comprises a hand silhouette segmentation module 410 formed of convolutional layers and a post-processing module 420.

[0060] The hand silhouette detector module 410 is configured for segmenting the hands of the human bodies present in an input image and generates mask images for left and/or right hand similarly to the body silhouette detector 60. The convolutional layers of the hand silhouette segmentation module 410 receives the human-related image feature tensors from the feature extractor 30 and constructs hand mask images with human body silhouettes.

[0061] The post-processing module 420 is configured for resizing the inferred hand mask images. The hand mask images may then be used for identifying the location and

shape of visible hands in an image. This information can be used for further analysis of each hand pose.

[0062] In one embodiment, the hand silhouette detector 70 can be merged with the body silhouette detector 60 and the merged detectors 60 and 70 can be trained together. The neural network layers in these merged detectors may be shared for more efficient computations.

[0063] Figure 6 illustrates one embodiment for the hand skeleton detector 80 which comprises a hand joint estimation module 510 comprising convolutional layers and a post-processing module 520.

[0064] The hand skeleton detector 80 receives an image of a hand, i.e. a hand image, and estimates hand joints in the hand image. The hand image may be any image of a hand such as an image not specified in the system. Alternatively, the hand image may be a hand image cropped from an input image data 20 using a hand region (or bounding box) detected by the hand silhouette detector 70.

[0065] The hand joint estimation module 510 can be designed with a similar architecture that combines the architecture of the feature extraction networks 110 and 120 and the architecture of the 2D body joint estimation networks 210. In one embodiment, the hand skeleton detector 80 can be designed to directly receive the human-related image feature tensors from the feature extractor 30.

[0066] The post-processing module 520 for hand pose estimation takes the estimated heat maps and decides the candidates of joint locations and constructs a hand skeleton.

[0067] Figure 7 illustrates one embodiment for the 3D body skeleton detector 90 which comprises a 3D body joint estimation module 610 comprising convolutional layers and a post-processing module 620.

[0068] The 3D body skeleton detector 90 is configured for estimating 3D coordinates of human body joints from a single image. The 3D body skeleton detector 90

receives human-related image feature tensors and estimates normalized 3D coordinates of a human body detected in an image. The post-processing module 620 is configured for mapping the normalized 3D locations into image and real-world spaces.

[0069] Figure 8 illustrates one embodiment for the facial keypoints detector 50 which comprises a facial keypoints estimation module 710 comprising convolutional layers, and a post-processing module 720.

[0070] The facial keypoints detector 50 receives a cropped facial image decided by the 2D body skeleton detector 40 which estimates rough location of facial keypoints such as eyes, ears, nose, and/or the like. The locations of more detailed keypoints such as contour points of eyes, upper and lower lips, chin, eyebrows, nose, etc. are estimated by the convolutional layers of the facial keypoints estimation module 710. Alignment of detected facial keypoints and/or outlier filtering may be performed by the post-processing module 720.

[0071] It should be understood that the same human-related image features determined by the feature extractor 30 are shared by at least some of the detectors 40-90 to infer the human pose information. In one embodiment, the feature extractor 30 determines all the human-related image features that can be obtained from an image and stores them at each neural network layer in a tensor form.

[0072] In one embodiment, the feature extractor 30 can be designed by explicitly defining feature descriptors such as scale-invariant feature transform (SIFT) and histogram of oriented gradients (HOG). Such a feature extractor pre-defines image features regardless of the dataset.

[0073] In one embodiment, the extractor 30 and the detectors 40-90 are each provided with at least one respective processor or processing unit, a respective communication unit and a respective memory. In another embodiment, at least two of the group consisting of the extractor 30 and the detectors 40-90 share a same processor, a same communication and/or a same memory. For example, the extractor 30 and the detectors 40-90 may share the same processor, the same communication unit and the same memory. In

this case, the extractor 30 and the detectors 40-90 may correspond to different modules executed by the processor of a computer machine such as a personal computer, a laptop, a tablet, a smart phone, etc.

[0074] While the above description refers to the system 10 comprising the detectors 40-90, it should be understood that the system 10 may comprise only one of the detectors 40-90. For example, the system 10 may comprise at least two of the detectors 40-90.

[0075] In one embodiment, the sharing of the same human-related image features between a plurality of detectors makes the analysis consistent and fast by minimizing computations for each detector.

