Computer Vision and Image Understanding 169 (2018) 62–74

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Ensemble convolutional neural networks for pose estimation

Yuki Kawana[b], Norimichi Ukita[*,1,a], Jia-Bin Huang[c], Ming-Hsuan Yang[d]

[a] Toyota Technological Institute, 2-12-1 Hisakata, Tempaku, Nagoya 468-8511 Japan
[b] Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara 630-0192, Japan
[c] Virginia Tech, 1185 Perry St, Room 430, Blacksburg, VA 24060, USA
[d] University of California, Merced, 5200 N. Lake Road, Merced, CA 95343, USA

ARTICLE INFO

Keywords:
Human pose estimation
Ensemble models
Pose modality

ABSTRACT

Human pose estimation is a challenging task due to significant appearance variations. An ensemble of models, each of which is optimized for a limited variety of poses, is capable of modeling a large variety of human body configurations.However, ensembling models is not a straightforward task due to the complex interdependence among noisy and ambiguous pose estimation predictions acquired by each model.We propose to capture this complex interdependence using a convolutional neural network. Our network achieves this interdependence representation using a combination of deep convolution and deconvolution layers for robust and accurate pose estimation. We evaluate the proposed ensemble model on publicly available datasets and show that our model compares favorably against baseline models and state-of-the-art methods.

1. Introduction

Human pose estimation is challenging due to the wide variety of appearances that can result from pose variations. One way to alleviate the complexity is to cluster a training dataset so that a set of expert models can be learned. Reducing the variation within each subset facilitates learning the expert model to accurately estimate the joint locations under a particular pose configuration. By combining each of the expert models from one of the heterogeneous variations (i.e., different types of pose variations), the ensemble of the expert models can capture complicated appearance variation. We call these heterogeneous expert models pose-modality (PM) models. For example, given an input image, the configurations of different body parts may be correctly localized using different PM models, e.g., PM models 1 and *N* correctly localize the right and left lower arms, respectively (see (b) Testing stage of Fig. 1).

In using the ensemble of PM models, it remains unclear how to determine a final estimation from diverse responses of PM models. Existing approaches combine the responses either by simply selecting the most confident response (Moghimi et al., 2016) or averaging over all the responses (Agostinelli et al., 2013; Ciresan et al., 2012; Krizhevsky et al., 2012). Such heuristics, however, do not capture the interdependency among the responses of PM models.

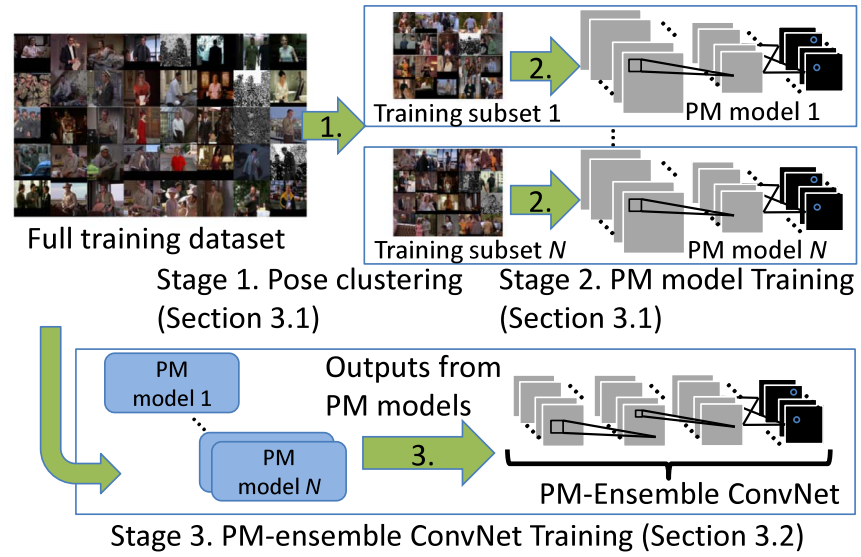
In this paper, we present a PM-ensemble (PME) model to infer body configurations by modeling the interdependence among the responses of PM models. As shown in Fig. 1(a), the model training process consists of three stages. At stage 1, the training samples are partitioned into subsets based on their similarity in a pose space. At stage 2, each PM model is trained using training samples from each cluster. At stage 3, we learn the PME model to incorporate all the responses to make the final estimation. Fig. 1(b) shows an example of the inference procedures. Given an input image, we use the learned PM models to localize body joint positions independently. As the PM models are trained with disjoint sets of training samples, the resultant joint heatmaps are typically diverse. Our PME model can selectively combine the correct pose predictions and merges them to the final estimation.

We note while it may be feasible to train a large network in an end-to-end fashion without pre-clustering, in practice it is rather challenging as it requires a large amount of manually annotated images, highly computational load and memory. The separated PM models can be trained in an efficient distributed manner which is better in terms of computational cost and memory capacity. In addition, it facilitates analyzing the network modules for pose estimation.

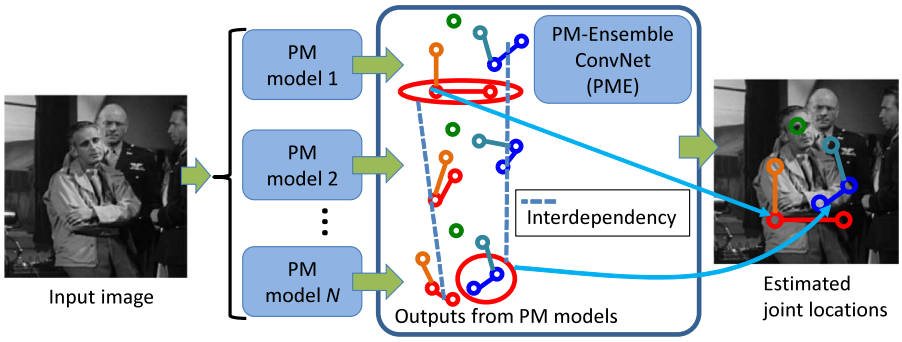
The main contributions of this work are as follows. First, we propose the PME model for human pose estimation that is capable of merging diverse responses from heterogeneous PM models (Section 3). We design the PME model so that (i) it can better model the interdependence among the diverse responses than previous clustering-based methods (Bourdev and Malik, 2009; Johnson and Everingham, 2010, 2011; Pfister et al., 2015; Sapp and Taskar, 2013) and (ii) it provides high-

* Corresponding author.
E-mail address: ukita@ieee.org (N. Ukita).
[1] He worked at Nara Institute of Science and Technology formerly.

https://doi.org/10.1016/j.cviu.2017.12.005
Received 24 July 2017; Received in revised form 9 November 2017; Accepted 26 December 2017
Available online 05 January 2018



(a) Training stage



(b) Testing stage

Fig. 1. Overview of the proposed ensemble model. (a) Training stage. At stages 1, 2, and 3, (1) dataset clustering based on pose similarity, (2) PM model training using each pose cluster, and (3) PM-ensemble ConvNet training for integrating the responses of the PM models are performed, respectively. (b) Testing stage. Given an input image, we first use the trained PM models to estimate the joint heatmaps. The PM-ensemble model then integrates the responses from all models to localize the joint positions.

precision joint localization without reducing its spatial resolution (Section 3.2). Second, we propose PM-dependent clustering of training images for individual PM modeling (Section 3.1). The clustering strategy is essential for making PM models heterogeneous, while other CNNs for ensembling (Agostinelli et al., 2013; Ciresan et al., 2012; Pfister et al., 2015; Sun et al., 2013) apply the same set of training data (i.e., with no clustering) to expert model(s). Third, we propose a two-stage training scheme to fine-tune each PM model for capturing a limited pose variety while avoiding overfitting (Section 3.1).

The novelty of this work lies also in a practical and efficient design for pose clusters using the ensemble network. While some components are known, it requires the proposed algorithmic design to integrate the modules. While it is feasible to train a large network without pre-clustering in the end-to-end fashion, in practice it is very challenging as it requires a large set of manually annotated images, considerable memory and computational costs and getting stuck in bad local minima. More importantly, we show the advantages of the proposed models over the end-to-end approaches.

- The proposed method can represent the interdependency among complex poses by a huge network consisting of multiple PM models and the ensemble model. Such a huge network cannot be implemented in the end-to-end fashion on limited memory on the GPU. But, the separated PM models and the ensemble model can be trained in an efficient, distributed manner in terms of computational cost and memory capacity. In fact, memory capacity on the GPU at our disposal (NVIDIA Titan X 12 GB) is fully occupied for training only the ensemble model for 10 PM models.
- Through analyzing the intermediate heatmaps of the PM models (e.g., how are they different? how is each one useful? how are they merged?), we can better understand each model and improve the overall performance.
- We also validate whether the performance can be improved by updating the weights of PM models with the ensemble model. Our results show that the performance remains the same or decreased in some cases. We attribute this due to the difficulty in end-to-end training by the current training scheme.

