Related Work

Deep Pose- Deep Neural Networks (DNNs):

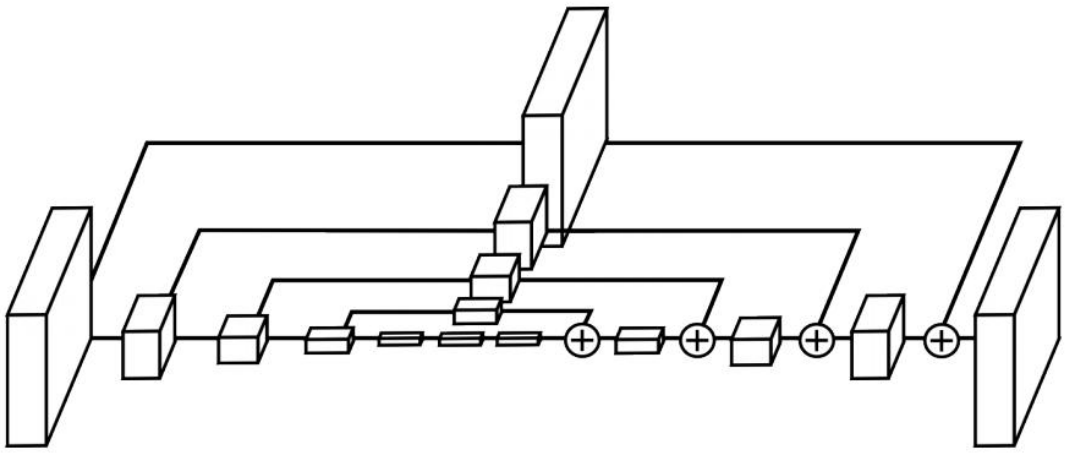
Input a image,output 2*k coordinates of joints' locations by DNNs.
Loss function:

$$\arg\min_{\theta} \sum_{(x,y)\in D_N} \sum_{i=1}^{k} ||\mathbf{y}_i - \psi_i(x;\theta)||_2^2$$

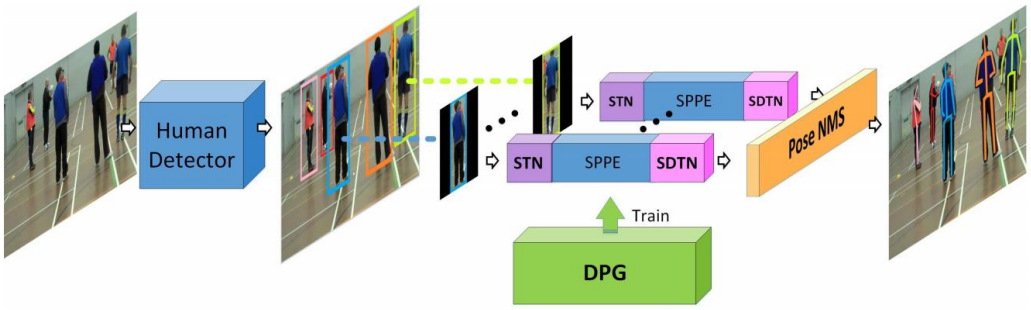
Staked Hourglass Network (SPPE)

A final pose estimate requires a coherent understanding of the full body. The person's orientation, the arrangement of their limbs, and the relationships of adjacent joints are among the many cues that are best recognized at different scales in the image. The hourglass is a simple, minimal design that has the capacity to capture all of these features and bring them together to output pixel-wise predictions.

Repeated bottom-up, top-down inference with stacked hourglasses integrate local and global cues.We can also do a intermediate supervision training.



Introduction



The human proposals obtained by human detector are fed into “Symmetric STN + SPPE” module and pose proposals will be generated individually. These proposals will be further refined by parametric Pose NMS. To augment the existing training samples, deep proposals generator is designed to largely augment training samples.

