Multi-source Deep Learning for Human Pose Estimation

Wanli Ouyang Xiao Chu Xiaogang Wang
Department of Electronic Engineering, The Chinese University of Hong Kong
wlouyang@ee.cuhk.edu.hk, xgwang@ee.cuhk.edu.hk

Abstract

Visual appearance score, appearance mixture type and deformation are three important information sources for human pose estimation. This paper proposes to build a multi-source deep model in order to extract non-linear representation from these different aspects of information sources. With the deep model, the global, high-order human body articulation patterns in these information sources are extracted for pose estimation. The task for estimating body locations and the task for human detection are jointly learned using a unified deep model. The proposed approach can be viewed as a post-processing of pose estimation results and can flexibly integrate with existing methods by taking their information sources as input. By extracting the non-linear representation from multiple information sources, the deep model outperforms state-of-the-art by up to 8.6 percent on three public benchmark datasets.

1. Introduction

Human pose estimation is the process of determining, from an image, the positions of human body parts such as the head, shoulder, elbow, wrist, hip, knee, and ankle. It is a fundamental problem in computer vision and has abundant important applications such as sports, action recognition, character animation, clinical analysis of gait pathologies, content-based video and image retrieval, and intelligent video surveillance. Despite many years of research [52, 54, 2, 40, 6, 57, 56], pose estimation remains a difficult problem. One of the most significant challenges in pose estimation is how to model the complex human articulation.

Many approaches have been used to handle the complex human articulation by using three information sources: mixture type, appearance score and deformation [57, 52, 54, 11, 58]. Influenced by human body articulation, clothing, occlusion etc., body part appearance varies. To handle this variation, the appearance of a part is clustered into multiple *mixture types* as shown in Fig. 1 . For each mixture type of a part, a part template is learned to capture its appearance. Then the *appearance scores* (log-likelihoods) of body parts being at different locations are obtained by convolving the part templates with the visual features of the input image, e.g. HOG [7]. The appearance scores are inaccurate for well-locating body parts because the part template is imperfect. Therefore, the *deformations* (relative locations) among body parts are used as for encoding likely pairwise poses; for example, the head should not be far from the neck.

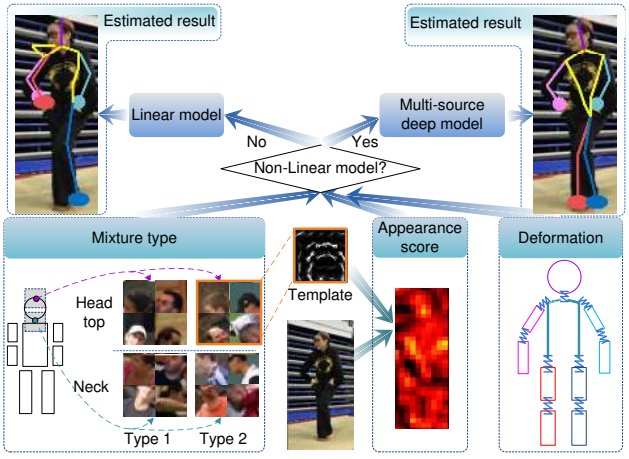

Figure 1. The motivation of this paper in using multi-source deep model for constructing the non-linear representation from three information sources: mixture type, appearance score and deformation. Best viewed in color.

Existing approaches use log-linear models with pairwise potentials of these three information sources [52, 54, 40, 57, 56] to determine whether an estimated location is correct. However, these information sources are not log-linearly correlated when choosing the correct candidate. For the example in Fig. 1, linear models may find that the estimated result on the left and the result on the right have the same deformation score because they simply linearly add local deformation cost. While it is obvious for human to find that the result on the left is not reasonable. Similar situations also occur for mixture type and appearance score. Therefore, it is desirable to construct the non-linear representation that identifies reasonable configurations of deformation, appearance score and mixture type.

In order to construct useful representation from multiple information sources for pose estimation, a model should

satisfy certain properties. First, the model should capture the global, complex relationships among body parts. For the example in Fig. 1, the result on the left is unreasonable because of its global configuration in arm, torso, and leg. Second, since reasonable configuration is a very abstract concept while the information sources are less abstract concepts, the model should construct more abstract representation from the less abstract representation. Third, since different information sources describe different aspects of human pose and have different statistical properties, the model should learn useful representation from these sources and fuse them into a joint representation for pose estimation. The multi-source deep architecture we propose satisfies the above requirement.

There are three contributions of this paper.

1. We propose a deep architecture to construct the non-linear representation from different aspects of information sources. To the best of our knowledge, this paper is the first to use deep model for pose estimation.
2. The body articulation patterns (global and more abstract representations) are captured by the deep model from the information sources (local and less abstract representations). For each information source, more abstract representation at the higher layer is composed by the less abstract representation of all body parts in the lower layer. Then representations of all information sources in the higher layer are fused for pose estimation.
3. Both the task for detecting human and the task for estimating body locations are jointly learned using a single deep model. Joint learning of these tasks with a shared representation improves pose estimation accuracy.

2. Related work

Human pose estimation. Pose estimation is considered as holistic recognition in [15, 33, 34]. On the other hand, many recent works use local body parts [52, 54, 11, 58, 9, 13, 48, 40, 2, 42, 21, 46, 55, 41, 1] in order to handle the many degrees of freedom in body part articulation. Since the first work in [57], some approaches [52, 54, 11, 58, 9] have clustered part appearance into mixture types as shown in Fig. 1. There are also approaches that warp the part template by flexible sizes and orientations [13, 48, 40, 2, 42, 21, 46, 55]. The appearance score, rotation, size, and location used in these approaches can be treated as multiple information sources and used by our deep model for pose estimation.