2. Related work

Pictorial structure models (PSMs). PSMs have been applied to human pose estimation (Andriluka et al., 2009; Bourdev et al., 2010; Dantone et al., 2013; Eichner et al., 2009, 2012; Ferrari et al., 2009; Puwein et al., 2014; Ramakrishna et al., 2014; Sapp et al., 2011) because of their ability for efficient and global optimization. Many extensions have been proposed to improve PSMs, e.g., discriminative training (Felzenszwalb et al., 2010; Yang and Ramanan, 2013), graphical models with loops (Bergtholdt et al., 2010; Tran and Forsyth, 2010), coarse-to-fine and hierarchical modeling (Sapp et al., 2010; Sun and Savarese, 2011; Tian et al., 2012), appearance learning between parts

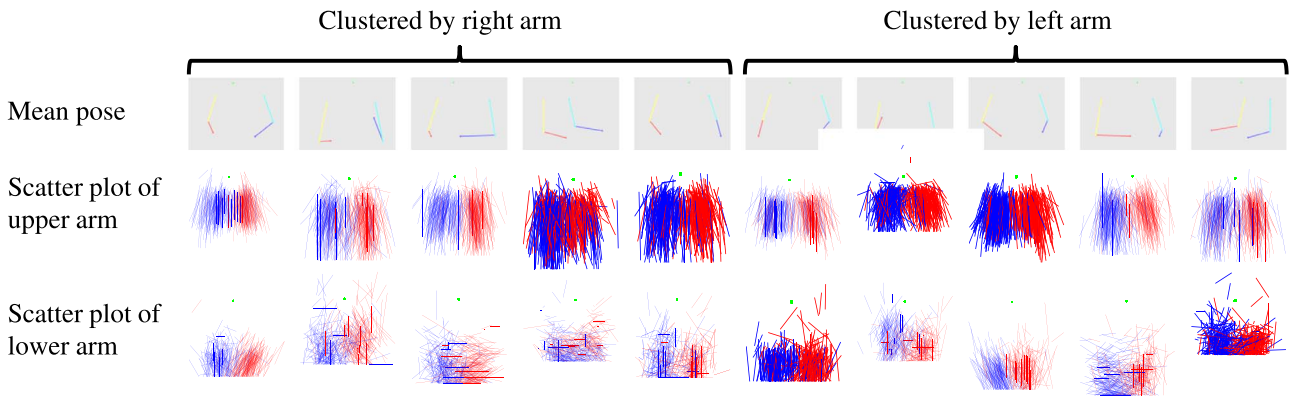


Fig. 2. Visualization of the mean poses and the scatter plots of the arm poses from each cluster. Each column represents each cluster. The first five columns show clusters obtained by clustering pose based on right arm positions. The rest five columns are for the left arm. While the configurations of the upper arm are similar within each cluster, the lower arm configurations are significantly different. The PM models trained using the training samples in the respective cluster may thus have diverse estimations for a given image.

(Chen and Yuille, 2014; Ukita, 2012), and a conditional random field with a dense graph representation in Kiefel and Gehler (2014). While global optimality of the PSM is attractive, its ability to represent complex relations among parts is limited compared to deep neural networks.

ConvNet-based pose estimation. ConvNets have recently been applied to pose estimation. Chen and Yuille (2014) use a ConvNet to learn the appearance of parts within a PSM framework. In addition to appearance modeling, a ConvNet can also model the distribution of joint locations. For example, a ConvNet can directly estimate the joint locations (Toshev and Szegedy, 2014) or estimate the pixel-wise likelihood of each joint location as a heatmap (Tompson et al., 2015; 2014). Recent approaches explore sequential structured estimation to iteratively improve the joint locations (Carreira et al., 2016; Ramakrishna et al., 2014; Singh et al., 2015; Wei et al., 2016). Pfister et al. (2015) extend the ConvNet for pose estimation in still images to video by combining warped responses across multiple frames using optical flow. Our approach builds upon such state-of-the-art models. Specifically, we use Pfister et al. (2015) and Wei et al. (2016) as our PM model for upper-body and full-body pose models, respectively. In contrast to existing work that focuses on improving the performance of one single model, our goal is to develop an ensemble method that can merge responses from multiple PM models.

Multi-modality of human poses. In Ouyang et al. (2014), different types of visual cues such as part appearance and geometric deformation between parts are integrated into a neural network for human pose estimation. Non-maxima suppression is extended in order to integrate multiple pose hypotheses in Burgos-Artizzu et al. (2013). However, only one single model is trained using all the training data. On the other hand, Johnson and Everingham (2010, 2011) and Sapp and Taskar (2013) train multiple models with clustered training data. Similar to Johnson and Everingham (2010, 2011); Sapp and Taskar (2013), we also cluster the training dataset for learning PM models. The main difference lies in that our approach integrates the outputs of all PM models based on the interdependency among the models, while Johnson and Everingham (2010, 2011); Sapp and Taskar (2013) select only the model with the highest confidence. Poselets (Bourdev and Malik, 2009) also adopt pose clustering to localize multiple target body configurations. While our PM models are trained in a similar way, our primary focus is on how to effectively combine the model responses rather than localization of each body part using one single model.

Ensemble of neural networks. Model ensembling is widely used in machine learning and recently in the context of ConvNets. Ciresan et al. (2012) apply ConvNets multiple times and average over their estimations for image classification. Agostinelli et al. (2013) address image denoising by weighted-average over the estimations from multiple ConvNets, each of which is trained to remove a particular type of images noise (e.g., Gaussian, speckle). In the context of face verification, Sun et al. (2013) show that using additional neural network layers to combine multiple ConvNets can further improve the recognition accuracy.

The ensemble of neural networks has also been used for human pose estimation. Pfister et al. (2015) merge the estimated body configurations from adjacent video frames using a convolutional layer. In each frame, the body configuration is estimated by the same model. Here, using a single convolution layer as an ensemble model may be sufficient for merging *similar* pose estimations (as the pose configurations in the adjacent video frames are estimated by the same model). However, our PM models may generate *diverse* pose estimations as the PM models are trained with disjoint sets of training samples. The simple ensemble method in Pfister et al. (2015) may fail to capture this diversity.

3. Ensemble model for human pose estimation

3.1. PM models

The main idea in PM modeling is to estimate a particular body configuration with high accuracy (even at the expense of false localization for other types of body configurations). We show in Fig. 2 several examples of such pose configurations clustered based on the arms. Each PM model is fine-tuned over the respective clustered training samples.

With each trained PM model, we obtain the heatmap of each joint location given an input image. As demonstrated in Pfister et al. (2015), the multi-modality (i.e., high confidence at multiple spatial locations) in the heatmap allows us to better capture the ambiguity of the estimated joint locations compared to directly regressing the 2D joint coordinates. In our PME model, therefore, heatmaps are fed into the PME[2] because the interdependence among joint locations estimated by multiple PM models is complex and ambiguous.

For effective PM modeling, we discuss two important aspects: (1) data clustering strategy and (2) model fine-tuning.

Clustering of training data. Clustering pose samples using a full-body configuration (as done in (Johnson and Everingham, 2010, 2011)) produces a larger variation in each pose cluster. The large variation in the cluster prevents the PM model from learning particular body configurations. To facilitate the PM model learning, our strategy for pose clustering is to use a *partial* body region as a pose feature. In our implementation, the pose feature is computed from the configuration of arms, i.e., shoulder, elbow, and wrist. This is because the pose

[2] While pose estimation models used in our experiments (i.e., (Pfister et al., 2015) and (Wei et al., 2016)) produce the heatmap of each joint location, many other models directly infer the joint (x, y) locations. A heatmap can be produced from the joint location with the Gaussian distribution centered on the location.

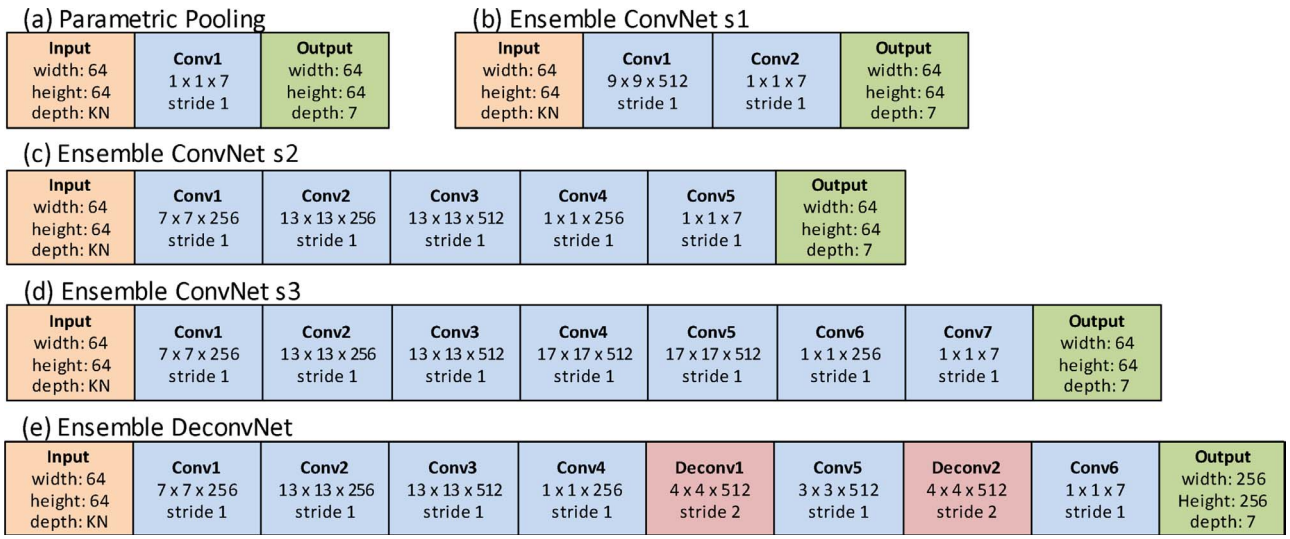


Fig. 3. (a) The parametric pooling by Pfister et al. (2015) for upper-body pose estimation. (b) PME s1, (c) PME s2, and (d) PME s3. (e) PME ConvNet + DeconvNet (PME-D). K denotes the number of PM models. Here, "conv" and "deconv" in each box represents convolutional and deconvolutional layers, respectively. In the box of the convolutional and deconvolutional layers, the first two numbers in the second row of the box represents height and width of the filter size and the third number the size of the output feature map. The last row shows the stride of the filter.

configuration of the arms is larger than that of other body parts and so appropriate for making PM models more heterogeneous. To obtain tighter clusters, we apply pose clustering to all training data independently using each of the left and right arms. Given K clusters in each arm, we obtain $2K$ clusters in total. Our experiments suggest that pose clustering using each of the left and right arms achieves improved accuracy than pose clustering using the both arms. We attribute this to the smaller variation of human poses in a partial body region.