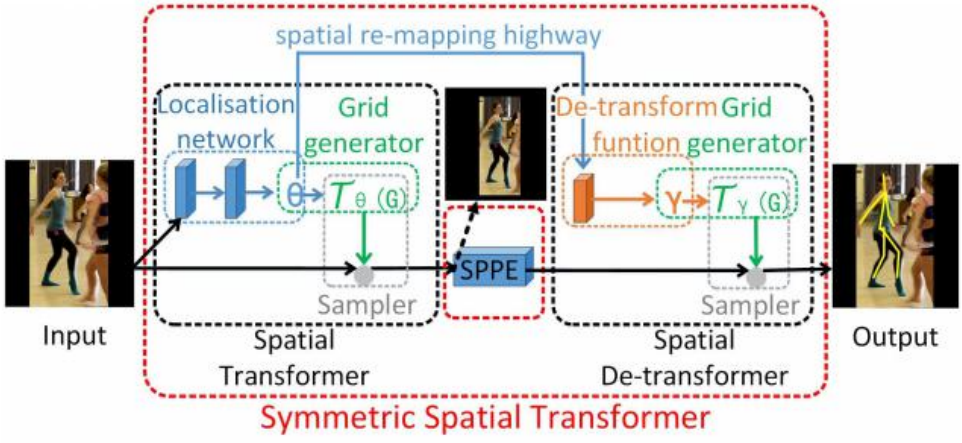
Implementing Details

Proposal Extension

In order to guarantee the entire person region to be extracted, human proposal is extended with 20% along both height and width direction.

STN and SDTN

STN(spatial transformer network). SDTN is the inverse procedure of STN.



Spatial Re-mapping Highway

Jointly optimize STN and SDTN by back propagation technique.

$$\begin{bmatrix}\gamma_1 & \gamma_2\end{bmatrix} = \begin{bmatrix}\theta_1 & \theta_2\end{bmatrix}^{-1}$$

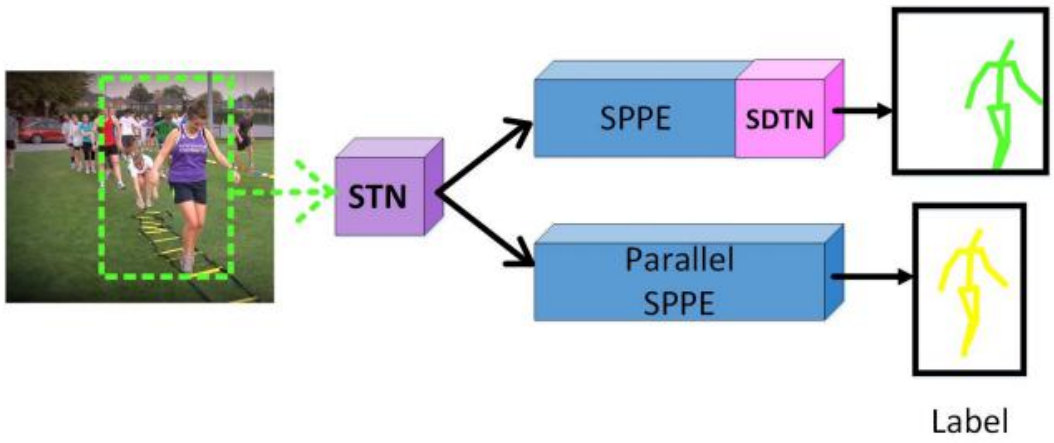
$$\gamma_3 = -1 \times \begin{bmatrix}\gamma_1 & \gamma_2\end{bmatrix} \theta_3$$

$$\frac{\partial J(W,b)}{\partial \begin{bmatrix}\theta_1 & \theta_2\end{bmatrix}} = \frac{\partial J(W,b)}{\partial \begin{bmatrix}\gamma_1 & \gamma_2\end{bmatrix}} \times \frac{\partial \begin{bmatrix}\gamma_1 & \gamma_2\end{bmatrix}}{\partial \begin{bmatrix}\theta_1 & \theta_2\end{bmatrix}} + \frac{\partial J(W,b)}{\partial \gamma_3} \times \frac{\partial \gamma_3}{\partial \begin{bmatrix}\gamma_1 & \gamma_2\end{bmatrix}} \times \frac{\partial \begin{bmatrix}\gamma_1 & \gamma_2\end{bmatrix}}{\partial \begin{bmatrix}\theta_1 & \theta_2\end{bmatrix}}$$

$$\frac{\partial J(W,b)}{\partial \theta_3} = \frac{\partial J(W,b)}{\partial \gamma_3} \times \frac{\partial \gamma_3}{\partial \theta_3}$$

