Bi-directional Graph Structure Information Model for Multi-Person Pose Estimation

Jing Wang*, Ze Peng*, Pei Lv, Junyi Sun, Bing Zhou and Mingliang Xu

Abstract—The performance of Multi-person pose estimation has been greatly improved in recent years, especially with the development of convolutional neural network. However, there still exists a lot of challenging cases, such as crowded scene and ambiguous overlapping, which cannot be well addressed. In this paper, we propose a novel bi-directional graph structure information model (BGSIM) for multi-person pose estimation to relieve above problems. More specifically, the BGSIM is contained in a multi-stage network architecture with two branches. The first branch predicts the confidence maps of joints and uses a geometrical transform kernel to propagate information between neighboring joints at the confidence level. The second branch leverages the BGSIM to encode rich contextual information and to infer the occlusion relationship among different joints. In the BSSIM, we dynamically determine the joint point with highest response of the confidence maps as the base point of passing message. The base point propagates information to other joints (especially for the occluded joints) via unrolled tree structures, this ensures reliable information which is transmitted among different body parts and facilitates the prediction of joints for multi-person. Based on our proposed model, we achieve an average precision of 62.9 on the COCO Keypoint Challenge dataset and 77.6 on the MPII (multi-person) dataset. Moreover Compared with other state-of-art methods, our method can achieve highly promising results on our selected multi-person dataset without extra training.

Index Terms—Multi-person, confidence maps, geometrical transform kernel.

I. INTRODUCTION

HUMAN pose estimation from still image is a challenging problem in computer vision. It is key to many visual tasks, such as action recognition [1], [2], [3], person re-identification [4], and human-computer interaction [5]. Human pose estimation contains single-person pose estimation and multiple-person pose estimation and recognizing the poses of multiple persons is a lot more challenging than recognizing the pose of a single person in an image [6], [7], [8], [9], [10]. So many researchers are still focus on this problem. Recently, significant advancements have been achieved in this research field by deep convolutional neural networks (DCNN) [11], [12]. Some works [13], [14], [15], [16] use human detectors to obtain all the bounding boxes of individuals in an image and then perform single-person pose estimation in each of the detected bounding boxes. However, these top-down approaches do not work very well in complex or crowded environments, where individuals cannot be detected accurately. Meanwhile, other researchers use bottom-up approaches [17], [18], [19], [20] directly predict all joints and assemble them into different individuals. Although great progress has been made by this kind of method, there still exist a lot of challenging cases, such as crowded scene, self occlusion and other occlusion, which cannot be well predicted. The main reasons lie at two points: 1) the joints are wrongly connected due to too many people, 2) there exists some ambiguous overlapping when a joint of an individual is obscured by the same one of another individual.

To address above problems, we propose a novel bi-directional graph structure information model (BGSIM) for multi-person pose estimation. The BGSIM is contained in a network architecture which has three stages. In our method, we don’t directly predict all body joints, we divide the full body pose estimation into three subsets \{head, shoulders\}, \{elbows, hips\} and \{wrists, knees, ankles\}. So we first predicts confidence maps of head and shoulders in first stage, because heads and shoulders are easier to predict among all body parts due to their relatively fixed position. Then the elbows and hips with relatively flexibility are predicted in the second stage. The third stage will consider the full body joints. Each stage has two branches in our network architecture. The first branch learns confidence maps of each body joint based on basic feature which is captured by the VGG [21]. More importantly, this part also implements a geometrical transform kernel [22] with a convolution layer to shift the response between neighboring joints of a person at the confidence map level. Compared with those work on feature level [22], [23], we can capture more explicit location information of the joints.

Based on the confidence maps estimated in the first branch, the second branch BGSIM explicitly address those occluded joints. Different from traditional tree structure model, BGSIM
is a graph structure model for multi-person pose estimation. It can model the interactions between unconnected body parts, and the interactions between human body and objects. These interactions are important cues for occlusion reasoning to capture rich contextual information. As the proposed BGSIM is a graph containing loops, we perform inference on tree structures unrolled by the BGSIM. Because one joint point with the highest response of confidence maps contains most reliable information, we dynamically determine it as the base point in the BGSIM. And the base point propagates information to other joints (especially for the occluded joints) via the tree structures. This ensures we can encode richer contextual information containing some invisible and occluded joints to achieve good performance for multi-person pose estimation.

Our proposed method achieves state-of-the-art results in the COCO and MPII benchmark datasets. We also construct a new dataset that contains 5899 images of multiple people with different degrees of occlusions. Some image samples are shown in Figure 1. On this dataset, our method can achieve a greater improvement in accuracy compared with the existing methods.

