ADORE: An Adaptive Holons Representation Framework for Human Pose Estimation

Le Dong, *Member, IEEE,* Xiuyuan Chen, Ran Wang,
Qianni Zhang and Ebroul Izquierdo, *Senior Member, IEEE*

Abstract—In this paper, the problem of human pose estimation in a 2D still image is addressed. A framework called ADORE (Adaptive Holons Representation) adaptively taking advantages of local and global cues is proposed to improve the pose estimation accuracy. In particular, ADORE is made up of two components: 1) the holons part, Independent Losses Pose Nets (ILPNs) is designed to first infer joints location on the global level; 2) the adaptive part, Convolutional Local Detectors (CLDs) is proposed to subsequently detect the joints in the potential regions generated by ILPN. Pose estimation is formulated as a classification problem towards body joints in ILPN which consists of two independent loss layers that respectively instruct the learning of x and y coordinates of a joint. Experimental results on two challenging benchmark tasks demonstrate that our proposed framework is more efficient than other deep models while remains desirable performance.

Index Terms—pose estimation, convolutional neural network, holons, adaptivity

1 Introduction

HUMAN pose estimation is the task of locating the human joints or body parts in an image. Occlusions, appearance variations, in-plane and out-plane rotations, small and barely visible joints make human pose estimation long be a significant challenge in computer vision. The dominant approaches to human pose estimation are based on Pictorial Structures [1], [2] which represents an object as a collection of parts with spatial constraints in between. The Pictorial Structures based methods [3], [4], [5], [6], [7], [8] commonly adopt the local detector to accomplish pose estimation. These local methods have been successfully applied to articulated pose estimation. However, the local detector has to passively scan the whole image to search the possible locations for a particular joint, and this procedure is obviously inefficient. Besides, local detection usually results in many false positives which require predefined spatial constraints to percolate a reasonable pose, while the design of the spatial constraints largely depends on individuals' experience. Recent work DeepPose [9] and DS-CNN [10] formulate the pose estimation as a regression problem and successfully applies DNN (Deep Neural Network) to provide holistc reasoning. However the direct mapping from image to body pose vector is highly non-linear, which thus makes them suffer from difficulty in training and inaccuracy in the high-precision region.

Inspired from the concept of "Holons" that Ken Wilber described in [11]. "Holons" are autonomous, self-reliant units that possess a degree of independence and handle contingencies without asking higher authorities for instructions. These holons are also simultaneously subject to control from one or more of these higher authorities. The first property ensures that holons are stable forms that are able to withstand disturbances, while the latter property signifies that they are intermediate forms, providing a context for the proper functionality for the larger whole. This paper proposes an Adaptive Holons Representation (ADORE) for human pose estimation. Here, ADORE contains two aspects:

1) ADORE is a holons model, i.e., each submodel independently works while their integration plays a new role. For example, in Figure 1, ILPNs locate joints independently while their integration forms a complete holons pose. At the same time, we represent the location of each joint by coordinate x and y, which share the same convolutional layers and are independent of one another at fully-connected layers.
2) ADORE is adaptive, i,e, the model can be adaptively embedded with some other complementary models. For example, in Figure 1, local detectors can be adaptively embedded in the ADORE framework to refine

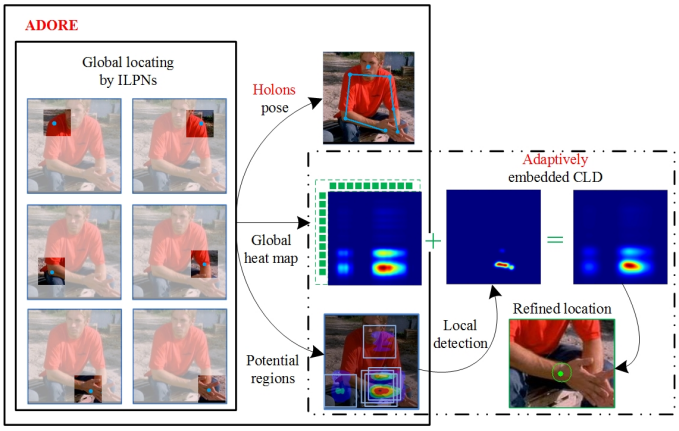


Fig. 1. ADORE via the fusion of ILPN and CLD.

- *L. Dong, X. Chen, R. Wang, are with the University of Electronic Science and Technology of China, Chengdu 611731, China. E-mail: ledong@uestc.edu.cn.*
- *E. Izquierdo is with the Queen Mary, University of London, London E1 4NS, U.K.*

the holons pose.

Each part of ADORE is a self-organized system for low-level tasks which can be processed efficiently. Besides, ADORE is an open system which can conveniently accommodate other complementary models. Current commonly adopted local detectors can work independently, whereas they do not meet the definition of ADORE. For each local detector, an additional spatial model is necessary to provide the global spatial information. Hence, from this perspective, local detectors can not even be holons.

With the elegant "Holons" design, ADORE successfully overcomes the restrictions that previous detection based models [4], [8], [12], [13] and DeepPose [9] suffer from. Specifically, ADORE consists of two components: the holons part, ILPN (Independent Losses Pose Net) for a global estimation and the adaptive part, CLD (Convolutional Local Detector) for a local complement. It should be pointed out that ILPN is the heart of ADORE and provides the framework the holons-ability and adaptability, while the CLD may be replaced with any other local detection models.

The mechanism of ADORE is presented in Figure 1. In ADORE, we do not need to artificially stipulate the spatial constraints, ILPN can directly learn the spatial constraints, instead.

ILPN formulates pose estimation as a classification problem which is easier to be solved than regression formulation proposed by DeepPose [9] (Section 4.2 introduces the details about classification vs. regression problem). ILPN is a DNN model in which the output layer is split into two independent parts and each part is followed by a softmax loss layer. The two independent loss layers separately instruct one joint's x and y coordinates learning. The output of ILPN forms a heat-map indicating the possibility distribution of the joint position in the image. Instead of directly mapping an image to body pose vector like DeepPose [9], one ILPN is trained for one joint, and the human pose is estimated by multiple ILPNs. As ILPN taking the whole image as input, it does make use of the global information of the whole image to infer the joint location. Experimental results demonstrate that ILPN achieves higher precision and easier to train than DeepPose [9].

In addition to ILPN, we subsequently employ CLD to refine the decision on location. Instead of applying inefficient sliding windows to search the whole image as other detection methods [12], [13] commonly do, we use the heat-map produced by ILPN as a filter to reduce the search space and hence accelerate the detection. Then, the heat-map generated by CLD will be fused with the one produced by ILPN to compute the final joint location.

Experiments on two benchmarks demonstrate that ILPN alone can provide desirable performance, and ADORE achieves more competitive performance due to the adaptive training of CLD. Generally, the contributions of this paper are three folds:

1) We formulate the pose estimation as a classification problem and separately predict each joint location, which makes the task easier to achieve;

2)We do not explicitly design a spatial constraint to express the body parts' relationship. Instead, it is learned from data, which is more adaptive to different image conditions;

3) The ADORE can be duly applied to different tasks. If the task is sensitive to time consumption, the ILPN alone can be competent enough to achieve desirable accuracy efficiently. If high precision is required, the CLD embedded ADORE can be adopted to provide satisfactory performance.

