A Multi-Stage Convolution Machine with Scaling and Dilation for Human Pose Estimation

Yali Nie[1], Jaehwan Lee[1], Sook Yoon[2] and Dong Sun Park[3*]
[1]Dept. of Electronics Engineering, Chonbuk National University
Jeonju, South Korea
[e-mail: yali_nie@hotmail.com]
[2]Dept. of Computer Engineering, Mokpo National University
Mokpo, South Korea
[e-mail:syoon@mokpo.ac.kr]
[3]IT Convergence Research Center, Chonbuk National University
Jeonju, South Korea
[e-mail: dspark@jbnu.ac.kr]
*Corresponding author: Dong Sun Park[3]

Received March 30, 2018; revised June 15, 2018; revised September 16, 2018; revised November 20, 2018; accepted December 23, 2018; published June 30, 2019

Abstract

Vision-based Human Pose Estimation has been considered as one of challenging research subjects due to problems including confounding background clutter, diversity of human appearances and illumination changes in scenes. To tackle these problems, we propose to use a new multi-stage convolution machine for estimating human pose. To provide better heatmap prediction of body joints, the proposed machine repeatedly produces multiple predictions according to stages with receptive field large enough for learning the long-range spatial relationship. And stages are composed of various modules according to their strategic purposes. Pyramid stacking module and dilation module are used to handle problem of human pose at multiple scales. Their multi-scale information from different receptive fields are fused with concatenation, which can catch more contextual information from different features. And spatial and channel information of a given input are converted to gating factors by squeezing the feature maps to a single numeric value based on its importance in order to give each of the network channels different weights. Compared with other ConvNet-based architectures, we demonstrated that our proposed architecture achieved higher accuracy on experiments using standard benchmarks of LSP and MPII pose datasets.

Keywords: CNN, Human pose estimation, Multi-stage, Pyramid stacking, Dilation, Gating.

This paper was supported by research funds of Chonbuk National University in 2017. This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(No.NRF-2016R1D1A1B03936269).

http://doi.org/10.3837/tiis.2019.06.023 **ISSN : 1976-7277**

1. Introduction

Human Pose Estimation (HPE) is a daunting task in computer vision. HPE requires to recognize and locate the key points of person in an image, which serves as a fundamental research topic for many visual applications like clothing parsing, human re-identification, movie making, human-computer interaction, gesture recognition, activity understanding and human tracking. There are many challenging issues in human pose estimation, such as cluttered background, motion blur, occlusion and self-occlusion, varying illumination and foreshortening, which hamper that human keypoints cannot be well localized. Traditional approaches on pose estimation have tackled these challenges with hand-crafted features and graphical models.

Since AlexNet [2] was introduced, ConvNets have gained immense popularity due to significant improvements of accuracy achieved by their use in the image classification. Consequently, some state-of-the-art human pose estimation methods have been proposed by using complex ConvNet architectures and they [1,4] perform considerably well in human pose datasets. And, as CNNs architectures have been modeled deeper and deeper, it also becomes more difficult to manually design structures for each layer. As one of methods to deal with this problem, various building blocks with a fixed structure called modules have been introduced to build some higher level layers effectively.

Since it is necessary to well extract the spatial context information for better performance of pose estimation, most of state-of-the-art ConvNet-based methods, like CPM [1], have been trying using multiple stages with different receptive field. As the number of stages with different receptive fields is increased, the performance of the system becomes better. However, it has made the system more complex in architecture and in computation.

Based on convolution pose machine (CPM) [1], we propose a Convolution Machine of Pose which incorporates with multiple stages, multiple layer based feature maps, various receptive fields and scalable region of interest. To reduce the number of stages, which can make a running time faster, we use a feature map structure stacked by pyramid stacking modules. In addition, our proposed architecture is specified in tandem with recent module-based higher level construction technique to build up our architecture more simply. Our proposed multi-stage convolution machine involves scaling and dilation on heatmaps of body joint regression with intermediate supervision by using additional well-known modules as building blocks. Fire module is used to reduce complexity of the proposed architecture.

The contributions of this paper are summarized as follows:

(a). We modularize the architecture to improve a multi-stage network for better human pose estimation.

(b). We introduce pyramid stacking modules to efficiently assemble feature maps from layers with various receptive fields for better pose estimation.

(c). We introduce gate modules to control weighting of feature maps for better pose estimation.

(d). We introduce fire modules to reduce the number of architecture parameters which becomes more cumbersome for better pose estimation.

(e). We introduce dilation modules which can learn multi-scale information for each body part for better pose estimation.

In our multi-stage network, we make the receptive file large enough to learn long-range spatial relationships. Intermediate supervision is applied to produce intermediate confidence maps and refine score maps.

2.Related Work

Overview of prior work. Traditional human pose estimation techniques have been usually based on pictorial structure models such as tree-structured graphical models [5, 6, 7]and hierarchical models [8, 9]. They have been used to encode the relationships among body parts broadly. But all these approaches were built on hand-crafted features. Recently, human pose estimation has achieved significant progresses by introducing CNNs for learning feature representation better [3, 4, 10, 11, 12, 13, 14, 15]. For example, DeepPose [3] is the first CNNs proposed to solve human pose estimation problem with cascade convolution network regressed by Toshev et al. If the initial estimate is very far from the ground truth, DeepPose will get low accuracy in high precision. A graphical model based on predicting heatmaps of keypoints [4] solves the problem later. State-of-the-art performances are achieved by Convolutional Pose Machine (CPM) [1] and stacked hourglass network [14]. CPM [1] propose a multi-stage architecture to incorporate the inference of the spatial relationship in body parts. Newell et al. [14] use Hourglass network, also known as conv-deconv structure, which repeatedly use spooling down and up sampling process to learn the spatial distribution.

Relation to fire module. Fire module concept comes from SqueezeNet [16], which is a small CNN architecture and achieves AlexNet-level accuracy on ImageNet with 50x fewer parameters. The original fire module consists of two layers: a squeeze layer and an expand layer. A squeeze layer is comprised of 1×1convolution filters to help to limit the number of input channels. And an expand layer is comprised of $1 \times 1$and 3×3 convolution filters to help to compensate limitation from the squeeze layer. The fire module used in the proposed architecture is a variant of the original. It includes two original fire modules: a normal original fire module keeping going into a further deep process and another taking a bypass additionally for reusing feature.

Relation to dilation module. In the dilated convolution of [17],the same filter can be applied at different receptive fields using different dilation factors by skipping a certain number of input values when applying the filter to a convolutional layer. It can support expansion of the receptive field without loss of resolution and additional training. And in [18], pyramid scene parsing network (PSPNet) with a pyramid pooling module is proposed to exploit the global context information at different ranges by using feature maps in different levels generated by pyramid pooling. Deeply inspired by [17] and [18], we design our dilation module which employ dilation convolution in parallel to capture multi-scale context by adopting multiple dilation factors.