[0076] Figure 9 is a block diagram illustrating an exemplary processing module 800 for executing the above described pose information extraction from an image, in accordance with some embodiments. The processing module 800 typically includes one or more Computer Processing Units (CPUs) and/or Graphic Processing Units (GPUs) 802 for executing modules or programs and/or instructions stored in memory 804 and thereby performing processing operations, memory 804, and one or more communication buses 806 for interconnecting these components. The communication buses 806 optionally include circuitry (sometimes called a chipset) that interconnects and controls communications between system components. The memory 804 includes high-speed random access memory, such as DRAM, SRAM, DDR RAM or other random access solid state memory devices, and may include non-volatile memory, such as one or more magnetic disk storage devices, optical disk storage devices, flash memory devices, or other non-volatile solid state storage devices. The memory 804 optionally includes one or more storage devices remotely located from the CPU(s) 802. The memory 804, or alternately the non-volatile memory device(s) within the memory 804, comprises a non-transitory computer readable storage medium. In some embodiments, the memory 804, or the computer readable storage medium of the memory 804 stores the following programs, modules, and data structures, or a subset thereof:

a feature extraction module 810 for extracting human-related image features from an image;

a 2D body skeleton detection module 812 for estimating 2D body joint positions;

a body silhouette detection module 814 for identifying and segmenting body silhouettes;

a hand silhouette detection module 816 for identifying and segmenting hand silhouettes;

a 3D body skeleton detection module 818 for estimating 3D body joint positions;

a facial keypoints detection module 820 for estimating facial keypoint positions: and

a hand skeleton detection module 822 for estimating hand joint positions.

[0077] Each of the above identified elements may be stored in one or more of the previously mentioned memory devices, and corresponds to a set of instructions for performing a function described above. The above identified modules or programs (i.e., sets of instructions) need not be implemented as separate software programs, procedures or modules, and thus various subsets of these modules may be combined or otherwise re-arranged in various embodiments. In some embodiments, the memory 804 may store a subset of the modules and data structures identified above. Furthermore, the memory 804 may store additional modules and data structures not described above.

[0078] Although it shows a processing module 800, Figure 9 is intended more as functional description of the various features which may be present in a management module than as a structural schematic of the embodiments described herein. In practice, and as recognized by those of ordinary skill in the art, items shown separately could be combined and some items could be separated.

[0079] The embodiments of the invention described above are intended to be exemplary only. The scope of the invention is therefore intended to be limited solely by the scope of the appended claims.

I/WE CLAIM:

1. A system for extracting human pose information from an image, comprising:

a feature extractor for extracting human-related image features from the image, the feature extractor being connectable to a database comprising a dataset of reference images and provided with a first convolutional neural network (CNN) architecture including a first plurality of CNN layers, each convolutional layer applies convolutional operation to its input data using trained kernel weights; and

at least one of the following modules:

a 2D body skeleton detector for determining 2D body skeleton information from the human-related image features;

a body silhouette detector for determining body silhouette information from the human-related image features;

a hand silhouette detector for determining hand silhouette detector from the human-related image features;

a hand skeleton detector for determining hand skeleton from the human-related image features;

a 3D body skeleton detector for determining 3D body skeleton from the human-related image features; and

a facial keypoints detector for determining facial keypoints from the human-related image features,

wherein each one of the 2D body skeleton detector, the body silhouette detector, the hand silhouette detector, the hand skeleton detector, the 3D body skeleton detector and the facial keypoints detector is provided with a second convolutional neural network (CNN) architecture including a second plurality of CNN layers.

2. The system of claim 2, wherein the feature extractor comprises:

a low-level feature extractor for extracting low-level features from the image; and

an intermediate feature extractor for extracting intermediate features, the low-level features and the intermediate features forming together the human-related image features.

3. The system of claim 1 or 2, wherein at least one of the first and second architecture comprises a deep CNN architecture.

4. The system of any one of claims 1 to 3, wherein one of the first and second CNN layers comprise lightweight layers.

5. A method for extracting human pose information from an image, comprising:

receiving an image;

extracting human-related image features from the image using a feature extractor, the feature extractor being connectable to a database comprising a dataset of reference images and provided with a first convolutional neural network (CNN) architecture including a first plurality of CNN layers, each convolutional layer applies convolutional operation to its input data using trained kernel weights; and

determining the human pose information using at least one of the following modules:

a 2D body skeleton detector for determining 2D body skeleton information from the human-related image features;

a body silhouette detector for determining body silhouette information from the human-related image features;

a hand silhouette detector for determining hand silhouette detector from the human-related image features;

a hand skeleton detector for determining hand skeleton from the human-related image features;

a 3D body skeleton detector for determining 3D body skeleton from the human-related image features; and

a facial keypoints detector for determining facial keypoints from the human-related image features,

wherein each one of the 2D body skeleton detector, the body silhouette detector, the hand silhouette detector, the hand skeleton detector, the 3D body skeleton detector and the facial keypoints detector is provided with a second convolutional neural network (CNN) architecture including a second plurality of CNN layers.

