2016 23rd International Conference on Pattern Recognition (ICPR)
Cancún Center, Cancún, México, December 4-8, 2016

Human Pose Estimation based on Human Limbs

Guoqiang Liang*, Xuguang Lan*, Jiang Wang†, Nanning Zheng*
*Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, China
†Institute for Deep Learning, Baidu, Sunnyvale, USA
gqliang@stu.xjtu.edu.cn, {xglan, nnzheng}@mail.xjtu.edu.cn, wangjiangb@gmail.com

***Abstract*—Modeling the relationship among human joints is one of the most important components in human pose estimation. Previous methods usually define this relationship as geometric constraints on the relative location of two neighboring joints. In this definition, the local image appearance of the region connecting two neighboring joints is ignored. In fact, this image appearance, called human limb, plays an important role in human joint localization in human visual system. To make full use of this local image appearance, we propose to solve a new task: human limb detection. We combine it with human joint localization in one deep convolutional neural network. After getting coarse results, we employ a graphical model to remove false positive detections. Besides, shallow and deep features are combined in this model. We evaluate our method on the FLIC and LSP datasets. The experiments results show the effectiveness of our method.**

Keywords*—*Human Pose estimation; Limbs Detection;ConvNet, Graphical model

I. Introduction

Human pose estimation is the task of estimating the spatial location of human parts from a 2D monocular image. This task is one of the fundamental tasks in computer vision and has wide applications in various computer vision systems, such as action recognition, human computer interface, and activity detection. Great improvements have been obtained in recent years, especially after the rise of Convolutional Neural Networks (ConvNets). However, it is still a challenging problem due to large variability of human pose, camera view and occlusion among different human parts.

Most of previous methods in pose estimation are based on deformable part model [1], [2], [5]-[13]. In deformable part models, human body is represented by a collection of physiologically inspired parts, which are human limbs or joints. A graphical model over parts is defined with nodes representing parts and edges encoding constraints between pairwise human parts. A fter this seminal work [1], a wide variety of features and relation models have been proposed [2], [5]-[13]. However, limited by the hand-crafted features and tree-based graphical models, the pose estimation accuracy was far from satisfactory. Thanks to the powerful learning capacity of ConvNets [15] and much larger human pose estimation datasets [11]-[13], ConvNets have been used to learn better representation and the joint relationship [16]-[21]. These ConvNet based models have achieved much better performance over traditional methods.

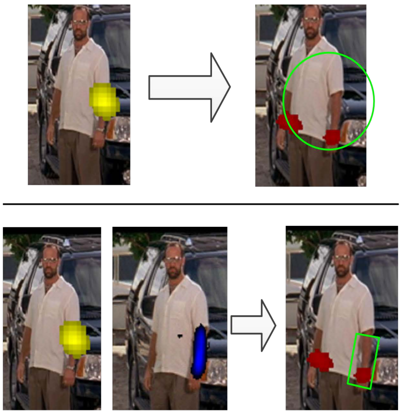


Fig. 1. Comparison between our method and previous methods. Top: Using joints produces constraints only on the relative distance. Bottom: Combining limbs and joints generates better candidate space.

In most of the current methods, the relations between two human parts is defined as constraints on their relative location and orientation, such as the Gaussian model in [5]-[13] and the conditional probability of joints' location used in [20], [21]. For joints with high flexibility, like wrist, this constraint is too weak to remove false positive detections near to the reference joint. As shown in the top of Fig. 1, there are two wrist candidates in the similar range of the elbow joint. Only using relative distance constraints cannot differentiate these two candidates. However, they can be distinguished by combining the detection of elbow and lower arm as shown in the bottom of Fig. 1. The limb detection result augments the location constraints between joints, and it improves the possibility of finding the true joint location. Using human limbs, which model the local image appearance between two neighbor joints, can reduce candidate space of joints and remove the false positive detection that cannot be filtered by the relative location constraints. Previously, human limbs are used as parts [1]-[6] or contextual information for joints detection [7]-[13]. In these works, human limbs are detected by segmentation cues [5], [6] or simple hand-crafted feature [8], [9], [12]. However, these hand-crafted models cannot extract invariant information of human limbs due to their variety caused by the camera view, occlusion, clothes and poses of humans.