We use the K-means clustering algorithm with a feature vector $\mathbf{f}_p$ of the form:

$$\mathbf{f}_p = (\sin\theta_1, \cos\theta_1, ..., \sin\theta_N, \cos\theta_N), \tag{1}$$

where N is the number of joints in a feature vector (i.e., $N = 2$ in our experiments) and θ denotes an angle between adjacent body parts. For example, θ_1 is an angle between two line segments defined by a neck, a shoulder, and an elbow, and θ_2 defined by a shoulder, an elbow, and a wrist.

The mean poses and scatter plots of the lower and upper arms in each cluster for both arms with $K = 5$ are shown in Fig. 2. It can be seen that tight clusters can be obtained in the lower arms as well as in the upper arms.

While this pose clustering is required to make PM models with tighter pose clusters, human poses near the boundaries of neighboring clusters are similar to each other. By distributing these pose data to the neighboring clusters with overlapped samples, the PME can represent more complex interdependency among the PM models. The amount of overlap is determined based on the tradeoff between the tightness of clusters and the number of training samples in each cluster. The effect of overlapped samples is discussed in Section 4.6.

Training PM models. The PM model training procedure at stage 2 is shown in Fig. 1. As the number of training images for each PM model is reduced through clustering, the model training may be prone to overfitting compared with the one with all training images. Although we do expect that a PM model overfits to its pose cluster, a small number of training images often result in excessive overfitting in deep neural networks. Even when sufficient training images are given, the training images that are not in the pose cluster can still provide useful information for the PM model (e.g., the local appearance of each body joint).

As a result, we use two approaches in training each PM model: fine-tuning and dropout. We first pre-train a *generic* pose estimation model using the entire training dataset. We then fine-tune the PM model from the initial pre-trained model using training images in each cluster. To further alleviate overfitting, we apply a variant of dropout (Huang et al., 2015; Tompson et al., 2015) for regularizing the training. We find that this is essential to prevent excessive overfitting and boost the generalization performance of a PM model trained on a small number of images in each cluster.

3.2. PM-ensemble convnet

PM ensembling. We now present the PM-ensemble model to merge the responses from multiple PM models, as shown in Fig. 1. In the training stage (stage 3 in Fig. 1), we train the PME model to minimize the loss between an estimated heatmap and a ground-truth heatmap for each joint k over training data $\mathbf{N}$. Denote a set of training images and its ground-truth joint locations as $\{X, \mathbf{y}\}$. We minimize the network weights W by

$$\min_{W} \sum_{(X,y)\in N} \sum_{i,j,k} \|H_{ijk}(X, W) - H'_{ijk}(\mathbf{y})\|^2. \tag{2}$$

Here, $H_{ijk}(X, W)$ denotes a likelihood on image coordinates (i, j) of kth channel in the estimated heatmap given the network parameters, W, of the PME. The heatmap $H'_{ijk}(\mathbf{y})$ is the ground-truth joint location likelihood as in Pfister et al. (2015).

Fig. 3 shows the architecture used in the proposed PME. Unlike the simple parametric pooling (Pfister et al., 2015) (Fig. 3 (a)), we use a deeper model to encode the complex relationships among the diverse responses of the PM models. Compared to the model in Pfister et al. (2015) that uses a convolutional filter with kernel size 1×1, we use larger spatial kernels of 7×7, 9×9, 13×13 and 17×17 to capture spatial relation among PM models. We show three variants of the PMEs s1, s2, and s3 in Fig. 3(b), (c), and (d), respectively.

Similar to other heatmap-based methods, we obtain the human pose from the fused heatmaps by finding x-y coordinates with the max values.

Deconvolution for improved localization. While the PME model is capable of integrating the diverse responses from PM models, its performance depends on the spatial resolution of the heatmaps of the PM models. For example, the spatial resolution of each heatmap is lower than an input image due to the cascade of pooling layers in pose

Table 1
Comparison to the state-of-the-art pose estimation algorithms in terms of PCP on the FLIC dataset (Sapp and Taskar, 2013). Bold: the best, underline: second best performance. See Fig. 11 also for more detailed comparison with more state-of-the-art methods (Newell et al., 2016; Wei et al., 2016).

Method	U. Arms	L. Arms	Mean
Baseline Pfister et al. (2015)	96.9	85.7	91.3
Yang et al. (2016)	**98.1**	<u>89.5</u>	<u>93.8</u>
Chen and Yuille (2014)	97.0	86.8	91.9
Tompson et al. (2014)	93.7	80.9	87.3
Tompson et al. (2015)	90.0	78.5	84.3
Sapp and Taskar (2013)	84.4	52.1	68.3
PME s3 (Ours)	<u>97.8</u>	**90.2**	**94.0**

Table 2
Comparison to the state-of-the-art pose estimation algorithms in terms of PCP on the BBC pose dataset (Charles et al., 2014). Note here we only use still images as inputs. Bold: the best, underline: second best performance.

Method	U. Arms	L. Arms	Mean
Baseline Pfister et al. (2015)	75.4	75.1	75.3
Charles et al. (2014)	<u>89.1</u>	<u>75.8</u>	<u>82.5</u>
Yang and Ramanan (2013)	88.7	73.9	81.3
Buehler et al. (2011)	87.2	74.4	80.8
PME s3 (Ours)	**89.3**	**77.2**	**83.3**

estimation by ConvNets (e.g., (Pfister et al., 2015) and (Wei et al., 2016), which are used in our experiments). Such a low-resolution heatmap does not allow accurate pose estimation in the original resolution.

To address this issue, we add deconvolutional layers to the PME, as shown in Fig. 3(e). The effect of the deconvolution layers has been demonstrated in other problems such as image segmentation (Noh et al., 2015) and image synthesis (Dosovitskiy et al., 2015; Goodfellow et al., 2014). We alternate the convolution and deconvolution layers in the proposed model and find that this design generates a smoother heatmap. We call the PME model with deconvolution layers *the PME-D*.

4. Experiments

4.1. Implementation details

We use the models in Pfister et al. (2015) and Wei et al. (2016) as our upper-body and full-body PM models, respectively. In all experiments, the entire training dataset is partitioned into $K = 5$ clusters for the pose configuration of each arm. The total number of clusters (i.e., the number of PM models) is 10; 5 for the right arm and 5 for the left arm.

The ConvNet of the PME model consists of 13 convolution, 12 activation, and 2 pooling layers. After independently fine-tuning each PM model, we fixed the weights of the PM models and only update the weights of the PME. We avoided end-to-end learning with all PM models and the PME model because it is practically difficult to optimize a huge network consisting of all the models due to difficulty in avoiding local minima as well as due to a memory issue. Actually, the deep PME models (i.e., PME s3 and PME-D) for the full body could be trained only if the batch size was eight even in Titan X 12 GB. The proposed method was implemented with Jia et al., 2014 in accordance with two baselines (Pfister et al., 2015; Wei et al., 2016), but its current version has no function for a distributed memory usage. However, we consider the modularity of the models to be one of advantages for independent and efficient learning of a huge network. While we investigated the effect of end-to-end learning with the shallowest PME model (i.e., PME s1) with fewer clusters (i.e., $K = 3$ clusters)[3], further investigation for complex models should be important future work.

4.2. Datasets

We validate the performance of the proposed ensemble method using publicly available datasets: FLIC (Sapp and Taskar, 2013), BBC pose (Charles et al., 2014), LSP (Johnson and Everingham, 2011), MPII (Andriluka et al., 2014) datasets.

In the FLIC-full dataset (Sapp and Taskar, 2013), images are collected from 30 Hollywood movies. The upper-body joint positions are annotated. For training, we use the FLIC-plus dataset (Tompson et al., 2014) which is a subset of the FLIC-full dataset with around 17 K training images. For testing, we use a standard test set of 1000 images. As there are multiple people in several images of the FLIC dataset, we use the ground-truth torso box to crop out an image of a target person for evaluation.

In the BBC pose dataset (Charles et al., 2014), images are collected from 20 videos from the BBC. The training frames are annotated in a semi-automatic manner using the pose estimator of Buehler et al. (2011). In our experiments, we use about 600 K frames for training and 1000 images for testing.

While the FLIC and BBC datasets include upper-body human pose data, the LSP and MPII datasets provide data for the full-body human pose. The LSP dataset (Johnson and Everingham, 2011) consists of 1000 training and 1000 testing images collected from the Internet. In addition, extra 10,000 images are also given for a training purpose in the LSP extended dataset. The MPII human pose dataset (Andriluka et al., 2014) contains around 40 K human poses observed in 25 K images.

4.3. Evaluation protocols

We adopt three metrics for evaluation: (1) Percentage of Correct Parts (PCP), (2) Probability of Correct Keypoint (PCK), and (3) Percentage of Detected Joints (PDJ).

For PCP Ferrari et al. (2008), each body part is represented as a line segment between its two joints. The estimated location of a body part is considered as correct when both of the two joints locate within a certain fraction α of the length of the limb from their ground-truth locations. We set $\alpha = 0.5$ in all the experiments. We evaluate the PMEs using the strict PCP metric (Ferrari et al., 2008; Pishchulin et al., 2012) with person-centric annotations (Eichner and Ferrari, 2012).

In PCK (Yang and Ramanan, 2013), the detection is correct if the distance between the estimated and the ground-truth joints is less than a certain fraction $\beta = 0.5$ of the full-body size.