The contributions of our work are summarized as follows:

- We propose a BGSIM to encode rich contextual information and to infer occlusion relationship among different body parts.
- We determine the base point dynamically in BGSIM and use it as the root node for unrolled tree structures to propagate information to other joints.
- We involve the BGSIM into a network architecture with three stages to capture more explicit location information of body joints at the confidence level for multi-person pose estimation.

II. RELATED WORK

Automatic human pose estimation is crucial to intelligent surveillance systems, especially for human tracking and action recognition. Image-based human pose estimation in real-world environments remains a challenging task, and its current performance is still far from satisfactory due to occlusion and multi-person problems in complex scenes.

These problems are usually solved by using pictorial structures [24], [25] or graphical models [26]. The Pictorial Structure Model [25] defines some pairwise connections of independent parts to represent the correlations among different body parts. However, the tree-structured model is “oversimplified”. Accordingly, many researchers proposed highly complex tree-structured models and the graph structure model [27]. Tian et al. [28] proposed a hierarchical spatial model that captured an exponential number of poses with a compact mixture representation on each part. L. Pishchulin et al. [20] proposed a typical approach to articulate pose estimation by combining the spatial modeling of the human body with the appearance modeling of body parts. Meanwhile, other approaches [28], [29], [30], [31], [32], [33], [34], [35], [36] used the tree-structured or graphical model to solve the problems of human keypoints estimation and predicted the keypoint locations based on handcrafted features (e.g., SIFT and HOG). With the rapid development of deep convolutional neural networks (CNN), recent studies [15], [22], [37], [38], [11], [39], [40], [41], [10], [23] have made great process in this research area.

Single-Person Pose Estimation

Early studies on human pose estimation mainly focused on single-person pose estimation due to its simplicity. Toshev et al. introduced CNN to solve the pose estimation problem in DeepPose [37], which cascaded CNN pose regressors to deal with the body joints. Wei et al. [10] and Newell et al. [7] found that deep CNNs performed effectively in human pose estimation as well as proposed a multi-stage architecture to solve the pose estimation problem. Newell et al. compiled several hourglass modules and proposed a U-shape architecture for human pose prediction. X. Chu et al. [22] designed a structured feature-learning architecture to infer the correlations among different body parts at the feature level for human pose estimation. Although CNN can effectively extract feature information, the occlusion relationship among different body parts have not been fully explored.

Multi-Person Pose Estimation

The above approaches show a promising performance in single-person pose estimation, but real-world images usually contain multiple persons. Multi-person pose estimation is a challenging task due to occlusion, the unpredictable interactions among different persons, and the diversity of human gestures. Some researchers applied bottom-up approaches to deal with this problem. For instance, DeepCut [20] directly predicted all body part candidates, categorized them into different classes, and partitioned them into the corresponding people. DeeperCut [19] employed image-conditioned pairwise terms to achieve better pose estimation performance based on ResNet [42]. Cao et al. [17] mapped the relationship among different keypoints into part affinity fields (PAFs) and assembled the detected keypoints into different poses of people. In contrast to bottom-up approaches, top-down approaches initially located each person and then independently analyzed their joints. For example, Mask-RCNN [15] initially predicted the human bounding boxes, and then cropped out the feature map of these bounding boxes to predict the human keypoints. Papandreou et al. [13] predicted both the dense heatmaps and offsets of each keypoint type from each bounding box by using a fully convolutional resnet, and then fused these heatmaps with offsets to obtain the final predicted location of keypoints. RMPE [16] predicted human pose by using the stacked hourglass model and then adjusted the imperfect proposal by combining STN, SPPE, and SDTN to overcome the location errors caused by the detection.

Nevertheless, these methods have not good enough performance in cases of multiple targets or serious occlusion. To address such limitations, we propose the BGSIM in our network architecture that can fully explore the connections among different body parts.
is analyzed by a convolutional network initialized with the first 10 layers of VGG-19 \cite{21}. This convolutional network generates the basic feature information $F=g(I)$, where $I$ denotes the input image and $g$ is a nonlinear function. Then, we use a multi-stage network architecture to handle the basic feature information $F$. The network has three stages, and each stage contains two branches: JOINTS and BGSIM. JOINTS predicts the confidence maps of joints. More specifically, we use a convolutional network to predict basic confidence maps of joints $c(x) = p(x | \sigma)$ of body part locations as defined in \cite{43}, where $x \in R^2$ denotes the joint location and $\sigma$ denotes the types of joints. BGSIM can explicitly address some occluded joints based on the confidence maps $c(x)$. If we are unable to capture information about heads or shoulders in the first stage, the second and third stage are still effective. The BGSIM can infer the information about heads and shoulders by using existing information of elbows and hips which is obtained in the second stage. More details about JOINTS and BGSIM will be described in Section \textbf{III-A} and Section \textbf{III-B}.