In this paper, we propose a deep learning based architecture for human pose estimation by integrating limbs detection and joints localization. First, we design limbs detector to estimate the region of limbs from images. The limbs detector generates one per-pixel likelihood map for each limb. Because of the close relevance between joints and limbs, we integrate their detection in a single ConvNet. Then we filter the coarse detection by a graphical model. This graphical model defines appearance and distance constraints among joints and limbs and is implemented as another ConvNet. Finally, we connect these two ConvNets into one ConvNet, where information from both deep layer and shallow layer are combined to improve localization accuracy. Moreover, we use some deconvolution layer to increase the resolution of final predicted maps. These layers eliminate the influence of pooling layers and improve the performance.

In the experiments, we evaluate the proposed method on the two widely used datasets: Frame Labeled in Cinema (FLIC) dataset [11] and Leeds Sports Pose (LSP) dataset [12], [13]. The experiments results show the effectiveness of our method.

II. RELATED WORK

Due to the wide applications, human pose estimation is a very popular research area. There are many excellent methods, such as mixture of parts model [8], Pictorial Structures (PS) based model [2], [6], [11]. For a detailed review, please refer to [14]. In this paper, we just review the most related works on modeling human limb and CNN based human pose estimation.

A. Limb Modeling

Detecting and modeling limb region is critical component for appearance based relation model. In many early works, human limbs are used as human parts inspired by human visual system. In the PS framework [1], [2], human limbs are modeled as a rigid oriented rectangle, whose position and angle determine human pose. Since the limbs are modeled as bars [3], [4], limbs are detected by finding parallel edges. Compared with edge-based models, image segmentation methods are used to determine the limbs' location in [5], [6]. In these models, it is needed to estimate several parameters for a limb, such as orientation, location, length and width. This is impractical in realistic images due to foreshorten and variation of view.

Sapp et al. [7] defined a human part as two joints instead of one limb, which are at the end of the limb. Compared with fixed length and limited orientation bin in PS model, Sapp's model can denote nearly any angle between parts and finely discretized limb length. Because the two joints model is more suitable for parts detection, it has been a principal model for human pose estimation [8]-[11]. Yang et al. [8] add another joint at the middle point of the limb to capture the appearance of the middle area in a limb. This addition of joints makes the model cover more contents of human body, so [8] outperforms previous methods. Compared with the equal model for joints and limbs, Wang et al. [9] use combined parts to model the middle area of limbs, which are generated from an appearance-based latent SVM.

In realistic images, limb region varies greatly both in its shape and appearance. As a result, it is tough to completely capture limb's shape and appearance by one joint even with mixture models. However, the shape information is critical for parts detection in human vision system. In this paper, we model entire limb region as a wide line connecting these two joints. This representation could capture richer label and shape information, especially when part of the line are occluded.

B. ConvNet based pose estimation

In recent years, ConvNets have achieved huge success in many computer vision tasks [15]. Due to this success, many ConvNet based models have been developed for human pose estimation and achieved state of the art performance [16]-[21].

Chen et al. [16] use a ConvNet to extract appearance and type score, and then combine these in the DPM framework. The large improvement of experiment performance proves that features extracted by ConvNets are more effective than hand-crafted feature. Toshev et al. [17] cascade two ConvNets to directly regress human joints coordinates from images. Although the performance is improved a lot compared with traditional methods, their method performs poorly in high precision metric. This result shows the difficulty of directly learning the mapping. Fan et al. [18] integrate local part appearance and holistic view of each part for accurate human pose estimation. Their method shows significant improvement in low precision.

In contrast, [19]-[21] design a different framework in which ConvNets are used to generate a discrete heat-map for each human joint. Jain et al. [19] use a single ConvNet to map local window to a binary output for each joint. After obtaining the raw detection, a weak high-level spatial model is used to enforce the global consistency. Because of better adaption to pose estimation, their method achieves better performance in high precision metric. Following [19], [20] transforms the MRF-based graphical model as a ConvNet. To eliminate the effect of pooling layers, [21] employs another ConvNet to estimate the location offset given the previous predicted region. In essence, these methods are per-pixel classification problems with large contextual information. This frame reduces learning difficulty and obtains success.

Different from the above works, we aim at modeling the human joints and human limbs together and promote human pose estimation by considering limb region. Our main contributions include:

We propose to extract limb region by ConvNet explicitly and complete the detection of joints and limbs in one ConvNet.

