IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 21, 2020, accepted March 6, 2020, date of publication March 13, 2020, date of current version March 25, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2980565

A Hybrid Neural Network for Graph-Based Human Pose Estimation From 2D Images

HUYNH THE VU[1], RICHARDT H. WILKINSON[1], MARGARET LECH[1], AND EVA CHENG[2]
[1]School of Engineering, RMIT University, Melbourne, VIC 3001, Australia
[2]School of Electrical and Data Engineering, University of Technology Sydney, Sydney, NSW 2007, Australia

Corresponding author: Richardt H. Wilkinson (richardt.wilkinson@rmit.edu.au)

ABSTRACT This paper investigates the problem of human pose estimation (HPE) from single 2-dimensional (2D) still images using a convolutional neural network (CNN). The aim was to train the CNN to analyze a 2D input image of a person to determine the person's pose. The CNN output was given in the form of a tree-structured graph of interconnected nodes representing 2D image coordinates of the person's body joints. A new data-driven tree-based model for HPE was validated and compared to the traditional anatomy-based tree-based structures. The effect of the number of nodes in anatomy-based tree-based structures on the accuracy of HPE was examined. The tree-based techniques were compared with non-tree-based methods using a common HPE framework and a benchmark dataset. As a result of this investigation, a new hybrid two-stage approach to the HPE estimation was proposed. In the first stage, a non-tree-based network was used to generate approximate results that were then passed for further refinement to the second, tree-based stage. Experimental results showed that both of the proposed methods, the data-driven tree-based model (TD_26) and the hybrid model (H_26_2B) lead to very similar results, obtaining 1% higher HPE accuracy compared to the benchmark anatomy-based model (TA_26) and 3% higher accuracy compared to the non-tree-based benchmark (NT_26_A). The best overall HPE results were obtained using the anatomy-based benchmark with the number of nodes increased from 26 to 50, which also significantly increased the computational cost.

INDEX TERMS Convolutional neural networks, human pose estimation, graph structures.

I. INTRODUCTION

Human pose estimation (HPE) from 2-dimensional (2D) images is the process of determining 2D locations of body parts (or joints) within the image array. This research field is an important building block for a variety of applications, e.g. human activity recognition for computer vision or human-computer interaction. Human pose estimation can be applied in surveillance systems to detect suspicious or abnormal human behavior, in clinical diagnosis to analyze human gait, in training of sportsmen to ensure correct posture, or in generation of naturalistic cartoon or computer game animations. Estimation of the human pose from a static 2D image can be formulated as a structured prediction problem in which the outputs (locations of joints) maintain a specific spatial relationship. In contrast to object detection, where the focus is on learning an accurate object location, human pose estimation requires both accurate localization of the body parts, as well as determining the correct relationship between the detected body parts. Assuming that this relationship can be described as a set of relative distances between body parts, it is noteworthy that these distances are not fixed, as they can vary depending on the given pose. Therefore, the process of determining the relationship between articulated body parts is a highly challenging task. Another important challenge of the HPE is created by the presence of occlusions between body parts. This means that some body parts can be masked by other ones, or by surrounding objects and therefore, making the HPE even more difficult. In addition, low contrast, cluttered background, variations in scene lighting and color scheme can also have a significant effect on HPE accuracy.

Recent approaches to HPE have successfully applied deep convolutional neural networks (CNNs). Due to their complex multi-layered structures, CNNs require a relatively large number of labeled (training) images to generate well-performing models [1], [2]. Given that the available

The associate editor coordinating the review of this manuscript and approving it for publication was Qiang Lai.

datasets frequently provide only a relatively small, fixed amount of labeled images, a common approach is to apply data augmentation techniques such as image rotations, to increase the number of training images and reduce the problem of over-fitting. Deeper and more complex CNN structures are more likely to reach higher levels of data generalization and discrimination capacity. Examples of such high-performing, and very complex neural network designs (with several blocks of neural networks stuck together) are given in [3]–[5]. These designs were shown to increase the accuracy of HPE. However, the data and computational costs were extremely high, making the use of graphic processing units (GPUs) paramount. To move away from increasing the CNN depth and complexity, a number of studies have proposed to integrate ''prior knowledge'' into CNNs to model structural information [2], [6]. These approaches offered low computational and training data requirements, while maintaining relatively high HPE accuracy.

Inspired by the advantages of the ''prior-knowledge'' methods, this study investigates the integration of structured graphs representing the human pose with the CNNs. An integration of a tree-based structure with a CNN to model human poses was originally proposed by Chu *et al.* [2]. The structure was based on human anatomy and included 26 nodes corresponding to 26 body joints. It was validated on the Leeds Sports Pose (LSP) dataset [7]. The purpose of the CNN was to learn the structural dependencies between feature maps of the body joints. However, the study did not conclusively show that this particular tree-based structure or the applied number of nodes were optimal. Yang *et al.* [8] used CNNs to model structural information of body parts using a non-tree-based iterative (loopy) model. However, like in the tree-based case, the optimality of this representation was not investigated. This study addresses this gap and investigates the optimality of CNN-based approaches using both, tree-based and non-tree-based models for HPE.

This study is an extended version of the original work presented at the International Conference on Digital Image Computing: Techniques and Applications (DICTA) 2017 [9]. While the conference paper introduced a new data-driven (as apposed to an anatomy-based) tree-based structure for HPE, the original contribution of the current study is a novel approach that combines non-tree-based and tree-based networks into a single hybrid network for HPE.

The remaining parts of this paper are organized as follows: Section 2 provides a brief review of previous related studies, Section 3 describes the methods used, Section 4 includes a discussion on experimental results, and the paper is concluded in Section 5.

II. PREVIOUS RELATED STUDIES

A. TREE-BASED MODELS FOR HPE

Tree-based models of human poses were first proposed by Felzenszwalb and Huttenlocher [10] and have been used in part-based approaches to model pairwise relationships between adjacent human body parts. To capture a larger range of pose variations, a global mixture of trees [11], or a mixture of local parts for each tree-based node [12] were introduced. One disadvantage of the tree-based representation was the inability to model complex poses, as only the pairwise interactions between nearby parts were captured. To solve this problem, Wang *et al.* [13] proposed a non-tree-based structure (or a loopy graph) to model high-order relationships between body parts. However, the loopy graphs used approximate inference, which lost the exact inference benefits given by the tree-based structures. This limitation was overcome by the hierarchical tree-based structure with latent nodes introduced by Tian *et al.* [14].