\subsection{JOINTS}

As mentioned above, we need a reliable body part detection model to detect confidence maps of all the joints of a person before analyzing his or her pose. We use a deep, fully-convolutional human body part detection model that draws on powerful ideas regarding semantic segmentation, object classification, and human pose estimation.

Our convolutional network model is initialized with the first 10 layers of VGG-19 \cite{21} and is finetuned by changing the parameters of the remain layers (the names, number of output, and coefficient of learning rate). The branch JOINTS in the network architecture is to predict confidence maps of joints. It also has three stages, and the convolution parameters of JOINTS is different at each stage. Figure \textbf{3} shows specific convolution parameters of JOINTS at each stage. All 1*1 kernel size at each stage is to reduce dimension. From Figure \textbf{3} we can see that the deeper the number of stage, the larger the convolution kernel size, because the bigger kernel size can obtain bigger receive field which is helpful to predict the position of joints.

In addition, for JOINTS, we implement geometrical transform kernel by a convolutional layer. We can use geometrical transform kernel to shift confidence maps of neighboring joints. In order to illustrate the process of information passing, an example is shown in Figure \textbf{4}. The confidence maps of elbow and wrist are estimated by the convolutional network and the confidence map of wrist has high response(as shown in Figure \textbf{4}d)). However, the elbow is occluded by other object, we cannot obtain its confidence map directly. It is not suitable to shift confidence map of wrist to elbow, since there is a spatial mismatch between the two joints. Instead we shift the confidence map of the wrist towards the elbow through geometrical transform kernel and then add the transformed confidence map to that of the elbow. The high confidence map of the elbow is obtained finally.

Although we can obtain all confidence maps of each body part, it is still uncertain whether these body parts belong to the same person or some joints with the same type among multiple people are overlapping. So some body parts may be incorrectly connected. Our network architecture is robust to...
shooting angles and occlusion similar to [44]. Especially, in our network architecture we constrain the relationship among different body parts according to the BGSIM proposed in Section III-B to achieve highly accurate location results, and BGSIM is proven to be highly robust to shooting angles and occlusion.

B. BGSIM

Considering the tree-structured model is simple and cannot solve occlusion problem, we propose BGSIM based on graph-structured model. At each stage, our proposed BGSIM selects the joint with the highest response of confidence maps as the base point (the orange joint point in Figure 5) in this stage. Then we use the base point as the root node of a series of tree structures unrolled by our BGSIM to propagate information to other joints, especially for the occluded joints. As is shown in Figure 6, the left elbow of the woman is completely occluded by other object; the right side of Figure 6 is correct joint connection after utilizing BGSIM to compute the confidence map of the occluded joint.

In the BGSIM, we denote $G = (V, \varepsilon)$ be the $N$-node graph, where $V$ denotes the set of body parts, $\varepsilon$ denotes the set of constraint edges and $N = |V|$ is the number of joints. The position of the $i$ joint point is denoted as $p_i = \{x_i, y_i\}$. The set of pairwise spatial relationships are denoted as $t = \{t_{ij}, t_{ji} | (i, j) \in \varepsilon\}$, where $t_{ij}, t_{ji} \in \{1, \cdots, T_{ij}\}$ denotes the relative position between joint $i$ and $j$. $\phi = (p, t, o)$ is the pose configuration, $p = \{p_i\}$ and $o = \{o_i\}$, $o_i \in \{0, 1, 2\}$ is the occlusion state ("0" for visibility, "1" for self-occlusion, and "2" for occlusion by other objects). We can encode richer information through bi-directional information propagation between joint $i$ and joint $j$.

Given an input image $I$, the goal of the BGSIM is to maximize the posterior as follows

$$P(\phi | I) \propto \sum_{i \in V} U(I, p_i, o_i) + \sum_{(i, j) \in \varepsilon_K} R^K(t_{ij}, t_{ji}, o_i, o_j) + \sum_{(m, n) \in \varepsilon_C} R^C(t_{mn}, t_{nm}, o_m, o_n)$$ \hspace{1cm} (1)

where $\varepsilon_K$ denotes the kinematic constraints among body parts,
\( \varepsilon_C \) denotes the set of additional constraints among body parts that are not physically connected, and \( \varepsilon = \varepsilon_K \cup \varepsilon_C \) denotes the full set of constraints. For simplicity, we call them kinetic edges (\( \varepsilon_K \)) and contextual edges (\( \varepsilon_C \)) respectively. In addition, \( U(I, p_i, o_j) \) is the local part appearance score considering the occlusion state, while \( R^K(t_{ij}, t_{ji}, o_i, o_j) \) and \( R^C(t_{mn}, t_{nm}, o_m, o_n) \) represent the deformation (kinetic and contextual respectively) scores with occlusion coherence. The score function of BGSIM is formulated as:

\[
S = \frac{F(I, \phi)}{|F(I, \phi)| + |F'(I, \phi)|} + \frac{F'(I, \phi)}{|F(I, \phi)| + |F'(I, \phi)|}
\]  