The PDJ metric (Sapp and Taskar, 2013) aims to evaluate the performance of the model under different precision through normalizing the location precision by the diagonal length of the torso.

4.4. Comparison to the state-of-the-art methods

Upper-body pose estimation. Tables 1 and 2 show the quantitative comparisons against several state-of-the-art pose estimation algorithms on the FLIC and BBC pose datasets, respectively. In both tables, the baseline model Pfister et al. (2015) has the same ConvNet architecture as our PM model and is trained using all the training data.

The proposed PME s3 outperforms all existing methods in FLIC dataset in terms of the mean PCP. Our results compare favorably against the state-of-the-art algorithms, particularly on the lower arm.

We also compare the PME model with the state-of-the-art approaches in BBC pose dataset, as shown in Table 2. Similar to the results of the FLIC dataset, our method compares favorably against the state-of-

[3] See *Effect of End-to-end Learning* in Section 4.6.

Table 3
PCK-0.2 evaluation on the LSP dataset. Each result is obtained on different training datasets specified in the brackets. The best score obtained on each dataset is colored by bold in each column. Methods marked with (*) and (**) were trained on "MPII, LSP, and LSP-extended datasets" and "LSP and LSP-extended datasets", respectively. In the former case, the score of PME s3 (Outs) is colored by underline if it is greater than or equal to the baseline (Wei et al., 2016).

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean
Pishchulin et al. (2016) (*)	97.0	91.0	83.8	78.1	91.0	86.7	82.0	87.1
Insafutdinov et al. (2016) (*)	97.4	92.7	87.5	84.4	91.5	89.9	87.2	90.1
Chu et al. (2017) (*)	98.1	93.7	89.3	86.9	93.4	94.0	92.5	92.6
Chou et al. (2017) (*)	98.2	94.9	**92.2**	**89.5**	94.2	**95.0**	**94.1**	**94.0**
Chen et al. (2017) (*)	**98.5**	94.0	89.8	87.5	93.9	94.1	93.0	93.1
Baseline Wei et al. (2016) (*)	97.8	92.5	87.0	83.9	91.5	90.8	89.9	90.5
PME s3 (Ours) (*)	97.6	**95.3**	87.9	81.2	**97.8**	90.8	87.8	91.2
Yu et al. (2016) (**)	87.2	88.2	82.4	76.3	91.4	85.8	78.7	84.3
Baseline Wei et al. (2016) (**)	**96.9**	**97.1**	80.4	75.1	86.5	83.2	**81.0**	84.3
PME s3 (Ours) (**)	92.0	92.0	**87.3**	**77.8**	**97.4**	**87.4**	77.1	**87.3**

Table 4
PCKh-0.5 evaluation on the MPII dataset. Methods marked with (*) were trained on "MPII, LSP, and LSP-extended datasets", while others were trained with only "MPII". Bold scores mean the best ones in each column. The score of PME s3 (Outs) is colored by underline if it is greater than the baseline (Wei et al., 2016).

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean
Pishchulin et al. (2016) (*)	94.1	90.2	83.4	77.3	82.6	75.7	68.6	82.4
Lifshitz et al. (2016)	97.8	93.3	85.7	80.4	85.3	76.6	70.2	85.0
Gkioxari et al. (2016)	96.2	93.1	86.7	82.1	85.2	81.4	74.1	86.1
Insafutdinov et al. (2016)	96.8	95.2	89.3	84.4	88.4	83.4	78.0	88.5
Bulat and Tzimiropoulos (2016)	97.9	95.1	89.9	85.3	89.4	85.7	81.7	89.7
Newell et al. (2016)	98.2	96.3	91.2	87.1	90.1	87.4	83.6	90.9
Chu et al. (2017)	**98.5**	96.3	91.9	88.1	90.6	88.0	85.0	91.5
Chou et al. (2017)	98.2	**96.8**	92.2	88.0	**91.3**	89.1	84.9	91.8
Chen et al. (2017)	98.1	96.5	**92.5**	**88.5**	90.2	**89.6**	**86.0**	**91.9**
Baseline Wei et al. (2016) (*)	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5
PME s3 (Ours)	97.7	95.8	90.1	85.6	88.8	84.8	81.7	89.6

the-art methods.

Full-body pose estimation. Table 3 shows the comparative evaluation results using PCK on the LSP. Unlike the results shown in Tables 1 and 2, each result is obtained with a different training dataset in Table 3. Comparing with other methods that use the same set of training data, our PME model demonstrates competitive performance. In particular, our PME outperforms its baseline (Wei et al., 2016) in the mean score for both sets of training data.

We also evaluates the proposed method on a larger full-body pose dataset, the MPII human pose dataset. Table 4 shows the comparative evaluation results using PCKh. While very recent models (Chen et al., 2017; Chou et al., 2017; Chu et al., 2017) are better than our PME, it outperforms the baseline (Wei et al., 2016), which was trained with more data.

4.5. Comparison to other ensemble approaches

We investigate the importance of ensembling approach on both FLIC and BBC pose datasets in Table 5. Compared to the simple ensemble approach such as averaging over multiple outputs (average pooling in Table 5) and parametric pooling using a single convolution layer used in the baseline (Pfister et al., 2015), we can see that the PME model outperforms these conventional ensemble approaches by a large margin. This suggests that a deeper ConvNet architecture with a cascade of convolutional layers with large kernels is critical to merge diverse estimations from PM models and represent the spatial contexts of body joints.

For the FLIC dataset, PME s3 outperforms the baseline (Pfister et al., 2015) by 2.8% in PCP[4] The improvement over Pfister et al. (2015) on the lower arm is particularly significant, with 4.5% improvement in PCP. In BBC pose dataset, PME s3 improves the baseline by 14.5% in PCP (upper arm).

Table 5
PCP for different ensemble approaches on the FLIC dataset and the BBC pose dataset. PCP on upper and lower arms, which are difficult to be localized, are shown. Bold: the best, underline: second best performance.

	FLIC			BBC pose		
Method	U. Arms	L. Arms	Mean	U. Arms	L. Arms	Mean
Baseline Pfister et al. (2015)	96.6	85.7	91.2	75.4	75.1	75.3
Average Pooling	96.7	86.7	91.7	75.6	73.6	74.6
Parametric Pooling Pfister et al. (2015)	96.9	87.2	92.1	74.6	73.8	74.2
PME s1	97.0	87.5	92.3	74.5	72.5	73.0
PME s2	97.7	89.8	93.8	83.1	76.6	79.9
PME s3	**97.8**	**90.2**	**94.0**	**89.3**	**77.2**	**83.3**

Improvement cases. We show several qualitative examples in Fig. 4. Fig. 4 (a) and (d) are the results of the average pooling, and results shown in (b) and (e) are obtained from PME s3. While the same set of heatmaps is provided to the average pooling and the PME s3, the average pooling fails to correctly localize the right wrist in (a) due to the self-occlusion (occluded by the left arm). In (d), the average pooling confused the left elbow with the right one possibly because the self-occlusion of the left wrist gives a negative impact on localizing the left elbow. On the other hand, our method successfully localizes the right wrist in (b) and the left elbow in (e). Fig. 4 (c) and (f) visualize the five heatmaps of the joints mislocalized by the average pooling. It is clear that several peaks are observed in the heatmaps for both images. The distributions of the peaks differ between the heatmaps, which demonstrate the heterogeneous properties of the PM models. It can also be seen that erroneous peaks are observed; strong peaks at the right elbow in (c) and (f). These erroneous peaks resulted in mislocalization in the average pooling. The correctly-localized joints are suggested by the

[4] See Section 4.6 for comparisons among the variants.

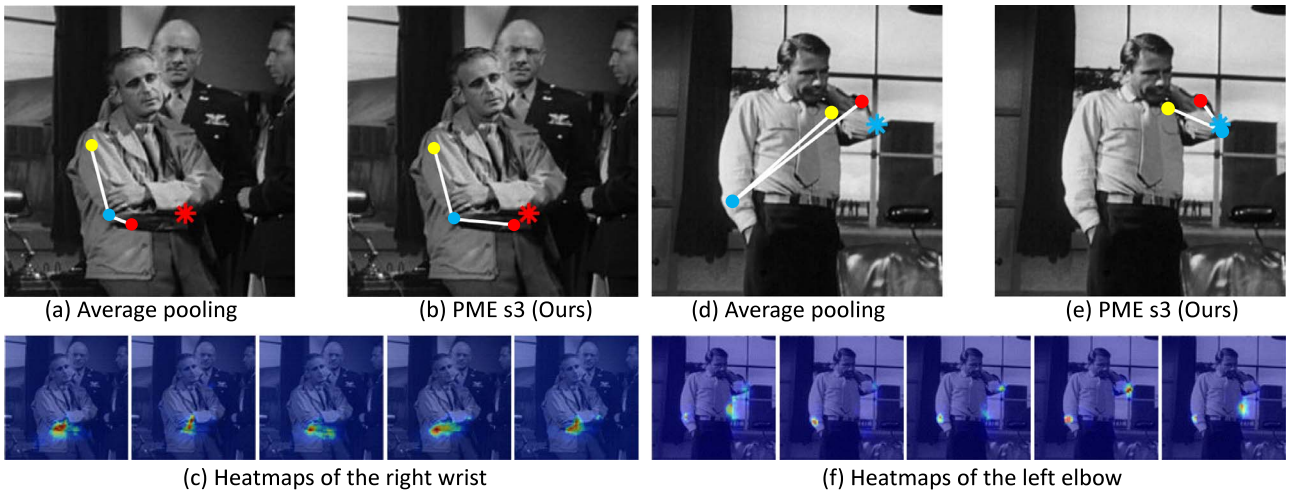


Fig. 4. Two improvement cases by the PME model. The positions of the estimated wrist, elbow, and shoulder are indicated by red, blue, and yellow circles, respectively. For presentation clarity, only one arm is visualized. The ground truth of joints mislocated by the average pooling (i.e., the right wrist and the left elbow in upper and lower examples, respectively) are indicated by stars and the heatmaps of these joints are also shown. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

PME model where each PM model focuses on a specific pose configuration such as the joint locations shown in Fig. 4 (b) and (e).