In existing pose estimation approaches, the pair-wise part deformation relationships are arranged in tree models [52, 54, 2, 40, 57], multi-tree model [55], or loopy models [56, 53, 10]. Tree models allow for efficient and exact inference but are insufficient in modeling the complex relationships among body parts. Hence, tree models often suffer from double counting; for example, given the position of a torso, the positions of two legs are independent and often respond to the same visual cue. Loopy models allow more complex relationships among parts, but require approximate inference. Our deep architecture models the complex relationships among parts and is computationally efficient in both training and testing.

Deep learning. Since the breakthrough in deep learning initiated by G. Hinton in [18, 19], deep learning is gaining more and more attention. Bengio [3] proved that existing commonly used machine learning tools such as SVM and Boosting are shallow models, and they may require many more computational elements, potentially exponentially more (with respect to input size), than deep models whose depth is matched to the task. Deep architecture is found to yield better data representation, for example, in terms of classification error [25], invariance to input transformations [16], or modeling multi-modal data [35]. Deep learning has achieved spectacular progress in computer vision [45, 20, 26, 36, 23, 59, 12, 43, 39, 50, 49, 60, 38, 37, 30, 32, 31, 29, 61, 28, 51]. Recent progress on deep learning is reviewed in [4]. Krizhevsky *et al.* [23] proposed a large-scale deep convolutional network [27] with breakthrough on the large-scale ImageNet object recognition dataset [8], attaining a significant gap compared with existing approaches that use shallow models, and bringing high impact to research in computer vision. Our approaches in [38, 39, 37, 29] learns feature learning, translational deformation, and occlusion relationship in pedestrian detection; the approach in [50] learns relational filter pairs in face verification. To the best of our knowledge, however, deep model for human pose estimation has not yet been explored.

Our work is inspired by multi-modality models that learn from multiple modalities such as audio, visual, text data [35, 47, 17]. In contrast to these works, we investigate multi-source learning from single modality, which is image data in pose estimation.

3. Pictorial structure model for pose estimation

The model introduced in this section is used to provide our deep model with information sources. Pictorial structure model considers human body parts as nodes tied together in a conditional random field. Let l_p for $p = 1, \ldots, P$ be the configuration of the pth part. The posterior of a configuration of parts $\mathbb{L} = \{l_p | p = 1 \ldots P\}$ given an image I is:

$$P(\mathbb{L}|I) \propto \exp\Big(\sum_{p=1}^{P} \phi(I|l_p)) + \sum_{(p,q)\in E} \psi(l_p, l_q)\Big). \quad (1)$$

$\psi(l_p, l_q)$ is the pair-wise term that models the geometric deformation constraint on the pth and qth parts; for example, head shall not be too far from torso. The edge set denoted by E is arranged in tree models [52, 54, 2, 40, 57, 11]

or loopy models [56, 10, 53].

$\phi(I|l_p)$ is the unary term that models the appearance of l_p. The appearance varies as body articulates. To model this variation, $l_p = \{s_p, \theta_p, z_p\}$ and $\phi(I|l_p)$ specifies the part appearance warped by size s_p, orientation θ_p at location z_p in [2, 40, 13]. Alternatively, Yang and Ramanan propose to use appearance mixture type t_p for approximating the variation in rotation θ_p and size s_p in [57]. In this model, $l_p = \{t_p, z_p\}$ and $\phi(I|l_p)$ specifies the part appearance with mixture type t_p at location z_p. The appearance of a part is clustered into multiple appearance mixture types as shown in Fig. 1. The overall model in [57, 58] is as follows:

$$P(\mathbb{L}|I) \propto \exp\big(S(I, \mathbf{t}, \mathbf{z})\big), \tag{2}$$

$$\text{where } S(I, \mathbf{t}, \mathbf{z}) = S_c(\mathbf{t}) + \sum_{p,q} S_d(\mathbf{t}, \mathbf{z}, p, q) + \sum_{p} S_a(I, t_p, z_p),$$

$$S_c(\mathbf{t}) = \sum_{p} b_p^{t_p} + \sum_{p,q} b_{p,q}^{t_p,t_q}, \tag{3}$$

$$S_d(\mathbf{t}, \mathbf{z}, p, q) = \mathbf{w}_{p,q}^{t_p,t_q\,\mathrm{T}} d(z_p - z_q), \tag{4}$$

$$S_a(I, t_p, z_p) = \mathbf{w}_p^{t_p\,\mathrm{T}} f(I, z_p). \tag{5}$$

- $S_c(\mathbf{t})$ is the pair-wise compatibility term that models the compatibility/co-occurrence of mixture types.
- $S_d(\mathbf{t}, \mathbf{z}, p, q)$ is the pair-wise deformation term that models the geometric deformation constraints on the pth and qth parts. $d(z_p - z_q) = [dx, dy, dx^2 dy^2]^{\mathrm{T}}$.
- $S_a(I, t_p, z_p)$ is the unary appearance term that computes the score of placing a template $w_p^{t_p}$ at location z_p of the HOG feature map for image I, denoted by $f(I, z_p)$.

Linear SVM is used to learn the linear weights $\mathbf{w}_p^{t_p}$, $\mathbf{w}_{p,q}^{t_p,t_q}$ and compatibility biases $b_p^{t_p}$, $b_{p,q}^{t_p,t_q}$. The model in Eq.(2)-(5) is used in many approaches, with different implementations on edge set, part size, and part locations [52, 54, 57, 10, 11, 58].

4. The multi-source deep model

An overview of our framework in the testing stage is shown in Fig. 2. In this framework, an existing approach is used to generate candidate body locations with conservative thresholding. In the experiment, the existing approach is the off-the-shelf approach in [58]. A multi-source deep model is then applied to a candidate of all body locations in order to determine whether its body locations are correct. Simultaneously, the body locations of this candidate is estimated.