(2)

\( F(I, \phi) \) denotes score function with forward information flow (we define that the forward direction is from the base point to the end joint point of human body and the backward direction is opposite), \( F'(I, \phi) \) denotes score function with backward information flow. The calculation of \( F(I, \phi) \) and \( F'(I, \phi) \) is same, so we only analyze the \( F(I, \phi) \)

\[
F(I, \phi) = \sum_{i \in V} U(I, p_i, o_i) + \sum_{(i,j) \in \varepsilon_K} R^K(t_{ij}, t_{ji}, o_i, o_j) + \sum_{(m,n) \in \varepsilon_C} R^C(t_{mn}, t_{nm}, o_m, o_n)
\]  

(3)

the three terms on the right side of Equation (3) handle occlusion from different aspects. The first term considers occlusion “local”, i.e., whether each local part is visible, other-occluded or self-occluded. The second term encourages the occlusion states among physically connected parts to be continuous, while the third term reflects the self-occlusion relationship between physically non-connected parts.

The part appearance score is represented as

\[
U(I, p_i, o_i) = w_i f(i|I(p_i); \theta) \cdot L_{(0,1)}(o_i) + b_i(o_i)
\]  

(4)

where \( w_i \) is appearance weight parameter, \( f(i|I(p_i); \theta) \) is the image evidence based on the local image \( I(p_i) \), \( \theta \) denotes the parameters of convolutional neural network, \( L_{(0,1)}(o_i) \) is an indicator function as below:

\[
L_{(0,1)}(o_i) = \begin{cases} 1, & \text{if } o_i = 0, 1 \\ 0, & \text{if } o_i = 2 \end{cases}
\]  

(5)

The kinetic pairwise score is formulated as

\[
R^K(I, p_i, t_{ij}, t_{ji}, o_i, o_j) = w_{ij} t_{ij}^T \cdot \mu(p_j - p_i - r_{ij}^T)
+ w_{ij} t_{ij} t_{ji}^T \cdot \mu(p_j - p_i - r_{ij}^T)
+ w_{ji} f(t_{ji}|I(p_j); \theta) \cdot L_{(0,1)}(o_j)
+ w_{ji} t_{ji} t_{ij}^T \cdot \mu(p_j - p_i - r_{ij}^T)
+ O_{ij}(o_i, o_j)
\]  

(6)

where \( u(\Delta_p = [\Delta x - \Delta y]) = [\Delta x, \Delta x^2, \Delta y, \Delta y^2]^T \) are the deformation features, \( r_{ij}^T \) is the mean relative position of special relationship type \( t_{ij} \), \( w_{ij}, w_{ji} \) are the weight parameters of special relationship types, and \( O_{ij}(o_i, o_j) \) is the bias term which encodes the occlusion coherence between part \( i \) and part \( j \).

The contextual pairwise score is formulated as

\[
R^C(I, p_m, p_n, t_{mn}, t_{nm}, o_m, o_n) = w_{mn} t_{mn} \cdot \mu(p_m - p_n - r_{mn}^T)
+ w_{mn} t_{mn} t_{nm}^T \cdot \mu(p_m - p_n - r_{mn}^T)
+ O_{mn}(o_m, o_n)
\]  

(7)

where all the terms on the right hand side are in accordance with that in Equation (6). The bias term \( O_{mn}(o_m, o_n) \) reflects long range spatial constraints as well as occlusion between non-adjacent parts \( m \) and \( n \).

In fact, we unroll the BGSIM formed at each stage into a tree model to generate the pose hypothesis and then rescroe the candidate pose configurations with BGSIM. Afterward, we introduce the inference of the proposed model in the forward information flow and the inference of the proposed model in the backward information flow follows the same operation.

We simply the symbols of the generic score function in Equation (3) as follows

\[
F(I, \phi) = \sum_{i \in V} U_i(\phi) + \sum_{(i,j) \in \varepsilon_K} R^K_{ij}(\phi) + \sum_{(i,j) \in \varepsilon_C} R^C_{ij}(\phi)
\]  

(8)

1) Model Unrolling: for any part \( i \) with an out-degree(number of connections pointing to the other parts) of \( \nu_i > 1 \), we generate \( \nu_i - 1 \) virtual parts and unroll the contextual edges to form a computation tree similar to as shown in Figure 7(d). The unrolled tree model is used to generate pose configurations.