2 Related work

The Pictorial Structures (PS) invented by Fischler and Elschlager [1] is a general and powerful model for object structure representation. Felzenszwalb and Huttenlocher [2] restricted the object representation to be an acyclic multi-part model (or tree model) and provided an efficient algorithm for the energy minimization problem. Since then, a series of object/body parts estimation approaches were developed based on Pictorial Structures. Related studies have been mainly focused on modeling parts appearance and spatial constraints among these parts. For better performance, some works [14], [15], [16] integrated automatic foreground segmentation to pose estimation, while parts-based models [3], [4], [8], [17] trained a discriminative part detector to search the potential location of each part. Parts-based models usually result in many false detections, because each part detector only detects a single body part and hence, the information among multiple parts is lost. Yang and Ramanan [5], [6] proposed a model named mixtures-of-parts which successfully captures the contextual co-occurrence relationship among multiple parts. However, such model requires the tree-structure as the spatial constraint to achieve efficient optimization, which results in limited capacity due to the inherent restriction of the tree-structure. Therefore, non-tree models [18], [19], [20], [21] were designed to handle self-occlusion. Semi-global approaches were proposed to achieve higher order part relationships rather than pairwise spatial constraint. Representative works of such approach includes Armlet [22] for upper body pose estimation and Poselet [7], [8], [23], [24] for full body. Furthermore, Sapp et al. [25] proposed a multimodal decomposable models called MODEC to exploit both local and holistic cues. [26] reformulated human pose estimation to a segmentation-like problem and proposed the fields of parts model in which more local image evidence can be easier added than PS.

Aforementioned works either rely on hand-crafted features or fail to make use of the contextual information. To address such constraints, researchers are concentrated on employing the prevalent deep learning models to improve the pose estimation performance. Bourdev et al. [27] replaced HoG with deep pose representation to get deep Poselets. In [9], DNNs were adopted to achieve both holistic reasoning on full images and localization fine-tuning on sub-images. In [28], a novel joint binary codes learning method is proposed to combine image feature to latent semantic feature with minimum encoding loss. Furthermore, DS-CNN [10] proposed by Fan et al. jointly trained the holistic network and local network to improve the accuracy. However, both the DeepPose and DS-CNN formulate the human pose estimation as a joint regression problem, which is difficult to be done. To solve this problem, this paper formulates pose estimation as classification towards coordinate. Ouyang et al. [29] tried to combine DNNs with

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2707477, IEEE Transactions on Circuits and Systems for Video Technology

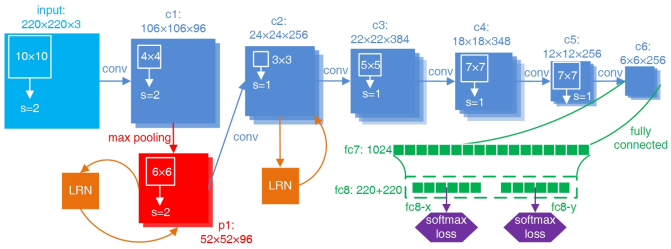


Fig. 2. The architecture of ILPN. An image with 3 color channels is presented as the input, which is convolved with 96 different 1st layer filters (white), each of size 10×10, using stride $s = 2$ in both x and y. The resulting feature maps are then: (i) pooled (max within 4×4 regions, using stride $s = 2$) and (ii) Local Response Normalized (LRN) among each five adjacent feature maps. Similar operations are repeated in layers 2,3,4,5,6. The last two layers are fully-connected, taking features from the top convolutional layer as input. The final layer is two independent softmax loss functions. This figure is best viewed in color.

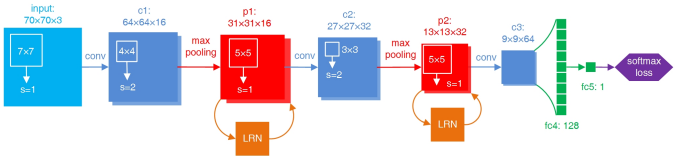


Fig. 3. Formulation of the CLD. An image patch with 3 color channels is presented as the input, which is convolved with 16 different 1st layer filters (white), each of size 7×7, using stride $s = 1$ in both x and y. The resulting feature maps are then: (i) pooled (max within 4×4 regions, using stride $s = 2$) and (ii) Local Response Normalized (LRN) among each five adjacent feature maps. Similar operations are repeated in layers 2,3. The last two layers are fully-connected, taking features from the top convolutional layer as input. The final layer is a softmax loss function. This figure is best viewed in color.

the computer vision prior knowledge, which fed the visual information computed by [6] to a deep network to perform non-linear inference. The methods in [12], [13] combined the convolutional part detectors and spatial model to improve the performance. Chen [30] et al. argued that image dependent relations are helpful to model highly variable poses and proposed a CNN detector model to capture image dependent pairwise relations. In [31], a discriminative representation is proposed by discovering key information of the input data for human action recognition. The method in [32], an effective approach is proposed to decrease collisions in the synopsis through reducing the sizes of moving objects. Yuan [33] et al. propose a novel action recognition method that simultaneously learns middle-level representation and classifier by jointly training a multinomial logistic regression model and a discriminative dictionary. By contrast, our ILPN captures global relations which refer to the spatial relationship between joints on the whole. In order to capture the global relations, we trained ILPN to directly detect the coordinate of joints on whole image. Besides, ILPN was designed with just one pooling layer to decrease the loss of spatial information.

3 Adaptive Holons Representation Framework

A body is regarded as a collection of joints, that is, once the joints' locations are determined, the body pose can be estimated. Here, we use pose vector $L = (\cdots, \boldsymbol{l}_i, \cdots)^{\mathrm{T}}, i \in \{1, \cdots, K\}$ to represent a human pose supported by K joints, where $\boldsymbol{l}_i = (x_i, y_i)$ is the coordinate of the i^{th} joint. A labeled sample is denoted as (D, L), and a estimated sample is denoted as $(D, \hat{L})$, where D is the image data, L and $\hat{L}$ represents the ground truth and estimated pose vector respectively. Specifically, we train a model for each joint, and $\hat{\boldsymbol{l}}_i$ is estimated by the corresponding i^{th} model. All these K models are with the same neural network structure.

3.1 Data augmentation and preprocessing

We take different strategies to preprocess the training data and the test data. In order to enrich the diversity of training samples, data augmentation is applied to the training dataset.

In this work, an image is measured in the one-based coordinates system ordered from left to right (x) and top to bottom (y). An original image is rotated with a random degree, which forces the network to be more robust to in-plane rotation. Then, the annotated human regions are cropped from both the original and rotated images with three bounding boxes of different sizes. A cropped image and its corresponding label are denoted as (D^c, L^c). The coordinate of the i^{th} joint in the cropped image is computed according to Eq. (1):

$$\boldsymbol{l}_i^c = \boldsymbol{l}_i - \boldsymbol{c} \tag{1}$$

where $\boldsymbol{c} \in \mathbb{R}^2$ is the coordinate of the top left point of the bounding box.

Next, the gained 6 images are all resized to $m \times n$ pixels (in this paper, $m = n = 220$) through the bicubic convolution algorithm [34]. A resized image is denoted as (D^r, L^r). Since the cropping sizes are different, the resized images present multi-scale poses. Finally, the resized images are all flipped horizontally. After the data augmentation, the amount of training images is 12 times of the original inputs.