6. The method of claim 5, wherein the feature extractor comprises:

a low-level feature extractor for extracting low-level features from the image; and

an intermediate feature extractor for extracting intermediate features, the low-level features and the intermediate features forming together the human-related image features.

7. The method of claim 5 or 6, wherein at least one of the first and second architecture comprises a deep CNN architecture.

8. The method of any one of claims 5 to 7, wherein one of the first and second CNN layers comprise lightweight layers.

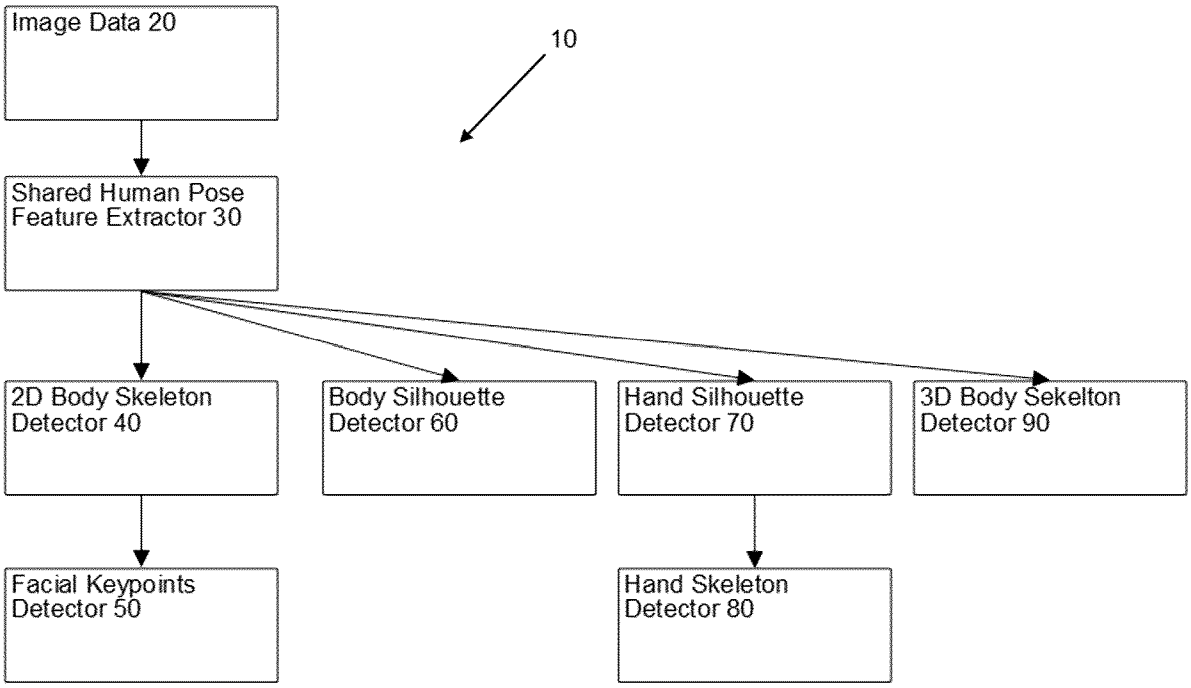
Image Data 20
10
Shared Human Pose Feature Extractor 30
2D Body Skeleton Detector 40
Body Silhouette Detector 60
Hand Silhouette Detector 70
3D Body Sekelton Detector 90
Facial Keypoints Detector 50
Hand Skeleton Detector 80

