Athlete pose estimation by a global-local network

Jihye Hwang, Sungheon Park and Nojun Kwak

Graduate School of Convergence Science and Technology, Seoul National University, Seoul, Korea

hjh881120@gmail.com, sungheonpark@snu.ac.kr, nojunk@snu.ac.kr

Abstract

Analyzing joint movements of an athlete helps to improve the pose of the athlete. Human pose estimation (HPE) algorithms regress the locations of parts such as wrists, ankles and knees. In this paper, we propose a network that combines global and local information for HPE using a 2D image. Unlike previous works that have used global or local information separately, we use the combined information to enhance the performance of HPE. General information from a global network is used as an input to a local network to refine the location of a part using a variety of regions. The global network is based on ResNet-101 [6] and trained to regress a heatmap representing parts' locations. The output features from the global network are used as input features for the local network. The local network learns spatial information using position sensitive score maps [11]. Through the end-to-end learning, the global network is affected by the local information. We demonstrate that the proposed HPE method is efficient on the LSP and UCF sports datasets.

1. Introduction

In sports, not only managing the condition of an athlete, but also analysis of the data after the game is important to improve the competence of the athlete. Especially, methods for video or image data analysis have long been used to analyze the ability of an athlete. In the past, sports experts analyzed athletics videos through their domain knowledge. This analysis is highly subjective and requires a lot of feedback to get justified. Thus, an auxiliary objective analysis is needed.

Recently, a lot of high quality sports media datasets are available and we can access a significant volume of analyzed data. In the analysis of sports datasets, various computer vision methods can be utilized. Especially, a pose estimation method helps to correct athlete's posture. Because sports are environmentally constrained and have a lot of rough motions, obtaining 3D information is quite challenging and 2D data obtained by a single camera is commonly used.

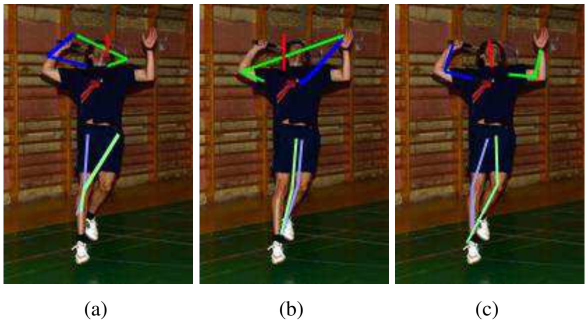


Figure 1. Comparison of HPE results from different algorithms. (a) local network, (b) global network, (c) global-local network (proposed). Red, Blue and Green line are head, right arm and left arm respectively. Light blue is right leg and light green is left leg. Resnet-101 is used as a base network.

Previous works on HPE can be categorized based on whether an algorithm uses general (global) information or partial (local) information [9, 17, 7, 18, 13, 2, 14, 3, 16, 20, 15]. In case of using local information, most of the previous works used contextual relations between local parts. Pictorial structures [5], which have been successful on images where all the limbs are visible, learn the pairwise geometric relations between parts using a tree model. However, this tree-based model has a limitation in expressing complex human pose and the proposed hand-crafted features are sensitive to noise. Fan et al. [4] adapted the local part appearance using the dual-source deep convolutional neural network (CNN). They take a set of image patches extracted by a region proposal method such as Edgebox [21] and train the local appearance by considering their holistic views in the full body.

CNNs have shown outstanding results for HPE and many CNN-based researches used global information [17, 7, 18, 13, 2]. Because CNN is robust to variations and can extract good representations without using hand-crafted lowlevel features, these methods have outperformed in estimating the pose using the full image. Among these, DeepPose [17] applied holistic way regression to refine the joint locations around the initial estimation. Deepercut [7] introduced a strong body part detector and used a very deep network trained with image-conditioned pairwise terms between body parts. Recently, many methods iteratively train the same convolutional networks, each of which taking the features from the previous stage [18, 13]. Stacked hourglass [13] repeated bottom-up, top-down processing and Wei et al. [18] concatenated the same structure as a pose machine. Those deeper networks have a large receptive field to explicate global context.

Figure 1(a) and (b) are exemplary results that use only local or global networks. On the other hand, Fig. 1(c) is a result by the proposed method which combines global and local networks into one. In the figure, we can see that the proposed local-global network performs better than the other two. As can be inferred from this example, because humans have very diverse poses, using both global and local information is needed for accurate pose estimation.