Parallel SPPE

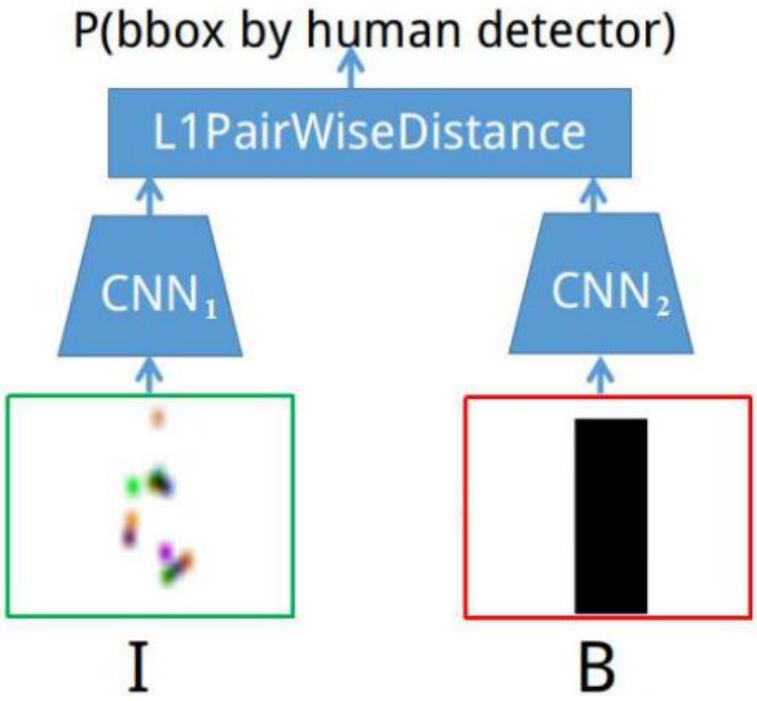
We minimize the sum error of parallel SPPE and symmetric STN together. In testing phase, we discard parallel SPPE and forward in Symmetric STN only. In experiment section, we verify that this additional scheme improve the system indeed.



Deep Proposals Generator

The quantity of training data is not sufficient for training a good model. Therefore, proper augmentation of the training data should be considered, where the augmented data should follow same distribution as human detector’s output. So A novel deep proposals generator is proposed to produce a large amount of high quality augmented samples.

We randomly sample a huge number of region proposals. The network will select part of them as augmented samples according to their output scores (possibilities). In this way, a large number of high quality training samples can be obtained. In particular, box proposal is represented by the proposal indicator image that assigns pixels inside proposal as black and outside ones as white. Pose is captured by labeling different joints with different colors.



Parametric Pose NMS

This part intends to eliminate redundant pose for similar poses or redundant detections.
Firstly, most confident pose is selected as reference, and some poses close to it are eliminated by an appropriate elimination criterion.

As before, the pose Pi with m joints is denoted as $\{\langle k_i^1, c_i^1\rangle, \ldots, \langle k_i^m, c_i^m\rangle\}$, where k_i^j and c_i^j are the j^{th} location and confidence score of joints.

The condition means ‘if k_j^n near k_i^n’.

$$H_{Sim}(P_i, P_j|\sigma_2) = \sum_n \exp[-\frac{(k_i^n - k_j^n)^2}{\sigma_2}]$$

$$d(P_i, P_j|\Lambda) = K_{Sim}(P_i, P_j|\sigma_1) + \lambda H_{Sim}(P_i, P_j|\sigma_2)$$

And if $d(P_i, P_j|\Lambda)$ is smaller than threshold then we can eliminate Pi when Pj existed.