We design a graphical model to capture the relations among joints and neighbor joints.

Our method outperforms the baseline [20] and achieves comparable performance to the state of the art.

III. MODEL

First, we illustrate our notations. In this paper, a human body is represented as a set of human joints $\boldsymbol{U}$ and human limbs $\boldsymbol{E}$. And $N = |\boldsymbol{U}|, M = |\boldsymbol{E}|$ is the number of human joints and human limbs respectively. We represent a pixel coordinate as a two dimension vector $\mathbf{x} \in \{1, \dots H\} \times \{1, \dots W\} \subset R^2$, where H and W are the height and width of input images. We use $p_u(\mathbf{x})$ to denote the likelihood that the image patch centered at pixel coordinate $\mathbf{x}$ belongs to joint $u \in \mathbf{U}$. We use $p_{uv}(\mathbf{x})$ to denote the likelihood for the human limb $uv \in \mathbf{E}$, whose endpoints are joint u and joint v. For simplicity, we use p_u to denote the whole likelihood $p_u(\mathbf{x}), \forall \mathbf{x} \in \{1, \dots H\} \times \{1, \dots W\}$.

In this paper, we attempt to parse human limbs to promote the accuracy of human joints localization, especially the joints with higher flexibility, such as elbows and wrists. In other words, we focus on the detection of arms and legs, which severely affect the precision of joints localization. To improve

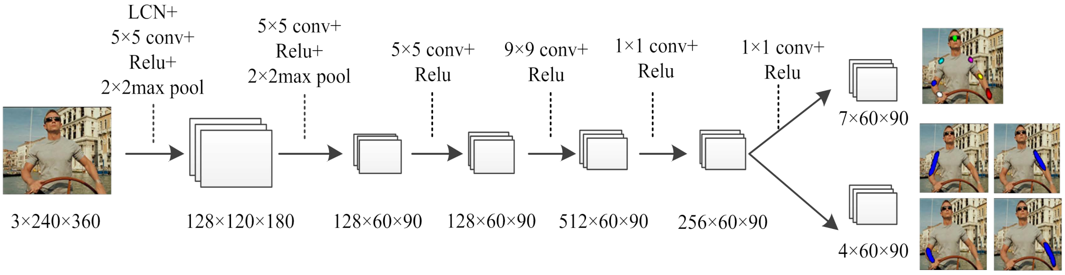


Fig. 2. Joint and limb detector with overlapping contexts (model used on Flic dataset)

the robustness of detection to limbs variations, we divide a human limb into two small straight limbs at an elbow or a knee joint. In other words, we use lower arm, upper arm, lower leg and upper leg for whole body detection. Compared with whole limbs, the shape of human limbs within this definition are similar to cylinders, as shown in Fig. 2. Because of less variation of limb's shape, it is much easier to detect them than to detect the whole limbs.

Like most of the methods for pose estimation, our method consists of two stages. In the first stage, a ConvNet is used to compute $N + M$ heat-maps for N human joints and M human limbs. In the second stage, we employ a graphical model to remove the false positive detection, which is implemented by another ConvNet. Finally, we connect them and combine the information both from shallow layer and deep layer.

A. Human joints and limbs detector

Generally, human joints are in the proximity of limbs. So they share similar contextual information, such as similar color and texture, which constitutes the constraints for their localization. Considering this close relevance between joints and limbs, we combine their detection in one ConvNet. In this ConvNet, several layers are shared to extract common features for these two tasks. After obtaining the same features, several individual layers are used to seek their unique properties to distinguish limbs from joints. This design takes full advantage of the close relevance as well as their individual property.

Due to the excellent performance and high efficiency of Tompson's model [20], we use a similar ConvNet. Different from [20], we add a new branch in the last layer of the ConvNet. This branch is to search the unique property of human limbs. Moreover, we use one resolution bank rather than multiple resolution banks. The resulting ConvNet for FLIC dataset is shown in Fig. 2. The network for LSP just varies in the size of input images and output heat-maps. This ConvNet in Fig. 2 takes a RGB image as input and outputs $N + M$ heat-maps. The heat-map of the joint u and a limb $uv \in \mathbf{E}$ describes its per-pixel likelihood p_u, p_{uv} separately. Due to the presence of two pooling layers, the resolution of heat maps is a quarter of that of input images. Firstly, an image gets through the local contrast normalization (LCN) layer. Then, the LCN image is input to six convolution layers and two max pooling layers. Each of the last two convolution layers in the network simulates a fully-connected layer for a target input patch size, which is typically a much smaller context than the input image. Refer to [20] for more details.