2) Pose Selection: The goal of our graphical model is to maximize the score in Equation (8), i.e.,

\[
\phi_m = \arg \max_{\phi \in \mathcal{X}} \sum_{i \in V} U_i(\phi) + \sum_{(i,j) \in \varepsilon_K} R^K_{ij}(\phi) + \sum_{(i,j) \in \varepsilon_C} R^C_{ij}(\phi)
\]  

(9)

where \( \phi_m \) is the optimal pose configuration with the proposed BGSIM. Instead of propagating information on a loopy graph,
Fig. 8. The response of some occluded joints is strengthened by optimizing the three stages, and our method makes the response highly centralized and accurate.

(a) Confidence maps of our method  
(b) Confidence maps of traditional method

Fig. 9. AP performance on the COCO validation set in (a), (b), and (c).

we propagate information on the unrolled tree structure to generate root hypothesis. This allows us to employ dynamic programming to propagate information from leaf nodes to the root node efficiently.

The optimization over the unrolled model is formulated as

$$
\phi'_m = \arg \max_{X'} \left\{ \sum_{i \in V} U_i(\phi) + \sum_{m \in V'} U_k(\phi) + \sum_{(i,j) \in E} R^K_{ij}(\phi) + \sum_{(m,n) \in E_C} R^C_{mn}(\phi) \right\} 
$$

where $\phi'_m$ is the optimal pose configuration of the unrolled tree-structured model. This is equivalent to adding part appearance weights to those nodes that have more connections.

Suppose the number of possible root hypotheses to be $H$ in the test image. We sort them according to root score and select top $H_\sigma/H$ is the ratio of hypothesis selection. And we assume that the optimal hypothesis is included in the selected top-$\sigma$ hypotheses of the unrolled configurations.

3) Backtracking and Rescoring: As soon as the top-$\sigma$ hypotheses of root nodes are determined, the corresponding pose configurations can be obtained by backtracking directly from the root node to the leaf nodes. We only backtrack the child node from the real parent node (e.g., node 1 is backtracked from node 2 rather than node 6 in Figure 7 (e)) because the parent near the root node has less freedom of movement and is more reliable. We recompute the score of the pose configurations with a graphical model and rerank the hypothesis to obtain the optimal pose configuration.

IV. EXPERIMENTAL RESULTS

Our model is only trained on the COCO 2014 dataset and evaluated on two public benchmarks for person pose estimation, namely, the (1) COCO 2014 dataset and the (2) MPII
human multi-person dataset. These two datasets contain many real-world challenging scenes, such as crowding, occlusion, and contact. Our approach demonstrates the same performance and even outperforms the other state-of-the-art methods on the COCO and MPII datasets, and shows its best performance in our own multi-person “occlusion” dataset with different levels of occlusion.

In our method, the first stage predicts the heads and shoulders, the second stage predicts the elbows and hips, and the third stage predicts the wrists, knees, and ankles. Each stage shares the output information of the previous stage as the input to optimize the joints at this stage. Our method shows a very significant improvement in responding to some partially or severely occluded joints in a cascading way. Compared with traditional CNN architecture (Figure 8 (b)), the response of occluded joints in our architecture is strengthened by optimizing the three stages, and our method makes the response highly centralized and accurate. In our experiment, we use the convolutions with a kernel size of $7 \times 7$. In fact, the larger kernel size has a larger receptive field, that can obtain rich global information, which is especially helpful when dealing with occluded joints.

A. Results on the COCO dataset

The training, validation, and testing sets of the COCO dataset contain over 200K images and 250K persons-instances-points labels with keypoints. Each person is annotated with 15 body joints for the full body. We apply the following object keypoint similarity (OKS) \[46\] defined by the COCO evaluation:

$$OKS = \frac{\sum_i \left[ \exp \left( -\frac{d_i^2}{2\sigma^2} \right) \delta (v_i > 0) \right]}{\sum_i \left[ \delta (v_i > 0) \right]} \quad (11)$$

The OKS is analogous to the IoU in object detection. By using OKS, we can compute the mean average precision (AP) and average recall (AR) as competing metrics. Table II shows the mAP of our method with different OKS thresholds and compares the mAP performance of our method with that of other approaches. For $AP^{L}$, our method outperforms the other methods and achieves 70.2% AP. Figure 9 shows a breakdown of the errors of our method on the COCO validation set. The area under each curve is shown in brackets in the legend. C75 denotes PR at IoU(Intersection over Union)=.75, and area under curve corresponds to $AP^{IoU=.75}$. Loc represents PR at IoU=.10(localization errors ignored, but not duplicate detections) and all remaining settings use IoU=.1. Sim denotes PR after supercategory false positives (fps) are removed and Oth represents PR after all class confusions are removed. BG denotes PR after all background (and class confusion) fps are removed and FN represents PR after all remaining errors are removed (trivially AP=1). In the case of all people, AP at IoU=.75 is .715 and perfect localization would increase AP to .897. Interesting, removing all class confusions (both within supercategory and across supercategories) would not raise AP. Removing background fp would only raise AP slightly to .901. In summary, Most of the false positive results are dominated by imperfect localization rather than background confusion, thereby indicating a large space for improvement in capturing spatial dependencies than in recognizing the appearance of body parts. Figure 10 shows some experimental results on COCO dataset.