As above, there are four hyperparameters.

Since exhaustive search in the 4D space is intractable, we optimize two parameters by fixing other two parameters in an iterative manner. Once convergence, the elimination criterion will be determined.

Results

Results on MPII

	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total	time [s/frame]
Iqbal&Gall [31]	58.4	53.9	44.5	35.0	42.2	36.7	31.1	43.1	10
DeeperCut [16]	78.4	72.5	60.2	51.0	57.2	52.0	45.4	59.5	485
ours, 2-stack SPPE	85.5	81.5	71.3	62.1	65.0	62.7	56.5	69.2	**0.8**
ours, 8-stack SPPE	**87.7**	**85.1**	**77.3**	**68.5**	**73.5**	**70.9**	**63.8**	**75.3**	**0.8**

Table 1. Results on MPII multi-person full test set (mAP). Note that the 2-stack SPPE and 8-stack SPPE have the same depth but different structure. For more details please refer to [21]

Compared to the previous stateof-the-art results, our method could achieve significant improvements over the previous method by 16 mAP. Notably, we achieve an average accuracy of 70 mAP on identifying difficult joints such as wrists, elbows, ankles, and knees, which is 18 mAP higher than the most recent stateof-the-art result. In terms of computational speed, our method is also much faster than the part-based method that can achieve the state-of-the-art performance[16]. The whole procedure is speed up 600 times.

Results on WAF

Only evaluate method on WAF test set by using our model trained on MPII datasets.

	Head	Shoulder	Elbow	Wrist	Total
Chen&Yuille [6]	83.3	56.1	46.3	35.5	55.3
DeepCut [24]	76.6	80.8	73.7	73.6	76.2
DeeperCut [16]	92.6	81.1	75.7	78.8	82.0
ours, 2-stack SPPE	**95.6**	**83.5**	**88.3**	**81.6**	**87.3**

Results on MSCOCO Keypoints dataset

Team	**AP**	AP^{50}	AP^{75}	AP^{M}	AP^{L}
CMU-Pose[5]	**61.8**	**84.9**	**67.5**	57.1	**68.2**
G-RMI	60.5	82.2	66.2	57.6	66.6
DL-61	54.4	75.3	50.9	**58.3**	54.3
R4D	51.4	75	55.9	47.4	56.7
umich_vl	46	74.6	48.4	38.8	55.6
Caltech	40.2	65.2	41.9	34.9	49.2
ours, 8-stack SPPE	57.2	81.0	64	53.6	63.3

Ablation Studies

	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
RMPE, 2-stack SPPE	**85.5**	**81.5**	**71.3**	**62.1**	**65.0**	**62.7**	**56.5**	**69.2**
rm SSTN+parallel SPPE	81.5	78.8	67.3	58.9	61.6	58.1	49.7	65.1
rm parallel SPPE only	82.1	79.8	69.0	59.3	61.0	59.6	52.8	66.2
rm DPG	81.8	77.8	66.9	57.4	60.4	53.9	40.1	62.6
rm PoseNMS	76.7	73.9	64.5	56.5	59.7	58.2	52.7	63.2
PoseNMS [6]	82.1	78.8	69.1	60.3	63.1	61.1	55.1	67.1
straight forward two-steps	71.4	68.3	58.8	51.4	49.9	52.3	50.2	57.5

We observe apprarent performance degradation without parallel SPPE, which implies that parallel SPPE with single person image labels would strongly encourage the STN to extract single person regions with minimized errors. Without DPG, the whole SSTN module has much less data to learn during training phase, where the degraded performance occurs.

The mAP drops significantly if the parametric Pose NMS is removed.The state-of-the-art pose NMS algorithm is used to replace our parametric Pose NMS, where the result is given in the picture. This scheme performs less effective compared to ours, since the parameter learning is missing.