One direct approach with which to train a multi-source model is to train a deep model over the concatenated information sources as shown in Fig. 3(a). This approach is limited because information sources with different statistical properties are mixed in the first hidden layer. A better solution is to have their high-level representations constructed before they are mixed. Therefore, we use the architecture as shown in Fig. 3(b), in which each information source

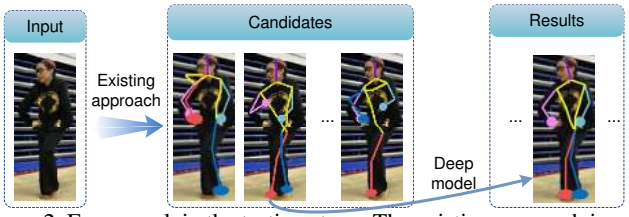

Figure 2. Framework in the testing stage. The existing approach is used to generate multiple candidate locations. A candidate is used as the input to a deep model to determine whether the candidate is correct and estimate body locations. Best viewed in color.

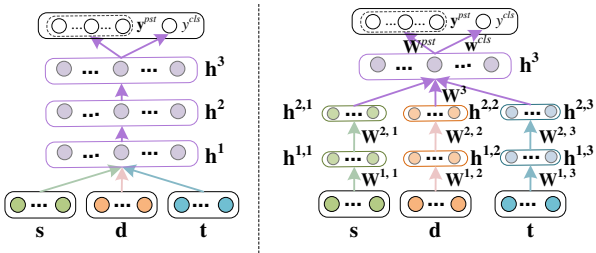

Figure 3. Direct use of deep model (a) and the deep architecture we propose (b) for part score $\mathbf{s}$, deformation $\mathbf{d}$ and mixture type $\mathbf{t}$. Best viewed in color.

is connected to two layers for constructing high level representation individually. High-level representations of different information sources are then fused using other two layers for pose estimation.

4.1. Inference

The mixture type information $\mathbf{t}$ in Fig. 3 is taken from the $\mathbf{t}$ in (3). The relative positions among parts, denoted by $\mathbf{d}$, comes from the deformation information $d(z_p - z_q)$ in (4). The appearance scores, denoted by $\mathbf{s}$, is obtained from the unary appearance term in (5). In our experiment, $\mathbf{s}$, $\mathbf{t}$, and $\mathbf{d}$ are obtained using the approach in [58]. At the inference stage, the model is as follows:

$$\mathbf{h}^{1,1} = a(\mathbf{s}^{\mathrm{T}}\mathbf{W}^{1,1} + \mathbf{b}^{1,1}), \tag{6}$$

$$\mathbf{h}^{1,2} = a(\mathbf{d}^{\mathrm{T}}\mathbf{W}^{1,2} + \mathbf{b}^{1,2}), \tag{7}$$

$$\mathbf{h}^{1,3} = a(\mathbf{t}^{\mathrm{T}}\mathbf{W}^{1,3} + \mathbf{b}^{1,3}), \tag{8}$$

$$\mathbf{h}^{2,u} = a(\mathbf{h}^{1,u\,\mathrm{T}}\mathbf{W}^{2,u} + \mathbf{b}^{2,u}), u = 1, 2, 3, \tag{9}$$

$$\mathbf{h}^{2} = [\mathbf{h}^{2,1\mathrm{T}}\ \mathbf{h}^{2,2\mathrm{T}}\ \mathbf{h}^{2,3\mathrm{T}}]^{\mathrm{T}}, \tag{10}$$

$$\mathbf{h}^{3} = a(\mathbf{h}^{2\mathrm{T}}\mathbf{W}^{3} + \mathbf{b}^{2}), \tag{11}$$

$$\tilde{y}^{cls} = \sigma(\mathbf{h}^{3\mathrm{T}}\mathbf{w}^{cls} + b^{cls}), \tag{12}$$

$$\tilde{\mathbf{y}}^{pst} = \mathbf{h}^{3\mathrm{T}}\mathbf{W}^{pst} + \mathbf{b}^{pst}. \tag{13}$$

- $\sigma(x) = (1 + \exp(-x))^{-1}$ is the sigmoid function.
- $a(*)$ is the point-wise non-linear activation function, for which sigmoid function can be used.
- $\mathbf{W}^*$ and $\mathbf{w}^{cls}$ connect nodes between adjacent layers.
- $\mathbf{b}^*$ and b^{cls} are biases.

- $\mathbf{h}^*$ are hidden nodes in different layers used for extracting non-linear representations from $\mathbf{s}$, $\mathbf{d}$, and $\mathbf{t}$.
- $\tilde{y}^{cls}$ is the estimated label indicating whether the candidate of body locations is correct. For pose estimatio of single human, the candidate with the largest $\tilde{y}^{cls}$ is used as the final output in our experiments.
- $\tilde{\mathbf{y}}^{pst}$ contains the estimated part locations.

Through the first two separate layers in Eq. (6)-(9), each information source has its individual representation constructed. Then all high-order representations are combined by two layers in Eq. (11)-(13).