Relation to gate (scaling) module. Squeeze-and-Excitation Network (SENet) [19] which is the winner of ImageNet 2017 classification task introduces a new block called Squeeze-and-Excitation (SE) block to improve channel interdependencies at almost no computational cost. Their basic idea let activation maps learn a weight vector and give each channel different weightings. There are two steps in a SE block: Squeeze and Excitation. Squeeze use a global average pooling to squeeze global spatial information into channel descriptors. Excitation adapts recalibration. Briefly, a SE block uses one simple gating mechanism with a sigmoid activation and the final output is obtained by rescaling. The core idea of SENets is to learn the feature weighting according to the loss in the network. **Fig. 1** shows our proposed architecture with five gates: one big gate and four small gates. These gates are designed to control the weighting of each feature map from previous stages by using a concept of SE block. The network can adaptively adjust the weighting of each feature map by concatenating scaled feature maps additionally to feature map of a convolutional block. It can achieve better results by handling feature maps with different weightings. We can get the global information by the big gate and local information by the small gates.

3.Method Architecture

3.1 Our Architecture

Many state-of-the-art human pose estimation methods have been proposed by using complex ConvNet architectures and they [1, 14] perform considerably well in human pose datasets. Especially, the famous CPM [1] is based on multi-stage convolutional networks that can capture information in large receptive field and each of its stage uses supervision to train avoiding the problems of difficult optimization. The authors of CPM made a comparison of performance across different numbers of stages and showed that the performance of CPM increases monotonically up to 5 stages. To lever the parameters and accuracy, we chose 4 stages to train our model. Every stage produces 15 joint confidence maps (heatmaps) for each still image and then the heatmaps are sent into next stage. And confidence maps are created by putting Gaussian peaks at the location of each body part’s ground truth.

Fig. 1 shows an overview of our proposed network architecture, based on multiple stages like CPM. It consists of one *PRE-STAGE* and other four *STAGE*s, where G is a large gate working on global information, and g is a small gate working on local information. These gates (G and g) are operated based on SENet [19]. We apply gates (G and g) to every stage in order to control the importance of feature map weights. For more details, we will explain in the following *STAGE* part.

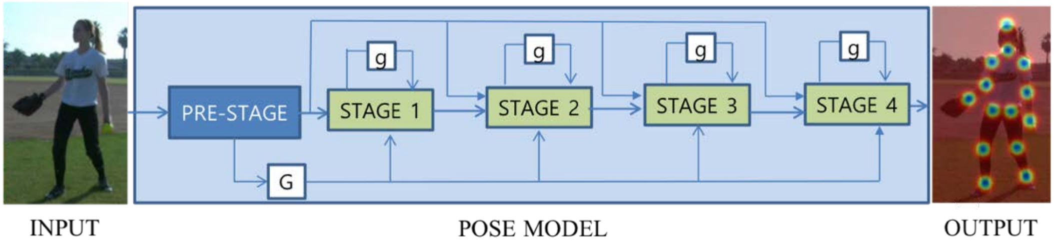


Fig. 1. Overview of the Proposed Architecture

3.2 PRE-STAGE

PRE-STAGE consists of one normal convolution layer tightly followed by ReLU and three pyramid stacking modules, as shown in **Fig. 2**(a) – (b). *PRE-STAGE* outputs two kinds of feature maps:$F_l$andF_h. F_l is a feature map from layers with smaller receptive field. It is obtained from an input image through a convolutional layer followed by ReLU and given to big gate (G). And then, F_l is converted to$F_h$through a series of pyramid stacking modules. F_h is a feature map from layers with larger receptive field and connected to small gate (g) and the next stage. We show our pyramid stacking module detail in **Fig. 2**(c). A pyramid stacking module is composed of a pooling layer and convolution layer with stride of 2 in parallel, heavily inspired by [29]. It converts a feature map of i^{th} layer($i - feature\ map$) into afeature map of $(\mathrm{i}+1)^{\mathrm{th}}$ layer ($(i+1) - feature\ map$) . ($i - feature\ map$) is downsampled by a pooling layer, represented in $i' - feature\ map$. Meanwhile, it is also convolved with stride 2 and becomes($(i+1)^{*} - feature\ map$), as shown in Fig. 2 (b)-(c) in yellow and green colors, respectively. And these two feature maps are concatenated into a feature map($(i+1) - feature\ map$).

($(i+1) - feature\ map$) contains feature map information of i^{th} layerand $(i+1)^{th}$ layertogether while it becomes a half in the spatial domain but becomes double in the channel domain.The i^{th}feature map with size of$m \times n$and l channels becomes $(i+1)^{th}$ feature map with size of $\frac{m}{2} \times \frac{n}{2}$ and 2l channels. Simply, through a series of pyramid stacking modules, a series of feature maps is stacked by iteratively applying these pooling and convolution operations as shown in Fig. 2.

We use Rectified Linear Units (ReLU) for faster training and apply dropout to prevent from overfitting. We also take Stochastic Gradient Descent (SGD) to make a back propagation.