B. Results on the MPII dataset

The MPII Human Pose dataset includes around 25000 images containing over 40000 persons with annotated body joints. We use the toolkit provided in [20] to measure the Average Precision (mAP) of all body parts based on the PCKh threshold. Table II compares the mAP performance of our method with that of other state-of-the-art methods. Compared with [17] and RMPE, our method shows the similar performance for some body parts, such as heads or shoulders. However, for other body parts that are easily occluded or show a large variation in the limb orientation (e.g., elbows or wrists), our method outperforms these two state-of-the-art methods. Unlike top-down methods such as RMPE, our method does not rely on human detection. Our method also shows the same performance as RMPE for single-person detection and can effectively detect multiple targets. Figure 11 shows some experimental results on MPII dataset.

C. Results on our own dataset

Our dataset contains 5899 multi-person and high occlusion images that are selected from the AI challenger [50], COCO,
|                  | Head | Shoulder | Elbow | Wrist | Hip   | Knee  | Ankle | Total |
|------------------|------|----------|-------|-------|-------|-------|-------|-------|
| Iqbal & Gall, ECCV16 [47] | 58.4 | 53.9     | 44.5  | 35.0  | 42.2  | 36.7  | 31.1  | 43.1  |
| DeeperCut, EVVC16 [19]    | 78.4 | 72.5     | 60.2  | 51.0  | 57.2  | 52.0  | 45.4  | 59.5  |
| Levinkov et al., CVPR17 [48] | 89.8 | 85.2     | 71.8  | 59.6  | 71.1  | 63.0  | 53.5  | 70.6  |
| Insafutdinov et al., CVPR17 [49] | 88.8 | 87.0     | 75.9  | 64.9  | 74.2  | 68.8  | 60.5  | 74.3  |
| CMU-Pose [17]            | 91.2 | 87.6     | 77.7  | 66.8  | 75.4  | 68.9  | 61.7  | 75.6  |
| RMPE [16]                | 88.4 | 86.5     | 78.6  | 70.4  | 74.4  | 73.0  | 68.5  | 76.7  |
| **Ours**                 | 90.7 | **87.7** | **80.2** | **71.1** | 74.8  | 72.9  | **66.3** | **77.6** |

Fig. 12. Experimental results on our own dataset, which images are selected from the AI challenger, COCO, MPII, and LSP public datasets.
our method to perform well without any extra training.

Table III shows some experimental results on our own dataset. The AP (Average Precision) is used to evaluate the performance of our method. The table shows that our method outperforms CMU-Pose on average for all joints. The reason is that our method incorporates rich contextual information among all joints.

V. CONCLUSION

In this paper, we follow the bottom-up pipeline for multi-person pose estimation and a new bi-directional graph structure information model (BGSIM) is presented to address the occluded keypoints. More specifically, our proposed BGSIM is involved into a multi-stage network architecture to encode rich contextual information among all joints. In addition, we dynamically select the base point with the highest response of confidence maps in BGSIM and take it as root node to propagate information to other joints, especially for the occluded joints, via tree structures which is unraveled by our proposed BGSIM. Our method achieves state-of-the-art results on the COCO keypoint localization, as well as the best results on our own dataset without any extra training.

REFERENCES

[1] Y. Song, Y. T. Zheng, S. Tang, X. Zhou, Y. Zhang, S. Lin, and T. S. Chua, “Localized multiple kernel learning for realistic human action recognition in videos,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 21, no. 9, pp. 1193–1202, 2011.

[2] G. Yu, J. Yuan, and Z. Liu, “Propagative hough voting for human activity detection and recognition,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 25, no. 1, pp. 87–96, 2015.

[3] Z. Zhou, X. Chen, Y. C. Chung, Z. He, T. X. Han, and J. M. Keller, “Activity analysis, summarization, and visualization for indoor human activity monitoring,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 18, no. 11, pp. 1498–1508, 2008.

[4] J. Zhu, H. Zeng, S. Liao, Z. Lei, C. Cai, and L. X. Zheng, “Deep hybrid similarity learning for person re-identification,” IEEE Transactions on Circuits and Systems for Video Technology, vol. PP, no. 99, pp. 1–1, 2017.