FIGURE 1

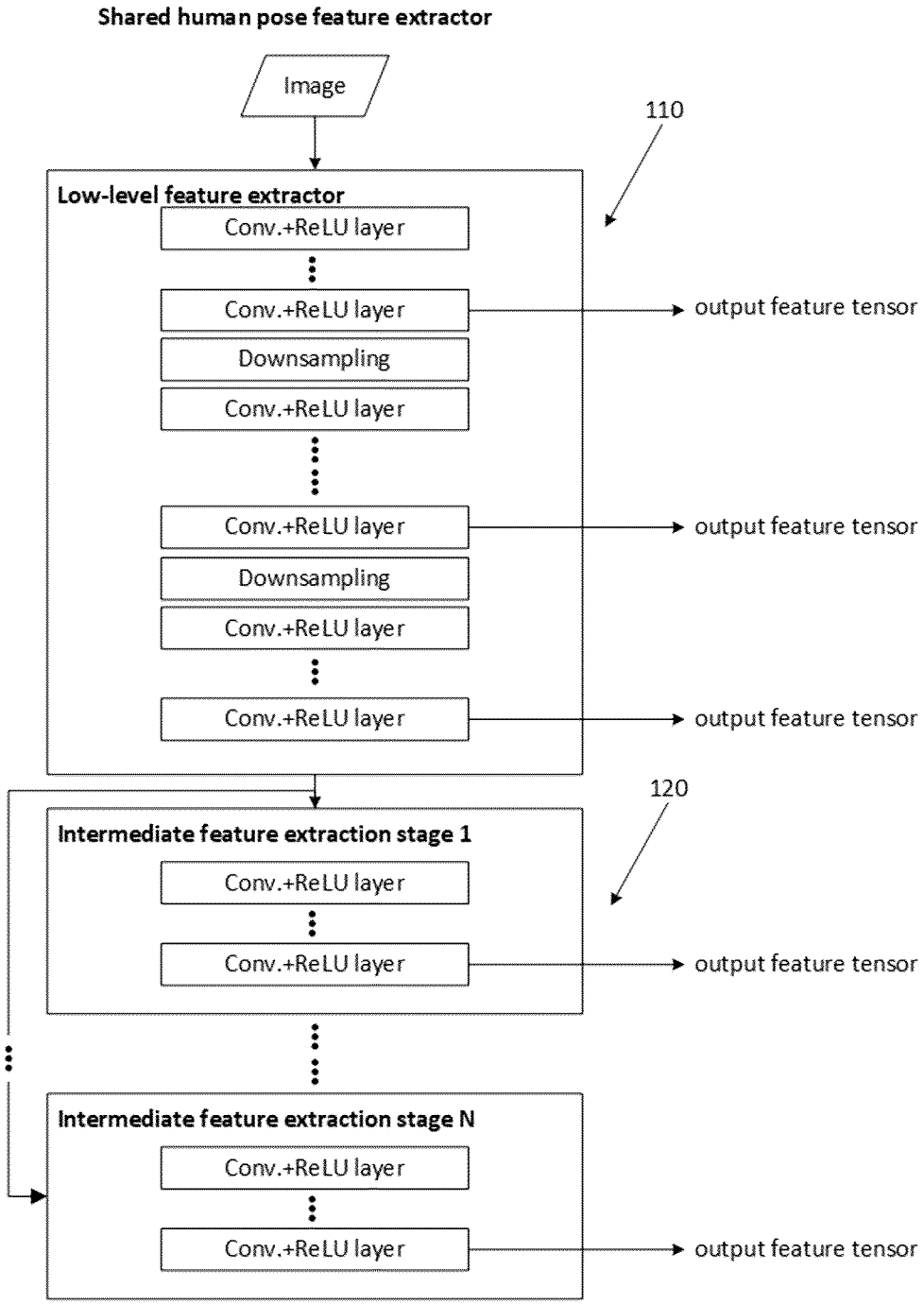
Shared human pose feature extractor
Image
110
Low-level feature extractor
Conv.+ReLU layer
Conv.+ReLU layer
output feature tensor
Downsampling
Conv.+ReLU layer
Conv.+ReLU layer
output feature tensor
Downsampling
Conv.+ReLU layer
Conv.+ReLU layer
output feature tensor
120
Intermediate feature extraction stage 1
Conv.+ReLU layer
Conv.+ReLU layer
output feature tensor
Intermediate feature extraction stage N
Conv.+ReLU layer
Conv.+ReLU layer
output feature tensor