As for the test data, the object is cropped via just one bounding box, and then resized to the same size in the same way as the training data. It deserves to be mentioned that, in test phase, the mapping from D^c to D^r needs to be recorded for subsequent postprocessing:

$$D^r = f(D^c) \tag{2}$$

Contrary to above process, postprocessing reverses the estimated joint to the original image by

$$\begin{aligned} \widehat{\boldsymbol{l}^c} &= f^{-1}(\widehat{\boldsymbol{l}^r}) \\ \widehat{\boldsymbol{l}_i} &= \widehat{\boldsymbol{l}_i^r} + \hat{\boldsymbol{c}} \end{aligned} \tag{3}$$

Note that postprocessing is required in test phase only, because training is based on the resized image coordinate system.

3.2 Independent losses pose net

A DNN with two independent loss layers is designed for globally locating a particular joint, which is called Independent Losses Pose Net (ILPN). Compared with the local detector based methods that train a detector for each part or

joint to search for the object in an image, ILPN makes use of the contextual information of the whole image to decide the location of a joint.

Specifically, ILPN formulates the pose estimation problem as classification towards body joints, i.e., for an image, the position of each joint is a class and the goal is to figure out which class the joint belongs to. Since a position is determined by two coordinates in a 2D image, the classification of a position is transfered to the classification of two coordinates. By contrast, the classification formulation is easier than the regression formulation proposed by DeepPose [9], because classification formulation restricts outputs to be integers in the range of image size, while regression may output any real numbers.

The architecture of ILPN is shown in Figure 2. ILPN takes a 220×220 RGB image and its corresponding joint coordinate as input $(D^r, \boldsymbol{l}_i)$ to output a heat-map for the i^{th} joint. ILPN comprises 8 learnable layers: 6 convolutional layers and 2 fully-connected layers. Each of these layers is composed of a linear matrix or vector multiplication with learned bias, followed by an element-wise Rectified Linear Units (ReLU) [35]. Specially, just as discussed in [12], the pooling operation leads to a loss of spatial precision. Therefore, our ILPN contains only one max pooling layer following the first convolutional layer, which helps to reduce the computational complexity and improve the translation invariance with a little loss of joint information. Meanwhile, in order to improve the generalization [36] of ILPN, we implement the Local Response Normalization (LRN) after the max pooling layer and the second convolutional layer.

In ILPN, we set two fully-connected layers, i.e., $fc7$ (the 7^{th} layer) and $fc8$ (the 8^{th} layer), as illustrated in Figure 2. The output layer $fc8$ is split into $fc8$-x and $fc8$-y, each of which has 220 neurons representing the 220 pixels of the corresponding coordinate x or y. Such special design allows the network to separately compute the possible value of the x and the y coordinates of the joint. In detail, the output value of the n^{th} neuron in the $fc8$-x (or $fc8$-y) indicates the confidence that the model believes a joint's x (or y) coordinate is n. The two neurons with max values are selected from the independent $fc8$-x and $fc8$-y respectively, and is set as the final coordinate of the corresponding joint. Taking the $fc8$-x as an example, the softmax loss layer maps the output of $fc8$-x to a probability distribution using the softmax function:

$$\begin{aligned} p(\boldsymbol{z}) &= \{p_1(\boldsymbol{z}), \cdots, p_n(\boldsymbol{z}), \cdots, p_N(\boldsymbol{z})\} \\ p_n(\boldsymbol{z}) &= \frac{e^{z_n}}{\sum_{j=1}^{N} e^{z_j}} \end{aligned} \tag{4}$$

where $\boldsymbol{z} = (z_1, \cdots, z_N)$ is the output vector of $fc8$-x, N is the neuron number in $fc8$-x, and p_n is the probablity of the joint's x coordinate being n. If the ground truth is k, we need to maximize $p_k(\boldsymbol{z})$, namely, minimize the following loss function:

$$loss(\boldsymbol{z}, k) = -\log p_k(\boldsymbol{z}) = \log\left(\sum_{j=1}^{N} e^{z_j}\right) - z_k \tag{5}$$

To achieve the function minimization, the softmax loss layer should back propagate the gradient of the loss function to $fc8$-x with respect to each neuron:

$$\begin{aligned} \frac{\partial loss(\boldsymbol{z}, k)}{\partial z_n} &= p_k(\boldsymbol{z}) - \delta_{nk} \\ \delta_{nk} &= \begin{cases} 1 & n = k \\ 0 & n \neq k \end{cases} \end{aligned} \tag{6}$$

The softmax loss layer after $fc8$-y does the same operation as that of $fc8$-x independently. Once the two loss layers' gradients are computed and back propagated to $fc8$-x and $fc8$-y, the entire $fc8$ continues the backpropagation as a regular fully-connected layer in a typical DNN.

From the aforementioned formulation of ILPN, we observe that the critical design is the last softmax loss layer. ILPN holds two independent loss layers to separately instruct the learning of the joint's x and y coordinates, with the deductive inference that: 1) the back propagation procedure for learning x and y should be independent, otherwise the incorrect prediction of x may result in penalty on the prediction of y; 2) independent learning x and y improves the generalization ability of ILPN, e.g., adding a softmax loss layer $fc8$-z makes ILPN be able to predict the coordinate in 3D space. Besides, we do not train two distinct models separately for x and y. Because x and y mutually determine the position of a single joint, they share the joint features learned by previous layers. Therefore, only the output layer is divided into two parts and two softmax layers are set for independent training.

3.3 Convolutional local detector

We train local detectors based on the convolutional networks to compensate the global ILPNs to refine the location precision. The convolutional local detector (CLD) shown in Figure 3 consists of one input layer, four hidden layers and a binary classifier as output layer. The input layer reads a 70×70 pixels RGB image patch and its corresponding binary label $l_i = 0$ or $l_i = 1$ as input. The first two hidden layers are the typical patterns used in the convolutional network for feature learning: a convolutional layer followed by a max pooling layer and LRN. The third hidden layer is a convolutional layer. The last hidden layer is a fully-connected layer. All the neurons in the hidden layers use the ReLU as the activation function. The output layer contains only one neuron with a logistic activation function as a simple binary classifier.

Since the local visual cues have rare spatial information, we train a local detector to detect a left joint as well as its corresponding right joint. For example, if we are given a shoulder without any other contexts, it is hard to tell weather this is the left or right shoulder, whereas ILPN can provide such spatial information. The trained CLD is used to detect the potential region computed based on ILPN outputs (see Section 3.4 for details about the potential region). The local detection results form a heat-map indicating the possibility distribution of the joint position in an image. The heat-map produced by the CLD will then be combined with the heat-map generated by ILPN to jointly decide the joint location. It should be pointed out that other local detection methods can similarly cooperate with ILPN to give desirable performance, while our proposed CLD is proven straightforward and powerful.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2707477, IEEE Transactions on Circuits and Systems for Video Technology

3.4 The fusion of ILPN and CLD

The proposed ADORE is flexible to different applications. In the case that time consumption is prior to precision, we use ILPN alone:

$$(\hat{x}_i, \hat{y}_i) = (\arg\max(\boldsymbol{o}_x), \arg\max(\boldsymbol{o}_y)) \tag{7}$$

where $\boldsymbol{o}_x$ is a $n \times 1$ vector produced by $fc8$-x and $\boldsymbol{o}_y$ is a $m \times 1$ vector produced by $fc8$-y, and $\arg\max(\cdot)$ returns the index of the max value of the given vector.