4.2. Training method

Denote the parameter set for the model in Eq. (6)-(13) by λ, $\lambda = \{\mathbf{W}^*, \mathbf{w}^{cls}, \mathbf{b}^*, b^{cls}\}$. The objective function $J(\lambda)$ for backpropagating error derivatives is as follows:

$$\begin{aligned} J(\lambda) =& \sum_n \Big(J_1(y_n^{cls}, \tilde{y}_n^{cls}) + y_n^{cls} J_2(\mathbf{y}_n^{pst}, \tilde{\mathbf{y}}_n^{pst}) \Big) \\ &+ J_3(\mathbf{W}^*, \mathbf{w}^{cls}), \qquad (14) \\ J_1(y_n^{cls}, \tilde{y}_n^{cls}) =& - y_n^{cls} \log(\tilde{y}_n^{cls}) - (1 - y_n^{cls}) \log(1 - \tilde{y}_n^{cls}), \\ J_2(\mathbf{y}_n^{pst}, \tilde{\mathbf{y}}_n^{pst}) =& ||\mathbf{y}_n^{pst} - \tilde{\mathbf{y}}_n^{pst}||^2, \\ J_3(\mathbf{W}^*, \mathbf{w}^{cls}) =& \sum_{i,j} |w_{i,j}^*| + \sum_i |w_i^{cls}|, \qquad (15) \end{aligned}$$

where $*_n$ denotes the nth sample, $n = 1, 2, \ldots N$.

- $\tilde{y}_n^{cls}$ and $\tilde{\mathbf{y}}_n^{pst}$ are computed using Eq. (6)-(13).
- $y_n^{cls} \in \{0, 1\}$ is the ground truth classification label indicating whether the current body location estimation is correct or not. Positive training samples have their part templates placed around annotated body locations. As in [58], negative training samples have their part templates placed on images without human. Therefore, y^{cls} can be used for human detection by considering it as an indicator on whether the rectangle covering body locations contains a human.
- $J_1(y_n^{cls}, \tilde{y}_n^{cls})$ is the cross-entropy error on classification.
- $\mathbf{y}_n^{pst}$ contains the ground truth body locations.
- $J_2(\mathbf{y}_n^{pst}, \tilde{\mathbf{y}}_n^{pst})$ is the sum of square error on body location estimation. Since negative background samples do not have ground truth body location, the y_n^{cls} is multiplied by J_2 in (14) to ensure that only positive samples are used to learn location estimation.
- $J_3(\mathbf{W}^*, \mathbf{w}^{cls})$ is the L_1 norm regularization term. $w_{i,j}^*$ is the (i, j)th element in $\mathbf{W}^*$ and w_i^{cls} is the ith and element in $\mathbf{w}^{cls}$. The information sources and hidden nodes may have different purpose. For example, a node in $\mathbf{h}^3$ may use the information source mixture type. Hence, $J_3(\mathbf{W}^*, \mathbf{w}^{cls})$ is used to encourage sparsity in the weights.

Body part location estimation and human detection are both learned through shared representation in this model. They are jointly learned because they are dependent tasks.

4.3. Analysis

The mixture type $\mathbf{t}$ is used as an example for analysis. In the layer-wise pre-training stage [18], $\mathbf{t}$ and hidden vector $\mathbf{h}^{1,3}$ are considered as a restricted Boltzmann machine with the following distribution:

$$p(\mathbf{t}, \mathbf{h}^{1,3}) \propto \exp(\mathbf{t}^{\mathrm{T}} \mathbf{W}^{1,3} \mathbf{h}^{1,3} + \mathbf{b}^{1,3^{\mathrm{T}}} \mathbf{h}^{1,3} + \mathbf{c}^{\mathrm{T}} \mathbf{t}). \quad (16)$$

Denote the jth column of $\mathbf{W}^{1,3}$ by $\mathbf{w}_{*,j}^{1,3}$. Denote the jth element of $\mathbf{b}^{1,3}$ by b_j. The marginal distribution $p(\mathbf{t})$ can be obtained as follows:

$$\begin{aligned} p(\mathbf{t}) =& \sum_{\mathbf{h}^{1,3}} p(\mathbf{t}, \mathbf{h}^{1,3}) \\ \propto& \sum_{\mathbf{h}^{1,3}} \exp(\mathbf{t}^{\mathrm{T}} \mathbf{W}^{1,3} \mathbf{h}^{1,3} + \mathbf{b}^{1,3^{\mathrm{T}}} \mathbf{h}^{1,3} + \mathbf{c}^{\mathrm{T}} \mathbf{t}) \qquad (17) \\ \propto& \exp(\mathbf{c}^{\mathrm{T}} \mathbf{t}) \prod_j \Big(1 + \exp(\mathbf{t}^{\mathrm{T}} \mathbf{w}_{*,j}^{1,3} + b_j) \Big) = \prod_j \phi_j(\mathbf{t}), \end{aligned}$$

where $\phi_j(\mathbf{t}) = 1 + \exp(\mathbf{t}^{\mathrm{T}} \mathbf{w}_{*,j}^{1,3} + b_j)$ and $\phi_j(\mathbf{t})$ is a fully connected graphical model because it cannot be factorized. $\phi_j(\mathbf{t})$ can be considered as a factor that explains $\mathbf{t}$ in factor graph [5, 24]. In pose estimation, $\phi_j(\mathbf{t})$ can be considered as a global pattern explaining the mixture type $\mathbf{t}$ for all parts. In both training and inference stages, every node in $\mathbf{h}^{1,3}$ is connected to the mixture types of all parts. Therefore, $\mathbf{h}^{1,3}$ nonlinearly extracts the global representation from $\mathbf{t}^{1,3}$. Similarly, the $\mathbf{h}^{2,3}$ extracts higher-level representation from $\mathbf{h}^{1,3}$. Therefore, the stack of hidden layers extracts global, high-level representation from the information source $\mathbf{t}$. The analysis to mixture type $\mathbf{t}$ is applicable to deformation and appearance score. As shown in Fig. 4, $\mathbf{h}^{2,3}$ captures the global articulation patterns of human body. One of the nodes in $\mathbf{h}^{2,3}$ has high response to people squat. Another node has high response to people standing upright. Yet another node concisely captures two clusters of pose patterns.