Figure 2

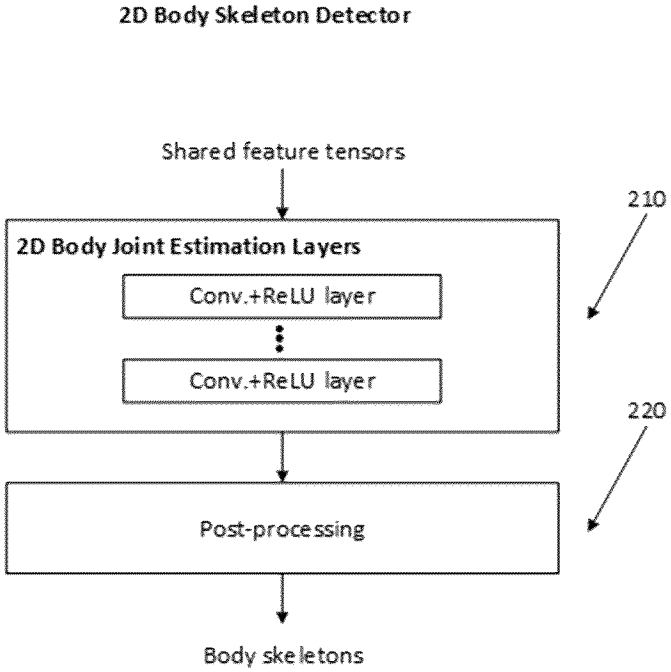

FIGURE 3

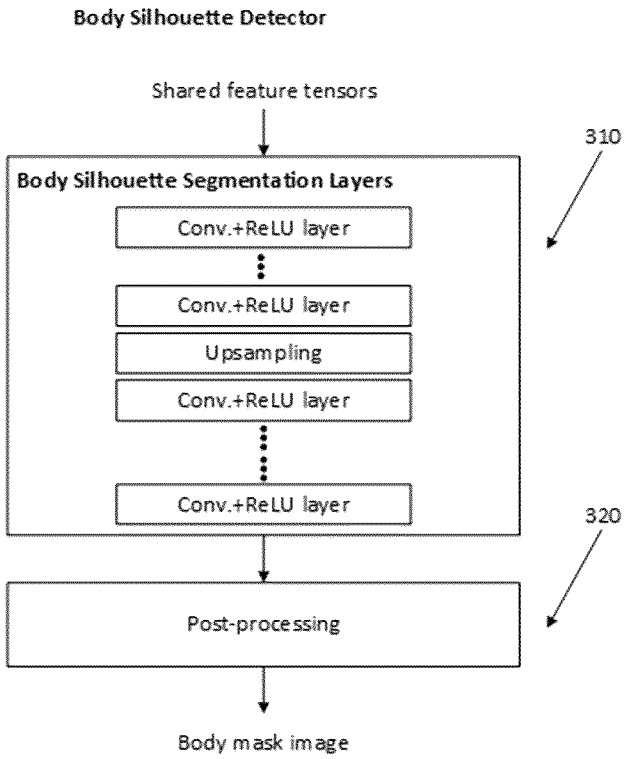

FIGURE 4

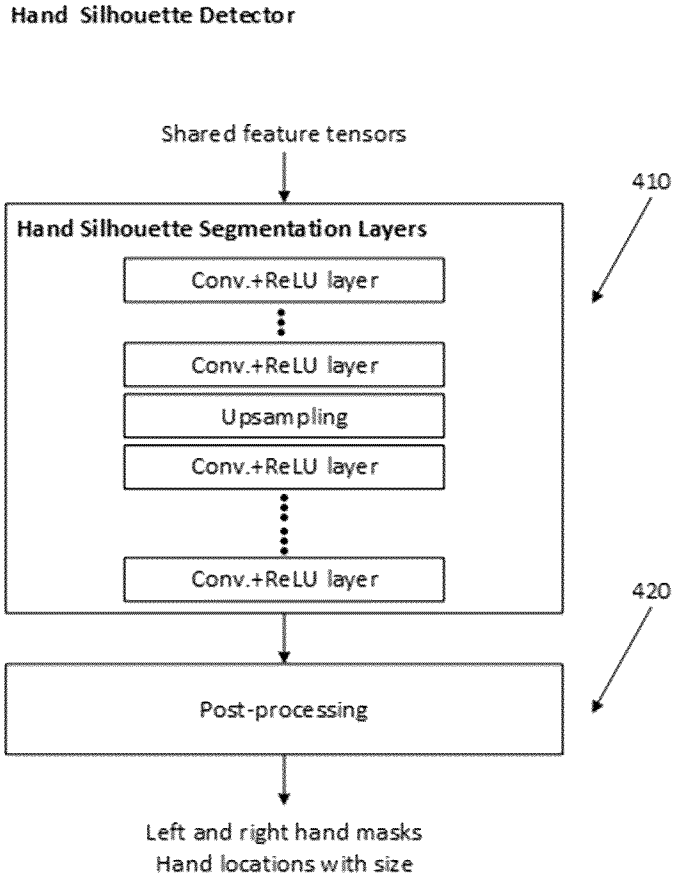
Hand Silhouette Detector
Shared feature tensors
410
Hand Silhouette Segmentation Layers
Conv.+ReLU layer
Conv.+ReLU layer
Upsampling
Conv.+ReLU layer
Conv.+ReLU layer
420
Post-processing
Left and right hand masks
Hand locations with size