In the scenarios emphasizing precision, ILPN needs CLD cooperating to provide desirable performance. In this case, ILPN should output a heat-map indicating the probability of the joint positions in the image, i.e., ILPN should read a $m \times n$ RGB image (in this paper, $m = n = 220$) to output a $m \times n$ ILPN heat-map $\boldsymbol{A}$:

$$\boldsymbol{A}_{m\times n} = \mathrm{G}(\begin{bmatrix} \boldsymbol{o}_x^T \\ \vdots \\ \boldsymbol{o}_x^T \end{bmatrix} \times [\boldsymbol{o}_y, \cdots, \boldsymbol{o}_y]) \tag{8}$$

where the vector $\boldsymbol{o}_x$ and $\boldsymbol{o}_y$ are both expanded to $m \times n$ matrices via self-replicating m and n times respectively, $\mathrm{G}(\cdot)$ is the Gaussian smooth operation. The example of ILPN heat-map is shown in Figure 1.

After the ILPN heat-map is produced, we first find the max value in the heat-map, and multiply it by a predefined threshold λ (between 0 and 1). The product is used to select the potential pixels of this joint. Specifically, scanning the ILPN heat-map, only those pixels with the values larger than the product are selected as potential pixels of this joint:

$$\begin{aligned} Pr &= \{a_1, \cdots, a_m, \cdots\} \\ a_m &= i \times \mathbb{1}(v_i \geq \lambda \cdot \max(\boldsymbol{A})) \end{aligned} \tag{9}$$

where the potential region Pr is a set of potential pixels, a_m is the m^{th} potential pixel, v_i is the value of i^{th} pixel in ILPN heat-map $\boldsymbol{A}$. The indicator function $\mathbb{1}(\cdot)$ returns 1 if the condition is satisfied, otherwise, returns 0.

Subsequently, the CLD is employed to accomplish joint detection in the potential region. The detection results form a $m \times n$ CLD heat-map of this joint. Finally, the ILPN heat-map and CLD heat-map are fused according to Eq. (10):

$$(\hat{x}_i, \hat{y}_i) = \arg\max(\beta \times \boldsymbol{A} + (1 - \beta) \times \boldsymbol{B}) \tag{10}$$

where $\boldsymbol{B}$ is the CLD heat-map smoothed by the Gaussian filter, $0 \leq \beta \leq 1$ is a super parameter indicating the reliability of ILPN.

Figure 4 shows an examplar of heat maps in ADORE which illustrates the procedure of coarse to fine. The first heat map is ILPN heat map which indicates the location probability of the right wrist in an image. The predicted location (red point) lies outside of the ground truth (red circle). After the ILPN heat map is smoothed, the probability distribution becomes continuous. In the smoothed ILPN heat map, there are two peaks that indicate the high probability locations. The highest peak (left) is the prediction of ILPN, but the lower peak (right) is a better location. CLD scans these high probability regions, and then, the smoothed CLD heat map is fused with smoothed ILPN heat map. In the final ADORE heat map, the highest peak becomes the right one which lies in the ground truth.

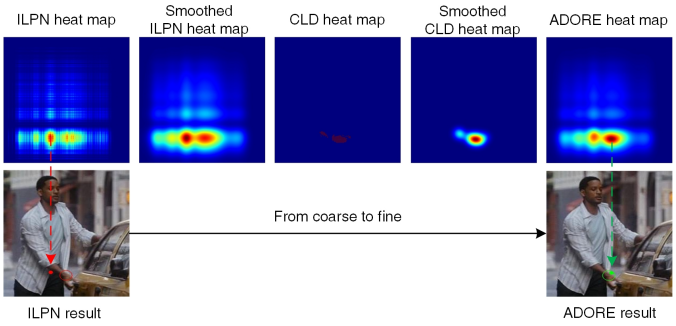


Fig. 4. Heat maps in ADORE. This figure is best viewed in color.

4 Experimental Evaluation

In this section, comprehensive experiments were conducted to evaluate ADORE and ILPN respectively. Particularly, Section 4.1 verifies the performance and adaptive ability of ADORE on the FLIC dataset, and Section 4.2 analyses the design of ILPN, including its generalization ability on the Buffy dataset and its inner structures.

Metric: In all experiments, we use the PDJ (Percent of Detected Joints) metric suggested by [9]: a joint is correctly predicted if the prediction is located within a circle centered at ground truth. The radius of this circle is computed through the torso diameter multiplying by a ratio. We report the percentage of correct predictions in the test-set with respect to different ratios. For a test-set of size M, ratio r and particular joint i, the accuracy is:

$$acc_i(r) = \frac{100}{M} \sum_{t=1}^{M} \mathbb{1}\left(\left\|\widehat{\boldsymbol{l}}_i^t - \boldsymbol{l}_i^t\right\|_2 \leq r \cdot \left\|\boldsymbol{l}_{lhip}^t - \boldsymbol{l}_{rsho}^t\right\|_2\right) \tag{11}$$

where $\widehat{\boldsymbol{l}}_i^t$ is the predicted i^{th} joint location on test sample t. $\boldsymbol{l}_{lhip}$ and $\boldsymbol{l}_{rsho}$ are the ground truth of left hip and right shoulder. We report $acc_i(r)$ for a range of r resulting in a curve that spans both the very tight and very loose regimes of joint localization.

4.1 Evaluation of ADORE

Dataset: the FLIC (Frames Labeled In Cinema) dataset [25] consists of 3987 training images and 1016 test images from Hollywood movies with actors in diverse poses and clothing. Each FLIC image contains more than one person while only one person is labeled. For each labeled human, 11 upper body joints (nose, eyes, shoulders, elbows, wrists, hips) are noted.

Implementation details: In the experiments, the training and testing inputs of the FLIC are all resized to 220×220 pixels. The threshold λ mentioned in Eq. (9) is set as 0.6. The reliability factor β on ILPN is 0.55. Both ILPN and CLD are implemented within Caffe [37] and our code will be publicly available soon. For each joint, it takes 10 hours to train ILPN and 3 hours to train CLD on a Pentium dual-core CPU and a Nvidia GTX750 GPU with 1G RAM.

Results comparison: The detection performance of our models and some other models (Fan et al. [10], Tompson et al. [13], Toshev et al. [9], Jain et al. [12], Sapp et al. [25],

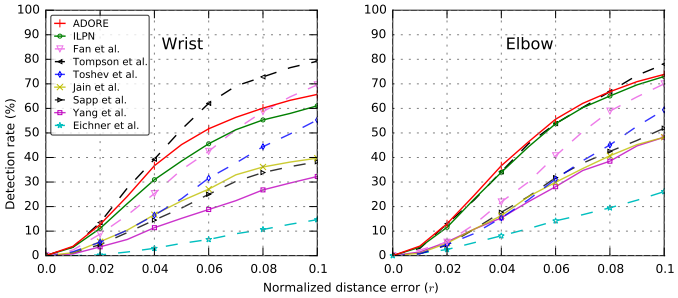


Fig. 5. Percentage of detected joints on the FLIC for two joints: elbow and wrist. This figure is best viewed in color.