We train the model in Fig. 2 by minimizing the Mean Square Error (MSE) distance between the predicted heat-maps and the ground truth heat-maps for all joints and limbs. The ground truth heat-map for each joint is a 2D Gaussian with a constant variance (σ = 1.5 pixels) and mean centered at the ground-truth location. However, the ground truth locations of limbs are not annotated in most of the datasets for human pose estimation. It is time-consuming and difficult to manually annotate human limbs accurately, since edges of limbs vary tremendously. Nevertheless, coarse labels of human limbs can be generated according to the location of joints. We use the following procedure to create these labels. For each limb, we start with a zero matrix, whose size is equal to the size of original images. Then we assign a fixed value to each pixel in the straight line, which connects the endpoints of the limb. Finally, this matrix is smoothed by a Gaussian filter. We regard the resulting matrix as the ground truth of this limb. This procedure is too simple to obtain accurate annotation, but it provides enough information where human limbs located. Our experiment shows it works well.

After obtaining all the ground truth heat-maps, we define the MSE distance as

$$E = \frac{1}{N}\sum_{i=1}^{N}\sum_{\mathbf{x}}\left\|\hat{p}_i(\mathbf{x}) - p_i(\mathbf{x})\right\|^2 + \frac{1}{M}\sum_{k=1}^{M}\sum_{\mathbf{x}}\left\|\hat{p}_k(\mathbf{x}) - p_k(\mathbf{x})\right\|^2 \quad (1)$$

where $\hat{p}_i(\mathbf{x})$ and $p_i(\mathbf{x})$ are the predicted and ground truth heat-maps at coordinate $\mathbf{x}$ respectively for the ith joint, $\hat{p}_k(\mathbf{x})$ and $p_k(\mathbf{x})$ for the kth limb similarly. In our experiments, the different between the MSE value of human limbs and joints can be adjusted automatically. So we use 1 for the balancing coefficient.

To train this ConvNet, we perform standard batched Stochastic Gradient Descent (SGD). We use Nesterov momentum as well as RMSPROP [22] to accelerate the leaning. L2 regulation and dropout on the input to each of 1×1 convolution layers are employed to reduce over-fitting. In the training, we also perform random perturbations of the input images (random flipping, rotating and scaling the images) to increase generalization performance. In the experiment, we found this random perturbation can improve the performance significantly by 2-5%, more in small distance error.

B. Graphical Model

We have employed a ConvNet accomplishing the detection of human joints and limb. In the experiment results, however, the predicted heat-maps still contain several unreasonable

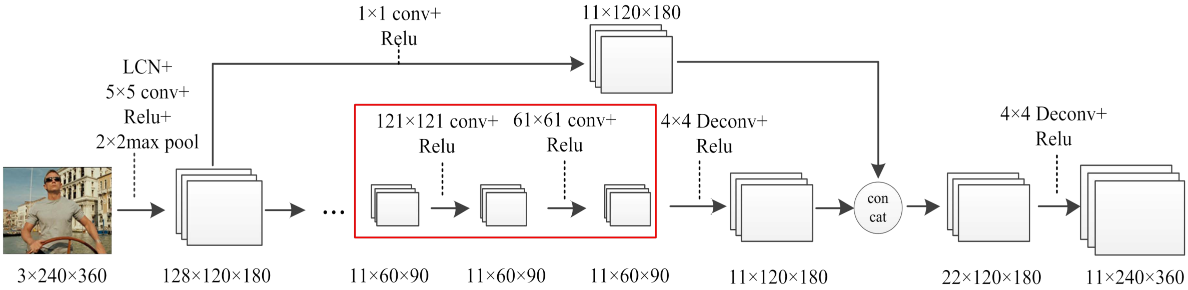


Fig. 3. Our full model for flic dataset

poses violating the constraints among limbs and joints, especially the heat-maps for wrist and elbow with higher flexibility. For example, the wrist joint lies far from the lower arm. This means the relations among joints and limbs are not completely captured by our detection model. According to our ConvNet in Fig. 1, the probability in each coordinate depends on a 64×64 local image region, which is not large enough to capture the whole body. As a result, an explicit model is needed for capturing these relations.