In our deep model, the first hidden layer has 200 hidden nodes, the second layer, i.e. $\mathbf{h}^2$ in Eq. (10) has 150 hidden nodes and the third layer, i.e. $\mathbf{h}^3$ in Eq. (11), has 100 hidden nodes. Since the dimensions of $\mathbf{s}$, $\mathbf{d}$, and $\mathbf{t}$ are small, training of the deep model is fast. Unlike loopy graphical models, the deep model is fast in the inference stage because it does not require loopy belief propagation or sampling. The extra testing time required by our deep model is less than 10 percent of the testing time required by the approach in [58].

5. Experimental results

The proposed approach is evaluated on three datasets: LSP [21], PARSE [44] and UIUC people [53]. The training procedure and training set are the same as [58]. Positive

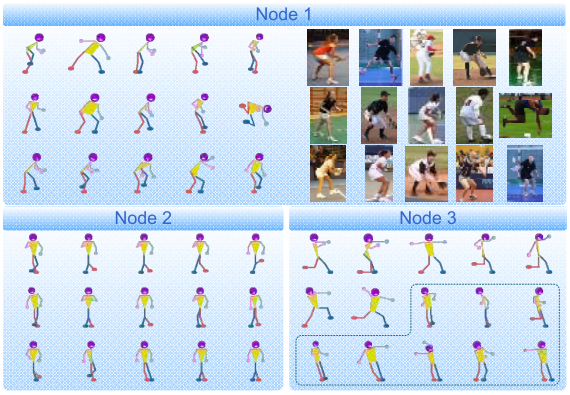

Figure 4. Visualization of mixture-type patterns extracted by hidden nodes in $\mathbf{h}^{2,3}$. We use the approach in [26] and visualize training samples with the largest responses on each hidden node. Samples with the highest responses are placed at the upper-left corner. Hidden node 1 has high response to people squat. Node 2 has high response to standing people. Node 3 has high response to two clusters of pose patterns. Best viewed in color.

training samples are constrained to have estimated part locations near the ground truth. Part of the training data is used for validation.

5.1. Evaluation criteria

In all experiments, we use the most popular criterion, which is the percentage of correctly localized parts (PCP) introduced in [14]. As stated in [42, 58], the PCP scoring metric has been implemented in different ways in different papers. These differences have two dimensions.

1. There are two ways to compute the final PCP score across the dataset. In the *single* way, only a single candidate (given by the maximum scoring candidate of an algorithm) for one image is used. The *match* way matches multiple candidates without penalizing false positives.
2. There are two definitions of a correct part localization. For the definition *both*, it requires both end points of a part (for example, end points wrist and elbow for the part lower arm) to be correct. For the definition *avg*, it requires only the average of the endpoints to be correct.

The paper in [54, 57, 52] used 'match+avg'. The paper in [40, 2, 42] used 'single+both', which is the strictest case and generally has lower PCP value. The paper in [10] provides results for 'match+both' and 'match+avg'. We follow [42, 40] and evaluate all approaches using the strictest 'single+both' criterion. This is used because of the following reasons:

1. For 'single' and 'match', as discussed in [58], the 'match' way gives unfair advantage to approaches that produce a large number of candidates because mismatched candidates (false positives) are not penalized.
2. For 'both' and 'avg', 'both' is better at describing the orientation of body parts and will facilitate the use of pose estimation for future applications. For example, in character animation, the rendering of a limb is possible only when both end points of the limb are correct.

We follow [11, 40] and use the observer-centric annotations for all approaches when we evaluate on the LSP dataset.

5.2. Overall experimental results

Table 1 shows the experimental results from the three datasets.

Pishchulin's approach in [40] used the LSP+PARSE training set when evaluated on the PARSE dataset and used the UIUC+LSP training set when evaluated on the UIUC dataset. To evaluate on the PARSE dataset, Pishchulin's approach [40] + [42] included LSP+PARSE and 2744 extra animated samples for training. Johnson's approach in [22] included 10,000 extra training samples when evaluated on the PARSE dataset. In all experiments, Andriluka's approach in [2], Yang and Ramanan's approach in [57, 58] and our approach are trained on the 1000 training images of the LSP dataset [21].

As shown in Table 1, our deep model obviously improves the pose estimation accuracy and outperforms all the state-of-the-art on these three datasets. Specifically, our approach is better in detecting legs, arms and head compared with existing approaches. The approach of Pishchulin *et al*. [42] is better than our approach in locating torso, possibly because the torso region is included in many poslets, which helps to increase the accuracy of their approach in locating torso.

Our approach is complementary to existing approaches because the information sources provided by these approaches can be used by our model to improve their results. Currently, our model uses the approach in [58] to obtain information sources. Compared with the approach in [58], our approach improves the pose estimation accuracy by 5.8% (62.8% vs. 68.6% PCP), 7.4% (63.6% vs. 71.0% PCP) and 8.6% (57.0% vs. 65.6% PCP) respectively on the LSP, PARSE and UIUC datasets. Fig. 5 shows the comparison between our approach (left) and the approach in [58] (right).

5.3. Results on different designs of deep models

In this section, we evaluate different designs of deep models. Yang and Ramanan's approach in [58] is used as the baseline because this approach is used by our model for obtaining information sources. To be concise, we only refer to the PCP results on the LSP dataset.

Depth of model is investigated in Table 2. The approach in [58] uses linear-SVM for combining information sources. We also trained a Kernel-SVM with RBF kernel for learning a non-linear model using the off-the-shelf tool Libsvm. The difference in PCP between Linear SVM and kernel-SVM is within 2% (62.8% vs. 64.2% on LSP). Bengio [3]

Table 1. Pose estimation results (PCP) on LSP [21], UIUC people [53] and PARSE [44].