FIGURE 5

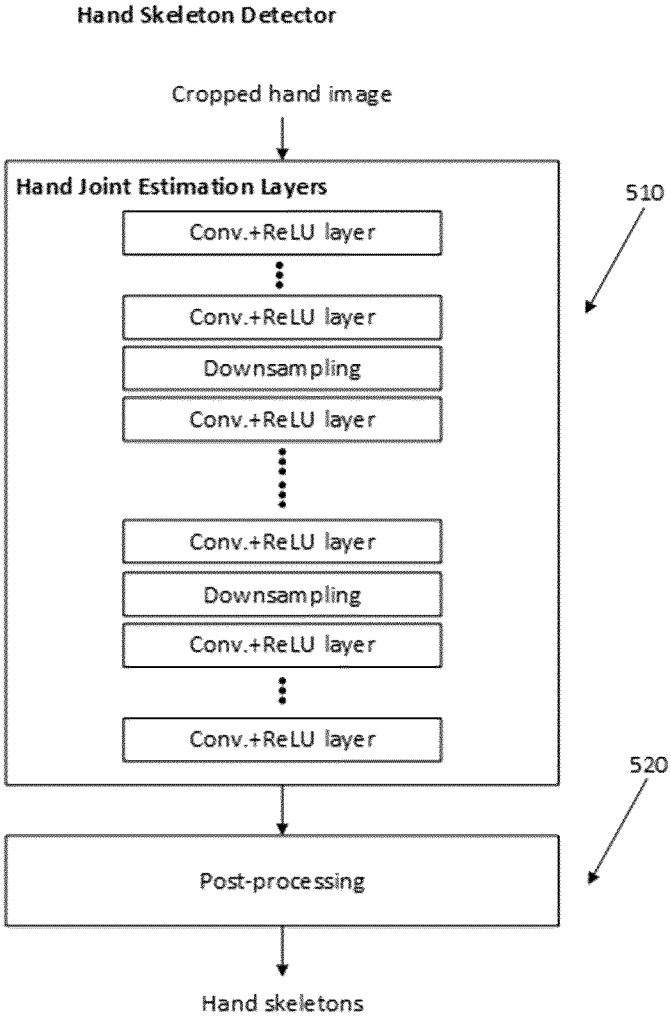
Hand Skeleton Detector
Cropped hand image
Hand Joint Estimation Layers
Conv.+ReLU layer
Conv.+ReLU layer
Downsampling
Conv.+ReLU layer
Conv.+ReLU layer
Downsampling
Conv.+ReLU layer
Conv.+ReLU layer
510
520
Post-processing
Hand skeletons

FIGURE 6

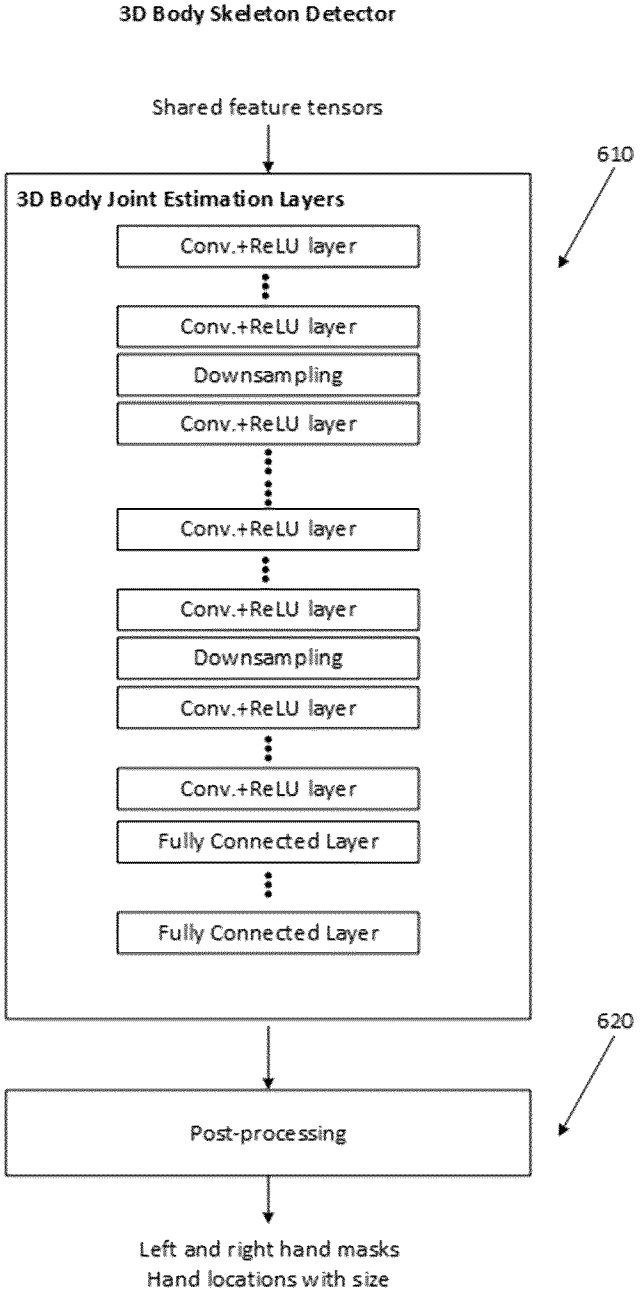
3D Body Skeleton Detector
Shared feature tensors
610
3D Body Joint Estimation Layers
Conv.+ReLU layer
Conv.+ReLU layer
Downsampling
Conv.+ReLU layer
Conv.+ReLU layer
Conv.+ReLU layer
Downsampling
Conv.+ReLU layer
Conv.+ReLU layer
Fully Connected Layer
Fully Connected Layer
620
Post-processing
Left and right hand masks
Hand locations with size