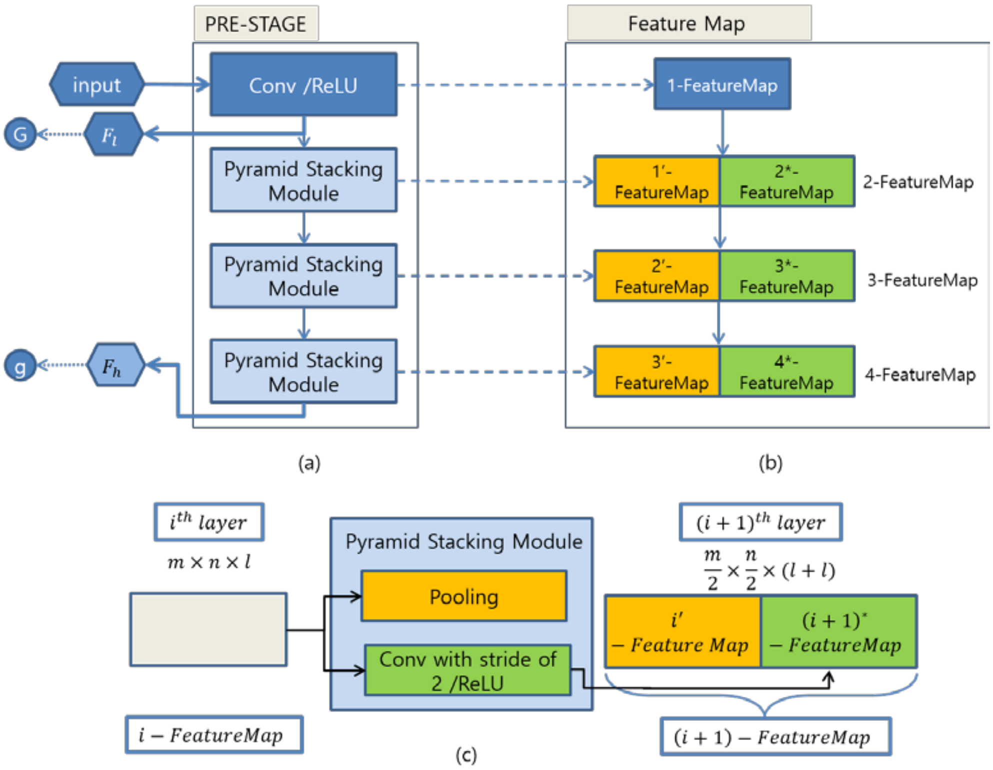


Fig. 2. Structure of PRE-STAGE

3.3 STAGE

Our architecture is based on multi-stages. And stages sequentially produce heatmaps which are refined across stages. Each *STAGE* consists of four fire modules, one dilation module and two convolution layers, as shown in Fig. 3. When i is not equal to 1, $STAGE\ i$ has three inputs: F_l and F_h from PRE-STAGE and an output of previous stage $STAGE$ (i-1). However, since the previous stage of $STAGE$ 1 is $PRE\text{-}STAGE$, $STAGE$ 1 has two inputs. In the figure,

$\otimes$ refers to concatenation, g represents small gate and Gmeans big gate. An output of $STAGE\ i$ is 15 heat-maps going to next new stage $STAGE\ (i+1)$.

Since our *PRE-STAGE* makes feature maps smaller in the spatial domain but larger in the channel domain, we deploy fire modules to reduce complexity in the channel domain, instead of using normal convolutional layers. As shown in **Fig. 4**, our fire module is a variant of the fire module in [16] and includes two original fire modules: a normal fire module and another one with additional bypass. The used fire module can use a skip to make a concatenation which can encourage feature reuse, keep going into a further deep process.

An input of each stage gated by a small gate g skips first three fire modules and is concatenated with an output of third fire module to go to next fourth fire module. And then F_l gated by a large gate $G$is concatenated with an output of fourth fire module. Through $g$and G gates, lower layer feature map information can be added to relatively higher layer feature map information.

Since we use multi-stage to refine the results, our model has a large number of layers in a deep architecture. To avoid the problem of vanishing gradients, we use intermediate supervision. Each stage of the architecture is trained to produce heatmaps repeatedly for locations of keypoints.

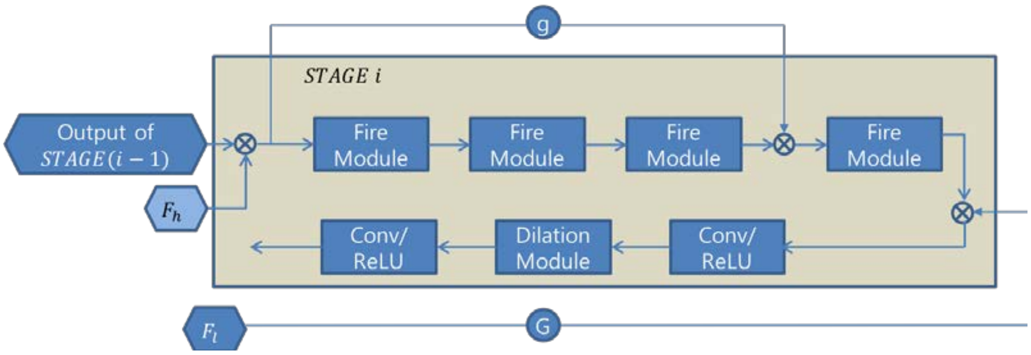


Fig. 3. Network Architecture of '*STAGE*' Module

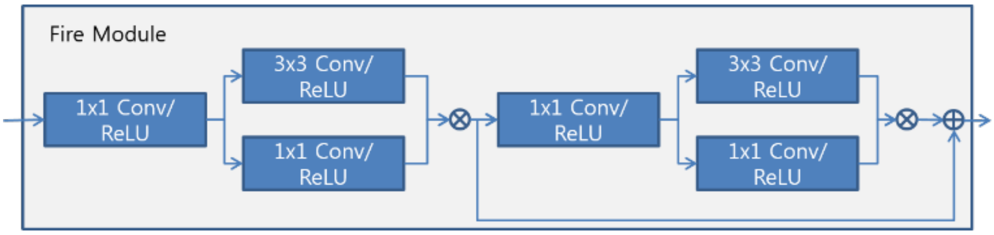


Fig. 4. Fire Module

Fig. 5 shows our dilation module that contains four parallel dilated convolutions with different dilations which are applied on top of the feature maps. It consists of four 3x3 convolutions with dilation factors 1, 2, 3 and 4 respectively. After dilated convolutions, they are fused together as the global prior. The original feature map also will be concatenated with them. After concatenating all the branches of features, one 1x1 convolution will be applied on them to generate 15 outputs.

Our dilation module has three advantages. Firstly, it allows us to enlarge the field of view of filters effectively incorporating multi-scale context. Secondly, by taking parallel dilated convolution layers, the network can capture multi-scale information. And thirdly, dilated convolutions will not lose resolution or coverage with exponential expansion of the receptive field. Without learning extra parameters, our dilation module is able to control the resolution where feature responses are computed with ConvNets.

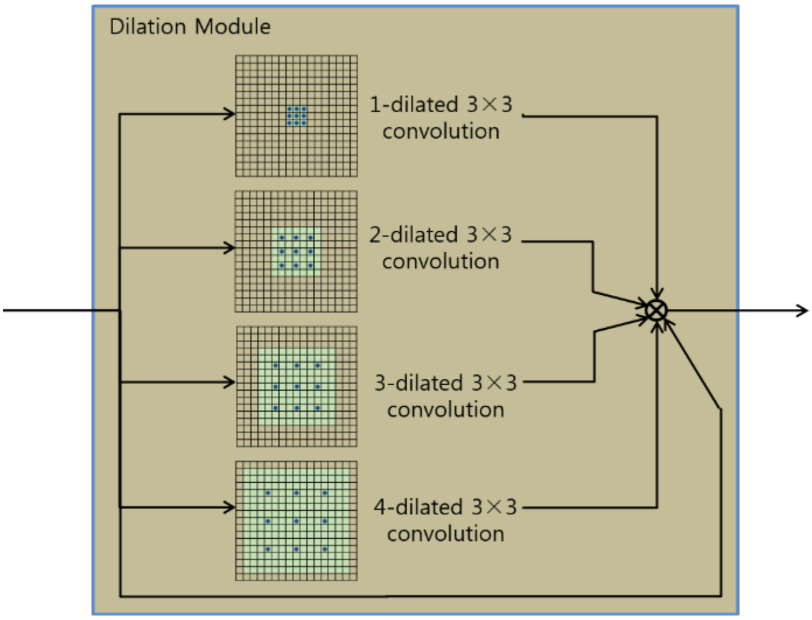

Dilation Module
1-dilated 3×3 convolution
2-dilated 3×3 convolution
3-dilated 3×3 convolution
4-dilated 3×3 convolution

Fig. 5. Dilation Module

4. Experimental Results and Analysis

4.1 Evaluation Metrics

In all experiments, we use most popular criterions which are the Percentage of Correctly estimated body Parts (PCP) and Probability of Correct Key point (PCK) [27, 28].