The majority of currently used tree-based structures for HPE are based on the anatomy of the human body [12], [14], [15]. Observable variables were used to train tree-based structures to model approximate positions of body joints [16]. Choi *et al.* [17] introduced two algorithms to automatically build latent tree-based structures from observations: the recursive grouping (RG) and the Chow-Liu Recursive Grouping (CLRG) algorithms. Using the CLRG algorithm, Wang and Li [16] trained different tree-based models using the LSP dataset, where the body joint positions played the role of observable variables. This study validated examples of these configurations (listed in Table 1) on the structured learning framework introduced in [2].

B. NON-TREE-BASED MODELS FOR HPE

Before the introduction of CNNs to HPE, several non-tree-based representations were proposed to extend the body parts modeling beyond pairwise links. Jiang *et al.* [18] combined tree-based and non-tree-based structures in a graph representation with strong (tree-based) edges to enforce arbitrary constraints and with weak (non-tree-based) edges to express the mutual exclusivity of inter-part occlusions and symmetric conditions. To further encapsulate the complexity of relations between body parts, Tran *et al.* [19] proposed a full-relation modeling of body parts by creating a comprehensive set of the body parts dependencies. Another, important representation was given by the hierarchical structure of body parts [13], [14] that included single rigid parts (e.g., torso, head, wrist) as well as parts that contained more than one rigid element. Finally, a number of recent studies applied CNNs to model the structural relationships between body parts using non-tree-based models [6], [8].

The non-tree-based modeling experiments for HPE described in this paper were based on methods proposed by Yang *et al.* [8] and Chu *et al.* [6]. Yang *et al.* incorporated prior knowledge into the CNN by jointly training it with deformable mixtures of body part models. The non-tree-based model was trained using the max-sum algorithm. Chu *et al.* [6], on the other hand, modeled the human pose as a non-tree-based structure using the sum-product algorithm [20]. In the case presented in this paper, the relationships between the body joints were estimated using the Conditional Random Field method combined with CNNs.

C. CNN-BASED POSE ESTIMATION APPROACHES

Recent CNN-based approaches apply deep convolutional neural networks (CNNs) to achieve higher expressive power. The idea is to train the network to map image parts into a number of locations denoting positions of body parts. These positions are iteratively refined through the hourglass encoding/decoding procedure. One of the main advantages over graphical models is the possibility of training the network to differentiate between many different poses even when some of the body parts are occluded. Although these techniques are compelling, the data and computational requirements are very high. Newell *et al.* [5] proposed the stacked hourglass network consisting of several coupled hourglass networks, functioning as a pose estimator. Wei *et al.* [3] introduced multi-stage convolutional pose machines with each stage containing receptive fields capturing local and global aspects of pose information. In another similar design, Chu *et al.* [21] added attention modules to each hourglass network to create a multi-context attention network. The current study investigates only the graphical tree and non-tree based approaches. In general, the graphical approaches do not perform as well as the CNN-based methods. However, they offer close to state-of-the-art performance at significantly lower computational and data costs.

III. METHODOLOGY

A. HUMAN POSE ESTIMATION FRAMEWORK

Similar to [8], the system uses a graph G = (V, E) to model human poses where, V denotes vertices or positions of body joints, and the edges $E \subseteq V \times V$ specify the spatial relationships between joints. Given an input image I, the full score $F(|)$ of a pose configuration is given as follows:

$$F(l, t|I; \theta, \omega) = \sum_{i \in V} \phi(l_i, t_i|I, \theta) + \sum_{i,j \in E} \psi(l_i, l_j, t_i, t_j|I, \omega_{i,j}^{t_i,t_j}) \tag{1}$$

where, θ and $\omega_{i,j}^{t_i,t_j}$ are model parameters, $K = |V|$ specifies the number of parts (nodes); $i \in \{1, \ldots K\}$ denotes the *ith* part; $l = \{l_i\}_{i=1}^{K}$ represent the pixel locations of parts; $t = \{t_i\}_{i=1}^{K}$ denote the mixture types of spatial relationships.

In the formula given by (1), the pose configuration $F(|)$ contains the part appearance term (or the unary term) $\phi(l_i, t_i|I, \theta)$ and the spatial relational term $\psi(l_i, l_j, t_i, t_j|I, \omega_{i,j}^{t_i,t_j})$. While the appearance term provides local confidence of the appearance of a part i located at l_i, the relational term models the spatial relationship of two neighboring parts i and j.

The experiments described in this study were based on the HPE system proposed in [2] where where the joint localization is formulated as a classification problem. It consisted of a pre-trained VGG16 image classification network [22] producing VGG16 features and a message passing network (MPN). The VGG16 network generated the appearance features while the MPN learned the spatial relationship features. In the VGG16 network structure [22], the *pool*4 and *pool*5 layers were removed to keep the prediction maps at a high resolution level. The sizes of the input images and the corresponding output feature maps were 336 × 336 pixels and 42 × 42 pixels respectively. In the message passing network, both tree-based and non-tree-based representations applied the sum-product algorithm. Denoting $m_{ij}(l_j, t_j)$ as a message sent from part i to part j and $u_i(l_i, t_i)$ as the belief of part i, the algorithm proceeds as follows:

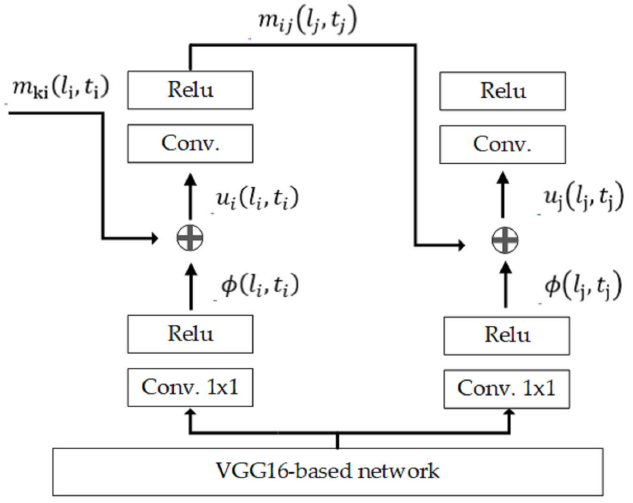

FIGURE 1. A message passing diagram from part *i* to part *j* within a CNN structure.

$$m_{ij}(l_j, t_j) \leftarrow \sum_{l_i, t_i} u_i(l_i, t_i) \otimes \omega_{i,j}^{t_i,t_j} \tag{2}$$

$$u_i(l_i, t_i) \leftarrow \phi(l_i, t_i) + \sum_{k \in N(i)} m_{ki}(l_i, t_i) \tag{3}$$