[5] H. P. Jain, A. Subramanian, S. Das, and A. Mittal, “Real-time upper-body human pose estimation using a depth camera,” in International Conference on Computer Vision/Computer Graphics Collaboration Techniques, 2011, pp. 227–238.

[6] L. Ladicky, P. H. S. Torr, and A. Zisserman, “Human pose estimation using a joint pixel-wise and part-wise formulation,” in Computer Vision and Pattern Recognition, 2013, pp. 3578–3585.

[7] A. Newell, K. Yang, and J. Deng, “Stacked hourglass networks for human pose estimation,” pp. 483–499, 2016.

[8] B. Sapp, A. Toshev, and B. Taskar, “Cascaded models for articulated pose estimation,” in European Conference on Computer Vision, 2010, pp. 406–420.

[9] M. Sun, P. Kohli, and J. Shotton, “Conditional regression forests for human pose estimation,” in IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 3394–3401.

[10] S. E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, “Convolutional pose machines,” pp. 4724–4732, 2016.

[11] K. Kang, W. Ouyang, H. Li, and X. Wang, “Object detection from video tubelets with convolutional neural networks,” in IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 817–825.

[12] K. Chen and W. Tao, “Once for all: a two-flow convolutional neural network for visual tracking,” IEEE Transactions on Circuits and Systems for Video Technology, vol. PP, no. 99, pp. 1–1, 2016.

[13] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy, “Towards accurate multi-person pose estimation in the wild,” pp. 3711–3719, 2017.

[14] S. Huang, M. Gong, and D. Tao, “A coarse-fine network for keypoint localization,” in IEEE International Conference on Computer Vision, 2017, pp. 3047–3056.

[15] K. He, G. Gkioxari, P. Dollr, and R. Girshick, “Mask r-cnn,” 2017.

[16] H. S. Fang, S. Xie, Y. W. Tai, and C. Lu, “Rmpc: Regional multi-person pose estimation,” pp. 2353–2362, 2016.

[17] Z. Cao, T. Simon, S. E. Wei, and Y. Sheikh, “Realtime multi-person 2d pose estimation using part affinity fields,” 2016.

[18] A. Newell, Z. Huang, and J. Deng, “Associative embedding end-to-end learning for joint detection and grouping,” 2016.

[19] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, “Deepercut: A deeper, stronger, and faster multi-person pose estimation model,” vol. 42, no. 5, pp. 34–50, 2016.

[20] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Geheb, and B. Schiele, “Deepercut: Joint subset partition and labeling for multi-person pose estimation,” Computer Science, vol. 2008, no. 1, pp. 4929–4937, 2015.

[21] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” Computer Science, 2014.

[22] X. Chu, W. Ouyang, H. Li, and X. Wang, “Structured feature learning for pose estimation,” in IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4715–4723.

[23] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang, “Learning feature pyramids for human pose estimation,” pp. 1290–1299, 2017.

[24] M. A. Fischler and R. A. Elschlager, “The representation and matching of pictorial structures,” IEEE Transactions on Computers, vol. C-22, no. 1, pp. 67–92, 1973.

[25] M. Andriluka, S. Roth, and B. Schiele, “Pictorial structures revisited: People detection and articulated pose estimation,” in Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, 2009, pp. 1014–1021.

[26] X. Chen and A. Yuille, “Articulated pose estimation by a graphical model with image dependent pairwise relations,” Eprint Arxiv, pp. 1736–1744, 2014.

[27] L. Fu, J. Zhang, and K. Huang, “Ogm: Occlusion relational graphical model for human pose estimation,” IEEE Transactions on Image Processing, vol. 26, no. 2, pp. 927–941, 2017.

[28] Y. Tian, C. L. Zitnick, and S. G. Narasimhan, “Exploring the spatial hierarchy of mixture models for human pose estimation,” in European Conference on Computer Vision, 2012, pp. 256–269.

[29] M. Dantone, J. Gall, C. Leistner, and L. V. Gool, “Human pose estimation using body parts dependent joint regressors,” in Computer Vision and Pattern Recognition, 2013, pp. 3041–3048.

[30] G. Gkioxari, P. Arbelaez, L. Bourdev, and J. Malik, “Articulated pose estimation using discriminative armlet classifiers,” in IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3342–3349.
A. Jain, J. Tompson, M. Andriluka, G. W. Taylor, and C. Bregler, “MoDeep: Multimodal decomposable models for human pose estimation,” in Computer Vision and Pattern Recognition, 2013, pp. 3674–3681.

F. Wang and Y. Li, “Beyond physical connections: Tree models in human pose estimation,” in Computer Vision and Pattern Recognition, 2013, pp. 596–603.