PCP measures the percentage of correctly localized body parts. This criteria proposed by [27] presents that one part is considered as correctly localized if its predicted endpoints are within 50% part length of the corresponding ground truth segment from their annotated locations.

PCK measures the probability of correct detection which falls within a tolerance range and the tolerance range is a fraction of torso size. To obtain the PCK evaluation, we need ground truth key point of the body joint and the predicted keypoint localization. PCK equation can be expressed as

$$\frac{\left\| l_i - \hat{l}_i \right\|_2}{\left\| l_{lhip} - l_{rsho} \right\|_2} \leq r \qquad (1)$$

where l_i and $\hat{l}_i$ are the ground truth location and the predicted location of the i^{th} keypoint respectively. l_{lhip} is left hip and l_{rsho} is right shoulder. The fraction of the person bounding box size r is bounded between 0 and 1.

4.2 Data

Our architecture predicts 14 human full-body keypoint locations. These 14 keypoints are head, neck, left shoulder, right shoulder, left elbow, right elbow, left wrist, right wrist, left hip, right hip, left knee, right knee, left ankle and right ankle respectively. We train our proposed method on two datasets: Leeds Sport Pose (LSP) Dataset and MPII Human Pose (MPII) Dataset.

Leeds Sports Pose (LSP) Dataset. This dataset consists of 11,000 images for training and 1000 images for testing. These images are collected from Flick searches with sport people.

MPII Human Pose Dataset. MPII dataset contains 28,821 annotated pose for training and 11,701 for testing. These images are gathered from YouTube videos. The annotated pose have 16 body joints, some of them are not present and some are in occlusion but can be predicted by the context information.

Data augmentation. We choose to augment the data on both MPII and LSP datasets. We use random rotation degrees in $[-40°,40°]$ and random rescaling in [0.7, 1.3] to make the model more robust for image changing. Input RGB images are cropped and centered on the main subject with one squared bounding box to keep people scale. Horizontal flipping is also applied to do data augmentation. In addition, the training images are resized to 368×368 pixels.

4.3 Experiments

We train the proposed model with an initial learning rate of$10^{-4}$with a modified version of caffe [30]. Our training maximum number of iterations is 1000000. We set the momentum 0.9 and weight decay 0.0005. The parameters are optimized by RMSprop [31] algorithm. Two Pascal TITAN GPUs with cuDNN v5.1are used to train the merged dataset of extended LSP and MPII. Both of these datasets provide the visibility of body parts and we use them as the supervision for occlusion signals during training. We test our model on LSP and MPII dataset. The input resolution of the images is 368×368 and it is dropped down to 46×46 at the end. The number of output features is 15.

Table 1 shows the complexity comparison of CPM [1] and the proposed method in the number of parameters, the number of operations, and model size. The numbers of parameters and operations in the proposed method are lower than CPM, since we use smaller kernel, some special modules and smaller number of stages. The proposed architecture is lighter than the original one. It has less complexity both in space and time than the original CPM.

Table 1. Comparisons of numbers of parameters and operations and model size

Method	#Parameters	#Operations(GFLOPS)	Model Size
CPM [1]	29.8M	63.2	119.4M
Ours	19.3M	40.7	77.1M

4.4 Results

To verify our strategies while consuming relatively less amount of time, we did a pilot study by using CPM with 3 stages as the baseline and training it for 100,000 iterations. We adopted structures of each strategic module into the CPM baseline to build its variants like SQ, PD and Gate. And **Tables 2** and **3** report their performances. Here SQ is a variant obtained by replacing a simple convolution layer of the baseline with a fire module. PD refers a variant with pyramid dilation convolution used in each stage of the baseline. The Gate (scaling) is a variant with five gates added to the baseline. All of methods were tested on LSP dataset. The result of our pilot study showed that each variant can work better for the human pose estimation than the CPM baseline.

In order to improve the performance in pose estimation, we deployed all of techniques, which had been verified through the pilot study, into the CPM baseline to build our own model.

Table 2. Performance comparison of CPM and its variants on the LSP dataset (PCK@0.2) in the pilot study

Methods	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
CPM [1]	94.9	65.9	58.0	51.7	61.8	54.5	53.2	62.9
SQ	95.4	70.0	62.5	56.5	66.2	57.4	54.2	66.0
PD	94.5	66.0	60.0	54.6	63.6	55.3	63.5	65.4
Gate	95.3	68.2	59.9	53.4	64.7	56.0	55.6	64.7

Table 3. Performance comparison of CPM and its variants on the LSP dataset (PCP) in the pilot study

Methods	Torso	Upper leg	Lower leg	Upper arm	Fore arm	Head	Total
CPM [1]	92.9	55.4	46.6	53.1	38.5	91.4	57.1
SQ	87.4	52.1	47.4	44.9	25.2	84.5	51.1
PD	93.2	57.5	47.1	52.9	37.5	90.6	57.4
Gate	95.0	58.4	48.3	54.5	38.7	90.8	58.6

We make comparisons with other approaches on the task of human pose estimation from a single image and the results are shown in **Fig. 6, 7** and **8** and **Table 4, 5** and **6**. In the tables, bold fonts indicate our performances. **Fig. 6** and **Table 4** and **5** show evaluation results on the LSP test set. It is clear that we achieve promising performance for the keypoint localization. In particular, our model performs better than the original method proposed by CPM [1]. For the most challenging body parts such as ankle and wrist, our approach achieves 1.7% and 2.4% improvement respectively as compared with CPM [1]. PCKh is one measure that takes a matching threshold as 50% of the head segment length. In addition, we improve the performance of all keypoint locations from 88.5 to 89.5 on MPII dataset as shown in **Fig. 7** and **Table 5**.