We employ a graphical model to capture these relations and integrate the detection results of human limbs and joints. For each joint or limb, we construct its probability at location $\mathbf{x}$ based on its score as well as the probability of other joints and limbs around $\mathbf{x}$. The filtered probability $\bar{p}_u$ of joint u at coordinate $\mathbf{x}$ is defined as

$$\bar{p}_u(\mathbf{x}) = \frac{1}{Z}\left(\sum_{v\in\mathbf{U}}\sum_{\mathbf{m}} p_v(\mathbf{x}+\mathbf{m})p_{u|v}(\mathbf{m}) + \sum_{ev\in E}\sum_{\mathbf{m}} p_{ev}(\mathbf{x}+\mathbf{m})p_{u|ev}(\mathbf{m})\right) \quad (2)$$

where $\mathbf{m} \in \{1,\dots K\} \times \{1,\dots K\} \subset \mathrm{R}^2$ represents a pixel coordinate, K is the size of the condition prior, p_v, p_{ev} are unary likelihood for arbitrary joint v and limb ev, $p_{u|v}(\mathbf{m})$ represents the conditional prior that joint u is in the location $\mathbf{m}$ when the joint v is in the center, similar for $p_{u|ev}(\mathbf{m})$ and Z is the partition function. According to [20], we can omit the partition function. These condition priors, which are learned from the dataset directly, model the relations among joints and limbs. The filtered probability for limbs is constructed similarly.

By using the matrix representation, (2) can be rewritten as

$$\bar{p}_u = \sum_{v\in\mathbf{U}} p_v * p_{u|v} + \sum_{ev\in E} p_{ev} * p_{u|ev} \quad (3)$$

where $*$ denotes convolution operation. So we can implement our graphical model in (3) as one convolution layer, whose kernel size is much larger than normal convolution layer to capture the whole image. However, convolution layer with large kernel will ignore tiny and local constraints, leading to inaccurate localization. To solve this problem, we cascade multiple convolution layers, whose kernel size are reduced gradually. In this way, each convolution layer operates on smaller and smaller image region. Thus, the joints relationship is modeled more and more accurately. Currently, we use two convolution layers for balancing the performance and efficiency. The part of Fig. 3 contained in the red box is the implementation of our graphical model. We train this graphical model by minimizing the MSE between the ground truth and the filtered heat-maps.

Compared with the graphical model in [20], our model has three differences. First, we add the detection results of human limbs to the graphical model, which can produce tighter constraints for the location of joints and limbs. Second, we cascade multiple convolution layers, whose kernel size are reduced gradually to model the relationship more and more accurate. Finally, we use sum instead of the product over different joints and limbs in the construction of filtered prediction. The sum operation makes the training easier to converge. Since sum operation is less sensitive to its factor, it can better represent the different role of human joints.

As shown in Fig. 3, we connect the network in Fig. 2 and graphical model into a unified model, where we combine information from shallow layer and deep layer by concatenating the two maps. Through this operation, more details in high resolution are introduced into the final features. This ConvNet works in higher resolution, which is better for accurate localization. For clear show, some middle layers of detection model are omitted and the first 60×90 layer is the final result of the detection model. To train the full network, we train the network in Fig. 2 and store the generated heat-maps of the training images. Then, we train the graphical model. Its input is results of the detection model and torso locations. The labels are the same as the first step. Finally, we combine these two networks and retrain them jointly, with parameters initialized by the parameters obtained in the last two training steps. Finally, we fine tune the full model in Fig. 3.

IV. Experiments

A. Dataset and Evaluation Metric

We evaluate the proposed method on FLIC dataset [11] and LSP dataset [12, 13], both of which consist of still RGB images with 2D ground-truth joint location. While most people in FLIC are front-facing standing up, human poses in LSP are much more various. Many images in the FLIC contain multiple persons, while only one is annotated. Therefore, an approximate torso bounding box is provided for the single labeled person in the scene. We incorporate this data by including an extra “torso-joint” to the input of the graphical model so that it can learn to select the correct feature activations in a cluttered scene.