Y. Wang, T. Duan, and Z. Liao, “Learning hierarchical poselets for human parsing,” in The IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, Co, Usa, 29-25 June, 2011, pp. 1705–1712.

B. Sapp, C. Jordan, and B. Taskar, “Adaptive pose priors for pictorial structures,” in Computer Vision and Pattern Recognition, 2010, pp. 422–429.

L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele, “Poselet conditioned pictorial structures,” in Computer Vision and Pattern Recognition, 2013, pp. 588–595.

A. Toshev and C. Szegedy, “DeepPose: Human pose estimation via deep neural networks,” pp. 1653–1660, 2013.

A. Jain, J. Tompson, Y. Lecun, and C. Bregler, “MoDeep: A deep learning framework using motion features for human pose estimation,” vol. 9004, pp. 302–315, 2014.

W. Ouyang, X. Zeng, X. Wang, S. Qiu, P. Luo, Y. Tian, H. Li, S. Yang, Z. Wang, and H. Li, “Deepid-net: Deformable deep convolutional neural networks for object detection,” IEEE Transactions on Pattern Analysis Machine Intelligence, vol. PP, no. 99, pp. 1–1, 2016.

A. Bulat and G. Tzimiropoulos, “Human pose estimation via convolutional part heatmap regression,” in European Conference on Computer Vision, 2016, pp. 717–732.

G. Gkioxari, A. Toshev, and N. Jaitly, “Chained predictions using convolutional neural networks,” in European Conference on Computer Vision, 2016, pp. 728–743.

K. Zhang, L. Zhang, M. H. Yang, and D. Zhang, “Fast tracking via spatio-temporal context learning,” Computer Science, 2013.

A. Jain, J. Tompson, M. Andriluka, G. W. Taylor, and C. Bregler, “Learning human pose estimation features with convolutional networks,” Computer Science, 2013.

S. C. Tatikonda and M. I. Jordan, “Loopy belief propagation and gibbs measures,” in Eighteenth Conference on Uncertainty in Artificial Intelligence, 2002, pp. 493–500.

T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollr, and C. L. Zitnick, “Microsoft coco: Common objects in context,” vol. 8693, pp. 740–755, 2014.

U. Iqbal and J. Gall, “Multi-person pose estimation with local joint-to-person associations,” pp. 627–642, 2016.

E. S. Levinkov, B. Andres, D. Uhrig, T. Wang, M. Omran, E. I. Ibarz, A. Kirillov, C. Rother, T. Brox, and B. Schiele, “Joint graph decomposition and node labeling: Problem, algorithms, applications,” 2016.

J. Wu, H. Zheng, B. Zhao, Y. Li, B. Yan, R. Liang, W. Wang, S. Zhou, G. Lin, and Y. Fu, “Ai challenger: A large-scale dataset for going deeper in image understanding,” 2017.

Jing Wang is a graduate student in the School of Information Engineering of Zhengzhou University, China. Her research interest is computer graphics and computer vision, particularly in deep learning.

Ze Peng is a M.S. in the Industrial Technology Research Institute of Zhengzhou University, China, and his research interest is deep learning and computer vision.

Junyi Sun is a M.S. in the Industrial Technology Research Institute of Zhengzhou University, China, and his research interest is deep learning and computer vision.

Pei Lv is an associate professor in School of Information Engineering, Zhengzhou University, China. His research interests include video analysis and crowd simulation. He received his Ph.D in 2013 from the State Key Lab of CAD&CG, Zhejiang University, China. He has authored more than 20 journal and conference papers in these areas, including IEEE TIP, IEEE TCYSVT, ACM MM, etc.

Bing Zhou is currently a professor at the School of Information Engineering, Zhengzhou University, Henan, China. He received the B.S. and M.S. degrees from Xian Jiao Tong University in 1986 and 1989, respectively, and the Ph.D. degree in Beihang University in 2003, all in computer science. His research interests cover video processing and understanding, surveillance, computer vision, multimedia applications.

Mingliang Xu is a full professor in the School of Information Engineering of Zhengzhou University, China, and currently is the director of CIISR (Center for Interdisciplinary Information Science Research) and the vice General Secretary of ACM SIGAI China. He received his Ph.D. degree in computer science and technology from the State Key Lab of CAD&CG at Zhejiang University, Hangzhou, China. He previously worked at the department of information science of NSFC (National Natural Science Foundation of China), Mar.2015-Feb.2016.

His current research interests include computer graphics, multimedia and artificial intelligence. He has authored more than 60 journal and conference papers in these areas, including ACM TOG, ACM TIST, IEEE TPAMI, IEEE TIP, IEEE TCSVT, ACM SIGGRAPH (Asia), ACM MM, ICCV, etc.