Yang et al. [5] and Eichner et al. [16]) on the FLIC test-set. More particular results on elbow and wrist are shown in Figure 5. With our adopted evaluation PDJ (Percent of Detected Joints) metric [9], joint estimation in a smaller distance ratio means getting higher detection precision, which is certainly a harder challenge. At the same time, wrist is harder to detect than elbow because wrist presents higher degree of freedom than elbow. From Figure 5, both ILPN and ADORE achieve competitive results. On comparison of wrist, ILPN achieves similar accuracy with Fan et al. [10] and is obviously superior to Toshev et al. [9], while ADORE is superior to Fan et al. [10] in most cases. Whereas all methods are inferior to Tompson et al. [13]. It is worth noting that Tompson et al. [13] jointly trains the part model and spatial model, and the part model detects by multi-scale sliding windows, while our gobal and local models are trained in parallel, and our accuracy is achieved without multi-scale sliding windows. On comparison of elbow accuracy, both ILPN and ADORE are superior to Fan et al. [10] and Toshev et al. [9] by a wide margin. ILPN shows similar curve with Tompson et al. [13], while ADORE holds a tight lead over Tompson et al. [13]. Furthermore, compared with ILPN, ADORE improves the wrist estimation; while since ILPN already achieves desirable performance on elbow estimation, ADORE shows further enhancement on this joint detection.

Fig. 6. Visualization of pose results in images from the FLIC. Each pose is represented as a stick figure, inferred from predicted joints. Different limbs in the same image are colored differently, and the same limb across different images has the same color.The red nodes represent the location of joints(ground truth). The first two rows are accurate results and the last row presents some false detections. This figure is best viewed in color.

Figure 6 presents some examples of the estimated joints through ADORE on the FLIC dataset(purple box represent the ground truth . From the top two rows, we observe that ADORE gives accurate detections under a variety of conditions. The bottom row gives some false detection examples, which shows that ADORE fails to detect joints blocked by other confusing objects, and it is somewhat difficult for ILPN to differentiate the left and right.

TABLE 1
Detection accuracy on the FLIC for all joints at different normalized distance error.

Ratio	0.02	0.04	0.06	0.08	0.1	Ratio	0.02	0.04	0.06	0.08	0.1
Left wrist						*Right wrist*					
ILPN	12.6	31.9	47.0	57.6	62.8	ILPN	9.6	30.0	44.2	53.0	59.3
Rise	+1.2	+5.7	**+6.1**	+4.4	+4.2	Rise	+2.0	+5.7	**+6.1**	+5.2	+5.1
+Local	13.8	37.6	53.1	62.0	67.0	+Local	11.6	35.7	50.3	58.2	64.4
Left elbow						*Right elbow*					
ILPN	10.8	33.3	52.9	64.0	72.1	ILPN	12.3	34.6	54.6	66.1	73.9
Rise	**+2.1**	+1.9	+1.3	+1.9	+0.8	Rise	+0.8	**+3.5**	+2.4	+1.6	+0.9
+Local	12.9	35.2	54.2	65.9	72.9	+Local	13.1	38.1	57.0	67.7	74.8
Left shoulder						*Right shoulder*					
ILPN	20.8	51.1	71.9	82.6	87.9	ILPN	20.6	56.0	76.4	85.6	90.4
Rise	+0.9	+2.9	**+4.5**	+4.0	+3.1	Rise	**+2.6**	+2.0	+2.5	+1.9	+1.1
+Local	21.7	54.0	76.4	86.6	91.0	+Local	23.2	58.0	78.9	87.5	91.5

Adaptive learning of CLD: Table 1 records the detection accuracy of ILPN and ADORE on the FLIC for six joints at different precisions. It is clearly observed that no matter what the joint it is, CLD brings obvious enhancement to ILPN, especially when the requirement for precision is strict and the joint is hard to detect (e.g., wrist is harder to detect than shoulder because wrist presents higher degree of freedom than shoulder). However, as discussed in Section 3.3, although the adoption of CLD brings complementary performance to ILPN, it compromises the time efficiency. Figure 7 presents some examples of refinement by CLDs based on ILPNs' results. From the Figure 7 we observe that ILPNs provide a quality initial pose, and CLDs further improve the precision based on the initial pose.

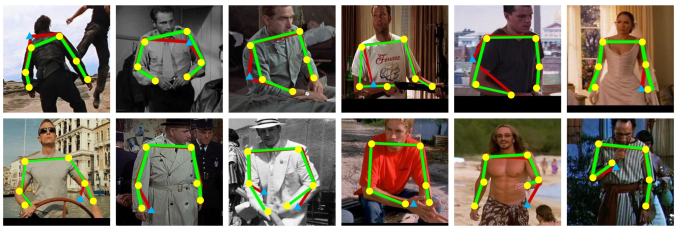


Fig. 7. Refinement by CLD. The yellow nodes represent the location of joints(ground truth). The blue triangles repersent the wrong predictions from ILPN. The red lines represent the predictions of ILPN, while green lines represent the predictions of ILPN+CLD. This figure is best viewed in color.

Parameters discussion: ADORE mainly contains three critical parameters: $\mathcal{N}(0, \sigma)$ for heat map smoothing in Equation 8, potential region threshold λ in Equation 9, and ILPN reliability parameter β in Equation 10. In the previous experiments, σ is set as 6 based on validation set. In order to investigate the effect of σ, we choose a set of values for σ, while keeping other parameters fixed and test in FLIC test set. The result is shown in Figure 8. Without Gaussian

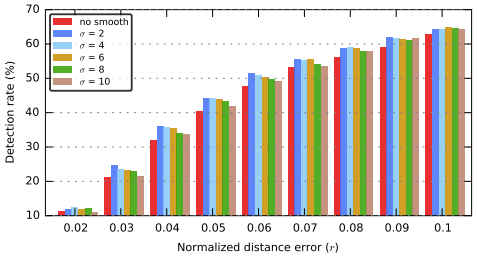


Fig. 8. Detection rate of right wrist on the FLIC with different σ. This figure is best viewed in color.

smooth, ADORE gives the worst performance. When σ is set as 2 or 4 or 6, ADORE achieves similar accuracy , which is higher than the situation when σ is set as 8 or 10.

For detection efficiency, in the previous experiments the threshold λ is set as 0.6, which remains less than 500 pixels in most cases. For example, when computing the potential pixels of the right wrist joint on 1016 test images, 357 pixels are selected as potential pixels in average for an image. Compared with the whole image space consisting of 48400 pixels, our potential search space is reduced 99.3%.

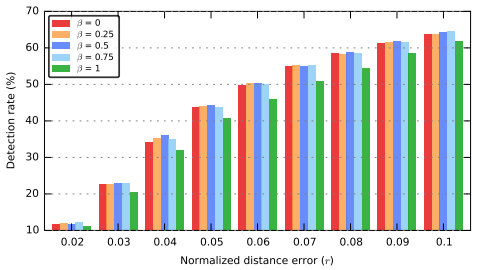


Fig. 9. Detection rate of right wrist on the FLIC with different β. This figure is best viewed in color.

Based on the somewhat subjective consideration that ILPN is a little more important than CLD, we set the reliability coefficient β in Eq. (10) as 0.55 in aforementioned experiments. To investigate how the parameter β affects the detection results, we conducted related experiments on the right wrist in the FLIC with different β and keeping other settings fixed. The result is shown in Figure 9. We observe that when $\beta = 1$ (i.e., CLD is totally abandoned and ILPN alone works), ADORE achieves the lowest accuracy, while when CLD is adaptively adopted, the estimation performance improves. However, we note that the reliability factor β shows subtle effects on the estimation performance. Recall that local detection is based on the results of ILPN ($\beta = 1$), which means that CLD already contains some contributions of ILPN, and hence, adjusting the reliability coefficient β between CLD and ILPN gives subtle performance difference. In total, when β is set around 0.5, the framework commonly gives the best performance.

