Multi-Person Pose Estimation via Learning Feature Integration

Xiaomei Hu1*, Jiahao Tian1,2, Huixiang Qiao3 and Minjie Wang1,2

1The Key Laboratory of Intelligent Manufacturing and Robotics, School of Mechatronic Engineering and Automation, Shanghai University, Shanghai, China
2Institute of Materials Technology and Engineering CAS, Ningbo315201, China
3School of Electronic Information and Control Engineering, Chang’an University, Xi’an 713100, Shaanxi, China

Email: *sufeimasohxm@163.com
{sufeimasohxm, tianjiahao_369, wangminjie920705}@163.com,
qiaohuixiang@outlook.com

Abstract. Human pose estimation is a fundamental but challenging task in computer vision. Although batch normalization is widely used for deep learning, feature extraction in deep convolutional neural networks (DCNN) is still not well explored. In this work, we propose a pose encoding module (PEM) to enhance the learning ability and generalization ability of feature extraction. Given input images, PEM integration instance normalization and batch normalization, combining them in an appropriate way to learn to capture and eliminate appearance changes while maintaining the distinction between learning features can be seen as an integration and adjustment of global information. In addition, we use a simple and efficient up-sampling strategy to recover high-resolution representations for predicting more accurate human keypoint heatmaps, which has achieved better performance than the average network in recent missions. We studied our approach on a standard benchmark for human pose estimation.

1. Introduction

Trying to give computers the ability to automatically understand human action information contained in images or videos has always been a hot topic in many machine learning related fields. Human pose estimation is an important basis for these tasks, and has wide applications in action recognition, human-computer interaction, human re-identification, audio-visual entertainment and other fields. Human pose estimation refers to the task of predicting person pose by locating body keypoints (head, shoulder, elbow, wrist, knee, ankle, etc.) from the image. We are concerned with the multi-person pose estimation problem in a single picture. Due to the complex variability of background, human appearance features and posture structure in natural pictures, this task faces many challenges. When the scene is extended to many people, the problem is further complicated.

Since the convolutional neural network [1] can automatically learn abstract features from big data, compared with the manual design features adopted in traditional human pose estimation algorithms, it can more accurately represent the real human appearance features. And the convolutional neural network model also has strong nonlinear mapping ability, which can more effectively realize the mapping from image features to human body posture.

In this work, we mainly focus on the convolutional neural network based method, and design a new network structure that learns feature integration. Our network architecture consists of two parts: feature encoding and pose parsing. The image feature encoding module based on global information
integration learns better features in the image, including instance-specific information that does not change in appearance. Our pose parsing module solves the problem of human joint positioning based on learning feature integration, and thus performs human pose estimation.

Our method has been applied to the COCO data set, and has achieved better results, resulting in higher joint positioning accuracy, and significantly reduced false detection and missed detection.

Figure 1. Overview of our framework.

2. The Proposed Approach

Our framework is shown in Figure 1. A two-step framework is applied: input an image of size W×H×3, adjust its size to the input size of CNN, and generate a set of human bounding boxes through the human body detector. The human body bounding box is input into the 'FEM+PPM' module, which sequentially predicts the positioning of each person’s keypoints, and generates k heatmaps $M_k$ of size $h \times w$ to indicate the positional confidence of the k-th keypoint. Finally, non-maximum suppression is performed to eliminate redundant poses and get the final human pose.

2.1. FEM and PPM

The stacked hourglass [1] is a popular pose estimation method that stacks eight hourglass modules, which are down-sampled and up-sampled with residual connections to enhance pose estimation performance. Inspired by this, Our network framework takes a similar structure. We found that using the existing good network structure for the down-sampling path and the simple up-sampling path is much better. Our network structure is based on the backbone network ResNet of the most commonly used image feature extraction, joins two sub-networks of FEM and PPM, and uses a normalization strategy different from the previous network to extract more detailed image features to predict more precise keypoint location.

The existing network for pose estimation is to calculate the mean and standard deviation for all pixels of all pictures in a batch. The feature specification method performs the following calculations:

$$
\hat{x}_i = \frac{1}{\sigma_i}(x_i - \mu_i)
$$

Where $x$ is the feature computed by the layer and $i$ is the index. In an RGB image, $i$ is a 4D vector with (N, C, H, W) ordering features, where N is the batch axis containing T images, C is the feature channel axis, and H and W are the height and width axes across the spatial dimension. The $\mu$ and $\sigma$ in (1) are the mean and standard deviation calculated by the following formula:

$$
\mu_i = \frac{1}{m} \sum_{k=1}^{m} x_k
$$

(2)
\[ \sigma_i = \sqrt{\frac{1}{m} \sum_{k \in S_i} (x_k - \mu_i)^2 + \varepsilon} \]  

(3)

\( \varepsilon \) is a small constant, \( S_i \) is the set of pixels for calculating the mean and standard deviation, and \( m \) is the size of the set. Batch normalization of pixels sharing the same channel index is normalized together, i.e., for each channel, batch normalization calculates \( \mu \) and \( \sigma \) along the \( (N, H, W) \) axis.