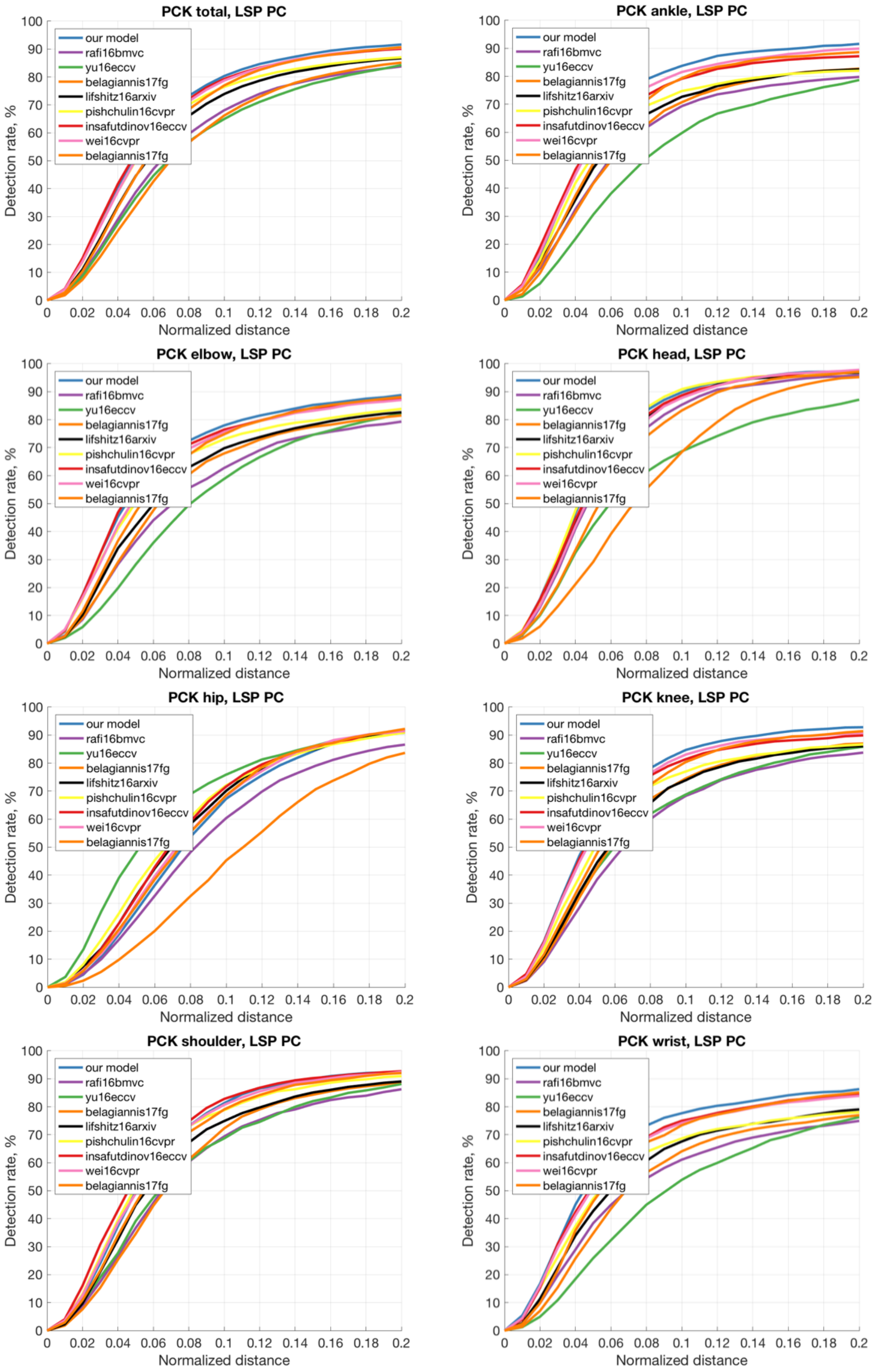


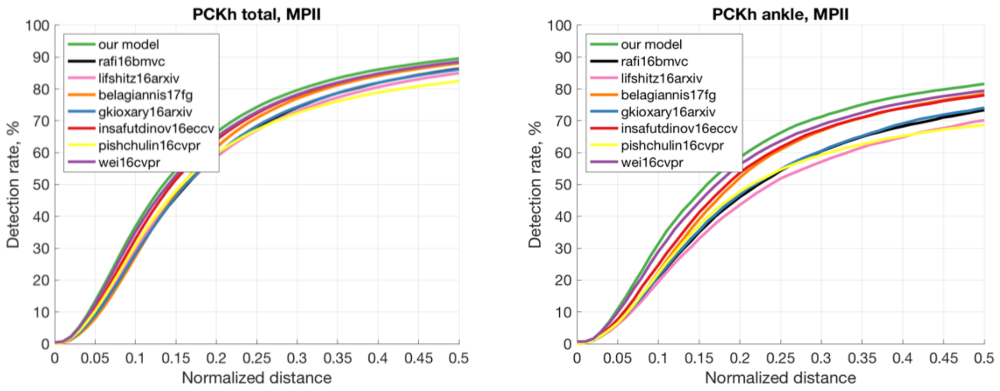
Fig. 6. PCK@0.2 comparison on LSP

Table 4. Performance comparison with previous works on the LSP dataset (PCK@0.2)

Methods	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
Rafi [20]	95.8	86.2	79.3	75.0	86.6	83.8	79.6	83.8
Yu [21]	87.2	88.2	82.4	76.3	91.4	85.8	78.7	84.3
Belagian [22]	95.2	89.0	81.5	77.0	83.7	87.0	82.8	85.2
Lifshitz [10]	96.8	89.0	82.7	79.1	90.9	86.0	82.5	86.7
Pishchu [23]	97.0	91.0	83.8	78.1	91.0	86.7	82.0	87.1
Insafutd [24]	97.4	92.7	87.5	84.4	91.5	89.9	87.2	90.1
CPM [1]	97.8	92.5	87.0	83.9	91.5	90.8	89.9	90.5
Bula [15]	97.2	92.1	88.1	85.2	92.2	91.4	88.7	90.7
Ours	**97.4**	**92.7**	**88.8**	**86.3**	**91.9**	**92.8**	**91.6**	**91.6**

Table 5. Performance comparison with previous works on the LSP dataset (PCP)

Methods	Torso	Upper leg	Lower leg	Upper arm	Fore arm	Head	Total
Rafi [20]	97.6	87.3	80.2	76.8	66.2	93.3	81.2
Belagia [22]	96.0	86.7	82.2	79.4	69.4	89.4	82.1
Lifshitz [10]	97.3	88.8	84.4	80.6	71.4	94.8	84.3
Pishchu [23]	97.0	88.8	82.0	82.4	71.8	95.8	84.3
Yu [21]	98.0	93.1	88.1	82.9	72.6	83.0	85.4
Insafutd [24]	97.0	90.6	86.9	86.1	79.5	95.4	87.8
CPM[1]	98.0	82.2	89.1	85.8	77.9	95.0	88.3
Bula [15]	97.7	92.4	89.3	86.7	79.7	95.2	88.9
Ours	**98.2**	**94.0**	**91.6**	**87.5**	**81.2**	**95.6**	**90.2**