4.2 Evaluation of ILPN

Time consumption: Table 2 compares ILPNs with [9] and [13] of the training time on FILC dataset and the running time of detecting all joints per image simultaneously. It

TABLE 2
Time consumption comparison

	ILPN	DeepPose	Tompson et al.	ADORE
device	2 core CPU + GTX750 GPU	12 core CPU	12 core CPU + Nvidia Titan X GPU	2 core CPU + GTX750 GPU
number of parameters	1.64 x 10^8	5.48 x 10^8	-	-
time of multiplication	1.58 x 10^8	7.41 x 10^8	-	-
training time	45 hours	17 days	420 hours	45hours + 12hours*joints number
running time	450ms	100ms	357ms	450ms+230ms

TABLE 3
Detection accuracy on the MPII for all joints, especially, the accuracy of shoulder is the average of left and right. The same as elbow, wrist, hip, knee and ankle.

Joints	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle
ILPN	67.4	60.8	44.7	32.1	44.8	40.8	33.4
Tompson *et.al.* [38]	**96.1**	**91.9**	**83.9**	**77.8**	**80.9**	**72.3**	**64.8**
Yang *et.al.* [6]	73.2	56.2	41.3	32.1	36.2	33.2	34.5
Pishchulin *et.al.* [8]	74.2	49.0	40.0	34.1	36.5	34.4	35.1
Gkioxari *et.al.* [22]	-	36.3	26.1	15.3	-	-	-
Sapp *et.al.* [25]	-	38.0	26.3	19.3	-	-	-

should pointed out that eventhough these three methods are tested on different hardwares, this comparison does make sense. Generally speaking, Nvidia Titan X GPU is much powerful than GTX750 (please refer to supplementary material for the detail GPU capability comparison). However, from Table 2 and Figure 5 we observed that ILPN achieves a desirable accuracy by training only 45 hours on 2 core cup + entry-level GPU, while DeepPose costs 17 days on 12 core CPU and [13] costs 420 hours on 12 core CPU and Nvidia Titan X GPU. Besides, the running time of ILPN is close to [13] even tests on the less powerful device. From a more intuitive perspective, the number of parameters and the times of multiplication in DeepPose are about 3.3 times and 4.7 times stronger than ILPN. But the training time and running time of ILPN are less than the corresponding proportion. It is generally considered that deep learning models rely heavily on powerful hardware, and our works push forward deep learning on the direction of resource-saving computation.

Results on MPII: MPII Human Pose dataset includes around $25K$ images containing over $40K$ people with annotated body joints. Each image was extracted from a YouTube video and provided with preceding and following un-annotated frames.

Considering that ILPN exploits global information of inputs, in our experiments, we make some simple processes with images of MPII. Specifically, from MPII, we randomly choose 10500 images containing all the 14 joints as training

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2017.2707477, IEEE Transactions on Circuits and Systems for Video Technology

set and select 1000 images containing joints location information form test set. And we get the final inputs of ILPN by clipping images of training set and test set according to the joints' location information. As to the evaluation on MPII, we adopt the PCKh metric proposed by [39].

In the experiments, the inputs of ILPN are all resized to 180*260 pixels. Considering the complexity of MPII dataset, we adjust the ILPN described in Section 3.2 to a network with eight convolutional layers, but retain the crucial independent classification outputs and losses structures. Moreover, we train all the 14 joints at a time, and this process takes about 70 hours on a dual-core CPU and GTX750 GPU.

The detection accuracies of ILPN and models (Tompson *et.al.* [38], Yang *et al.* [6], Pishchulin *et al.* [8], Gkioxari *et al.* [22], Sapp *et al.* [25]) on our test set for all joints are declared in Table 3. Apparently, compared with other models except [38] which is significant complex than our model, ILPN achieves better accuracies on most joints except head, wrist and ankle subtly inferior to Pishchulin. This is because ILPN uses only global information while Pishchulin utilizes rich local messages, which demonstrates the global features and local features have different application scenarios. Specifically, global features have advantages on joints that near the center of the body, *e.g.* in our experiments on MPII, the detection accuracy of ILPN on neck (77.2) achieve the better performance than head.

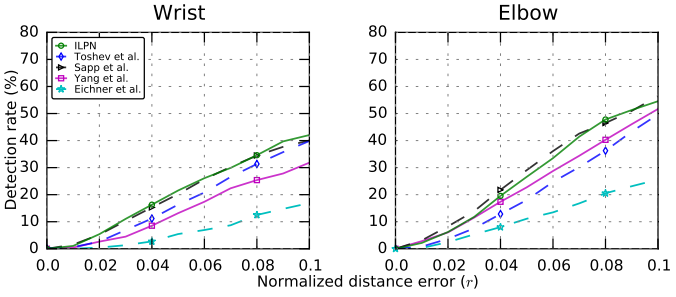


Fig. 10. Percentage of detected joints on the Buffy dataset for two joints: wrist and elbow. The ILPNs are trained on the FLIC dataset and tested on the Buffy dataset. This figure is best viewed in color.

Cross-dataset generalization: To evaluate the generalization properties of ILPN, we directly apply the ILPNs trained on the FLIC to the test portion of the the Buffy stickmen dataset [14] without fine-tuning. In each Buffy image, persons appear at a variety of scales, against highly cluttered background, and wear a variety of clothing. Figure 10 presents the detection rate of the trained ILPN on the Buffy test images. We compare ILPN with four other models (Toshev et al. [9], Sapp et al. [25], Yang et al. [5] and Eichner et al. [16]), and the results show that our approach can retain competitive performance, which demonstrates that ILPN has good generalization ability.

Classification vs. regression: In ILPN, pose estimation is implemented in the manner of classification rather than regression. Classification is able to constrain the output of ILPN in a reasonable scope, while regression allows the output to be infinite. Thus, regression is much harder to train when compared with classification. To directly compare the training difficulty of classification and regression, we replace the output layer of ILPN with a regression output

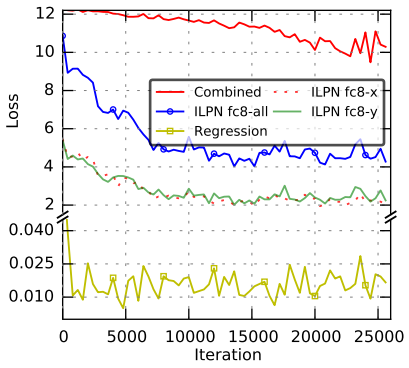


Fig. 11. The loss of classification vs. regression. This figure is best viewed in color.

TABLE 4
Detection accuracy on the FLIC for all joints

Joints	Head	Shoulder	Elbow	Wrist	Hip
Simultaneously	91.97	89.0	72.4	60.9	90.46
Independently	-	89.1	73.0	61.1	-

layer consisting of two neurons (which respectively denotes the joint's x and y coordinates). Besides, we replace the two independent softmax loss layers with a L_2 loss layer. When training in the manner of classification, the coordinate of the left wrist is directly used as its corresponding label, while when training in the manner of regression, the coordinate is normalized to $(0, 1]$ to avoid loss explosion due to the big output of regression. Furthermore, to verify the necessity of independence discussed in Sec. 3.2, we bring the combined loss into comparison. The combined loss fuction is a modified softmax loss that predicts the top 2 responsed neurons as the x and y.