For performance evaluation, we use the Percentage of Detected Joints (PDJ) suggested by Sapp et al. [11]. PDJ measures the performance using a curve of the percentage of correctly localized joints by varying localization precision. For fair comparison with prior works [17, 20], we use observer-

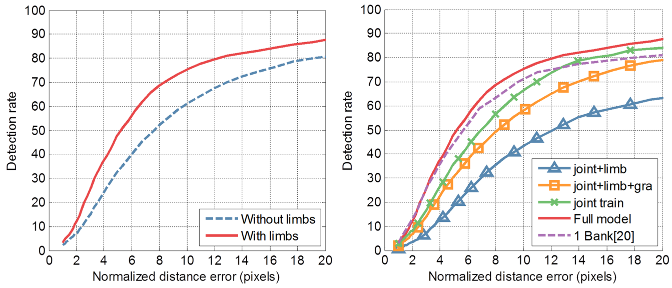
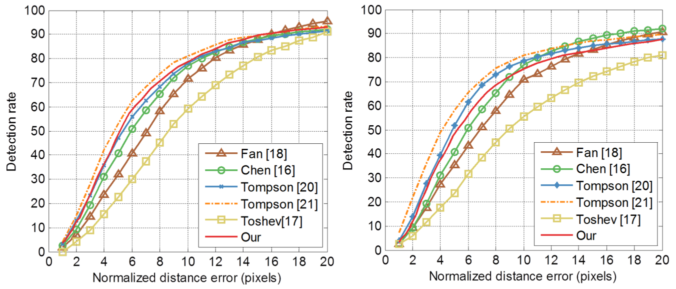


Fig. 4. Performance comparison on wrist joint of Flic dataset



(a) FLIC: Elbow (b) FLIC: Wrist

Fig. 5. PDJ comparison on FLIC dataset

centric coordinate (OC) on FLIC dataset and person-centric coordinates (PC) on LSP dataset.

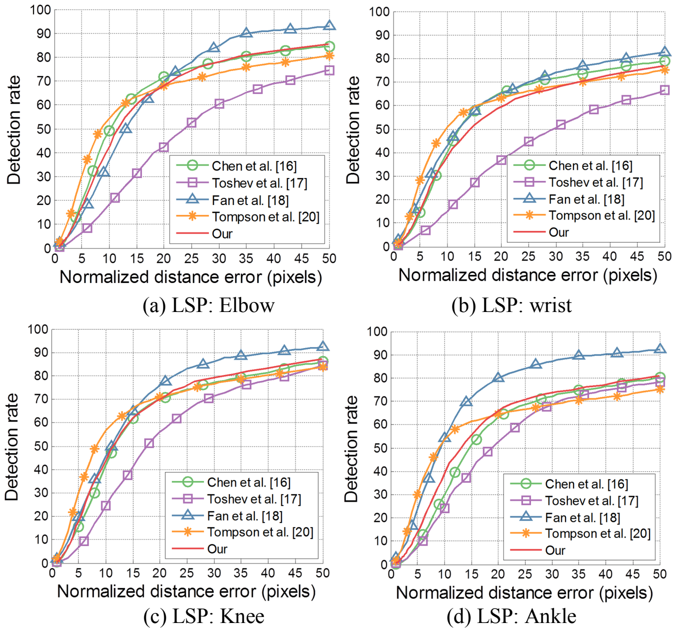
B. *Implementation detail*

For the FLIC dataset, we discard the left hip joint. Because of the tiny distance among the nose and eyes location, we use their average location as a face joint. As a result, we use 7 joints for the FLIC dataset. Besides, the image resolution in FLIC dataset is 480×720. If they are directly passed into the ConvNet, too large memory will be consumed. So we resize them to 240×360. The estimated coordinates are multiplied by 2 to calculate the PDJ curve in original resolution. For LSP dataset, we use all the 14 joints. For the joint without annotation, we use a ground truth heat-map with all zero value. Since the size of images in LSP dataset varies a lot, we crop or pad them into 256×256, then pass them into the ConvNet.

We implement our model with the open-source CNN library Caffe [23]. For layers which are not contained in current Caffe, like LCN layer and data layer with random perturbation, we implement these layers and will release them in the future. In training, all the convolution weights are initialized randomly and the learning rate are 4×10^{-4} and 5×10^{-5} for FLIC and LSP respectively. According to equation (1) in [24], the time complexity of our ConvNet is 5.3×10^{10}, little less than 5.7×10^{10} in [20]. On a 2 CPU workstation with a NVIDIA Tesla K40m GPU, training part detectors takes approximately 48 hours, the graphical model 24 hours, full model 10 hours. The test for a single image takes 136ms for Flic dataset.