Previous works have insisted that a receptive field is possible to explain spatial information as well as global information. However, the receptive fields of higher convolutional layers in a deep network cover large parts of the image instead of explaining spatial information and the fixed sizes of strides regulate the position of corresponding receptive fields. Thus, we propose the refine network that combines both the local and global information to estimate the sports pose as shown in Figure 2. The contributions of this work are as follows:

1) We adapted various region sizes and scales to detect the proposal regions (ROIs) directly.

2) The inputs to the local network are features concatenated with global features and heatmaps of the global network. This method is effective to refine local features guided by global features.

3) The combined global-local network can be trained end-to-end.

The proposed method consists of two parts: the global (general) network and the local (partial) network. In the global network module, we estimated heatmaps [1] of joints based on the ResNet-101. We adjusted the stride of the convolutional layers to increase the resolution of the heatmaps.

We trained the network using a pixel-wise L_2 loss function to create an estimation of heatmaps. Then, the feature maps of the final convolutional layer are concatenated by the heatmaps of global network, which are inputted to the local network to refine the location. In the local network, position-sensitive score maps are created to explain the spatial information on region of interest (ROI) as in [11], where region-based fully convoluitonal networks was proposed for object detection.

We test the proposed method on the Leeds sports pose (LSP) and UCF sports datasets. To show the effectiveness of the proposed approach, we compared two baseline algorithms: ResNet using local network loss and ResNet using global network loss. Also we compared a performance with recent pose estimation methods [19, 4]. On the LSP dataset, the average of percentage corrected keypoints (PCK) is 81.8% on a normalized threshold and the PCK of head position is 96.2% which is competitive to the state-of-the-art methods. The UCF sports dataset contains a variety of sports videos and we evaluated a qualitative results.

The rest of the paper is organized as follows. In Section 2, we present the proposed method and Section 2.1 and 2.2 describe how the global network and the local network are designed, respectively. In Section 3, we demonstrate the efficiency of the proposed method through the evaluation on the LSP and UCF datasets. Finally, we conclude in Section 4.

2. Method

We propose a global-local combined network for HPE: the first (global) network is a big deep network which estimates locations of parts using the global features, and the second (local) network is a small network which modifies the parts' locations using local information. In the first network, we regress the heatmaps of parts using a pixel-wise L_2 loss. The ground truth of the heatmaps are generated as Gaussian maps centered at the ground truth locations of body parts. The output of this network is fed into the small network to refine the location of each joint. Position-sensitive ROI pooling based on R-FCN [11] is applied to the small network. Detailed explanation on the global and local network is given in the following subsections.

2.1. Global Regression Network

In this module, we predict the position of each part using a wide range of global information. Because human pose estimation is a highly non-linear problem, it is difficult to directly regress the locations of the parts. Rather than directly regressing the position of the parts, we followed a simple alternative method which regressed a set of heatmaps centered at the visible target joints as in [1]. For the visible joints, the heatmaps are generated as Gaussian maps, each of which is centered at the location of the corresponding joint with standard deviation of 5.

To jointly regress the location of each part, we used the ResNet-101 model [6] which has a very large receptive field as shown in Figure 2. The large receptive field can explain the global context. We enlarged the size of the output feature maps of the network by adjusting the stride of the convolutional layers in order to increase the resolution of