Figure 11 shows that in the classification manner, the loss drops quickly in the first 10000 iterations, and then gradually being stable. However, in the regression manner, the loss randomly fluctuates in the first 26000 iterations, which means that the model has not yet to learn. The different loss performance indicates that the classification network is more effective to train than the regression network in the pose estimation problem. Unlike randomly fluctuated L_2 loss, The combined loss slowly goes down. Nonetheless, independent loss goes down significantly faster than combined classification loss. Besides, independent loss is more stable than the combined loss which starts fluctuating after 23000 iterations.

Simultaneously vs. independently: The detection accuracies of training single joint independently and training all joints simultaneously on the FLIC for each joints are declared in Table 4. In the experiment, we changed the single joint label of the input data to the label of all joints, then to train the nine joints such as head, shoulders, elbows, wrists and hips simultaneously. In Table 4, compared with training single joint independently, the method of training all joints simultaneously in the accuracy rate is a little lower. This is due to the limitations of hardware computing capacity and cannot fully identify the link between the

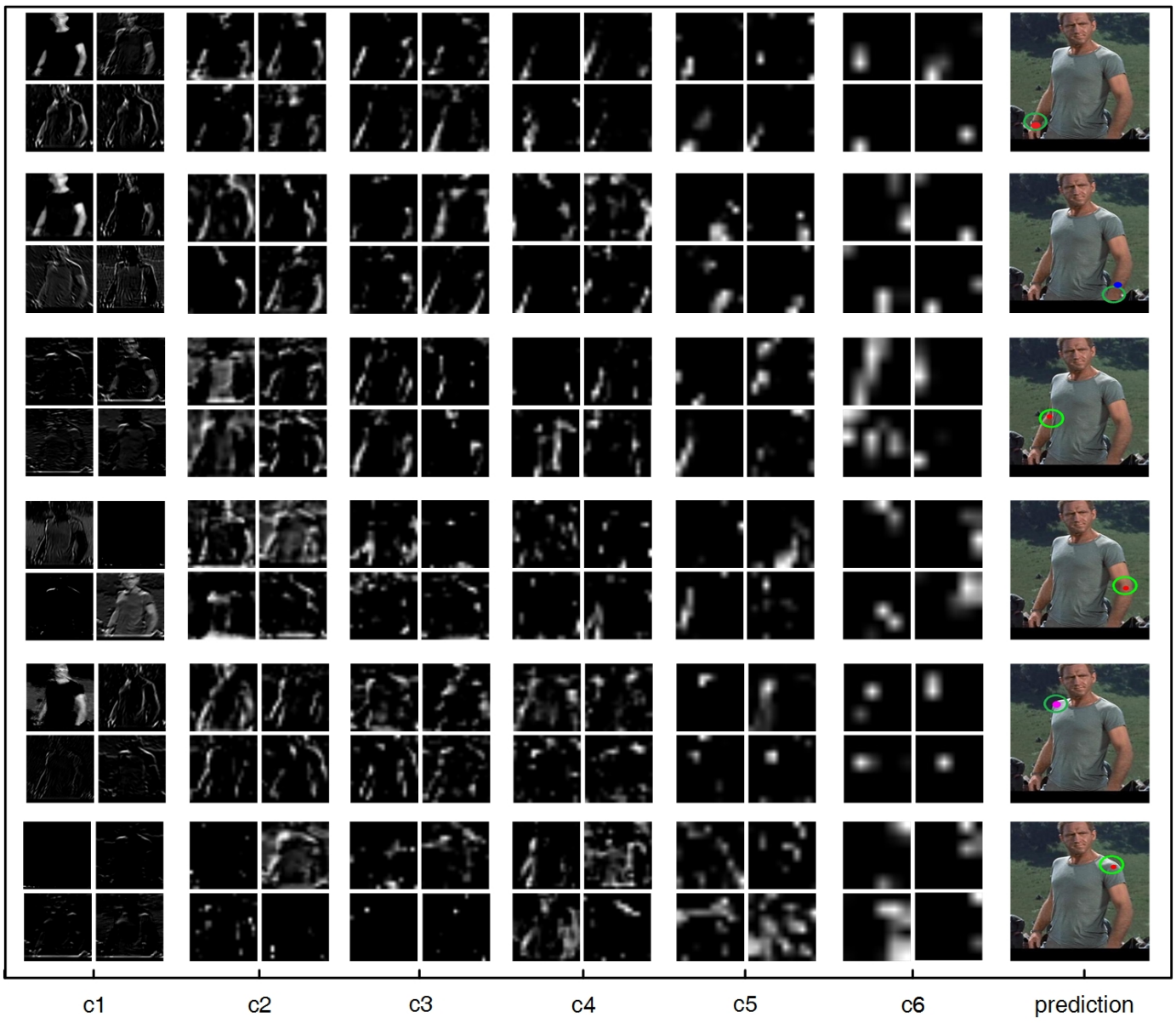


Fig. 12. Visualization of feature maps in ILPNs. For each convolutional layer (c1-c6), four feature maps are shown.

joints. But the training efficiency will be greatly improved. This demonstrates that ILPN can extract the link between joints all over the body and have the ability to train all joints simultaneously.

ILPN visualization: To analyse what features ILPNs have learned, we select one example image from the FLIC, run one forward-propagation, and visualize the feature maps learned in each convolutional layer. The feature maps and estimation results of right wrist, left wrist, right elbow, left elbow, right shoulder and left shoulder are shown in Figure 12. For each joint, 4 feature maps of each convolutional layer is shown. The prediction results are based on the person-centric viewpoint. Feature maps of the first and second layers present the contour of the people, which means the people is "sensed" by ILPN. In feature maps of layers 3 and 4, only related parts are retained while some unrelated parts are left out, e.g., the lower arms are reserved for the right and left wrist, and the upper arms are reserved for the right shoulder. The feature maps of layers 5 and 6 are clean whereas somewhat abstract. The white spot may indicate the relative position of corresponding joint. In Figure 12 we can observe that the feature extraction for a body pose is a procedure from global to local, which is consistent with the important features of the human visual perception pathway that researchers attempt to simulate [40].

5 Conclusion

In this paper, a framework named ADORE is proposed to adaptively address the pose estimation with the advantages of both global and local cues. ADORE is composed of two parts: ILPN for a global estimation and CLD for a subsequent complement on ILPN. Comprehensive experiments are conducted to evaluate the efficiency and flexibility of ADORE. The experimental reasults demonstrate that ILPN

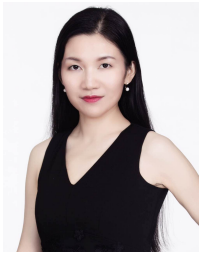
provides a reliable initial pose for CLD and CLD refines the initial pose in high-precision regions. ILPN may be improved to provide more precise structured output. In addition, more varieties of local detectors will be investigated to collaborate with ILPN for various pose estimation applications and general object locating problems.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant 61370149, in part by the Fundamental Research Funds for the Central Universities(ZYGX2013J083), in part by the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry and the open project of State Key Laboratory of virtual reality technology and system under Grant No. BUAA-VR-16KF-07.