Due to the variability of the data, the mean and standard deviation of each batch are unstable, which is equivalent to the introduction of noise. To this end, we add the instance normalization to the network. Unlike the former, instance normalization is to calculate the mean and standard deviation for all pixels of a single picture. The information is from its own picture, which can prevent the instance-specific mean and covariance conversion. From a certain perspective, it can be seen as an integration and adjustment of global information. It is also a more stable method for training. The instance normalized feature specification method is the same as (1), except that the calculation of \( \mu \) and \( \sigma \) is:

\[ \mu_i = \frac{1}{HW} \sum_{w=1}^{W} \sum_{h=1}^{H} x_{iwh} \]

(4)

\[ \sigma_i = \sqrt{\frac{1}{HW} \sum_{h=1}^{H} \sum_{w=1}^{W} (x_i - w\mu_i)^2} \]

(5)

Figure 2. Shows our FEM and our PPM strategy.

To this end, we propose a novel convolutional structure, FEM, that learns to capture and eliminate appearance changes while maintaining a distinction between learning features. Integrate instance normalization [2] (IN) and batch normalization (BN) into building blocks to enhance their learning and generalization capabilities. It has two attractive benefits that were not available in previous deep architectures. First, unlike the CNN structure that previously isolated IN and BN, FEM unifies them by delving into their learning characteristics. Combining them in an appropriate way can improve learning and generalization. Secondly, our FEM maintains shallow IN and BN features and higher BN features, inheriting the statistical characteristics of feature deviations at different depths of the network. As shown in the lower left of Figure 2, in order to preserve the image content information in the shallow layer, we replace the original BN layer with IN to obtain half of the features, and replace BN with the other half. These have produced our FEM.

Each time we perform a down-sampling operation, we double the number of channels, which can effectively reduce information loss. In addition, computing power is primarily allocated to the down-sampling unit rather than the up-sampling unit. This is reasonable because our goal is to extract more representative features during the down-sampling process and it is difficult to recover lost information.
during the up-sampling process. Therefore, it is more effective to increase the capacity of the down-sampling unit.

The PPM we designed is placed behind the down-sampling feature extraction, using a simpler up-sampling method after the last layer of resnet. As shown in Figure 2, the feature resolution conversion by PixelShuffle; the up-sampling layer consisting of convolution, batch normalization, ReLU activation and PixelShuffle; and the deconvolution layer combined by transposition convolution, batch normalization and ReLU activation. They are combined to form our pose parsing feature aggregation method. Finally, a 3x3 convolutional layer is added to generate a predicted heatmap for all k keypoints. For each up-sampling operation, we reduce the number of feature channels by half to ensure effective attention to features on the channel. As shown in Figure 3, using our network for pose estimation, it greatly facilitates key location and significantly improves results compared to previous papers [6].

![Image](image_url)

**Figure 3.** Human body key point positioning. On the left is the basenet predicted pose and keypoint heat map, and the right image is the result of our "FEM+PPM" improvement.

### 2.2. Loss Function

We simply return to the heatmap from the high resolution representation of the last exchange unit output, which is empirically effective. We describe the loss function L used to train the pose estimate. The loss L is defined by summing the L2 losses of the heatmaps of all keypoints, similar to the loss function in [1]. To detect k = 17 keypoints, k heat maps are generated after the last convolution. The loss compares the predicted heatmaps of all keypoints against the groundtruth heatmaps:

$$L = \frac{1}{k} \sum_{k=1}^{k} \sum_{x,y} \left\| S_p(x,y) - S_g(x,y) \right\|^2$$  \hspace{1cm} (6)

Here, $S_p(x,y)$ and $S_g(x,y)$ represent the prediction and groundtruth confidence map at the pixel position $(x, y)$ of the k-th keypoint, respectively. Same as the previous generation of the groundtruth heatmap, where the k-th keypoint groundtruth heatmap $S_p(x, y)$ is generated using 2D Gaussian centered on the keypoint position $(x, y)$. The standard deviation is 1 pixel. Figure 1 shows the predicted heatmap for some keypoints. We obtained the final recommendation using NMS with a threshold of 0.6.