Method	Torso	U.leg	L.leg	U.arm	L.arm	head	Total
LSP							
Andriluka *et al*. [2]	80.9	67.1	60.7	46.5	26.4	74.9	55.7
Yang&Ramanan [57]	81.0	69.5	65.9	53.5	35.8	76.8	60.7
Yang&Ramanan [58]	82.9	70.3	67.0	56.0	39.8	79.3	62.8
Pishchulin *et al*. [40]	**87.5**	75.7	68.0	54.2	33.9	78.1	62.9
Eichner& Ferrari [11]	86.2	74.3	69.3	56.5	37.4	80.1	64.3
Ours	85.8	**76.5**	**72.2**	**63.3**	**46.6**	**83.1**	**68.6**
PARSE							
Andriluka *et al*. [2]	86.3	66.3	60.0	54.6	35.6	72.7	59.2
Yang&Ramanan [57]	83.4	68.8	60.7	59.8	40.7	83.4	62.7
Yang&Ramanan [58]	82.9	68.8	60.5	63.4	42.4	82.4	63.6
Pishchulin *et al*. [42]	88.8	77.3	67.1	53.7	36.1	73.7	63.1
Pishchulin *et al*. [40]	**92.2**	74.6	63.7	54.9	39.8	70.7	62.9
[40]+[42]	90.7	80.0	70.0	59.3	37.1	77.6	66.1
Johnson& Everingham [22]	87.6	74.7	67.1	67.3	45.8	76.8	67.4
Ours	89.3	**78.0**	**72.0**	**67.8**	**47.8**	**89.3**	**71.0**
UIUC People							
Andriluka *et al*. [2]	88.3	64.0	50.6	42.3	21.3	81.8	52.6
Yang&Ramanan [57]	78.1	60.9	53.2	41.3	32.2	76.1	53.0
Yang&Ramanan [58]	81.8	65.0	55.1	46.8	37.7	79.8	57.0
Pishchulin *et al*. [40]	**91.5**	66.8	54.7	38.3	23.9	85.0	54.4
Wang *et al*. [56]	86.8	56.3	50.2	30.8	20.3	68.8	47.0
Ours	89.1	**72.9**	**62.4**	**56.3**	**47.6**	**89.1**	**65.6**

Table 2. Results (PCP) on investigating model depth.

Method	Torso	U.leg	L.leg	U.arm	L.arm	Head	Total
LSP							
[58]	82.9	70.3	67.0	56	39.8	79.3	62.8
Kernel SVM	81.9	72.2	67.6	58.8	42.8	77.5	64.2
1 hidden layer	84.9	73.9	69.5	57.5	42.9	50.7	62.3
2 hidden layers	85.0	74.6	70.7	61.2	45.2	82.2	67.1
Ours	85.8	76.5	72.2	63.3	46.6	83.1	68.6
PARSE							
[58]	82.9	68.8	60.5	63.4	42.4	82.4	63.6
Kernel SVM	81.0	67.8	61.2	63.2	44.1	78.0	63.2
1 hidden layer	84.4	71.2	63.2	62.4	44.4	70.2	63.7
2 hidden layers	85.9	74.4	68.3	64.6	46.3	85.4	67.9
Ours	89.3	78.0	72.0	67.8	47.8	89.3	71.0
UIUC							
[58]	81.8	65.0	55.1	46.8	37.7	79.8	57.0
Kernel SVM	82.2	65.0	54.9	50.2	43.1	80.6	58.9
1 hidden layer	83.0	65.6	55.9	50.6	42.3	79.8	59.2
2 hidden layers	84.2	68.4	59.3	53.0	45.3	83.4	62.0
Ours	89.1	72.9	62.3	56.3	47.6	89.1	65.6

proved that linear-SVM and kernel-SVM are shallow models. With the deep model, our approach performs better. As the number of hidden layers increases from *1 hidden layer* to *2 hidden layers*, the estimation accuracy increases from 62.3% to 67.1%. With PCP 68.6%, our final model in Fig. 3(b) uses three hidden layers and is better than SVM and deep models with fewer layers.

Table 3. Results (PCP) on investigating deep model structures.

Method	Torso	U.leg	L.leg	U.arm	L.arm	Head	Total
LSP							
DBN in Fig. 3(a)	82.9	73.2	69.5	59.8	43.8	79.2	65.5
Ours	85.8	76.5	72.2	63.3	46.6	83.1	68.6
PARSE							
DBN in Fig. 3(a)	82.0	70.0	64.6	62.9	46.3	80.5	65.0
Ours	89.3	78.0	72.0	67.8	47.8	89.3	71.0
UIUC							
DBN in Fig. 3(a)	87.4	68.4	58.3	52.2	44.3	84.6	61.8
Ours	89.1	72.9	62.3	56.3	47.6	89.1	65.6

Deep model structure design is investigated in Table 3. The DBN in Fig. 3(a) trains a three-layer deep model over the concatenated informations with three hidden layers. The model in 3(b) learns high-order representations individually. The model in 3(b) with PCP 68.6% is better in constructing the high-order representations and therefore has higher estimation accuracy compared with the DBN in Fig. 3(a) with PCP 65.5%.

Classification label and location learning is investigated in Table 4. There are two sets of labels to be estimated in our deep model: classification label y^{cls} and part positions $\mathbf{y}^{pst}$. In the experiments, we evaluate different ways of estimating these labels. The *Only* y^{cls} in Table 4, with PCP 63.7%, only estimates class label, with part location directly obtained by the approach in [58]. The *Only* $\mathbf{y}^{pst}$, with PCP 64.1%, only refines the part location, with class label directly obtained by the approach in [58]. *Separate* y^{cls}+$\mathbf{y}^{pst}$, with PCP 64.7%, uses two deep models for estimating y^{cls} and $\mathbf{y}^{pst}$ separately. It can be seen that both y^{cls} and $\mathbf{y}^{pst}$ are helpful for improving accuracy. Our model uses the single deep model to jointly learn both y^{cls} and $\mathbf{y}^{pst}$ (PCP 68.6%) and performs better than using two models to learn them separately (PCP 64.7%) because body location and the correctness of candidate body location are dependent.