REFERENCES

[1] M. A. Fischler and R. A. Elschlager, "The representation and matching of pictorial structures," *IEEE Transactions on Computers*, vol. 22, no. 1, pp. 67–92, 1973.
[2] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," *International Journal of Computer Vision*, vol. 61, no. 1, pp. 55–79, 2005.
[3] M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: people detection and articulated pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
[4] ——, "Discriminative appearance models for pictorial structures," *International Journal of Computer Vision*, vol. 99, no. 3, pp. 259–280, 2012.
[5] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.
[6] ——, "Articulated human detection with flexible mixtures of parts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2878–2890, 2013.
[7] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele, "Poselet conditioned pictorial structures," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.
[8] ——, "Strong appearance and expressive spatial models for human pose estimation," in *IEEE International Conference on Computer Vision*, 2013.
[9] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.
[10] F. Xiaochuan, Z. Kang, L. Yuewei, and W. Song, "Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.
[11] K. Wilber, *Sex, Ecology, Spirituality*. Shambhala Publications, 2000.
[12] A. Jain, J. Tompson, M. Andriluka, G. W. Taylor, and C. Bregler, "Learning human pose estimation features with convolutional networks," in *International Conference on Learning Representations*, 2014.
[13] J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in *Advances in Neural Information Processing Systems*, 2014.
[14] V. Ferrari, M. Marin-Jimenez, and A. Zisserman, "Progressive search space reduction for human pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.
[15] ——, "Pose search: retrieving people using their pose," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
[16] M. Eichner and V. Ferrari, "Better appearance models for pictorial structures," in *British Machine Vision Conference*, 2009.
[17] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained and multiscale and deformable part model," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.
[18] X. Ren, A. C. Berg, and J. Malik, "Recovering human body configurations using pairwise constraints between parts," in *IEEE International Conference on Computer Vision*, 2005.
[19] L. Sigal and M. J. Black, "Measure locally and reason globally: occlusion-sensitive articulated pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.
[20] H. Jiang and D. R. Martin, "Global pose estimation using non-tree models," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.
[21] M. Bergtholdt, J. Kappes, S. Schmidt, and C. Schnörr, "A study of parts-based object class detection using complete graphs," *International Journal of Computer Vision*, vol. 87, no. 1, pp. 93–117, 2010.
[22] G. Gkioxari, P. Arbeláez, L. Bourdev, and J. Malik, "Articulated pose estimation using discriminative armlet classifiers," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.
[23] L. Bourdev and J. Malik, "Poselets: body part detectors trained using 3d human pose annotations," in *IEEE International Conference on Computer Vision*, 2009.
[24] L. Zhao, X. Gao, D. Tao, and X. Li, "A deep structure for human pose estimation," *Signal Processing*, vol. 108, pp. 36–45, 2015.
[25] B. Sapp and B. Taskar, "Modec: Multimodal decomposable models for human pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.
[26] M. Kiefel and P. V. Gehler, "Human pose estimation with fields of parts," in *European Conference on Computer Vision*, 2014.
[27] L. Bourdev, F. Yang, and R. Fergus, "Deep poselets for human detection," *arXiv preprint arXiv:1407.0717*, 2014.
[28] X. Lu, X. Zheng, and X. Li, "Latent semantic minimal hashing for image retrieval," *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, vol. 26, no. 1, pp. 355–368, 2017.
[29] W. Ouyang, X. Chu, and X. Wang, "Multi-source deep learning for human pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.
[30] X. Chen and A. Yuille, "Articulated pose estimation by a graphical model with image dependent pairwise relations," in *Advances in Neural Information Processing Systems*, 2014.
[31] Y. Yuan, X. Zheng, and X. Lu, "A discriminative representation for human action recognition," *Pattern Recognition*, vol. 59, pp. 88–97, 2016.
[32] X. Li, Z. Wang, and X. Lu, "Surveillance video synopsis via scaling down objects," *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, vol. 25, no. 2, pp. 1–1, 2015.
[33] Y. Yuan, L. Qi, and X. Lu, "Action recognition by joint learning *," *Image and Vision Computing*, vol. 55, pp. 77–85, 2016.
[34] R. Keys, "Cubic convolution interpolation for digital image processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 6, pp. 1153–1160, 1981.
[35] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *International Conference on Machine Learning*, 2010.
[36] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classiffication with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012.
[37] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.
[38] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using convolutional networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.
[39] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," 2014.
[40] L. Dong, J. Su, and E. Izquierdo, "Scene-oriented hierarchical classification of blurry and noisy images," *IEEE Transactions on Image Processing*, vol. 25, no. 5, pp. 2534–2545, 2012.



Le Dong Le Dong received the PhD degree in Electronic Engineering and Computer Science from Queen Mary, University of London in 2009. She is an associate professor in University of Electronic Science and Technology of China. She is the coordinator of several National Natural Science Foundation Projects such as NSFC Surface Project, NSFC Youth Project, NSFC Important Research Project and so on. She has now published dozens of papers in International journals and conferences including several top journals and high level International Conference, such as TIP, ACM MM, TMM, PR, TCSVT, ICPR. She has served as a reviewer for several top journals and conferences. Now she is the Executive Chair of ACM SIGAI CHINA, the secretary-general of Vision And Learning Seminar(VALSE) and the executive secretary of next-generation of Sichuan Provincial Engineering Laboratory.

Xiuyuan Chen is currently a M.Sc student at School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China. His primary research interests is mainly in computer vision, particularly including image retrieval and image understanding.

Ran Wang is currently a M.Sc student at School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China. His primary research interests are mainly in computer vision and machine learning, particularly including pose estimation, behavior understanding, and deep learning.

Qianni Zhang received her M.Sc. degree in Internet signal processing in 2004 and the PhD degree in 2007, both from Queen Mary University of London. She is now working as a lecturer (associate professor) at the School of Electronic Engineering and Computer Science, Queen Mary University of London. Her research interests include multimedia processing, semantic inference and reasoning, machine learning, image understanding, 3D reconstruction and immersive environments. She has published over 30 technical papers and book chapters, and has actively contributed to several European funded research projects. She has served as a guest editor in a special issue in Journal of Multimedia and a reviewer in journals including IEEE TCSVT, Signal Processing: Image Communication, and various conferences and workshops including IEEE ICIP, ICASSP, ACM Multimedia, etc. She has served as an organiser, a session chair, or a member of technical program committee at several international conferences, workshops or special sessions.

Ebroul Izquierdo PhD, MSc, CEng, FIET, SMIEEE, MBMVA, is Chair of Multimedia and Computer Vision and head of the Multimedia and Vision Group in the school of Electronic Engineering and Computer Science at Queen Mary, University of London. He has been a senior researcher at the Heinrich-Hertz Institute for Communication Technology (HHI)), Berlin, Germany, and the Department of Electronic Systems Engineering of the University of Essex. Prof. Izquierdo is a Chartered Engineer, a Fellow member of The Institution of Engineering and Technology (IET), a senior member of the IEEE and a member of the British Machine Vision Association. He was a past chairman of the IET professional network on Information Engineering. He is a member of the Visual Signal Processing and Communications Technical Committee of the IEEE Circuits and Systems Society and member of the Multimedia Signal Processing technical committee of the IEEE. Prof. Izquierdo is or has been associated editor of the IEEE Transactions on Circuits and Systems for Video Technology (from 2002 to 2010), the IEEE Transactions on Multimedia (from 2010 to 2015). He is member of the editorial board of the EURASIP Journal on Image and Video processing (from 2004 to date), the Journal of Multimedia Tools and Applications (2008 to 2014) and the Journal of Multimedia (2009-2014), the Journal of Computer Engineering International (2008 to date) and the Infocommunications Journal (2008 to 1015). He has been guest editor of the Elsevier journal Signal Processing: Image Communication, The EURASIP Journal on Applied Signal Processing and the IEE Proceedings on Vision, Image and Signal Processing.