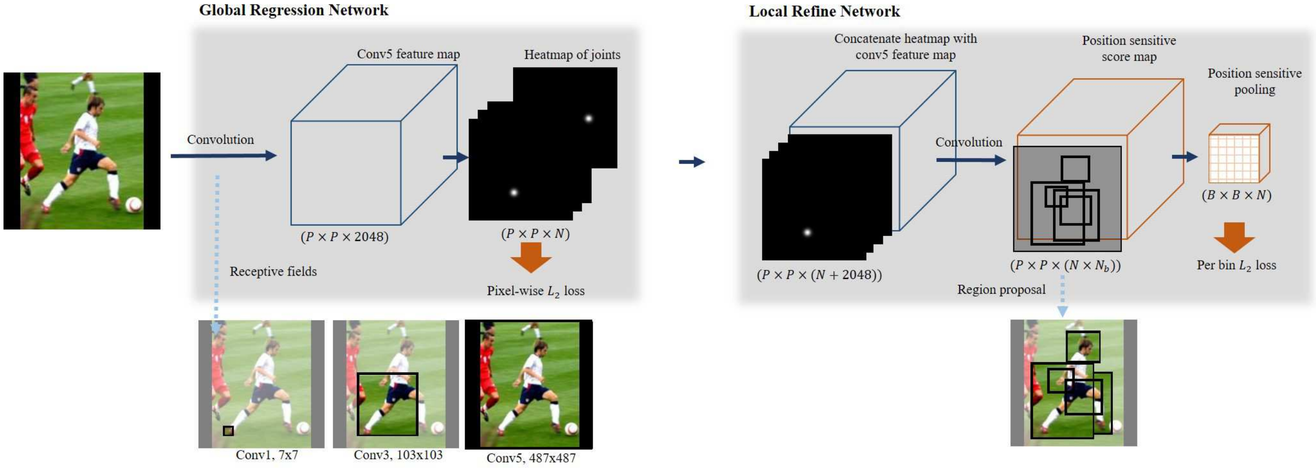


Figure 2. Overall architecture of the proposed method (global-local network). The output of the global network is used as an input to a local network to refine the location using a variety of region proposals. On the left, we show each receptive field of features after the corresponding convolution layer with its size (e.g., 7 × 7 for conv1 layer). On the right, several region proposals are shown.

the heatmaps. Specifically, we changed the stride of the *conv3*, *conv4* and *conv5* layers to 2,1 and 1 respectively, which yields the size of the output feature map (*conv5*) to be 14×14.

We trained the network with a pixel-wise L_2 loss, where the loss function is as follows:

$$L_g = \frac{1}{N} \sum_{n=1}^{N} \sum_{x,y} \left\| H_n(x,y) - \bar{H}_n(x,y) \right\|^2. \qquad (1)$$

Here, H_n means the predicted heatmap and $\bar{H}_n$ is the ground truth heatmaps. N is a number of parts and (x, y) means pixel locations of a heatmap.

2.2. Local Refine Network

We refine the location of the parts effectively using combined features and direct local evidences. Combined features were created from the global network, which is a concatenation of the output of *conv5* and the heatmaps. Those features represent the global features based on the regressed locations on the heatmaps.

Usually, lower convolutional layers can contain local information since the receptive fields cover small parts of the image. However, as passing convolutional layers more and more, the receptive fields of higher convolutional layers gets larger. At the bottom-left of Figure 2, receptive fields of different convolutional layers (*conv1*, *conv3*, and *conv5*) from the ResNet-101 are shown. It can be observed that the receptive field of the last layer of the ResNet-101 network covers almost the entire image. Moreover, since the convolutional layers have fixed sizes of strides, the corresponding receptive fields are regularly positioned. Different from the previous works, we exploited local evidences explicitly by extracting ROIs. ROIs have various scale, size and positions, and they contain a variety of combinations of body parts. These characteristics help to increase the expressiveness and generalization power of the network. At the bottom-right of Figure 2, the examples of ROIs using the Edgebox [21], which is one of the most popular methods in ROI extraction, are shown.

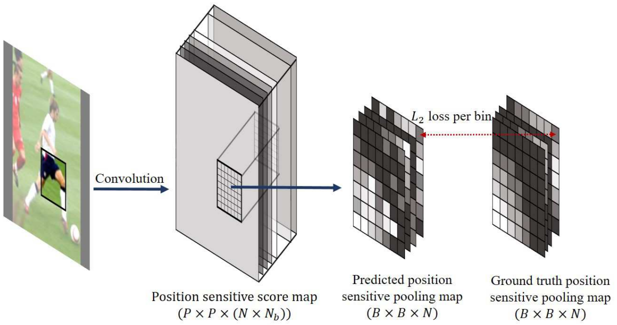

Convolution
L_2 loss per bin
Position sensitive score map $(P \times P \times (N \times N_b))$
Predicted position sensitive pooling map $(B \times B \times N)$
Ground truth position sensitive pooling map $(B \times B \times N)$

Figure 3. The position-sensitive model applied to our local network.

To learn the local information, we applied the position-sensitive score maps and the position-sensitive ROI pooling from R-FCN [11], which are the architectures for position-sensitive object detection. In R-FCN, each feature map in position-sensitive score maps contain information of a specific position. In our method, the position-sensitive score map has $N_b \times N$ channel to describe spatial information for each joint. Here, N is the number of parts and N_b is the number of bins to which ROIs are divided. Note that $N_b = B \times B$ in the figure. The position-sensitive ROI pooling is applied to the score maps to generate the feature maps that are used to locate the body parts as shown in Figure 3.

Layer name	Output size	Layer size
conv6	$P \times P$	$1 \times 1, 2048, 1$
conv7	$P \times P$	$1{\times}1, N_b{\times}N, 1$
PS-ROI pooling	$B \times B \times N$	

Table 1. The structure of the local refine network. The values in the layer size tap means (kernel, channels, stride). P is the output size of *conv6*, which depends on the stride of the global network. N is the number of parts and N_b is the number of bins.