Analysis. Our model extracts high-order representations of appearance, deformation and mixture types and better models their dependence at the top layer. For example, if the mixture types are upright upper- and lower-arms, the weighted combination of the locations of wrist and shoulder is a good estimation on the location of elbow. If the mixture types change, such estimation should change correspondingly. Such complex dependence cannot be modeled linearly and deep model is a better solution. When different information sources are extracted separately with the first several layers, the connections across sources are removed and the number of parameters is reduced. It helps to regularize optimization when training samples are limited. Existing methods only use y^{cls} for supervision, while we use both y^{pts} and y^{cls}. As shown in Fig. 4, refining y^{pts} does help to rectify incorrect part locations based on the high order prior model of body pose. Jointly learning

Table 4. PCP results on classification label and location learning.

Method	Torso	U.leg	L.leg	U.arm	L.arm	Head	Total
			LSP				
[58]	82.9	70.3	67.0	56.0	39.8	79.3	62.8
Only y^{cls}	82.0	71.5	68.0	57.6	42.0	77.2	63.7
Only $\mathbf{y}^{pst}$	80.4	72.0	68.0	59.2	42.8	76.8	64.1
Separate y^{cls}+$\mathbf{y}^{pst}$	81.1	72.8	69.0	59.5	43.0	77.7	64.7
Ours	85.8	76.5	72.2	63.3	46.6	83.1	68.6
			PARSE				
[58]	82.9	68.8	60.5	63.4	42.4	82.4	63.6
Only y^{cls}	81.0	69.8	66.1	60.5	43.9	76.1	63.8
Only $\mathbf{y}^{pst}$	80.5	71.2	65.4	62.2	44.4	79.5	64.6
Separate y^{cls}+$\mathbf{y}^{pst}$	83.4	73.7	67.6	64.4	47.1	82.0	67.1
Ours	89.3	78.0	72.0	67.8	47.8	89.3	71.0
			UIUC				
[58]	81.8	65.0	55.1	46.8	37.7	79.8	57.0
Only y^{cls}	85.4	68.8	59.3	49.2	40.5	83.4	60.4
Only $\mathbf{y}^{pst}$	82.6	66.6	58.3	52.2	44.7	81.8	60.8
Separate y^{cls}+$\mathbf{y}^{pst}$	87.9	69.6	60.3	53.0	44.3	85.4	62.8
Ours	89.1	72.9	62.3	56.3	47.6	89.1	65.6

y^{pst} and y^{cls} helps to find their shared representation under a multi-task learning framework, for which deep model is an ideal choice.

6. Conclusion

This paper has proposed a multi-source deep model for pose estimation. It non-linearly integrates three information sources: appearance score, deformation and appearance mixture type. These information sources are used for describing different aspects of the single modality data, which is the image data in our pose estimation approach. Extensive experimental comparisons on three public benchmark datasets show that the proposed model obviously improves the pose estimation accuracy and outperforms the state of the art. Since this model is a post-processing of information sources, it is very flexible in terms of integrating with existing approaches that use different information sources, features, or articulation models. Learning deep model from pixels for pose estimation and analyzing the influence of training data number will be the future work.

7. Acknowledgement

This work is supported by the General Research Fund sponsored by the Research Grants Council of Hong Kong (Project No. CUHK 417110, CUHK 417011, CUHK 429412), National Natural Science Foundation of China (91320101), Shenzhen Basic Research Program (JC201005270350A, JCYJ20120903092050890, JCYJ20120617114614438), and Guangdong Innovative Research Team Program (No.201001D0104648280).

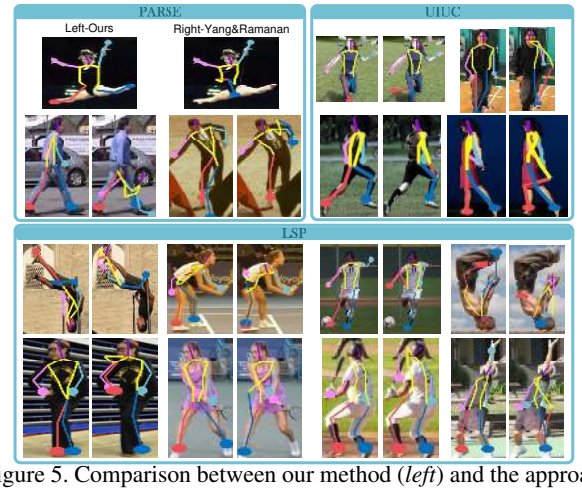

Figure 5. Comparison between our method (*left*) and the approach in [58] (*right*) on the LSP, PARSE and UIUC dataset. Our approach obtains more reasonable articulation patterns and is better in solving the double counting problem. Best viewed in color.

References

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014.

[2] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: people detection and articulated pose estimation. In *CVPR*, 2009.

[3] Y. Bengio. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–127, 2009.

[4] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Trans. PAMI*, 35(8):1798–1828, 2013.

[5] C. M. Bishop and N. M. Nasrabadi. *Pattern recognition and machine learning*. springer, 2006.

[6] L. Bourdev and J. Malik. Poselets: body part detectors trained using 3D human pose annotations. In *ICCV*, 2009.

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: a large-scale hierarchical image database. In *CVPR*, 2009.

[9] C. Desai and D. Ramanan. Detecting actions, poses, and objects with relational phraselets. In *ECCV*, 2012.

[10] K. Duan, D. Batra, and D. J. Crandall. A multi-layer composite model for human pose estimation. In *BMVC*, 2012.