A flowchart of the message passing procedure between two adjacent body parts i and j is illustrated in Figure 1. Starting at the bottom of the diagram and moving upward, the output features from the VGG16 network (replicated for each body part) are convolved with the convolution layer 1 × 1 (conv. 1 × 1) to obtain the corresponding appearance term (ϕ). The belief parameter of each body part feature (u) is then updated by adding the appearance term (ϕ) to messages $m_{ki}(l_i, t_i)$ coming from the neighboring parts, and sharing the same edge with the current part, as given by (3). This is followed by the convolution with the updated belief, to form the part message (m), as given by (2). It is worth noting that the tree-based and non-tree-based representations use different mechanisms to pass messages. Namely, the tree-based structures use a serial message passing scheme in which one message is passed at a time while the non-tree-based representations apply the flooding scheme where, messages are passed simultaneously across every link at each time [8].

B. TREE-BASED HPE FRAMEWORK

The tree-based framework for HPE used a sequential message passing scheme where messages were passed between

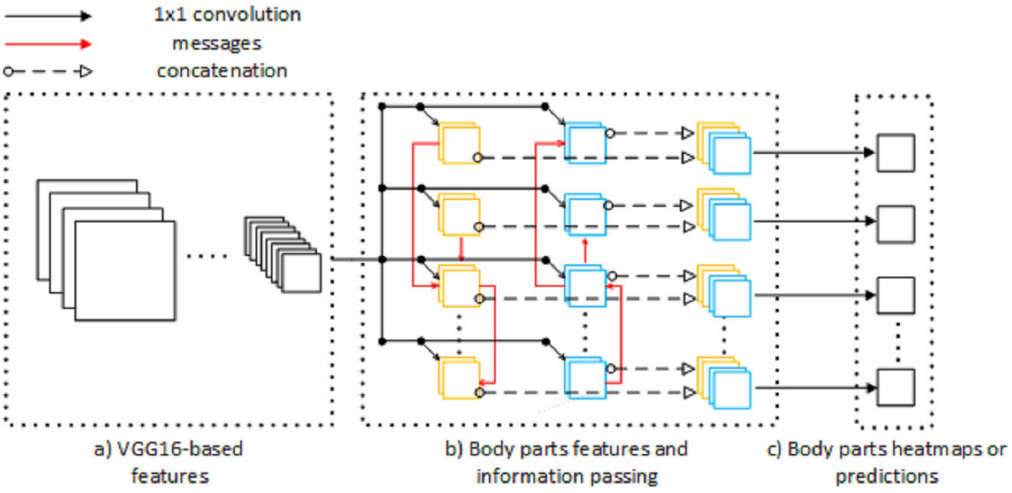

FIGURE 2. The tree-based HPE framework (adapted from [2]): (a) features obtained using layers similar to VGG16. (b) Body parts features and the refinement of these features by information passing. (c) Body parts heatmaps (predictions) of body parts: Yellow rectangles denote refined body parts features in the downward information passing direction; blue rectangles denote features refined in the upward information passing direction; red arrow-lines indicate the direction of information passing.

body part features in the order shown in Figure 2. The messages were passed within the tree-based structure in both upward and downward directions using geometric transform kernels [2]. The refined part-features obtained after message passing in upward and downward directions were concatenated and convolved with 1×1 convolutional layers to obtain part-detection heatmaps. The heatmaps represented spatial probability arrays showing the most likely positions of joints in a color-coded way.

In the tree-based experiments, each body joint was represented by a set of 128 feature maps. This number of feature maps was used for each joint as a compromise that gives good representative power of the network at a relatively low computational cost. Using a small number of feature maps would reduce the representative power while using a larger number of feature maps would increase the computational cost. All joints shared the *fconv*6 layer of the VGG16 network, which had 1024 feature channels. Feature maps of joints were passed from leaf nodes to the root node (upward direction) and from the root node to leaf nodes (downward direction). The refined feature maps in the upward direction were concatenated with those in the downward direction, generating 256 feature maps that were used to predict the final score map of a single joint.

C. NON-TREE-BASED HPE FRAMEWORK

The non-tree-based HPE framework uses the flooding message passing scheme where messages are passed simultaneously across every link. Suppose that, in a given graph structure the head and neck share the same edge, and so does the neck and left shoulder. This means that messages from the head to neck, neck to head, neck to left shoulder, and left shoulder to neck are being sent simultaneously. This scheme generates only approximate results, and the message passing procedure needs to be iterated a number of times for the training algorithm to converge to an acceptable solution [8].

The non-tree-based HPE framework applied in this study is shown in Figure 3. It uses the VGG16 network structure (with reference to the VGG16 weight layers proposed by [22]) to obtain appearance features for each body part. To learn the relationship between the appearance features, a non-tree based message passing network is used. It includes a cascade of two messaging layers, equivalent to two iterations of the message passing procedure. Figure 3b) demonstrates the belief u_1 and u_2 of each node after the first and the second iteration respectively. In each iteration, a node sends a message to its neighboring nodes simultaneously (denoted by solid lines in Figure 3b). If the network converges after n iterations, the achieved belief of each body part u_n is considered to be the final pose estimation.

D. PROPOSED HYBRID HPE FRAMEWORK

Another HPE approach tested in this study is a new hybrid framework consisting of combined tree-based and a non-tree-based representations as presented in Figure 4. This framework contains three main building blocks. The first block uses the VGG-based structure (with reference to the VGG16 weight layers proposed by [22]). The weights of this part of the network are generated during the pre-training process. During the training, the initial pre-trained weights are updated but at a lower speed (a tenth of the pre-training rate). The inputs to the first building block are training images of size $336 \times 336 \times 3$ pixels. The output features of the first building block (of size 42×42) are considered to be the appearance terms providing local confidence values for each body part. In the second building block, feature maps of each body part are updated and refined through two iterations of the non-tree-based message passing network. The belief outputs of