C. *Experimental Results*

First, we illustrate the impact of adding human limbs on human pose estimation. We train a new model without human limbs. The experiment results of them on FLIC dataset are shown in Fig. 4(a). Compared with the new model, the original model achieves better performance. The performance increase



(a) LSP: Elbow (b) LSP: wrist

(c) LSP: Knee (d) LSP: Ankle

Fig. 6. PDJ comparison on LSP. Note that [16] and [18] use OC coordinate

is about 5%-10%, larger in smaller distance error. This proves that human limbs play important role in human pose estimation. In Fig. 4(b), we show the performance of models with different components. The "joint+limb" is the result of our detector. The "joint+limb+gra" represents the result of our graphical model. As expected, our graphical model increases the performance by 5% to 15%. The "joint train" represents a model without the combination of shallow information and deep information. The "Full model" represents the complete model. These two curves illustrate combination of shallow and deep information adds an additional 2-10% detection rate. Besides, we compare our model with the 1 resolution bank model [20]. Our model outperforms the 1 resolution bank model even without shallow information.

Then, we compare our proposed method with several state of the art methods [16]-[18], [20], [21] on FLIC dataset. The PDJ curves of them at the elbows and wrists on FLIC dataset are shown in Fig. 5. From this figure, we can see that an improvement is obtained over the baseline [20], especially in large normalized distance error. What's more, we use 1 resolution banks rather than 3 banks [20]. This owes to that new constraints added by limbs can remove the remote false detections. In Fig. 5, our method surpasses all the methods except [21]. Their model can be cascaded with our model, which will improve the performance in low distance error.

Fig. 6 shows the PDJ curves of our method and [16]-[18], [20] at elbow, wrist, knee and ankle on LSP dataset. We can see that the proposed method outperforms all the comparison methods except Fan et al. [18]. Note that [16] and [18] use observer-centric coordinate, which leads to better results. The PDJ gain of the proposed model over Tompson et al. [20] in LSP is larger than that in FLIC. This shows that human limbs play a more important role in complex pose.

Finally, sample human pose estimation results on FLIC and LSP test-sets are shown in Fig. 7. Our model produces accurate

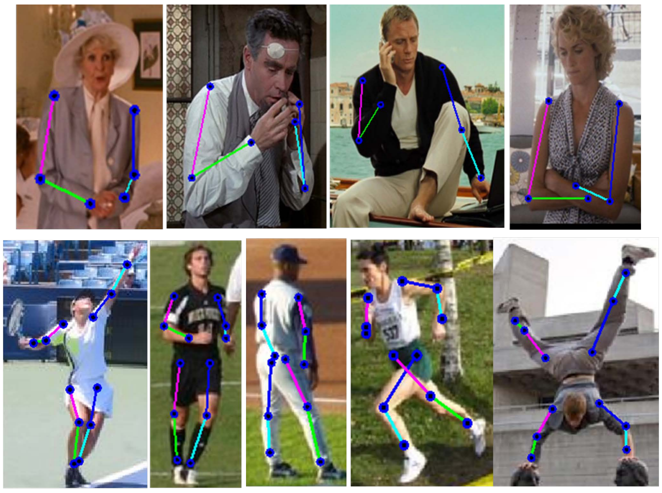


Fig. 7. our model's predicted joint location, Top Row: FLIC Test-Set, Bottom Row: LSP Test-Set

localization for all human joints, even though the pose is rare, like the last image of Fig. 7. In all experiments, our method improves the performance largely in big distance error. In failure cases, our predicted locations are in the limbs region instead of the background. We ascribe this to the use of limbs and a more accurate graphical model. The performance of the proposed method can be further improved using the techniques from other state-of-the-art models, such as fine tuning AlexNet in [18], multiple conditional priors, or image pyramid in [20].