More qualitative results on the FLIC (Sapp and Taskar, 2013), BBC (Charles et al., 2014), and LSP (Johnson and Everingham, 2011) datasets are shown in Figs. 6, 7, and 8, respectively. These figures show the ground-truth of a human pose, the pose estimation results of the proposed methods, and the heatmaps of one joint. The joint whose localization is failed in several methods is selected for heatmap visualization. It can be seen that peak distributions vary among the heatmaps. This variation results in difficulty in pose estimation by the pose-modality-ensemble (PME) model. The results of PME s3 and PME-D are better than PMEs s2 and s1 with shallow layers. In comparison between results for the upper-body and full bodies, further difficulty arises due to complex nature of a human pose in sports (e.g., self occlusion) in the full body estimation (i.e., LSP). Due to this difficulty, the heatmap variation in the LSP is larger than in the FLIC and the BBC pose. In samples shown in Fig. 8, our deeper models (i.e., PME s3 and PME-D) can cope with these difficulties and localize all joints well.

Failure cases. Two typical failure cases by the PME model are shown in Fig. 5. As shown in Fig. 5(a) and (d), the average pooling correctly localizes the right elbow and wrist, respectively, while PME s3 shown in (b) and (e) does not perform well. While the mislocalization in (b) is not severe, the estimated right elbow is far from its ground truth in (e). This significant error may be caused by the larger distributions of the heatmaps as the distributions are significantly larger than other heatmaps shown in Fig. 4. Such a set of large distributions of the heatmaps cannot be captured by the PME model and causes mislocalization, while several strong peaks around the right elbow may result in its correct localization the simple average pooling.

4.6. Detailed analysis

Computational cost. For full-body pose estimation, the computational time of the proposed method is slightly more than the baseline (Wei et al., 2016). Let T be the training time of the baseline. In our experiments, the training times for fine-tuning each PM model and the ensemble model were at most less than 0.05 T and 0.1 T. Since all PM models can be fine-tuned in parallel, our method needs additional 15% computational time in total for training. We believe this subtle increase in computational cost is acceptable to get 0.7% and 1.1% accuracy gains on the LSP and MPII datasets, respectively.

Effect of network depth and kernel size. We examine the effect of different depth and kernel size using three variants of PME: (1) PME s1 (two convolutional layers with 9×9 kernel), (2) PME s2 (five

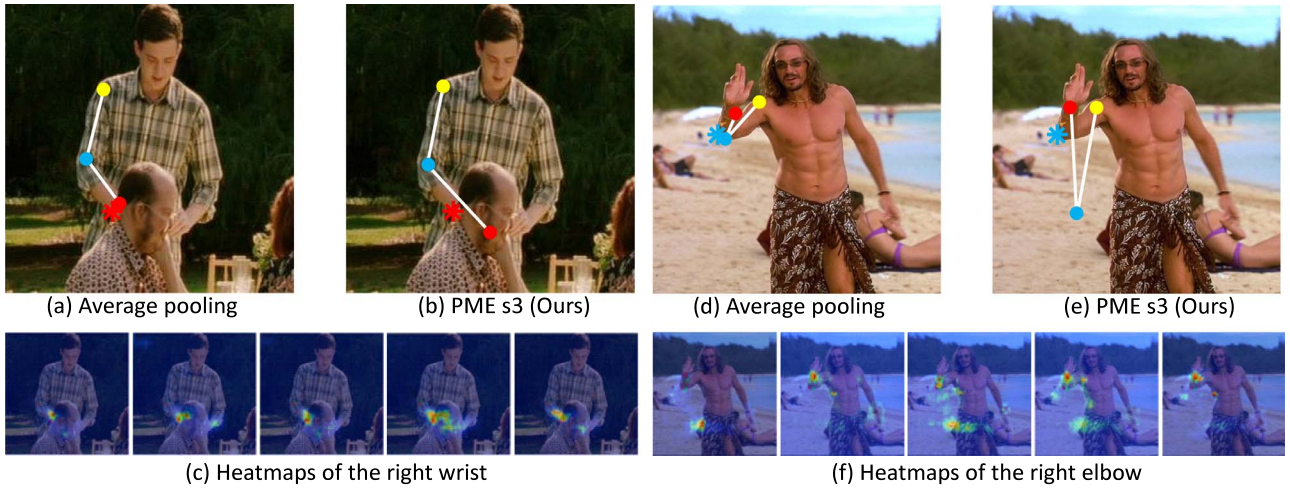


Fig. 5. Two failure cases by the PME model. The average pooling is able to localize the right wrist and elbow as shown in (a) and (d), respectively, whose heatmaps are shown in the figure, while the PME model in (b) and (e) does not perform well.

Fig. 6. Sample results on FLIC dataset. Heatmaps for each test image show those of a joint, in which the heterogeneous properties of different PM models can be observed, whose localization is failed in several methods.

convolutional layers with 7×7 and 13×13 kernels), and (3) PME s3 (seven convolutional layers with 7×7, 13×13, 17×17 kernels). As shown in Table 5, a deeper architecture with larger kernels have better performance. These results suggest that the network capacity is the essential to explicitly model interdependency among the responses of each PM model.

Effect of the number of PM models. In Fig. 9, we investigate the effect of the number of PM models in the ensemble using FLIC dataset. We

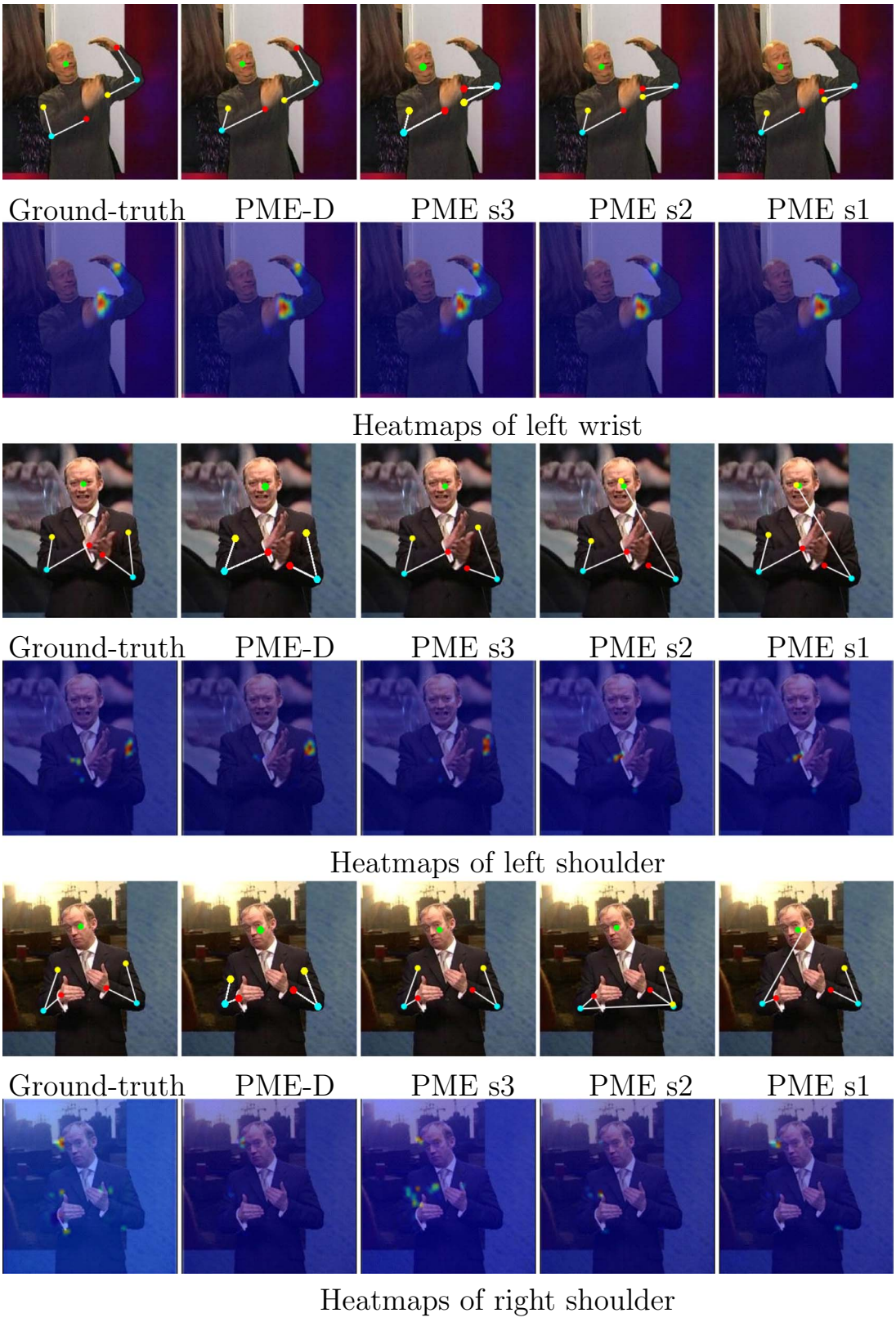


Fig. 7. Sample results on BBC pose dataset. Heatmaps in each row show those of a joint whose localization is failed in several methods.

plot the performance in PCP under different numbers of clusters $K = [1, 2, 3, 4, 5, 6]$, which is the number of PM models for one arm. As shown in Fig. 9, increasing the number of PM models improves the results. However, the improvement saturates when we use more than $K = 4$ models. We attribute this to the decreasing number of training samples in each cluster when we use more clusters. Based on the results, we use $K = 5$ in our experiments (10 models in total).