Table 1 is the structure of the proposed local network. The local network consists of one convolutional layer, one ReLu layer, and one position-sensitive score map layer. The output size P depends on the stride of the global network. We used average pooling as a method of the position-sensitive ROI pooling. The output value of b -th bin after the pooling is calculated as

$$r(b) = \frac{1}{E} \sum_{(x,y) \in bin(b)} C_b(x_0 + x, y_0 + y) \tag{2}$$

where E is the number of elements in a feature map that are inside the b-th bin, C_b is the value of the feature map that corresponds to the b-th bin, (x_0, y_0) is the offset of the top-left corner of an ROI, and (x, y) are the offsets in the b -th bin. .

To produce the ground truth for the local network, we applied the selective pooling to the region of the ground-truth heatmap that corresponds to the ROI. The average score value in each bin was trained using a L_2 loss where the loss function is as follows:

$$L_l = \frac{1}{N} \sum_{n=1}^{N} \left(\frac{1}{N_b} \sum_{b=1}^{N_b} \left\| r_n(b) - \bar{r}_n(b) \right\|^2 \right). \tag{3}$$

Here, $r_n(b)$ is the value after the selective pooling in the b -th bin of the n -th joint and $\bar{r}_n(b)$ is the corresponding ground truth heatmap.

Figure 4 shows a visualization of the results of position-sensitive ROI pooling on two region proposals. E_1 and E_2 are the region proposals extracted from Edgebox. The pooled features that are used to locate right ankle, right shoulder, and head are visualized for both region proposals. It is verified that the proposed local network successfully locate the joints in each region proposal. For example, because E_1 included the head and the shoulder but not the right ankle, the pooled features for the right ankle have low values while the values of the other joints are high at the position of the joints.

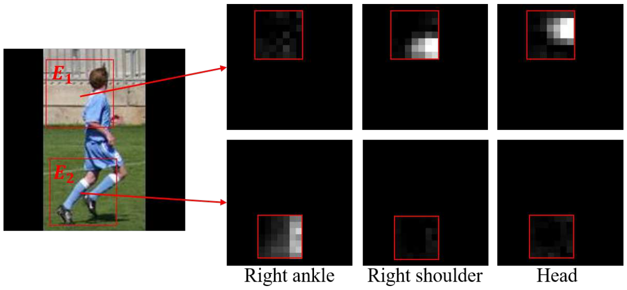


Figure 4. Visualization of position-sensitive score maps on two different ROIs, E_1 and E_2.

3. Experiments

3.1. Dataset

In order to assess the practical validity of the proposed method, experiments on the Leeds sports pose(LSP) dataset and the UCF sports dataset were performed.

The LSP dataset contains dynamic images of people in sports games gathered from Flickr. It contains various sports such as Badminton, Baseball, Gymnastics, Tennis and so on. The dataset consists of 11,000 training images including the extended data and 1,000 testing pose images which are annotated with person-centric (PC) 14 joint locations. The input image was adjusted to 224 × 224 pixels for training. We cropped each image such that a person is located at center and minimized the background. The images are zero padded to maintain the person in high resolution.

The UCF sports dataset is sports action dataset collected from broadcast television channels such as the BBC and ESPN. The dataset contains a wide range of sport actions and scenes such as golf swing and kicking. A total of 150 sequences are included in the dataset. Because UCF sports dataset have been usually used for sports action recognition, it does not have annotation of joints. Although we couldn't provide the quantitative results, we provided the qualitative results to demonstrate the applicability of the proposed method on sports videos. As in the LSP dataset, we adjusted the input image to 224 × 224 pixels.

3.2. Implementation details

We implemented our model using the open-source library Caffe [8]. We used a weight decay of 0.0005 and momentum of 0.9. For each iteration, the network accepted 1 image as an input. ROIs of the input image are generated via EdgeBox [21]. Among them, 20 ROIs are randomly selected and are fed to the local network. We used the ResNet-101 model pre-trained on ImageNet [10] as a base network. The proposed network was fine-tuned using a learning rate of 0.0001. Training is done in two stages. First, the global network is trained except the local refine network, and then, starting from the weights of the first stage, the full network