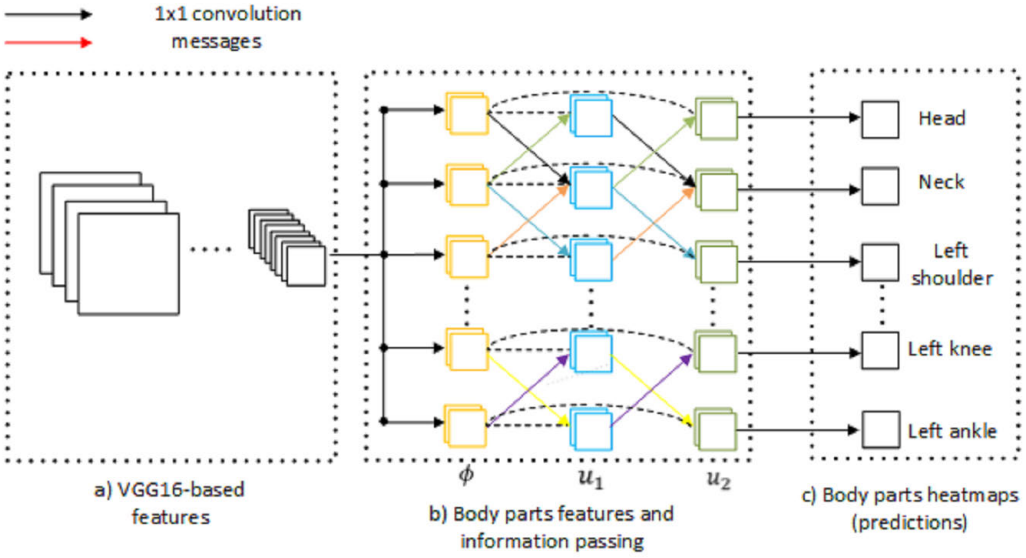

FIGURE 3. The non-tree-based HPE framework (adapted from [8]): (a) VGG16-based features obtained using layers similar to VGG16. (b) Body parts features and the refinement of these features by information passing. (c) Body parts heat maps (predictions).

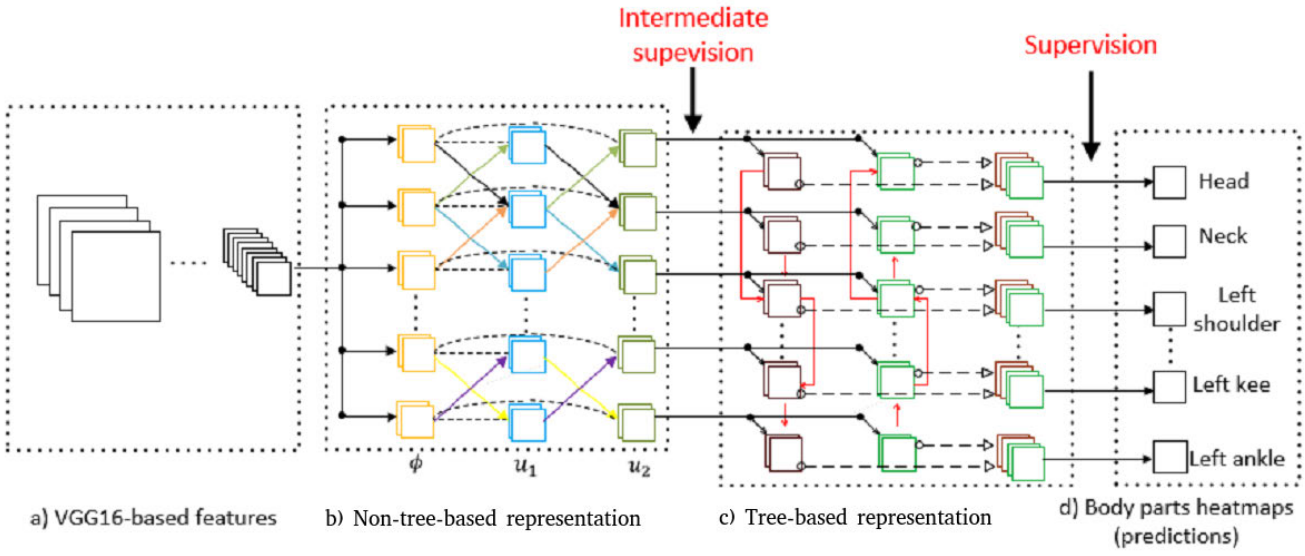

FIGURE 4. The hybrid (combined tree-based and non-tree-based) model.

the second block are used as the appearance features for the third building block, which is given as the tree-based message passing network proposed by [2]. These three building blocks are connected successively. The performance of the proposed framework was tested using both a single loss function and two loss functions.

E. HUMAN POSE MODELS

1) TESTED TREE-BASED MODELS

The data-driven tree-based models tested in this study were obtained by applying the Chow-Liu Recursive Grouping (CLRG) algorithm [16], [17] to the training dataset. The algorithm first grouped the observed nodes that were likely to be close to each other, and then followed a process of recursive grouping. During the recursive grouping stage, the distances between nodes were used to determine groups of sibling nodes and recursively build up a tree-based structure.

Table 1 shows the tested tree-based and non-tree-based configurations used for pose representations. The tree-based representations included both anatomy-based and data-driven models. The anatomy-based tree-based models included structures made out of different numbers of nodes (TA_14, TA_26, TA_30, TA_34_A, TA_34, TA_38, TA_50). The data-driven tree-based structures included 26-node models

TABLE 1. Human pose models tested in the HPE experiments.

Configurations	Tree-based, Non-tree-based or Hybrid	Anatomy-based or Data-driven	Number of nodes	Modification	Fig. number
TA_14	Tree-based	Anatomy-based	14		6
TA_26 [2]	Tree-based	Anatomy-based	26		6
TA_30	Tree-based	Anatomy-based	30		6
TA_34	Tree-based	Anatomy-based	34		6
TA_34_A	Tree-based	Anatomy-based	34		6
TA_38	Tree-based	Anatomy-based	38		6
TA_50	Tree-based	Anatomy-based	50		6
TD_26	Tree-based	Data-driven	26		5a
TD_26_C	Tree-based	Data-driven	26	Centroid	5b
NT_26_A [8]	Non-tree-based	-	26	2 iterations, no loopy connections between left and right body parts	7(a)
NT_26_B	Non-tree-based	-	26	2 iterations, 2 loopy connections between left and right body parts	7(b)
NT_26_C	Non-tree-based	-	26	2 iterations, 5 loopy connections between left and right body part	7(c)
H_26_1	Hybrid	Anatomy-based	26	A single loss functions	7(a),6
H_26_2A	Hybrid	Anatomy-based	26	Two loss functions	7(a),6
H_26_2B	Hybrid	Anatomy-based	26	Two loss functions and feature concatenation	7(a),6

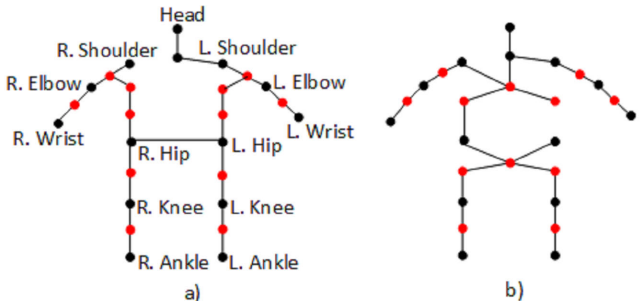

FIGURE 5. Data-driven tree-based configurations; (a) TD_26 and (b) TD_26_C.