While the effect of pose clusters is also empirically analyzed in Rogez et al. (2017), their results show that the optimum number of clusters is around 100 in the Human3.6M dataset. The difference

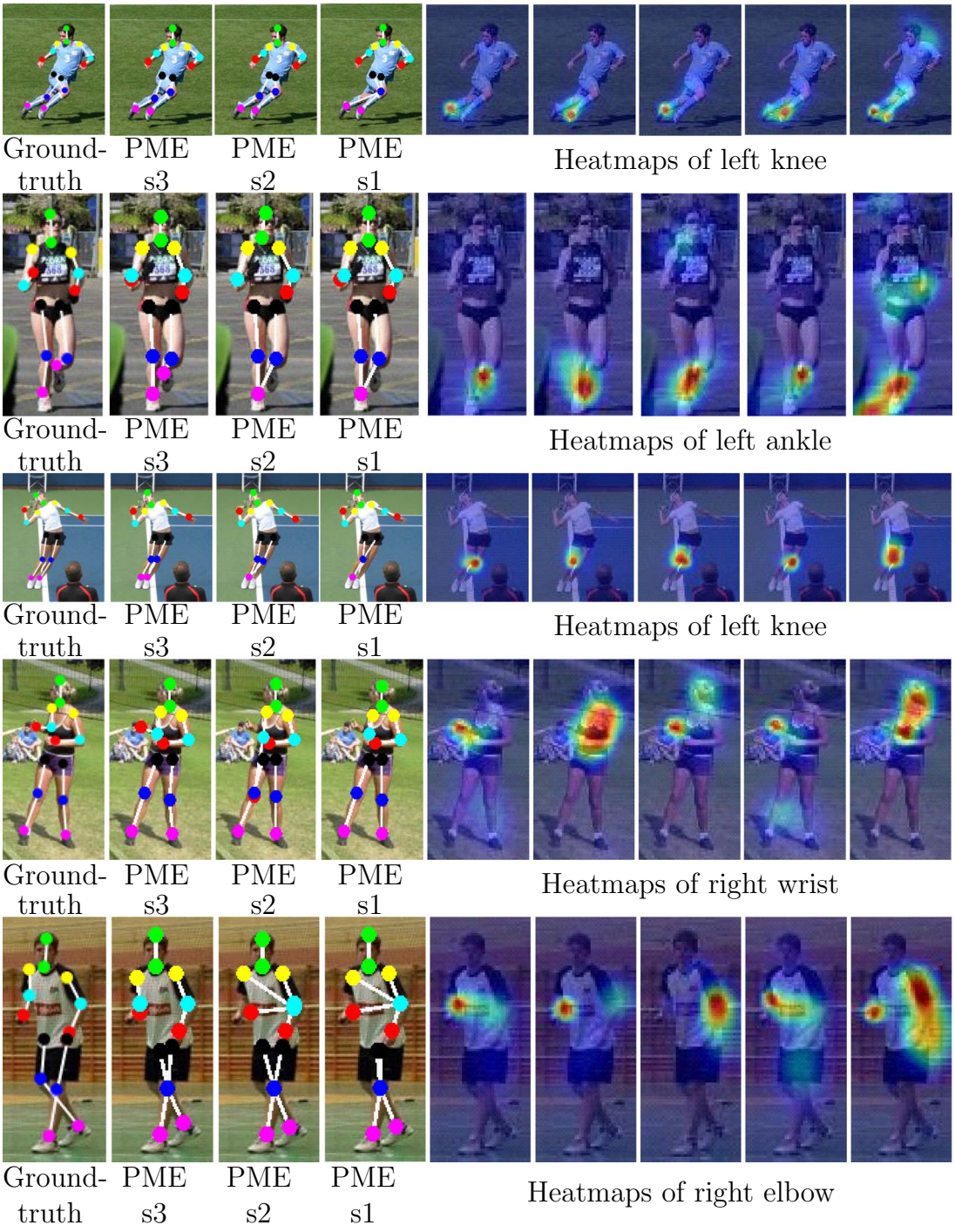


Fig. 8. Sample results on the LSP dataset. Heatmaps in each row show those of a joint whose localization is failed in several methods.

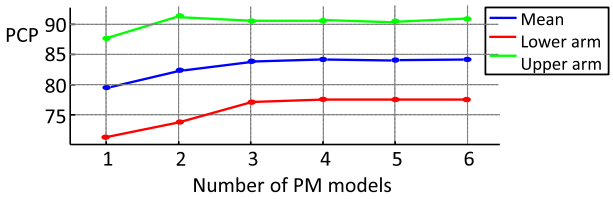


Fig. 9. Effect of the number of PM models. Increasing the number of the PM models improves the PCP performance.

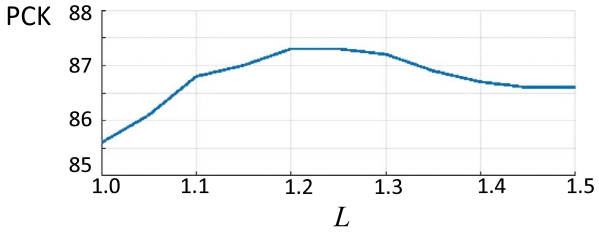


Fig. 10. PCK-0.2 results obtained from PM models trained by pose clusters having different overlaps. We evaluate L, which determines the overlap between neighboring pose clusters, between 1.0 and 1.5 with an interval of 0.05.

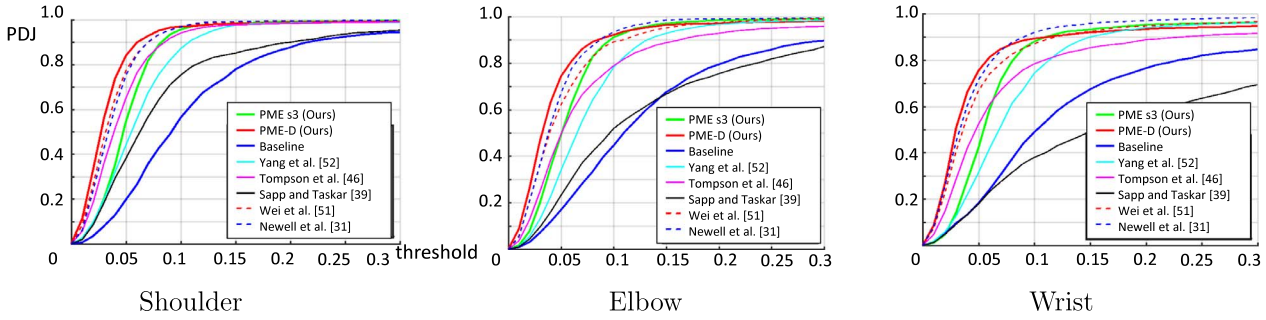


Fig. 11. PDJ curves comparison among PME s3, the PME-D, the baseline, and several state-of-the-art methods on the FLIC dataset.

between Rogez et al. (2017) and our case may be caused by the complexity in 3D pose estimation as well as the large number of data in the Human3.6M dataset. For a greater number of training images, in particular for complex full-body poses, more pose clusters (i.e., greater than $K=5$) may work better also in 2D human pose estimation.

Effect of overlaps among pose clusters. The proposed PME model is evaluated using pose clusters with different overlaps. Fig. 10 shows the results of PCK-0.2 evaluation with different overlaps among pose clusters on the LSP dataset. The overlaps are adjusted so that cluster c includes a pose sample d in another cluster if the distance between the centroid of c and d is less than $Ld_{\max}$. In Fig. 10, L is a parameter on the x-axis, and $d_{\max}$ is the distance between the centroid of c and the most distant pose data within c. Since we can see a peak around $L = 1.2$, $L = 1.2$ is used in all experiments for full-body pose estimation. Overall, the proposed algorithm performs better when a small amount of overlapped samples are included, and performs robustly within a reasonable range of overlaps.

Effect of deconvolution for improved localization. To show the effect of deconvolution layers, we evaluate the PME (PME s3), the PME-D, the baseline, and several state-of-the-art methods (Newell et al., 2016; Sapp and Taskar, 2013; Tompson et al., 2015; Wei et al., 2016; Yang et al., 2016) using the PDJ scores (Fig. 11). While the state-of-the-art methods deliver equal performance under a larger threshold (e.g., above 0.2), the proposed methods outperform others under a lower threshold (e.g., below 0.1). The PDJ scores at the threshold of 0.05 for wrist (a strict criterion for a challenging body part) are 75.4% (PME-D), 71.5% (Newell et al., 2016), 66.8% (Wei et al., 2016), and 18.2% (baseline). The proposed algorithms achieve 3.9% and 8.6% relative improvements over the best (Newell et al., 2016) and the second best (Wei et al., 2016) methods. For the elbow and shoulder, PME-D outperforms the (Newell et al., 2016) by 7.2% and 8.8%, respectively. The increase of spatial resolution by PME-D effectively reduces inaccurate estimates.

Table 6 shows the results for mean PDJ@0.05 as well as mean PDJ@2.0 as the advantage of our PME-D appears in strict thresholds. The results show that we should select PME-D or PME s3 depending on the application; if strict/loose joint detection is needed, PME-D/PMEs3 should be used.