	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
Local	74.8	63.9	44.7	29.7	66.6	47.9	28.3	50.8
Global	89.3	71.5	58.0	51.0	70.5	66.5	62.5	67.0
Global(14)-local	91.8	76.0	64.7	58.6	76.9	72.9	68.8	72.8
Global(14)-local*	92.3	79.1	69.2	62.9	80.8	76.0	71.5	76.0
Global(28)-local*	96.2	85.4	76.1	71.2	85.7	81.8	76.2	**81.8**
Fan[4] et al. CVPR'15	92.4	75.2	65.3	64.0	75.7	68.3	70.4	73.0
Yang [19] et al. CVPR'16	90.6	78.1	73.8	68.8	74.8	69.9	58.9	73.6

Table 2. PCK-based comparison on LSP. A threshold value was measured at 0.2 (@0.2). The mark * indicates weights from the additional fine-tuning step is used.

is trained to refine the heatmaps. For inference, we fed the input image of size 224 × 224, and then the output heatmap of the global network was used as a final result. We empirically determined the size of the bins used for the position-sensitive pooling as 7 × 7, i.e., $N_b = 49$.

3.3. Evaluation

Several methods are used to evaluate the performance of human pose estimation in the literature: Percentage of corrected parts (PCP) and Percentage of correct keypoints (PCK). PCP calculates the detection rate of limbs. A limb is considered as detected if the distance between the detected limb position and the ground truth limb position is smaller than half of the limb length. Due to the penalization of the short limbs in PCK, the other methods, in which normalization is done with respect to the human torso, are introduced. In this paper, the performance is reported in PCK. In PCK, the distance between the estimated and the ground truth joints is normalized with respect to the torso size of the target.

Results on the LSP dataset are shown in Table 2 and Figure 5. We compared the performance of the proposed global-local network (Global-local) with the case that only the global network is used (Global), the case that only the local network is used (Local), and recent human pose estimation methods [19, 4]. Global network was based on the ResNet-101 and used L_2 pixel-wise loss to regress heatmap. Local network was also based on the ResNet-101, and L_2 loss is used for the position-sensitive score maps. Yang et al. [19] proposed a combined network with the expressive deformable mixture of parts. Fan et al. [4] proposed a dual-source CNN without using any explicit graphical model. They used the local information in image patches. Unlike our method, they put the cropped image on input image. We compared those methods as representative methods of exploiting global information [19] and local information [4]. In the method tab of Table 2, the numbers in parentheses are the size of the output heatmap.

The performance of the local network and the global network were 50.8% and 67% in terms of PCK accuracy respectively. The accuracy of the proposed global-local network is 5.8% higher than that of the global network. From the results, we can see that using only local or only global feature is insufficient for expressing complex human poses. To boost the performance, we added intermediate fine-tuning step before training the global-local network. For the model, we added the deconvolution layer [12] after the last convolutional layer of the ResNet. The deconvolution layer upsamples the size of the feature maps to 224 × 224. Then, 224 × 224 heatmaps are generated, which are trained using the L_2 loss with the ground truth heatmaps. We found that using the weight trained from the intermediate fine-tuning step improves the performance. In Table 2, methods with the postfix (*) are the networks trained from the weights that comes from the intermediate step. It can be seen that the intermediate fine-tuning improves the PCK performance by 3.2%. As stated Section 2.1 that the stride of the convolutional layers inside the ResNet is adjusted to control the output heatmap size, we tested two different output heatmap sizes, 14 × 14 and 28 × 28 to show the importance of the output heatmap size. When the output heatmap size is doubled, PCK has been improved from 76.0% to 81.8% by 5.8%.

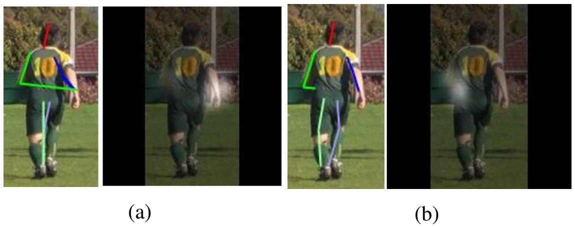


Figure 5. Compared result of global network and the proposed method. (a) Global network, (b) Global(14)-local network. For both (a) and (b), the left image shows the position of body parts and the right image is the original image overlapped with the heatmap of the left wrist.