V. Conclusion

This paper proposes a shared convolutional network to integrate human limbs detection and pose estimation to take advantage of the close relevance between these two tasks. After obtained coarse detection results, a graphical model is used to capture the relations among human joints and limbs, which is implemented by convolution layers. Moreover, we combine information from both a shallow layer and a deep layer to utilize the information from different resolutions. By testing on the FLIC and LSP dataset, our method significantly outperforms the baseline [20] in big distance error, especially on LSP dataset. This shows the effectiveness of limb-based pose detection deep learning model.

Although our graphical model can remove some false positive, it is still too simple to capture the variety of relationship among human joints and limbs. In the future, we expect to further improve the performance by designing more accurate model, like the mixture of relations model as in [8]. Besides, we can use much deeper ConvNet, like VGG16, to replace the simple ConvNet.

Acknowledgment

This work was supported in part by the National Key Research and Development Program of China under grant No. 2016YFB1000903, and NSFC No. 61573268.

References

[1] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," Int. J. Comput. Vis., vol. 61, no. 1, pp. 55–79, 2005.

[2] M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2009, pp. 1014–1021.

[3] S. Ioffe and D. A. Forsyth. "Probabilistic methods for finding people," Int. J. Comput. Vis., vol. 43, no. 1, pp. 45-68, 2001.

[4] X. Ren, A. C. Berg and J. Malik. "Recovering human body configurations using pairwise constraints between parts." In Proc IEEE Int'l Conf. Computer Vision, 2005, pp. 824-831.

[5] S. Johnson, M. Everingham. "Combining discriminative appearance and segmentation cues for articulated human pose estimation," In Proc IEEE Int'l Conf. Computer Vision Workshops, 2009, pp. 405-412.

[6] M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Ferrari, "2d articulated human pose estimation and retrieval in (almost) unconstrained still images," Int. J. Comput. Vis., vol. 99, no. 2, pp. 190–214, 2012.

[7] B. Sapp, D. Weiss, B. Taskar. "Parsing Human Motion with Stretchable Models," in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2011, pp. 1281-1288.

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

[9] F. Wang and Y. Li, "Beyond physical connections: Tree models in human pose estimation," in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2013, pp. 596–603.

[10] M. Dantone, J. Gall, C. Leistner, and L. Van Gool. "Body parts dependent joint regressors for human pose estimation in still Images," IEEE Trans. Pattern Anal. Mach. Intell., vol.36, no. 11, pp. 2131–2143, Nov. 2013.

[11] B. Sapp and B. Taskar, "Modec: Multimodal decomposable models for human pose estimation," in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2013, pp 3647-3681.

[12] S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation," BMVC 2010.

[13] S. Johnson and M. Everingham, "Learning Effective Human Pose Estimation from Inaccurate Annotation," in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2011, pp. 1465–1472.

[14] T. B. Moeslund, A. Hilton, V. Kruger, and L. Sigal, Visual Analysis of Humans: Looking at People. New York, NY, USA: Springer, 2011.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in NIPS, 2012.

[16] X. Chen and A. Yuille, "Articulated pose estimation by a g raphical model with image dependent pairwise relations," in Advances in neural information processing systems, 2014, pp 1097-1105.

[17] A. Toshev and C. Szegedy, "Deeppose: human pose estimation via deep neural networks," in CVPR, 2014, pp 1653-1660.

[18] X. Fan, K. Zheng, Y. Lin, S. Wang, "Combining local appearance and holistic view: dual-source deep neural networks for human pose estimation," in CVPR, 2015, pp. 1347-1355.

[19] A. Jain, J. Tompson, M. Andriluka, G. Taylor, and C. Bregler, "Learning human pose estimation features with convolutional networks," in International Conference on Learning Representations, 2014, pp. 1-14.

[20] J. Tompson, A. Jain, Y. LeCun, and C. Bregler. "Join training of a convolutional network and a graphical model for human pose estimation," in NIPS, 2014, pp. 1799-1807.

[21] J. Tompson, R. Goroshin, A. Jain, Y. LenCun, C. Bregler, "Efficient Object Localization Using Convolutional Networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2015, pp. 648-656..

[22] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," COURSERA: Neural Networks for Machine Learning 4.2, 2012.

[23] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, et al., "Caffe: Convolutional architecture for fast feature embedding," in Proc. ACM International conference on Multimedia, 2014, pp 675-678.

[24] K. He, J. Sun. "Convolutional Neural Networks at Constrained Time Cost," in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2015, pp 5353-5360.