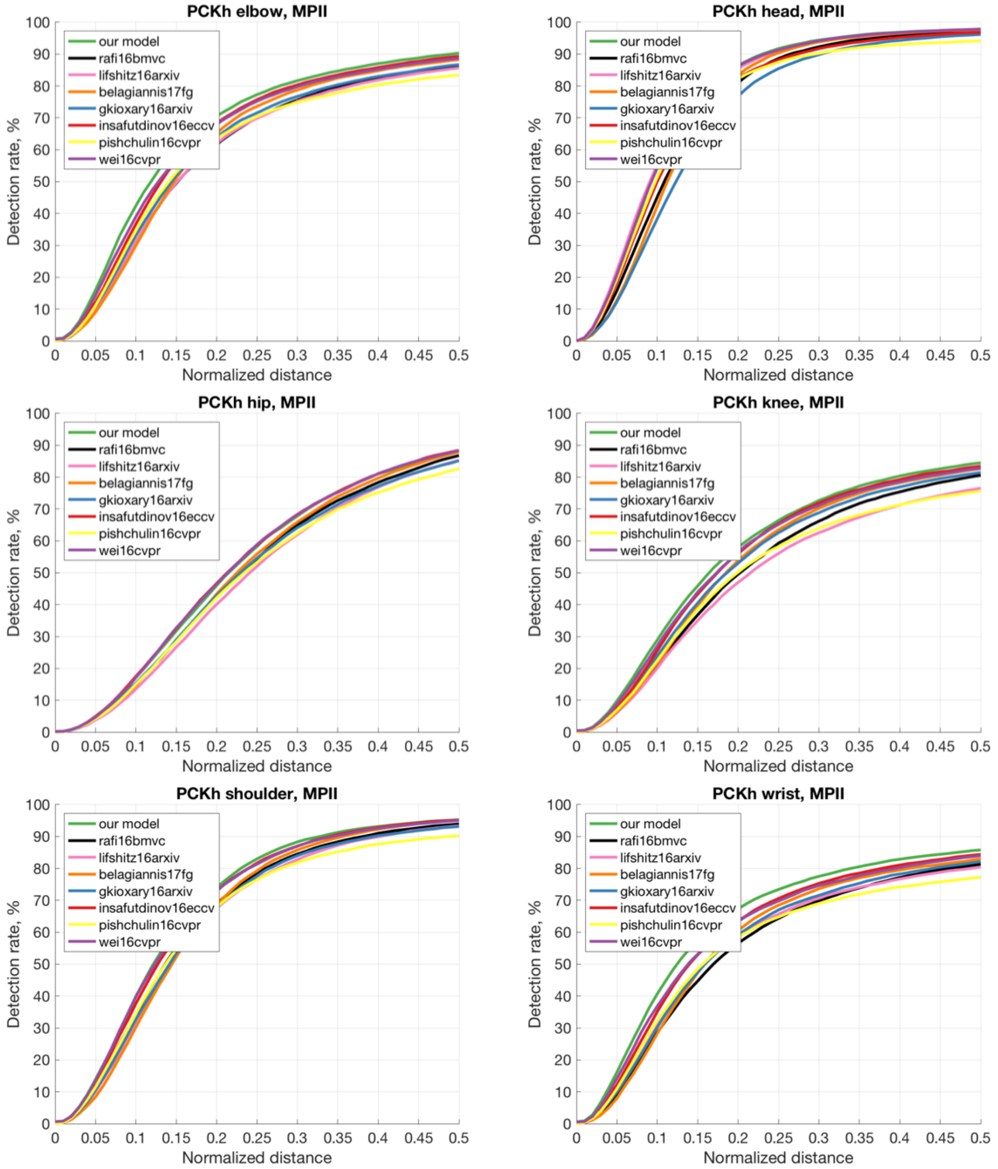


Fig. 7. PCKh comparison on MPII

Table 6. Performance comparison with previous works on the MPII dataset (PCKh)

Methods	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
Hu [26]	95.0	91.6	83.0	76.6	81.9	74.5	69.5	82.4
Pishchu [23]	94.1	90.2	83.4	77.3	82.6	75.7	68.6	82.4
Lifshitz [10]	97.8	93.3	85.7	80.4	85.3	76.6	70.2	85.0
Gkioxary [25]	96.2	93.1	86.7	82.1	85.2	81.4	74.1	86.1
Rafi [20]	97.2	93.9	86.4	81.3	86.8	80.6	73.4	86.3
Belagian [22]	97.7	95.0	88.2	83.0	87.9	82.6	78.4	88.1
Insafutd [24]	96.8	95.2	89.3	84.4	88.4	83.4	78.0	88.5
CPM [1]	97.8	90.5	88.7	84.0	88.4	82.8	79.4	88.5
Ours	**97.3**	**95.3**	**90.3**	**85.8**	**88.3**	**84.5**	**81.6**	**89.5**

And we make minute comparisons in three challenging parts of human body like ankle, knee and wrist between our methods and eight state-of-the-art deep ConvNets-based methods. As shown in **Fig. 8**, the proposed one outperforms all the others.