The model that shows the best performance is Global(28)-local* model of which PCK@0.2 is 81.8%. Note that the PCK of head regression shows superior results to the compared methods [19, 13]. Figure 6 shows the PCK curve according to the normalized distance of each part. It can be seen that the proposed Global(28)-local* model out-

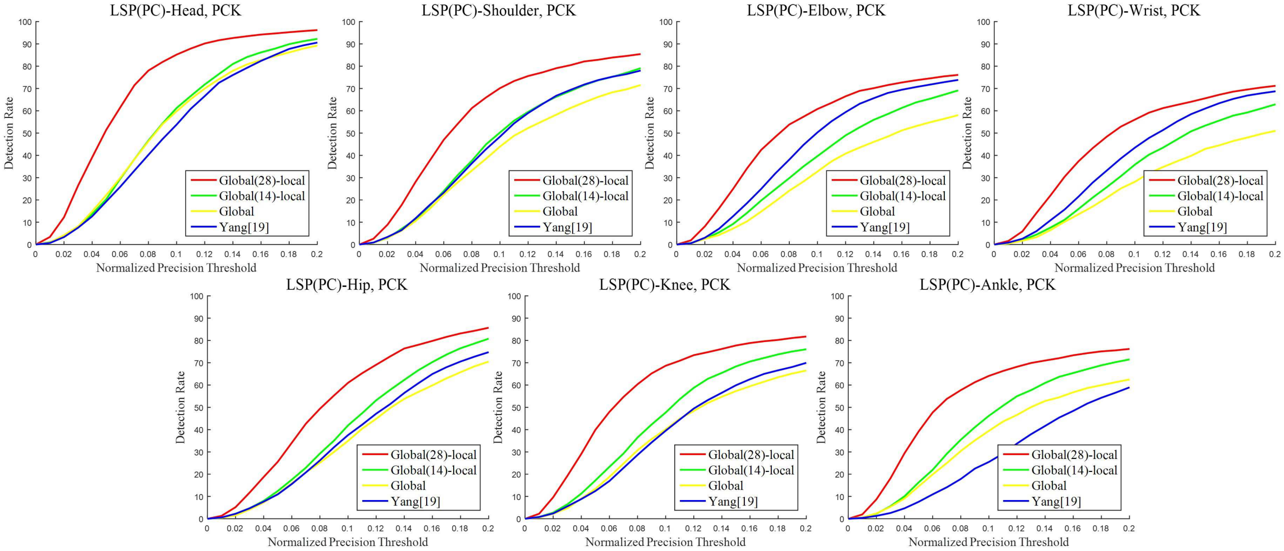


Figure 6. Quantitative typical results on the LSP dataset using the PCK. Our proposed method the highest performance especially on head part. The other parts tend to be similar.

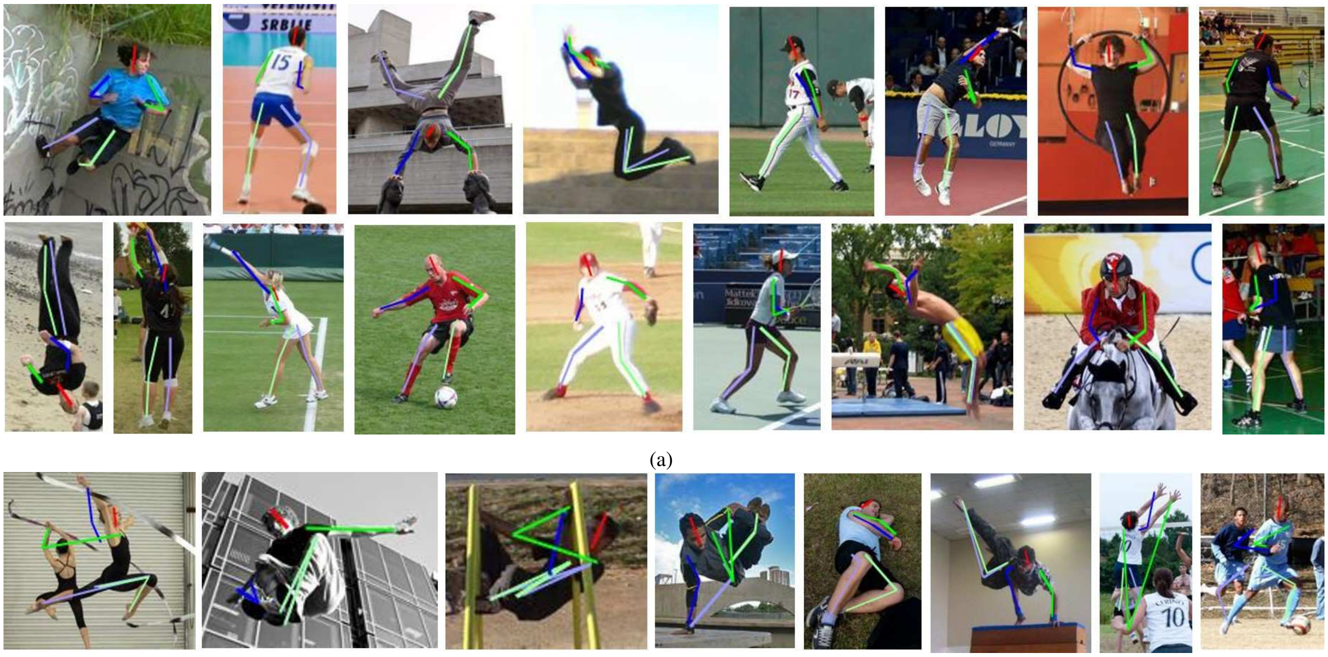
performs the other methods in estimating a variety of parts. Qualitative results from LSP dataset are shown in Figure 7.