FIGURE 7

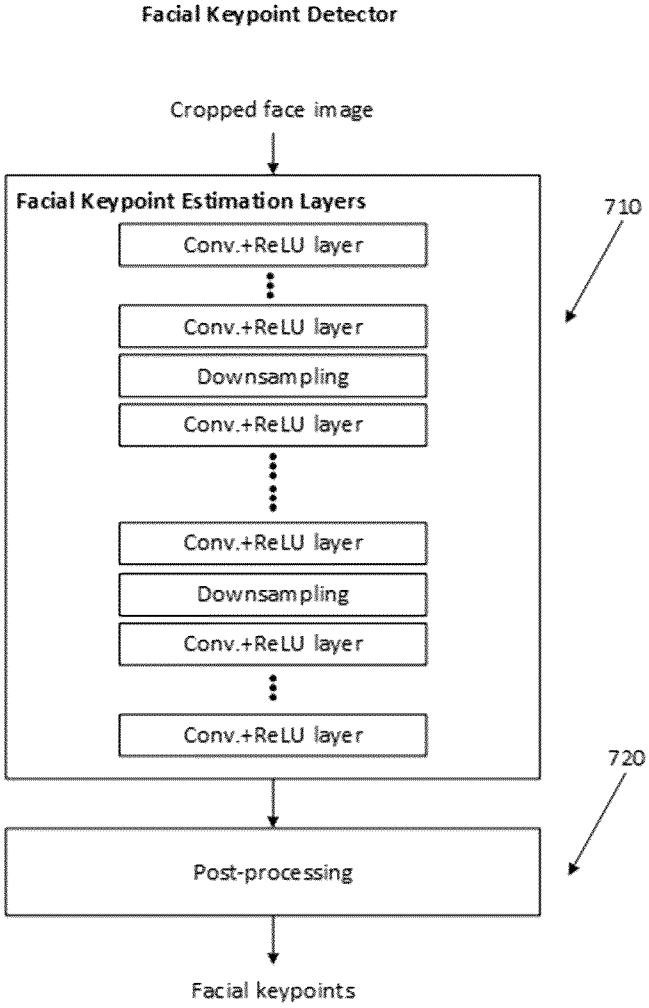
Facial Keypoint Detector
Cropped face image
Facial Keypoint Estimation Layers
Conv.+ReLU layer
Conv.+ReLU layer
Downsampling
Conv.+ReLU layer
Conv.+ReLU layer
Downsampling
Conv.+ReLU layer
Conv.+ReLU layer
710
720
Post-processing
Facial keypoints