To validate the effect of PME-D, typical examples of improvement are shown in Fig. 12 in which the local patches of the left shoulder are visualized. As can be seen in the figure, the joint locations inferred by PME-D (indicated by green stars in Fig. 12(b) and (d)) are closer to the ground-truth positions indicated by red stars than the baseline (indicated by green stars in Fig. 12(a) and (c)). While such a small improvement gives only a small impact on PCP and PCK with standard thresholds, its effect can be demonstrated by PDJ.

Table 6
PDJ scores on the FLIC dataset. Each score is the mean of three joints, a shoulder, an elbow, and a wrist. See Fig. 11 for the whole results.

	PDJ@0.05	PDJ@0.2
PME s3	48.7	97.6
PME-D	78.0	96.4
Baseline Pfister et al. (2015)	18.3	81.2
Newell et al. (2016)	68.8	98.4
Wei et al. (2016)	71.3	97.2

While PME-D works favorably against other methods in all thresholds as shown in Fig. 11 and Table 6, PME s3 is outperformed in lower thresholds (e.g., PDJ@0.05) by other methods including the baseline. This might be caused because of ambiguous features represented in only heatmaps given by PM models. As demonstrated in our experiments (e.g., Fig. 4), the PME model becomes relatively robust against the ambiguity of joints, which are occluded or have similar appearances with other objects, by integrating high confidence at multiple spatial locations in multiple PM models. This robustness might be obtained because not only of the data integration but also of rough localization given by the heatmaps, which have no image features. That is, PME is robust in higher thresholds at the sacrifice of localization precision in lower thresholds, while PME-D avoids this performance degradation by using deconvolution layers.

Effect of additional image features. To investigate the trade-off between robustness and precision described above, we fed image features also into our PME model. However, image features cannot be used in PME s3 due to a limited memory on a GPU (i.e., Titan X 12 GB in our experiments). This memory problem was avoided by evaluating our method in the shallowest PME model, PME s1, with fewer pose clusters (i.e., $K = 3$). This model is called PME s1*.

For experiments with the full-body model on the LSP, we used image features extracted by a sub-network for image-feature extraction in a PM model (i.e., the base model (Wei et al., 2016)). This sub-network consists of four convolution layers and three pooling layers. We call PME s1* using the image features PME $\widehat{\text{s1}}$*.

The results of PME $\widehat{\text{s1}}$* are shown in Table 7. It can be seen that PME $\widehat{\text{s1}}$* is better than PME s1* in the mean score as well as in many joints. This fact suggests the potential of the joint usage of the image features and heatmaps, while this too simple PME model, PME $\widehat{\text{s1}}$*, is not appropriate for evaluating the complete performance of our proposed scheme. We consider this issue to be an important research direction in order to explore further improvement with a larger computational resource.

Effect of end-to-end learning. While the PM and PME models were trained independently in all experiments shown before, we investigate the effect of end-to-end learning with the PME model and all PM models. Note again that this end-to-end learning is evaluated by the shallowest PME model, PME s1, with $K = 3$ clusters due to a memory issue, as described in Section 4.1. On the LSP dataset, mean PCK-0.2 scores of independent and end-to-end-learning schemes are 84.8 and 85.4, respectively, as shown in Table 7. Further analysis of improvement by this end-to-end learning scheme is also considered to be an important research direction.

Effect of pose clustering criteria. We have conducted additional experiments on the LSP, where a pose feature vector is represented by the configurations of legs as well as arms. That is, in addition to θ_1 and θ_2 for the upper body, θ_3 and θ_4 are used in the pose feature vector (i.e.,

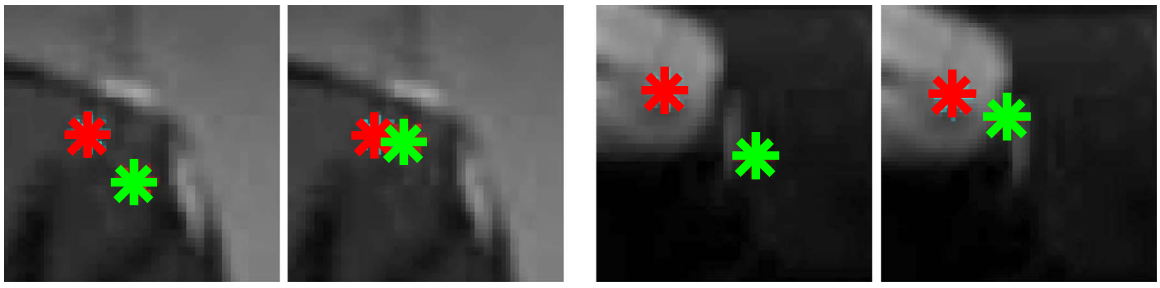


Fig. 12. Visualization of improvement from the baseline by PME-D. The ground-truth and estimated positions of the left shoulder are indicated by red and green stars, respectively. (a) and (c) show the results of the baseline, while (b) and (d) are those of PME-D that are closer to the ground truth than the baseline. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 7
PCK-0.2 evaluation for variants of our proposed method on the LSP dataset. All models were trained on the LSP and LSP-extended datasets. PME s1* differs from PME s1 so that PME s1* has only $K = 3$ pose clusters. PME $\widehat{\text{s1*}}$ uses image features in addition to heatmaps obtained from PM models as inputs for the PME model. PME $\widehat{\text{s1*}}$ is optimized by an end-to-end learning manner with all PM models and the PME model. The scores of PME $\widehat{\text{s1*}}$ and PME $\widetilde{\text{s1*}}$ are colored by underline in each column if they are above the score of PME s1*.

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean
PME s1*	89.3	90.3	85.3	74.1	95.5	85.1	73.8	84.8
PME $\widehat{\text{s1*}}$ (image)	90.9	90.8	85.3	76.4	95.3	86.4	76.0	85.9
PME $\widetilde{\text{s1*}}$ (E-to-E)	89.7	90.0	85.6	75.2	95.5	85.7	74.7	85.2

Table 8
PCK-0.2 evaluation for variants of our proposed method on the LSP dataset. The best score is colored by bold in each column. All models were trained on the LSP and LSP-extended datasets. PME s3' is different from PME s3 in terms of the components of a pose feature. In PME s3', the pose feature vector consists of legs as well as arms. s$\acute{3}$' differs from PME s3' in terms of a criterion used for pose clustering. Specifically, PME s$\acute{3}$' clusters the arm-and-leg pose features based on the Procrustes distance instead of the Euclid distance. PME s3" and PME s$\acute{3}$" respectively use the Euclid distance and the Procrustes distance for clustering, while both of them employs the full-body pose features.

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean
PME s3	92.0	92.0	87.3	77.8	**97.4**	87.4	77.1	87.3
PME s$\acute{3}$ (Procrustes)	92.6	91.8	87.4	78.2	**97.4**	87.7	77.6	87.5
PME s3' (partial arms-and-legs)	91.1	90.9	86.8	75.8	96.8	87.3	77.2	86.6
PME s$\acute{3}$' (partial arms-and-legs, Procrustes)	**92.9**	91.9	**87.8**	**78.4**	**97.4**	**88.0**	**78.1**	**87.8**
PME s3" (full-body)	91.0	91.3	85.7	74.8	97.0	86.9	74.1	85.8
PME s$\acute{3}$" (full-body, Procrustes)	92.6	**92.2**	87.0	77.2	97.1	86.9	76.9	87.1

Eq. (1)). In our experiments, θ_3 is an angle between two line segments defined by the mid point of two hips, a hip, and a knee, and θ_4 defined by a hip, a knee, and an ankle. PME s3 that uses this arm-and-leg pose feature, which is called PME s3', is expected to be able to represent the variation of full-body poses better than PME s3 using the pose feature with only arms. The experimental results show that the arm-and-leg pose feature cannot improve the performance when compared with the pose feature with only arms; mean accuracy: 86.6 (arms and legs) vs 87.3 (arms), as shown in Table 8. While the arm-and-leg pose feature may be better essentially, the pose variety in each of 10 PM models becomes large in contrast to the one with only arms. This variety makes it difficult to optimize the ensemble model.

For more rigid alignment of 2D pose annotations, it is known that the Procrustes analysis provides a more effective metric (Bourdev et al., 2010). PME s3' and PME s$\acute{3}$' respectively use the Euclid distance and the Procrustes distance for clustering, while both of them employ the arm-and-leg pose features. It can be seen that the Procrustes distance improves pose estimation accuracy in most joints: 86.6 (Euclid distance) vs 87.5 (Procrustes distance) on average. As a result, the model using the arm-and-leg pose feature becomes better than the one using the arm pose feature: 87.3 (arm) vs 87.5 (arms and legs).

All experiments shown above were conducted with our pose clustering using a *partial* body part, which is proposed in Section 3.1. In order to validate its effectiveness, we also conducted experiments with pose clustering using all body parts in the full body. While the proposed partial-part clustering uses each of the left and right arms/legs (i.e., $N = 4$ in Eq. (1) for the arm-and-leg pose feature), PME s3" and PME s$\acute{3}$" employ both left and right arms and legs (i.e., $N = 8$). PME s3" differs from PME s$\acute{3}$" so that PME s$\acute{3}$" uses the Procrustes distance. As shown in Table 8, the full-body pose feature is outperformed by the proposed partial-part clustering even though the Procrustes distance improves the performance.

5. Conclusions

In this paper, we propose the pose-modality-ensemble model for human pose estimation. Through training PM models with clustered training samples, we obtain heterogeneous PM models that are specialized to particular body configurations. The PME model is capable of merging diverse responses from the PM models. We demonstrate the effectiveness of PME model on public pose estimation datasets and show that the proposed method performs favorably against state-of-the-art methods and alternative model ensemble approaches.

Important future work includes (1) integration of heatmaps obtained from PM models and image features by the PME model and (2) an efficient end-to-end learning scheme with the PME model and all PM models.