(the TD_26 and TD_26_C configurations) in which data comes from the pose space of the LSP dataset [7]. The 26 nodes of the data-driven models included 14 original joints specified by the training dataset, and an additional 12 nodes representing joints formed at midpoints or centroids of the original 14 joints. Given two joint positions (x_1,y_1) and (x_2,y_2), the midpoint (x,y) can be obtained as given by (4). Similarly, given four joint positions (x_1,y_1), (x_2,y_2), (x_3,y_3), (x_4,y_4), the centroid-type joint (x,y) is calculated, as given by (5). The centroid-type joints were used only in the TD_26_C configuration, inspired by the tree-based representation described in [16].

$$(x, y) = \left(\frac{x_1 + x_2}{2}, \frac{y_1 + y_2}{2}\right) \tag{4}$$

$$(x, y) = \left(\frac{x_1 + x_2 + x_3 + x_4}{4}, \frac{y_1 + y_2 + y_3 + y_4}{4}\right) \tag{5}$$

The 26-node tree (TA_26 configuration) in Table 1 represents the tree-based structure proposed by [2]. In the case of the 14-node tree (TA_14), the average distance between neighboring joints was larger than a pre-defined kernel size of geometric transform kernels in [2], that affected the training of these kernels. To address this issue, this study introduced intermediate joints to reduce the distance between neighboring joints. The effect of the added joints (or tree-based nodes) was also investigated.

2) TESTED NON-TREE-BASED MODELS

As shown in Table 1, apart from different tree-based configurations, a number of non-tree-based configurations were tested and included the following configurations: NT_26_A, NT_26_B, and NT_26_C. As observed by Yang *et al.* [8], a cascade of two or three message passing layers was sufficient to produce good results. Therefore, all of the tested non-tree-based configurations contained only two message passing layers, which was equivalent to two iterations of the message passing procedure. The non-tree-based configuration NT_26_A [8] represents a basic structure of a human body, while the NT_26_B and the NT_26_C configurations introduce additional connections between left and right body parts. In contrast to the non-tree-based configurations proposed by [8] that applied the max-sum algorithm, the non-tree-based configurations proposed in this paper use the sum-product algorithm (similar to the one implemented in [6]).

3) PROPOSED NEW HYBRID MODELS

The proposed hybrid (combined tree-based and non-tree-based) configurations shown in Table 1 include H_26_1, H_26_2A and H_26_2B. The H_26_1 configuration contained a non-tree-based structure with two message passing layers followed by a tree-based structure with a single loss function applied to the whole network (Figure 4). The H_26_2A configuration has a similar structure to the previous configuration, except that instead of a single loss function, two loss functions were used, one for the non-tree-based part of the network and the other for the entire network. Moreover, instead of passing the output from the non-tree-based network to the tree-based network input, the input and output of the

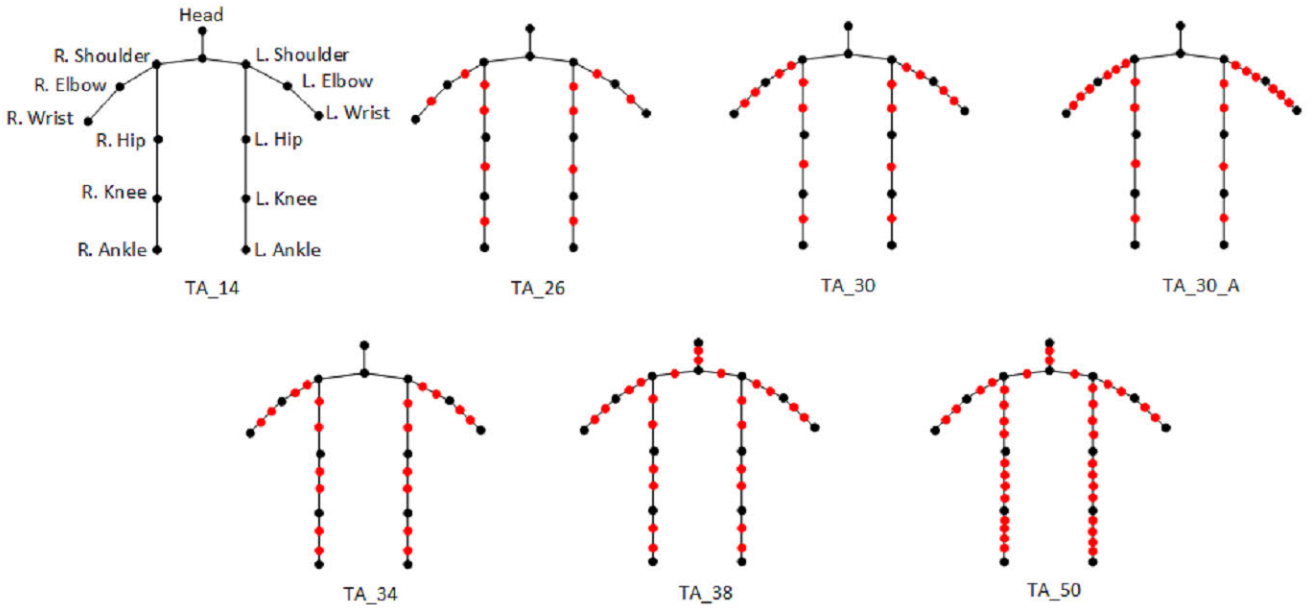

FIGURE 6. Anatomy-based tree-based configurations.

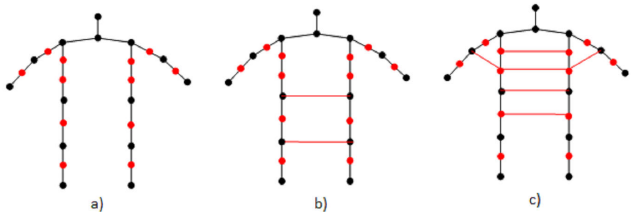

FIGURE 7. Non-tree-based configurations.

non-tree-based network in the H_26_2B configuration were concatenated to form a combined input to the tree-based network. This feature was inspired by the dense network proposed by [23], where all layers were connected to each other.