### 3. Experiment

#### 3.1. Dataset and Evaluation Protocol

In our experiments, we trained our keypoint model on the COCO keypoint dataset (without any external/extra data). The COCO data set contains more than 200,000 images and 250,000 individual instances with 17 keypoints. The COCO assessment defines object keypoint similarity (OKS) with the formula:
\[ \text{OKS} = \frac{\sum_i \exp \left( -\frac{d_i^2}{2s^2k^2} \right) \delta(v_i > 0)}{\sum_i \delta(v_i > 0)} \]  

The \( d_i \) are the Euclidean distances between each corresponding ground truth and detected keypoint and the \( v_i \) are the visibility flags of the ground truth, \( s \) is the object scale, and \( k_i \) is the constant of each keypoint that controls the attenuation. We used the average precision (AP) of 10 OKS thresholds as the evaluation indicators.

**Figure 4.** Precision-recall curves on COCO validation set across scales all, large and medium.

### 3.2. Result

**Results on the validation set** We used yolov3 as the human body detector. We report the results of our method and other advanced methods in Table 1. In addition, we present the recall-precision curves of our method for different scales all, large, medium in Figure 4. Our network input size is 320×256, reaching an AP score of 73.6, which is better than other methods with the same input size.

**Table 1.** Comparison of COCO val2017 sets. OHKM = online hard keypoints mining[3].

| Method                | Backbone         | Input size | AP  | AP50 | AP75 | AM  | AL  | AR  |
|----------------------|------------------|------------|-----|------|------|-----|-----|-----|
| 8-stageHourglass[1]  | 8-stageHourglass | 256×192    | 66.9|      |      |     |     |     |
| CPN[3]               | ResNet-50        | 256×192    | 68.6|      |      |     |     |     |
| CPN+OHKM[3]          | ResNet-50        | 256×192    | 69.4|      |      |     |     |     |
| RMPE[5]              | stage Hourglass  | 320×256    | 70.9| 88.0 | 79.2 | 69.0| 78.3| 76.4|
| Simple Baseline[4]   | ResNet-101       | 256×192    | 71.4| 89.3 | 79.3 | 68.1| 78.1| 77.1|
| Ours                 | ResNet-101       | 320×256    | 73.6| 90.5 | 80.8 | 69.9| 79.1| 77.2|

**Results on the test-dev set** As shown in Table 2, we implemented 67.7AP on test-dev only through a single model trained by COCO data, and is much better than other methods in all indicators.

**Table 2.** Comparison of COCO test-dev data sets.

| Method               | Backbone         | Input size | AP  | AP50 | AP75 | AM  | AL  | AR  |
|---------------------|------------------|------------|-----|------|------|-----|-----|-----|
| CMU Pose[6]         | —                | —          | 61.8| 84.9 | 67.5 | 57.1| 68.2| 66.5|
| Mask R-CNN[7]       | Res-50-FPN       | 353×257    | 63.1| 87.3 | 68.7 | 57.8| 71.4|     |
| G-RMI[8]            | ResNet-101       | 320×256    | 64.9| 85.5 | 71.3 | 62.3| 70.0| 69.7|
| PersonLab[9]        | ResNet-101       | —          | 65.5| 87.1 | 71.4 | 61.3| 71.5| 70.1|
| Ours                | ResNet-101       | 320×256    | **67.7**| **88.5**| **75.2**| **63.9**| **73.2**| **72.1**|
Finally, some of the pose estimation results generated by our method are shown in Figure 5. We can see our framework dealing with crowds and occlusions as well as effectively challenging postures.

4. Conclusion
In this paper, we present a network for multi-person pose estimation that produces accurate keypoint heatmaps. FEM learns to capture and eliminate appearance changes while maintaining a distinction between learning features to enhance their learning and generalization skills. In addition, a simple and effective up-sampling method is used for the extraction and localization of human keypoints.

Future work includes applying to other intensive predictive tasks such as semantic segmentation, object detection, face alignment, image conversion, our FEM and other feature extraction strategies are generic and can help with other tasks.

Figure 5. Qualitative results of some example images in the COCO datasets.

5. References
[1] Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European Conference on Computer Vision. pp. 483-499. Springer (2016)
[2] Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance normalization: The missing ingredient for fast stylization (2016)
[3] Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi person pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7103-7112 (2018)
[4] Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 466-481 (2018)
[5] Fang, H.S., Xie, S., Tai, Y.W., Lu, C.: Rmpe: Regional multi-person pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2334-2343 (2017)
[6] Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7291-7299 (2017)
[7] He, K., Gkioxari, G., Doll’ar, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961-2969 (2017)
[8] Papandreou, G., Zhu, T., Kanazawa, N., Toshev, A., Tompson, J., Bregler, C., Murphy, K.: Towards accurate multi-person pose estimation in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4903-4911 (2017)
[9] Papandreou G, Zhu T, Chen L C, et al. PersonLab: Person Pose Estimation and Instance Segmentation with a Bottom-Up, Part-Based, Geometric Embedding Model. In ECCV. 2018.