The proposed method included the local network to effectively learn local information. The role of the local network is to find the local information which cannot be inferred in the global network. Figure 5 is the example that shows the effect of the local network. The images that shows the results of all parts locations and the images that show the output heatmap of the left wrist are shown. Figure 5(a) is the results from Global model, and Figure 5(b) is the results from Global(14)-local model. In the case of Global model which aggregates the global information, the heatmap of the left wrist has high values around the right wrist which is visible in the image. On the other hand, in the case of Global(14)-local model which exploits the local information as well as the global one, it is possible to refine the heatmap even for the occluded part, and the position of the left wrist is correctly inferred as shown in Figure 5(b). Thus, we conclude that the global-local network is able to learn both global and local information.

Next, we tested the Global(28)-local* network model that had been trained using the LSP dataset on the UCF sports dataset and provide qualitative results. Figure 8 shows the results of representative frames of various videos such as kicking and skate-boarding. The network successfully locates parts even though it is not trained on the UCF-sports dataset. Especially the head, knee and shoulder positions were estimated with small amount of errors. Compared to LSP dataset, UCF sports dataset is more challenging since the dataset contains low resolution and blurry images as can be seen in Figure 8. The proposed network produces reliable results despite those challenging conditions. Lastly, we showed the failure cases of our methods in Figure 7 (b) . When a person is in a squatting position or a person is wearing loose clothes, it was difficult to locate the body parts. The proposed method also suffers from self-occlusions. In those case, it is difficult to regress the part's location correctly on a single frame. Using a tracking algorithm that contains temporal information can be a solution for the case. Our method performs slightly worse than the state-of-the-art. However, since the state-of-the-art methods are constructed as a repetitive structure or a very deep structure, our proposed method will work a little faster. Also, attaching our local refine network to other structures which has been previously proposed will be left for the future work.

4. Conclusion

We presented a network combining both global and local information to regress the human pose. Concatenating global features with heatmaps to refine locally and using several region proposals are the key features of our methods. The proposed architecture is a deep global network followed by a small local network, and the whole network is trained end-to-end. The combined network efficiently regressed body parts with high complexity such as veiled parts. As it is tested on the LSP and UCF sports dataset, the proposed method was demonstrated to produce good results in both image data and video. A future work will be attaching our local refine network to other networks that show the state-of-the-art performance



(a)

(b)

Figure 7. Qualitative results of our method on LSP dataset. (a) Successful results, (b) Failure results. Proposed method was successful in various poses. As like squatting pose, many joints had self-occlusion, then it made a failure result.

Acknowledgement

This work was supported by Ministry of Culture, Sports and Tourism (MCST) and Korea Creative Content Agency (KOCCA) in the Culture Technology (CT) Research Development Program 2017.

References

[1] A. Bulat and G. Tzimiropoulos. Human pose estimation via convolutional part heatmap regression. In *European Conference on Computer Vision*, pages 717–732. Springer, 2016. 4322

[2] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik. Human pose estimation with iterative error feedback. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4733–4742, 2016. 4321

[3] X. Chen and A. L. Yuille. Articulated pose estimation by a graphical model with image dependent pairwise relations. In *Advances in Neural Information Processing Systems*, pages 1736–1744, 2014. 4321

[4] X. Fan, K. Zheng, Y. Lin, and S. Wang. Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1347–1355, 2015. 4321, 4322, 4325

[5] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *International journal of computer vision*, 61(1):55–79, 2005. 4321

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 4321, 4322

[7] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *European Conference on Computer Vision*, pages 34–50. Springer, 2016. 4321, 4322

[8] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM, 2014. 4324

[9] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, volume 2, page 5, 2010. 4321

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 4324

[11] Y. Li, K. He, J. Sun, et al. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in Neural Information Processing Systems*, pages 379–387, 2016. 4321, 4322, 4323

[12] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015. 4325

[13] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499. Springer, 2016. 4321, 4322, 4325