F. DATABASE

The HPE experiments were conducted on the Leeds Sports Pose (LSP) benchmark dataset [7] containing 2000 images: 1000 images for training and 1000 images for testing. These images capture sports activities and are supplied with full-body annotations. The annotations used the Person Centric (PC) style, where the left/right sides of body parts were labeled according to the viewpoint of the person being depicted. The PC annotations were converted to the Observer Centric (OC) style following the approach used in [2]. In addition to the LSP dataset, the INRIA Person dataset images [24], which did not depict people, were used. The addition of these INRIA Person images provided ''negative'' training increasing the system robustness to noise.

G. PERFORMANCE MEASURES AND BENCHMARKS

The HPE performance was assessed using the strict Percentage of Correct Parts (strict PCP) measure [1]. PCP is a standard evaluation metric on several benchmarks, including the LSP dataset used in our study. This measure is consistent with the majority of similar publications using the LSP data. The strict PCP evaluates only a single highest-scoring estimation outcome for a given test image. A body part is considered as correctly classified if both of its endpoint-joints are located within 50% of the length of the ground-truth annotated endpoints. In the LSP dataset, each image contained only one annotated person. The experimental results were benchmarked against results obtained in [2] and [8].

IV. RESULTS

This section presents experimental results between tree-based, non-tree-based and hybrid models. In tree-based models, the result of different data-driven, anatomy-based models, as well as tree models with varied numbers of nodes are also demonstrated. The visual result of the anatomy-based model is also provided and discussed.

A. HPE RESULTS FOR TREE MODELS

1) A COMPARISON BETWEEN DATA-DRIVEN AND ANATOMY-BASED TREE MODELS

As shown in Table 2 and Figure 8, for the same set of joints, the proposed data-driven representation (TD_26) obtained 0.9% higher HPE accuracy compared to the benchmark anatomy-based representation (TA_26 [2]). These results demonstrate that given the same set of tree nodes, the way in which the nodes are connected can have a significant impact on learning the dependencies between joints. The higher accuracy of the data-driven representation can be attributed to the fact that, in contrast to the rigid predefined anatomy-based structures, the data-driven approach had the freedom to derive data-adaptive structures of inter-node dependencies.

TABLE 2. HPE accuracy in percentage (%) (using strict PCP evaluation protocol) for different tree-based configurations.

Configurations	Head	Torso	Upper arm	Lower arm	Upper leg	Lower leg	Mean HPE
TA_14	88	87.7	71.5	59.5	78.2	72.2	73.9
TA_26 ([2])	89.2	93.9	76.4	63.9	85.7	80.3	**79.6**
TA_30	88.6	93.6	77.5	65.9	85.9	81.1	80.3
TA_34_A	90.6	93.3	76.8	64.3	84.9	79.8	79.6
TA_34	89.5	92.6	76.6	65.5	87	82.3	80.5
TA_38	89.5	93	77.6	66.3	87	81.7	80.8
TA_50	90.5	94.1	76.6	65.7	87.8	82.9	81.1
TD_26	89	94.5	77	64.8	87.1	81.7	80.5
TD_26_C	87	93.6	74.4	63.5	85.7	80.6	79.1
NT_26_A [8]	88.6	93.5	74.1	59.7	84.2	79	**77.6**
NT_26_B	88.2	93.1	74.1	63.1	84.5	79.4	78.3
NT_26_C	87.4	94.3	74.6	62.3	84.6	79.8	78.4
H_26_1	88.0	92.7	74.6	62.7	84.7	79.4	78.35
H_26_2A	88.9	94.5	77.2	64.8	86.1	81.2	80.2
H_26_2B	89.3	94.8	77.8	65.5	85.9	81.4	80.5

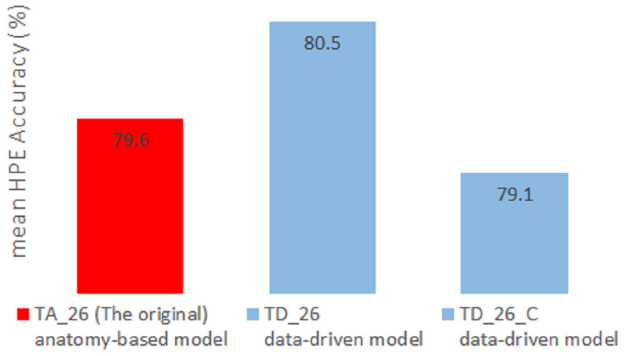

FIGURE 8. Mean HPE accuracy for anatomy-based and data-driven tree models.

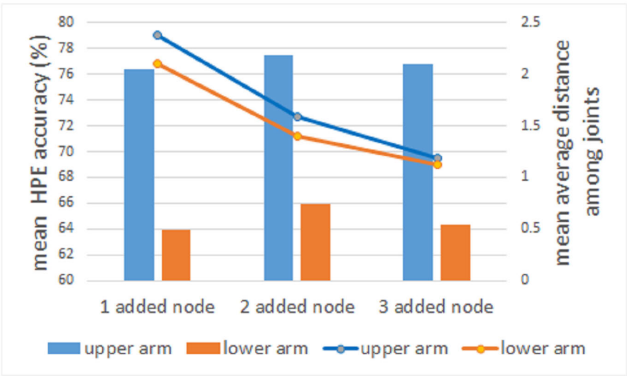

FIGURE 9. HPE accuracy of lower arm and upper arm for anatomy-based tree models with additional nodes.

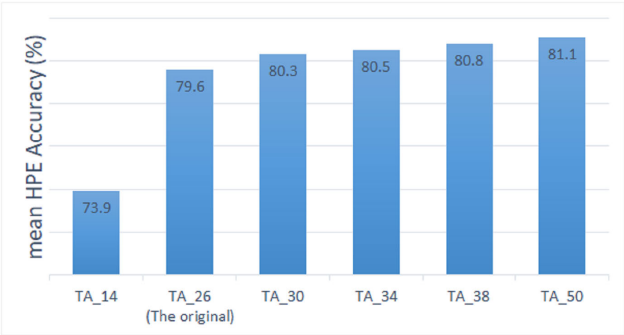

FIGURE 10. Mean HPE accuracy for anatomy-based tree-based models with different numbers of nodes.