FIGURE 8

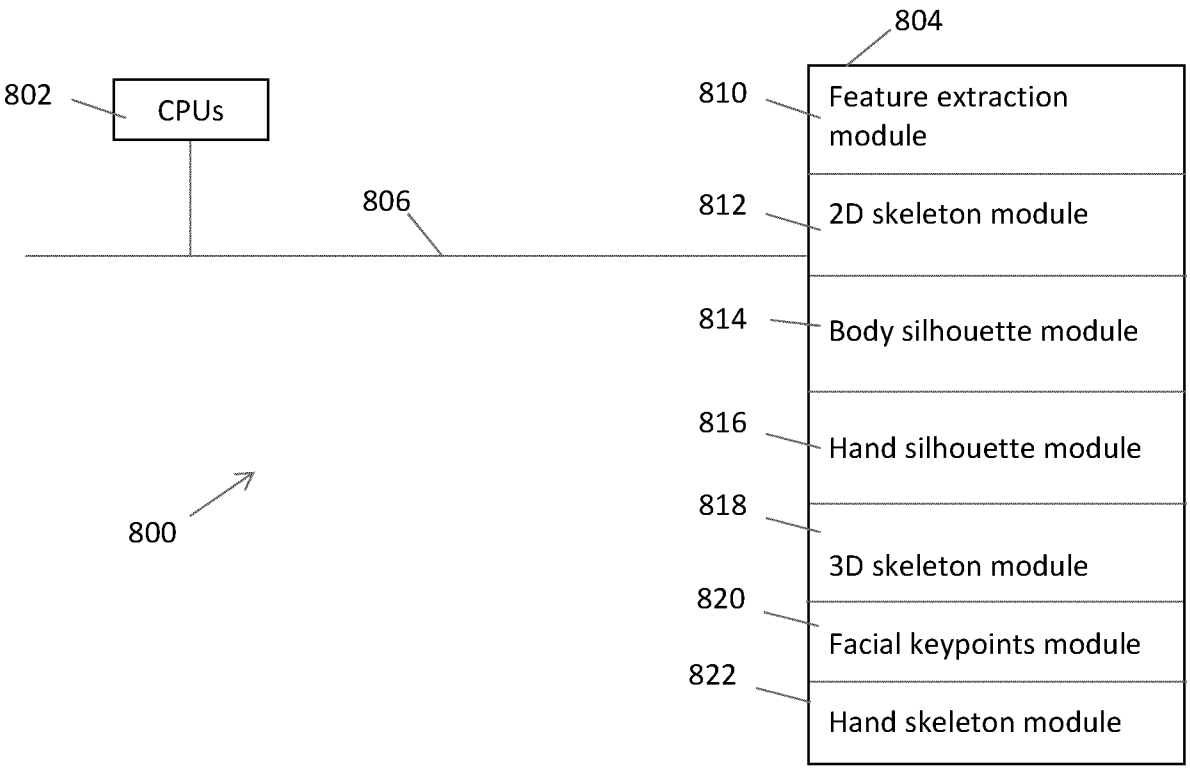
804
802
CPUs
810
Feature extraction module
806
812
2D skeleton module
814
Body silhouette module
816
Hand silhouette module
800
818
3D skeleton module
820
Facial keypoints module
822
Hand skeleton module

FIGURE 9

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CA2019/050887

A. CLASSIFICATION OF SUBJECT MATTER
IPC: ***A61B 5/103*** (2006.01) , ***G06N 3/04*** (2006.01) , ***G06T 1/40*** (2006.01) , ***G06T 7/00*** (2017.01)

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)
A61B 5/103 (2006.01) , ***G06N 3/04*** (2006.01) , ***G06T 1/40*** (2006.01) , ***G06T 7/00*** (2017.01)

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

Electronic database(s) consulted during the international search (name of database(s) and, where practicable, search terms used)
Databases: Google, Google Patents, Questel/Orbit, Canadian Patent Database/Intellect
KeyWords: human pose information feature extract* image neural network skeleton joint detector

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN104-346-607B (HU JING) 22 December 2017 (22-12-2017) - ABS, para.0001-0025	1-8
A	CN105-069-423A (PAN ZHENG et al.) 18 November 2015 (18-11-2015) - Whole doc; In particular: fig.6 and associated para.0104-111	
A	US8437506 B2 (WILLIAMS et al.) 7 May 2013 (07-05-2013) - Whole Doc	

☐ 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
"D"	document cited by the applicant in the international application
"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 05 September 2019 (05-09-2019)	Date of mailing of the international search report 05 September 2019 (05-09-2019)
Name and mailing address of the ISA/CA Canadian Intellectual Property Office Place du Portage I, C114 - 1st Floor, Box PCT 50 Victoria Street Gatineau, Quebec K1A 0C9 Facsimile No.: 819-953-2476	Authorized officer Allan Tam (819) 635-5764

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CA2019/050887

Patent Document Cited in Search Report	Publication Date	Patent Family Member(s)	Publication Date
CN104346607B	22 December 2017 (22-12-2017)	None	
CN105069423A	18 November 2015 (18-11-2015)	None	
US8437506B2	07 May 2013 (07-05-2013)	US2012056800A1	08 March 2012 (08-03-2012)
		US8437506B2	07 May 2013 (07-05-2013)
		CN102402288A	04 April 2012 (04-04-2012)
		CN102402288B	05 November 2014 (05-11-2014)
		US2013243255A1	19 September 2013 (19-09-2013)
		US8953844B2	10 February 2015 (10-02-2015)