References

Agostinelli, F., Anderson, M., Lee, H., 2013. Adaptive multi-column deep neural networks with application to robust image denoising. Proceedings of the Neural Information Processing Systems (NIPS).
Andriluka, M., Pishchulin, L., Gehler, P.V., Schiele, B., 2014. 2d human pose estimation: new benchmark and state of the art analysis. Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR).
Andriluka, M., Roth, S., Schiele, B., 2009. Pictorial structures revisited: people detection and articulated pose estimation. Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
Bergtholdt, M., Kappes, J.H., Schmidt, S., Schnörr, C., 2010. A study of parts-based object class detection using complete graphs. Int. J. Comput. Vis. (IJCV) 87 (1–2), 93–117.
Bourdev, L., Maji, S., Brox, T., Malik, J., 2010. Detecting people using mutually consistent poselet activations. Proceedings of the European Conference on Computer Vision (ECCV).

Bourdev, L., Malik, J., 2009. Poselets: body part detectors trained using 3d human pose annotations. Proceedings of the International Conference on Computer Vision (ICCV).
Buehler, P., Everingham, M., Huttenlocher, D.P., Zisserman, A., 2011. Upper body detection and tracking in extended signing sequences. Int. J. Comput. Vis. (IJCV) 95 (2), 180–197.
Bulat, A., Tzimiropoulos, G., 2016. Human pose estimation via convolutional part heatmap regression. Proceedings of the European Conference on Computer Vision (ECCV).
Burgos-Artizzu, X.P., Hall, D., Perona, P., Dollár, P., 2013. Merging pose estimates across space and time. Proceedings of the British Machine Vision Conference (BMVC). http://dx.doi.org/10.5244/C.27.58.
Carreira, J., Agrawal, P., Fragkiadaki, K., Malik, J., 2016. Human pose estimation with iterative error feedback. Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR).
Charles, J., Pfister, T., Everingham, M., Zisserman, A., 2014. Automatic and efficient human pose estimation for sign language videos. Int. J. Comput. Vis. 110 (1), 70–90.
Chen, X., Yuille, A., 2014. Articulated pose estimation by a graphical model with image dependent pairwise relations. Proceedings of the Neural Information Processing Systems (NIPS).
Chen, Y., Shen, C., Wei, X., Liu, L., Yang, J., 2017. Adversarial posenet: a structure-aware convolutional network for human pose estimation. Proceedings of the International Conference on Computer Vision (ICCV).
Chou, C.-J., Chien, J.-T., Chen, H.-T., 2017. Self adversarial training for human pose estimation. arXiv:1707.02439.
Chu, X., Yang, W., Ouyang, W., Ma, C., Yuille, A.L., Wang, X., 2017. Multi-context attention for human pose estimation. proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR).
Ciresan, D., Meier, U., Schmidhuber, J., 2012. Multi-column deep neural networks for image classification. Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR).
Dantone, M., Gall, J., Leistner, C., Gool, L.V., 2013. Human pose estimation using body parts dependent joint regressors. Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR).
Dosovitskiy, A., Springenberg, J.T., Brox, T., 2015. Learning to generate chairs with convolutional neural networks. Proceedings of the International Conference on Computer Vision (ICCV).
Eichner, M., Ferrari, V., 2012. Appearance sharing for collective human pose estimation. Proceedings of the Asian Conference on Computer Vision (ACCV).
Eichner, M., Ferrari, V., Zurich, S., 2009. Better appearance models for pictorial structures. Proceedings of the British Machine Vision Conference (BMVC).
Eichner, M., Marín-Jiménez, M.J., Zisserman, A., Ferrari, V., 2012. 2D articulated human pose estimation and retrieval in (almost) unconstrained still images. Int. J. Comput. Vis. (IJCV) 99 (2), 190–214. http://dx.doi.org/10.1007/s11263-012-0524-9.
Felzenszwalb, P.F., Girshick, R.B., McAllester, D.A., Ramanan, D., 2010. Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Anal. Mach. Intel. (PAMI) 32 (9), 1627–1645.
Ferrari, V., Marin-Jimenez, M., Zisserman, A., 2009. Pose search: retrieving people using their pose. Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR).
Ferrari, V., Marín-Jiménez, M.J., Zisserman, A., 2008. Progressive search space reduction for human pose estimation. Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR).
Gkioxari, G., Toshev, A., Jaitly, N., 2016. Chained predictions using convolutional neural networks. Proceedings of the European Conference on Computer Vision (ECCV).
Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y., 2014. Generative adversarial nets. Proceedings of the Neural Information Processing Systems (NIPS).
Huang, Y., Sun, X., Lu, M., Xu, M., 2015. Channel-max, channel-drop and stochastic max-pooling. Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR Workshops).
Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., Schiele, B., 2016. Deepercut: a deeper, stronger, and faster multi-person pose estimation model. Proceedings of the Neural Information Processing Systems (NIPS).
Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R.B., Guadarrama, S., Darrell, T., 2014. Caffe: Convolutional Architecture for Fast Feature Embedding. ACM Multimedia 675–678.
Johnson, S., Everingham, M., 2010. Clustered pose and nonlinear appearance models for human pose estimation. Proceedings of the British Machine Vision Conference (BMVC).
Johnson, S., Everingham, M., 2011. Learning effective human pose estimation from inaccurate annotation. Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR).
Kiefel, M., Gehler, P.V., 2014. Human pose estimation with fields of parts. Proceedings of the European Conference on Computer Vision (ECCV).
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Proceedings of the Neural Information Processing Systems (NIPS).
Lifshitz, I., Fetaya, E., Ullman, S., 2016. Human pose estimation using deep consensus voting. Proceedings of the European Conference on Computer Vision (ECCV).
Moghimi, M., Belongie, S., Saberian, M., Yang, J., Vasconcelos, N., Li, L.-J., 2016. Boosted convolutional neural networks. Proceedings of the British Machine Vision Conference (BMVC).
Newell, A., Yang, K., Deng, J., 2016. Stacked hourglass networks for human pose estimation. Proceedings of the European Conference on Computer Vision (ECCV).
Noh, H., Hong, S., Han, B., 2015. Learning deconvolution network for semantic segmentation. Proceedings of the International Conference on Computer Vision (ICCV).
Ouyang, W., Chu, X., Wang, X., 2014. Multi-source deep learning for human pose estimation. Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR).
Pfister, T., Charles, J., Zisserman, A., 2015. Flowing convnets for human pose estimation in videos. Proceedings of the International Conference on Computer Vision (ICCV).
Pishchulin, L., Insafutdinov, E., Tang, S., Andres, B., Andriluka, M., Gehler, P., Schiele, B., 2016. Deepcut: joint subset partition and labeling for multi person pose estimation. Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR).
Pishchulin, L., Jain, A., Andriluka, M., Thormählen, T., Schiele, B., 2012. Articulated people detection and pose estimation: reshaping the future. Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR).
Puwein, J., Ballan, L., Ziegler, R., Pollefeys, M., 2014. Foreground consistent human pose estimation using branch and bound. Proceedings of the European Conference on Computer Vision (ECCV).
Ramakrishna, V., Munoz, D., Hebert, M., Bagnell, J.A., Sheikh, Y., 2014. Pose machines: articulated pose estimation via inference machines. Proceedings of the European Conference on Computer Vision (ECCV).
Rogez, G., Weinzaepfel, P., Schmid, C., 2017. Lcr-net: localization-classification-regression for human pose. Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR).
Sapp, B., Taskar, B., 2013. Modec: multimodal decomposable models for human pose estimation. Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR).
Sapp, B., Toshev, A., Taskar, B., 2010. Cascaded models for articulated pose estimation. Proceedings of the European Conference on Computer Vision (ECCV).
Sapp, B., Weiss, D.J., Taskar, B., 2011. Parsing human motion with stretchable models. Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR).
Singh, S., Hoiem, D., Forsyth, D., 2015. Learning a sequential search for landmarks. Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR).
Sun, M., Savarese, S., 2011. Articulated part-based model for joint object detection and pose estimation. Proceedings of the International Conference on Computer Vision (ICCV).
Sun, Y., Wang, X., Tang, X., 2013. Hybrid deep learning for face verification. Proceedings of the International Conference on Computer Vision (ICCV).
Tian, Y., Zitnick, C.L., Narasimhan, S.G., 2012. Exploring the spatial hierarchy of mixture models for human pose estimation. Proceedings of the European Conference on Computer Vision (ECCV).
Tompson, J., Goroshin, R., Jain, A., LeCun, Y., Bregler, C., 2015. Efficient object localization using convolutional networks. Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR).
Tompson, J.J., Jain, A., LeCun, Y., Bregler, C., 2014. Joint training of a convolutional network and a graphical model for human pose estimation. Proceedings of the Neural Information Processing Systems (NIPS).
Toshev, A., Szegedy, C., 2014. Deeppose: human pose estimation via deep neural networks. Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR).
Tran, D., Forsyth, D., 2010. Improved human parsing with a full relational model. Proceedings of the European Conference on Computer Vision (ECCV).
Ukita, N., 2012. Articulated pose estimation with parts connectivity using discriminative local oriented contours. Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR).
Wei, S.-E., Ramakrishna, V., Kanade, T., Sheikh, Y., 2016. Convolutional pose machines. Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR).
Yang, W., Ouyang, W., Li, H., Wang, X., 2016. End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation. Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR).
Yang, Y., Ramanan, D., 2013. Articulated human detection with flexible mixtures of parts. IEEE Trans. Pattern Anal. Mach. Intel. (PAMI) 35 (12), 2878–2890.
Yu, X., Zhou, F., Chandraker, M., 2016. Deep deformation network for object landmark localization. Proceedings of the European Conference on Computer Vision (ECCV).