2) A COMPARISON BETWEEN DIFFERENT DATA-DRIVEN MODELS

In contrast to TD_26, the TD_26_C configuration has 2 tree nodes that represent centroid-type joints formed as centroids of a subset of existing joints. The TD_26_C representation (data-driven model) obtained a mean HPE accuracy 1.4% lower than the TD_26 representation (79.1% vs 80.5% in Table 2 and Figure 8). One of the possible explanations for the decreased HPE accuracy is the relatively large distance between some of the neighboring joints, which could be a problem when the distance between joints is estimated based on high-level features. For example, in the TD_26_C representation, the distance between some neighboring joints where the nodes are on the same edge, was larger than 9 pixels. Meanwhile, only two consecutive 7×7 geometry transform kernels (that was equivalent to one 9×9 kernel) were used to learn a deformation model for 2 neighboring joints. These 2 kernels were targeted for a high-level joint distance of less than 9 pixels. Therefore, it is possible that with a joint distance of more than 9 pixels, the kernels were unable to learn effectively, leading to the decreased HPE accuracy.

3) EFFECTS OF VARYING THE NUMBER OF TREE NODES

In order to find an optimal number of tree-based nodes, given the average distance (AD) between neighboring body joints, experiments with different numbers of joints on the upper and lower arms were conducted. Added joints (tree-based nodes) to a body part reduced the distance between neighboring joints (joints represented by nodes on the same tree-based edges). As illustrated in Figure 9, both the upper and lower arms achieved the highest HPE accuracy when the AD on the arm was approximately 1.5 whereas, the HPE accuracy decreased when the AD approached 1. As a kernel stride of 1 was used, when the average distance between two neighboring body joints was less than 1, the transform kernels between these two joints failed to learn. Therefore, when

FIGURE 11. Examples of the HPE estimation results using the LSP dataset [7]. a) Correct estimation results. b) Incorrect estimation results caused by the presence of multiple people. c) Incorrect estimation results caused by strong pose articulation and low image quality.

the AD approached the value of 1, transform kernels were not efficiently trained, leading to decreased HPE accuracy. On the other hand, with large AD values, the addition of intermediate joints generated more training data for a network, that resulted in an increased HPE accuracy. In conclusion, the experiments suggested that an AD value of 1.5, provided an average distance between neighboring joints that lead to optimal tree-based representations. The resulting optimal tree-based structure contained 50 nodes (the TA_50 configuration in Figure 6). It can be seen in Figure 10, that the 50-node representation obtained a mean HPE accuracy of 81.1%, that was 1.5% higher than the HPE accuracy obtained when using the original TA_26 representation (79.6%). With 4 added nodes for both lower legs (AD = 1.44), the lower leg accuracy was improved by 2.6% (from 80.3% to 82.9%); with 4 added nodes for both upper legs (AD = 1.4), the lower leg accuracy was improved by 2.1% (from 85.7% to 87.8%). Figure 10 shows that the lowest mean accuracy of 73.9% was achieved for the 14-node representation. This was an example of a tree having a small number of nodes with a resulting large number of neighboring joints that had an inter-joint distance larger than 9 pixels, creating a significant challenge for the geometric transform kernel.

4) HPE FROM IMAGES WITH MULTIPLE PEOPLE

Figure 11 shows examples of the human pose estimation for the LSP dataset using the model of the TD_26 configuration. Correct estimations are displayed in Figure 11a) whereas incorrect estimations are given in Figures 11b) and 11c). It can be observed that the incorrect estimations shown in Figures 11b) were most likely to occur for images of multiple people where body parts of one person were occluded by body parts of other people. The estimation errors in these cases were caused by the use of a very simple post-processing approach that did not account for the effects of multiple people [2]. As suggested in [4], this issue could be largely eliminated by applying more complex post-processing methods that include multi-person estimation. Figure 11c) illustrates examples of incorrect estimations resulting from either a very strong pose articulation, or very low image quality. These types of errors could be reduced by increasing the number of strong-pose and noisy images in the training set.

B. COMPARISON OF DIFFERENT NON-TREE-BASED MODELS

Table 2 shows the mean HPE accuracy of different non-tree-based configurations including NT_26_A, NT_26_B and NT_26_C. In comparison with the accuracy of the basic non-tree-based structure NT_26_A proposed by [8], that achieved a mean HPE accuracy of 77.6%, the NT_26_B and NT_26_C configurations proposed additional connections between the left and right body parts, resulting in higher accuracy of 78.3% and 78.4% respectively. This indicates that the number and type of connections between body parts

TABLE 3. HPE accuracy in percentage (%) for different CNN-based HPE methods. The LSP dataset [7] contains 1000 images for training and 1000 images for testing while the MPII [25] and LSP extended [26] datasets have 25000 and 10000 images respectively.

Methods	No Training Images	No Testing Images	HPE accuracy
Stacked hourglass network [5]	MPII[22000]	MPII[3000]	90.9%
Convolutional Pose Machine [3]	LSP[1000] + MPII[25000] + LSP extended[10000]	LSP[1000]	90.5%
Multi-context attention network [21]	LSP[1000] + MPII[25000] + LSP extended[10000]	LSP[1000]	92.6%
The proposed method	LSP[1000]	LSP[1000]	81.1%

can have a significant impact on the non-tree-based modeling outcomes.

C. COMPARISON OF DIFFERENT HYBRID MODELS

Table 2 shows the mean HPE accuracy for different hybrid models combining tree-based and non-tree-based structures. Since the depth of the combined network was significantly increased compared to a single network configuration, the system became prone to the vanishing gradient problem [5]. Therefore, it is understandable that the H_26_1 configuration with a single loss function obtained a low accuracy of 78.35%. However when intermediate supervision was applied by using two loss functions in both the H_26_2A and H_26_2B configuration, the accuracy was increased to 80.2% and 80.5% respectively, which was approximately 2% higher than the single-loss configuration. In addition, the concatenation of features from different layers in the H_26_2B configuration lead to 0.3% improvement in HPE accuracy, compared to the H_26_2A configuration (80.5% vs 80.2%). The hybrid configuration (H_26_2B) obtained an accuracy of 80.5%, that is nearly 1% higher than the HPE accuracy of either structure alone (i.e., the non-tree-based structure NT_26_A (77.6%) and the tree-based structure TA_26 (79.6%)).