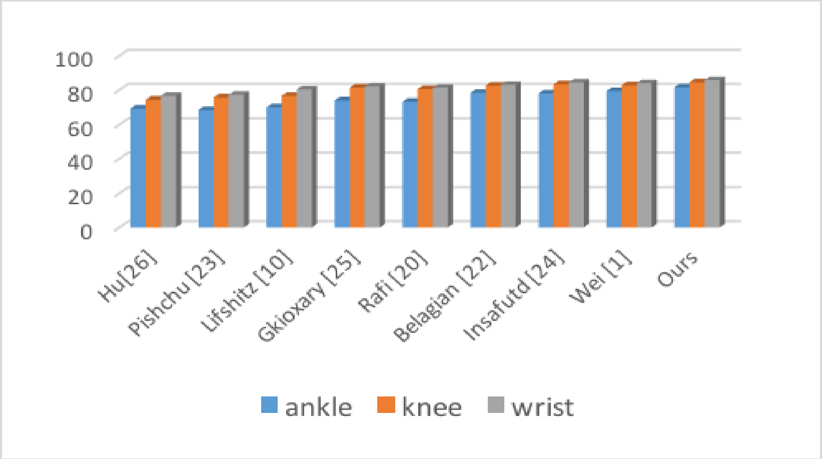


Fig. 8. Comparison of challenging parts on the MPII

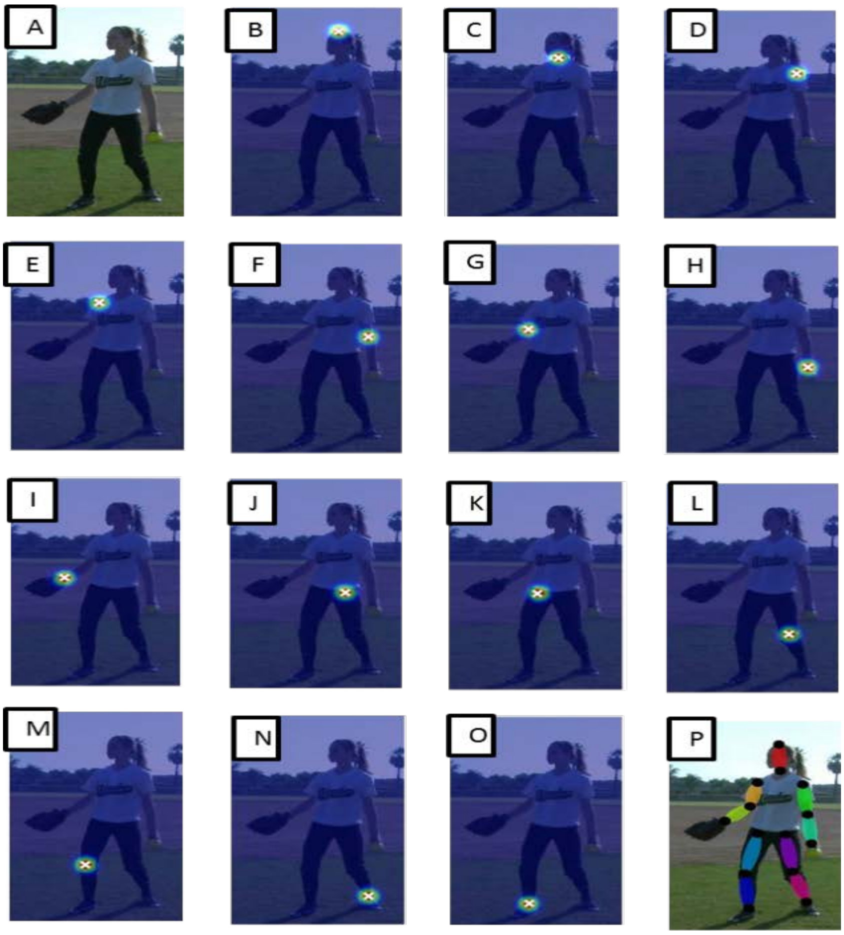


Fig. 9. Example of output from our pose estimation system: (A) an input image, (P) its final output of pose estimation and(C) – (O) heatmaps of each part

An example of results produced by our model is shown in Fig. 9. Given an input image (Fig. 9(A)), heatmaps corresponding to 14 body parts are obtained through the proposed model, as shown in Fig. 9(B) - (O). And then final output can be shown as Fig. 9(P).

Fig. 10 shows how heatmaps of most challenging parts like ankle, knee and wrist are changed according to stages. It shows that their accuracy improves and their heatmaps also become much clearer while increasing stages.

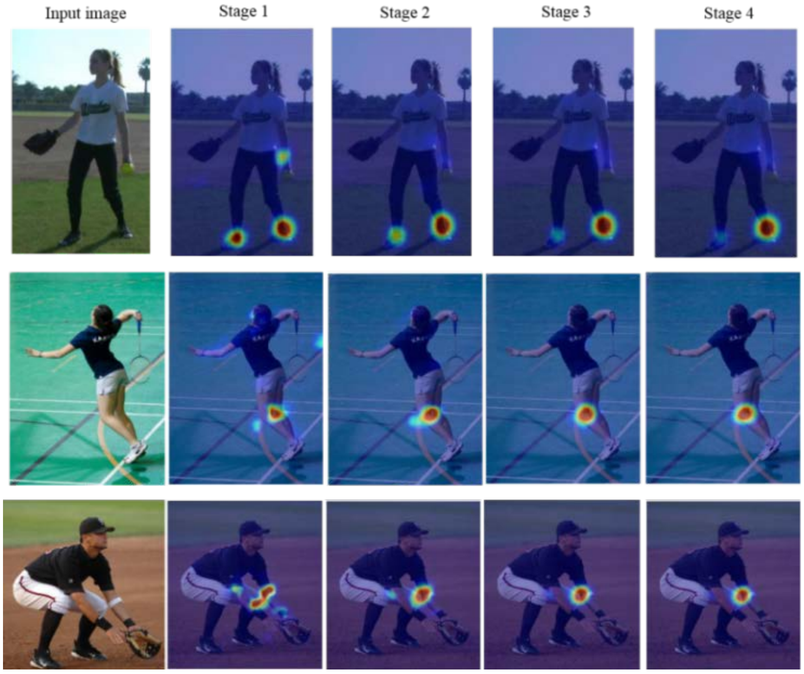


Fig. 10. Prediction of challenging keypoints in each stage

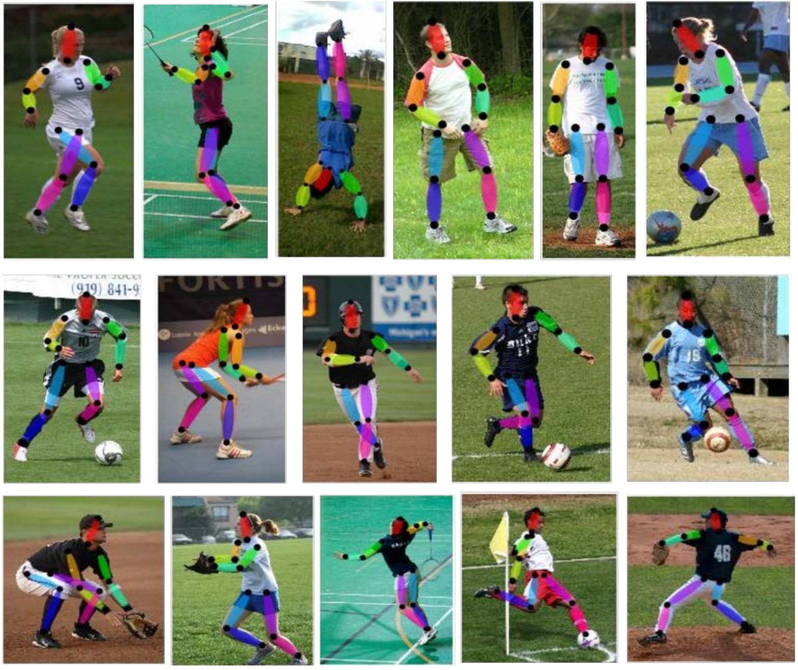


Fig. 11. Qualitative test example of pose on the LSP

Qualitative results are demonstrated in **Fig. 11**. We visualize some test pose examples on the LSP dataset based on our proposed methods. Our model can capture articulated pose with fused technologies to solve occlusion, blur and foreshortening problems.

5. Conclusion

We have shown how to improve the performance of human pose estimation (HPE) which is one of the most complex computer vision tasks. We proposed a HPE to improve accuracy and complexity in space and time of the original convolution pose machine (CPM). To propose an architecture showing better performance, we explored many different state-of-the-art structural models and adapted their concepts and modules. This method includes some modules such as fire module, pyramid stacking module, gating module, and dilation module to reduce the number of parameters in the proposed network and use multi-scale information in space and feature map, which is also based on a multi-stage structure. The experimental results show that the proposed HPE outperforms the original CPM and some other state-of-the-art approaches.

For further related research, it will be necessary to analyze the proposed method in more details based on its hyperparameters. In addition, we need to study the simplification of the proposed architecture to further reduce complexity in space and time. Furthermore, we hope that we can apply this method to multi-person pose estimation and real-time pose estimation. For gestures or facial landmark location, they also try to find the exact location. In the future, HPE can be applied to activity recognition to understand humans and their interactions with other people or objects.

References

[1] Wei, Shih-En, et al., “Convolutional pose machines,” in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. Article (CrossRef Link).