[11] M. Eichner and V. Ferrari. Appearance sharing for collective human pose estimation. In *ACCV*, 2012.

[12] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *IEEE Trans. PAMI*, 30:1915–1929, 2013.

[13] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61:55–79, 2005.

[14] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *CVPR*, 2008.

[15] G. Gkioxari, P. Arbeláez, L. Bourdev, and J. Malik. Articulated pose estimation using discriminative armlet classifiers. In *CVPR*, 2013.

[16] I. Goodfellow, H. Lee, Q. V. Le, A. Saxe, and A. Y. Ng. Measuring invariances in deep networks. In *NIPS*, 2009.

[17] M. Guillaumin, J. Verbeek, and C. Schmid. Multimodal semi-supervised learning for image classification. In *CVPR*, 2010.

[18] G. E. Hinton, S. Osindero, and Y. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554, 2006.

[19] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504 – 507, July 2006.

[20] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun. What is the best multi-stage architecture for object recognition? In *CVPR*, 2009.

[21] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010.

[22] S. Johnson and M. Everingham. Learning effective human pose estimation from inaccurate annotation. In *CVPR*, 2011.

[23] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[24] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Trans. Inf. Theory*, 47(2):498–519, 2001.

[25] H. Larochelle, Y. Bengio, J. Louradour, and P. Lamblin. Exploring strategies for training deep neural networks. *J. Machine Learning Research*, 10:1–40, 2009.

[26] Q. V. Le, M. Ranzato, R. Monga, M. Devin, K. Chen, G. S. Corrado, J. Dean, and A. Y. Ng. Building high-level features using large scale unsupervised learning. In *ICML*, 2012.

[27] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[28] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014.

[29] P. Luo, Y. Tian, X. Wang, and X. Tang. Switchable deep network for pedestrian detection. In *CVPR*, 2014.

[30] P. Luo, X. Wang, and X. Tang. Hierarchical face parsing via deep learning. In *CVPR*, 2012.

[31] P. Luo, X. Wang, and X. Tang. A deep sum-product architecture for robust facial attributes analysis. In *ICCV*, 2013.

[32] P. Luo, X. Wang, and X. Tang. Pedestrian parsing via deep decompositional neural network. In *ICCV*, 2013.

[33] G. Mori and J. Malik. Estimating human body configurations using shape context matching. In *ECCV*, 2002.

[34] G. Mori and J. Malik. Recovering 3D human body configurations using shape contexts. *IEEE Trans. PAMI*, 28(7):1052–1062, 2006.

[35] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng. Multimodal deep learning. In *ICML*, 2011.

[36] M. Norouzi, M. Ranjbar, and G. Mori. Stacks of convolutional restricted boltzmann machines for shift-invariant feature learning. In *CVPR*, 2009.

[37] W. Ouyang and X. Wang. A discriminative deep model for pedestrian detection with occlusion handling. In *CVPR*, 2012.

[38] W. Ouyang and X. Wang. Joint deep learning for pedestrian detection. In *ICCV*, 2013.

[39] W. Ouyang, X. Zeng, and X. Wang. Modeling mutual visibility relationship in pedestrian detection. In *CVPR*, 2013.

[40] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Poselet conditioned pictorial structures. In *CVPR*, 2013.

[41] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Strong appearance and expressive spatial models for human pose estimation. In *ICCV*, December 2013.

[42] L. Pishchulin, A. Jain, M. Andriluka, T. Thormahlen, and B. Schiele. Articulated people detection and pose estimation: Reshaping the future. In *CVPR*, 2012.

[43] H. Poon and P. Domingos. Sum-product networks: A new deep architecture. In *UAI*, 2011.

[44] D. Ramanan. Learning to parse images of articulated bodies. In *NIPS*, 2007.

[45] M. Ranzato, F. J. Huang, Y.-L. Boureau, and Y. Lecun. Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *CVPR*, 2007.

[46] B. Sapp, A. Toshev, and B. Taskar. Cascaded models for articulated pose estimation. In *ECCV*, 2010.

[47] N. Srivastava and R. Salakhutdinov. Multimodal learning with deep boltzmann machines. In *NIPS*, 2012.

[48] M. Sun and S. Savarese. Articulated part-based model for joint object detection and pose estimation. In *ICCV*, 2011.

[49] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *CVPR*, 2013.

[50] Y. Sun, X. Wang, and X. Tang. Hybrid deep learning for computing face similarities. In *ICCV*, 2013.

[51] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *CVPR*, 2014.

[52] Y. Tian, C. L. Zitnick, and S. G. Narasimhan. Exploring the spatial hierarchy of mixture models for human pose estimation. In *ECCV*, 2012.

[53] D. Tran and D. Forsyth. Improved human parsing with a full relational model. In *ECCV*, 2010.

[54] F. Wang and Y. Li. Beyond physical connections: Tree models in human pose estimation. In *CVPR*, 2013.

[55] Y. Wang and G. Mori. Multiple tree models for occlusion and spatial constraints in human pose estimation. In *ECCV*, 2008.

[56] Y. Wang, D. Tran, and Z. Liao. Learning hierarchical poselets for human parsing. In *CVPR*, 2011.

[57] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, 2011.

[58] Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures-of-parts. *IEEE Trans. PAMI*, To appear.

[59] M. D. Zeiler, G. W. Taylor, and R. Fergus. Adaptive deconvolutional networks for mid and high level feature learning. In *ICCV*, 2011.

[60] X. Zeng, W. Ouyang, and X. Wang. Multi-stage contextual deep learning for pedestrian detection. In *ICCV*, 2013.

[61] Z. Zhu, P. Luo, X. Wang, and X. Tang. Deep learning identity preserving face space. In *ICCV*, 2013.