D. COMPARISON WITH CNN-BASED METHODS

Table 3 compares the HPE accuracy with recent CNN-based HPE methods. In the Multi-context Attention [21] and Convolutional Pose Machine network [3], the MPII (25000 images) and LSP extended (10000 images) are added to the LSP training set (1000 images), generating a large amount of training data. The two networks stacked several individual networks and obtained HPE accuracies of 92.6% and 90.5% respectively. The Stacked Hourglass network [5], with eight networks stacked together, trained on MPII training set of 22000 images and achieved the HPE accuracy of 90.9% on 3000 images of MPII. On the other hand, the proposed method in this paper trained on the LSP training set of 1000 images and used much less computation resources, achieving an HPE accuracy of 81%.

V. CONCLUSION

This paper investigated the incorporation of prior knowledge into CNNs through graph structures including tree-based and non-tree-based models. It was observed that both of the proposed data-driven tree-based models and hybrid approaches obtained higher HPE accuracy compared to the benchmark anatomy-based and non-tree-based models. The best overall HPE results were obtained when using the anatomy-based benchmark with an increased number of nodes. Future work will investigate network designs with feature concatenation from different levels of network hierarchy to improve the feature representation.

ACKNOWLEDGMENT

The authors would like to thank Dr. W. Yang from the Chinese University of Hong Kong for valuable advice on the CNN based non-tree-based models, and Dr. X. Chu for her comments on the CNN based tree-based models.

REFERENCES

[1] X. Chen and A. L. Yuille, "Articulated pose estimation by a graphical model with image dependent pairwise relations," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1736–1744. [Online]. Available: http://dl.acm.org/citation.cfm?id=2968826.2969020

[2] X. Chu, W. Ouyang, H. Li, and X. Wang, "Structured feature learning for pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4715–4723.

[3] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4724–4732.

[4] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, and B. Schiele, "DeepCut: Joint subset partition and labeling for multi person pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4929–4937.

[5] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 483–499.

[6] X. Chu, W. Ouyang, H. Li, and X. Wang, "CRF-CNN: Modeling structured information in human pose estimation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 316–324. [Online]. Available: http://papers.nips.cc/paper/6278-crf-cnn-modeling-structured-information-in-human-pose-estimation.pdf

[7] S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation," in *Proc. Brit. Mach. Vis. Conf.*, 2010, vol. 2, no. 4, pp. 12.1–12.11.

[8] W. Yang, W. Ouyang, H. Li, and X. Wang, "End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3073–3082.

[9] H. T. Vu, R. H. Wilkinson, M. Lech, and E. Cheng, "A comparison between anatomy-based and data-driven tree models for human pose estimation," in *Proc. Int. Conf. Digit. Image Comput., Techn. Appl. (DICTA)*, Nov. 2017, pp. 1–7.

[10] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient matching of pictorial structures," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2000, pp. 66–73.

[11] Y. Wang and G. Mori, "Multiple tree models for occlusion and spatial constraints in human pose estimation," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2008, pp. 710–724.

[12] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2878–2890, Dec. 2013.

[13] Y. Wang, D. Tran, and Z. Liao, "Learning hierarchical poselets for human parsing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 1705–1712.
[14] Y. Tian, C. L. Zitnick, and S. G. Narasimhan, "Exploring the spatial hierarchy of mixture models for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2012, pp. 256–269.
[15] A. Jain, J. Tompson, M. Andriluka, G. W. Taylor, and C. Bregler, "Learning human pose estimation features with convolutional networks," in *Proc. Int. Conf. Learn. Represent.*, Apr. 2014, pp. 1–11. [Online]. Available: http://arxiv.org/abs/1312.7302
[16] F. Wang and Y. Li, "Beyond physical connections: Tree models in human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 596–603.
[17] M. J. Choi, V. Y. Tan, A. Anandkumar, and A. S. Willsky, "Learning latent tree graphical models," *J. Mach. Learn. Res.*, vol. 12, no. 5, pp. 1771–1812, 2011. [Online]. Available: http://dl.acm.org/citation.cfm?id=1953048.2021056
[18] H. Jiang and D. R. Martin, "Global pose estimation using non-tree models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.
[19] D. Tran and D. Forsyth, "Improved human parsing with a full relational model," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 227–240.
[20] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 498–519, Feb. 2001.
[21] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1831–1840.
[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, May 2015, pp. 1–14. [Online]. Available: http://arxiv.org/abs/1409.1556
[23] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.
[24] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 886–893.
[25] S. Johnson and M. Everingham, "Learning effective human pose estimation from inaccurate annotation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 1465–1472.
[26] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3686–3693.

HUYNH THE VU received the M.E. degree in electrical and computer engineering from RMIT University, Vietnam, and the Ph.D. degree in electrical and electronic engineering from RMIT University, Melbourne, Australia, where he is currently pursuing the Ph.D. degree with the School of Engineering. His research interests include human pose estimation, computer vision, and machine learning.

RICHARDT H. WILKINSON received the B.Eng., M.Eng., and Ph.D. degrees from the University of Stellenbosch, Stellenbosch, South Africa, in 1994, 1998, and 2004, respectively. He joined the Cape Peninsula University of Technology, Cape Town, South Africa, as a Postdoctoral Researcher, in 2005, after which, he was appointed as Senior Researcher, in 2007. In 2008, he was appointed as the Head of the Centre for Instrumentation Research and promoted to an Associate Professor, in 2011. He joined RMIT University, Melbourne, Australia, in 2012, where he is currently a Senior Lecturer with the School of Engineering. His research interests include multilevel power electronic converters, modulation techniques, Fourier techniques, digital signal processing, artificial intelligence, digital audio amplifiers, FPGA development, and embedded controller design. He is a Senior Member of the IEEE Signal Processing, Power Electronics, Industry Applications, Industrial Electronics, and Power Engineering Societies. He served as the Chapter Chair for the South Africa Section's Joint Industry Applications/Industrial Electronics/Power Electronics Chapter, from 2007 to 2012.

MARGARET LECH received the M.S. degree in physics from Maria Curie-Skłodowska University, Poland, and the Ph.D. degree in electrical engineering from the University of Melbourne, Australia. She is currently a Professor with the School of Engineering, RMIT University, Australia. Her research interests include psychoacoustic, speech and image processing, system modeling, and optimization.

EVA CHENG received the Ph.D. degree in telecommunications engineering from the University of Wollongong, Australia. She is currently a Senior Lecturer with the School of Electrical and Data Engineering, University of Technology Sydney, Australia. Her research interests in multimedia signal processing include areas such as 3D video/audio recording and reproduction, computer vision, and speech/audio processing.