[2] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Communications of the ACM*, vol. 60, no. 6, 84-90, 2017. Article (CrossRef Link).

[3] Toshev, Alexander, and Christian Szegedy, “Deeppose: Human pose estimation via deep neural networks,” in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014. Article (CrossRef Link).

[4] Tompson, Jonathan, et al., “Efficient object localization using convolutional networks,” in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015. Article (CrossRef Link).

[5] Andriluka, Mykhaylo, Stefan Roth, and Bernt Schiele, “Pictorial structures revisited: People detection and articulated pose estimation,” in *Proc. of Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE*, 2009. Article (CrossRef Link).

[6] Yang, Yi, and Deva Ramanan, “Articulated pose estimation with flexible mixtures-of-parts,” in *Proc. of Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE*, 2011. Article (CrossRef Link).

[7] Pishchulin, Leonid, et al., “Poselet conditioned pictorial structures” in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013. Article (CrossRef Link).

[8] Sun, Min, and Silvio Savarese, "Articulated part-based model for joint object detection and pose estimation," *Computer Vision (ICCV), 2011 IEEE International Conference on. IEEE*, 2011. Article (CrossRef Link).

[9] Tian, Yuandong, C. Lawrence Zitnick, and Srinivasa G. Narasimhan, "Exploring the spatial hierarchy of mixture models for human pose estimation," in *Proc. of European Conference on Computer Vision. Springer, Berlin, Heidelberg*, pp. 256-269, 2012. Article (CrossRef Link).

[10] Lifshitz, Ita, Ethan Fetaya, and Shimon Ullman, "Human pose estimation using deep consensus voting," in *Proc. of European Conference on Computer Vision. Springer International Publishing*, pp. 246-260, 2016. Article (CrossRef Link).

[11] Chu, Xiao, et al., "Structured feature learning for pose estimation," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. Article (CrossRef Link).

[12] Carreira, Joao, et al., "Human pose estimation with iterative error feedback," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. Article (CrossRef Link).

[13] Papandreou, George, et al., "Towards Accurate Multi-person Pose Estimation in the Wild," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. Article (CrossRef Link).

[14] Newell, Alejandro, Kaiyu Yang, and Jia Deng, "Stacked hourglass networks for human pose estimation," in *Proc. of European Conference on Computer Vision. Springer International Publishing*, pp. 483-499, 2016. Article (CrossRef Link).

[15] Bulat, Adrian, and Georgios Tzimiropoulos, "Human pose estimation via convolutional part heatmap regression," in *Proc. of European Conference on Computer Vision. Springer International Publishing*, pp. 717-732, 2016. Article (CrossRef Link).

[16] Iandola, Forrest N., et al., "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size," *arXiv preprint arXiv:1602.07360*, 2016. Article (CrossRef Link).

[17] Yu, Fisher, and VladlenKoltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015. Article (CrossRef Link).

[18] Zhao, Hengshuang, et al., "Pyramid scene parsing network," *arXiv preprint arXiv:1612.01105*, 2017. Article (CrossRef Link).

[19] Hu, Jie, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," *arXiv preprint arXiv:1709.01507*, 2018. Article (CrossRef Link).

[20] Rafi, Umer, et al., "An Efficient Convolutional Network for Human Pose Estimation," *BMVC*, Vol. 1, pp. 109.1-109.11, 2016. Article (CrossRef Link).

[21] Yu, Xiang, Feng Zhou, and Manmohan Chandraker, "Deep deformation network for object landmark localization," in *Proc. of European Conference on Compu*ter Vision. Springer International Publishing, pp. 52-70, 2016. Article (CrossRef Link).

[22] Belagiannis, Vasileios, and Andrew Zisserman, "Recurrent human pose estimation," in *Proc. of Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on. IEEE*, 2017. Article (CrossRef Link).

[23] Pishchulin, Leonid, et al., "Deepcut: Joint subset partition and labeling for multi person pose estimation," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. Article (CrossRef Link).

[24] Insafutdinov, Eldar, et al., "Deepercut: A deeper, stronger, and faster multi-person pose estimation model," in *Proc. of European Conference on Computer Vision*, pp. 34-50, 2016. Article (CrossRef Link).

[25] Gkioxari, Georgia, Alexander Toshev, and NavdeepJaitly, "Chained predictions using convolutional neural networks," in *Proc. of European Conference on Computer Vision. Springer, Cham*, pp. 728-743, 2016. Article (CrossRef Link).

[26] Hu, Peiyun, and Deva Ramanan, "Bottom-up and top-down reasoning with hierarchical rectified gaussians," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. Article (CrossRef Link).

[27] Ferrari, Vittorio, Manuel Marin-Jimenez, and Andrew Zisserman, "Progressive search space reduction for human pose estimation," *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE*, 2008. Article (CrossRef Link).

[28] Yang, Yi, and Deva Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 12, pp. 2878-2890, 2013. Article (CrossRef Link).

[29] Paszke, Adam, et al., "Enet: A deep neural network architecture for real-time semantic segmentation," *arXiv preprint arXiv:1606.02147*, 2016. Article (CrossRef Link).

[30] Jia, Yangqing, et al., "Caffe: Convolutional architecture for fast feature embedding," in *Proc. of the 22nd ACM international conference on Multimedia. ACM*, pp. 675-678, 2014. Article (CrossRef Link).

[31] Tieleman, Tijmen, and Geoffrey Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural networks for machine learning*, vol. 4, no. 2, pp. 26-31, 2012.

YaliNie received the B.S. degree in computer science from Jiangxi University of Finance and Economics, China in 2015, and M.S. degree in engineering from Chonbuk National University, Republic of Korea in 2018. Her research interests focus on image processing, machine learning and deep learning. E-mail: yali_nie@hotmail.com

Jaehwan Lee received the B.S. and M.S. degrees in engineering from Chonbuk National University, Jeonbuk, Korea, in 2012, and 2014, respectively. His research interests include image processing, pattern recognition, person re-identification, and machine learning. E-mail: dlwo6@jbnu.ac.kr

Sook Yoon received the B.S., M.S., and Ph.D. degrees in engineering from Chonbuk National University, Jeonbuk, Korea, in 1993, 1995, and 2003, respectively. Until June 2006, she conducted her postdoctoral research work in electrical engineering at the University of California, Berkeley. She is presently a professor at Department of Computer Engineering, Mokpo National University, Jeonnam, Korea. Her current research interests include image processing, pattern recognition, machine learning, and multimedia computing. E-mail: syoon@mokop.ac.kr

Dong Sun Park is a professor at the Chonbuk National University, Republic of Korea. He received his BS from Korea University, Republic of Korea in 1979, and MS and PhD degrees from the University of Missouri, United States in 1984 and 1990. He has published many papers in international conferences and journals. He is a member of IEEE Computer Society. His research interests include computer vision and artificial neural network, especially deep learning. E-mail: dspark@jbnu.ac.kr