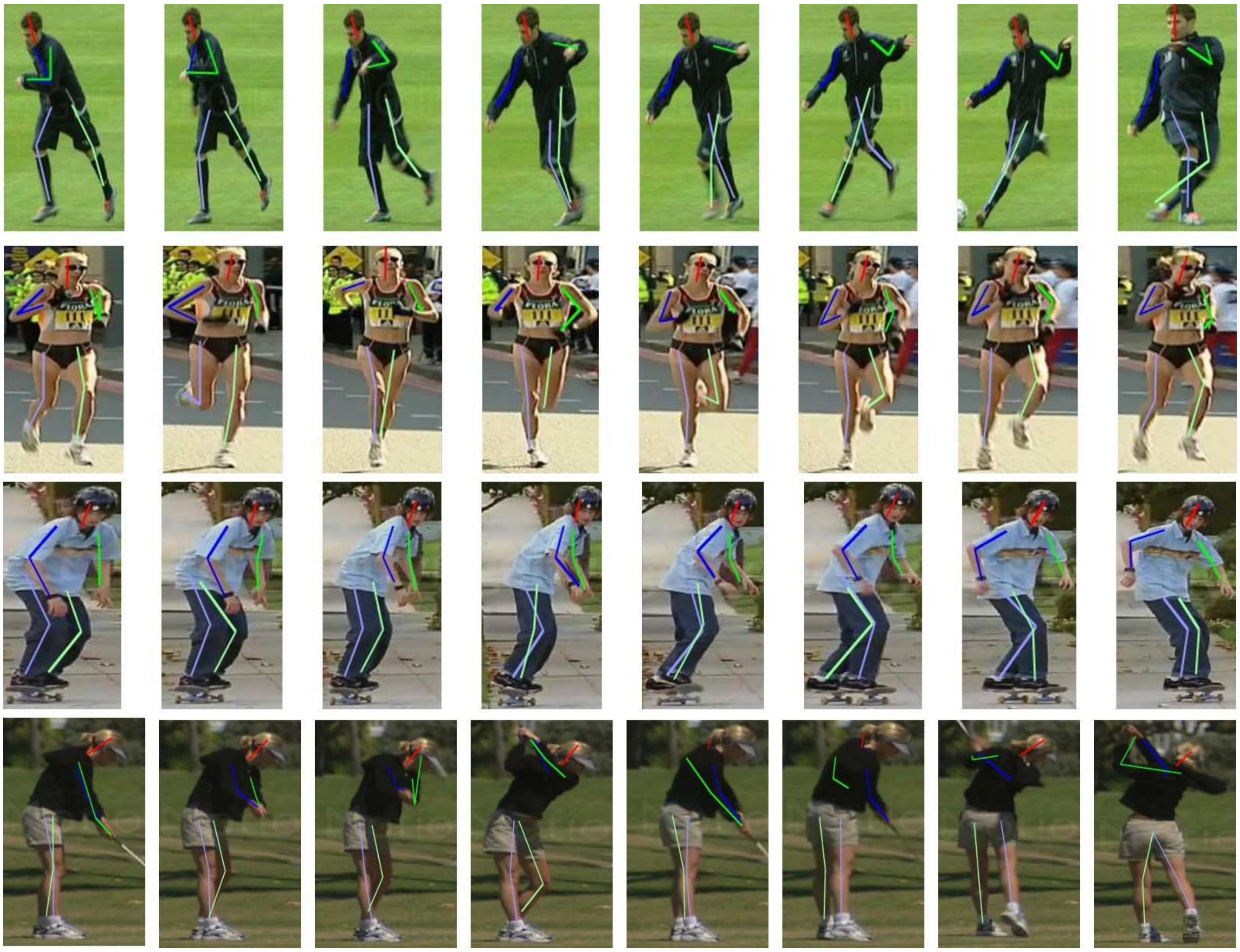


Figure 8. Qualitative results of our method on UCF dataset time-sequentially.

[14] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Strong appearance and expressive spatial models for human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3487–3494, 2013. 4321

[15] 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*, pages 256–269. Springer, 2012. 4321

[16] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler. Efficient object localization using convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 648–656, 2015. 4321

[17] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1653–1660, 2014. 4321, 4322

[18] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016. 4321, 4322

[19] W. Yang, W. Ouyang, H. Li, and X. Wang. End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3073–3082, 2016. 4322, 4325

[20] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1385–1392. IEEE, 2011. 4321

[21] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *European Conference on Computer Vision*, pages 391–405. Springer, 2014. 4321, 4323, 4324