International Journal of Advanced Smart Convergence Vol.7 No.2 95-100 (2018)
http://dx.doi.org/10.7236/IJASC.2018.7.2.95

IJASC 18-2-11

Pose Estimation with Binarized Multi-Scale Module

Yong-Gyun Choi, Sukho Lee

Department of Computer Engineering, Dongseo University, Busan, 47011, Korea
petrasuk@gmail.com

Abstract

In this paper, we propose a binarized multi-scale module to accelerate the speed of the pose estimating deep neural network. Recently, deep learning is also used for fine-tuned tasks such as pose estimation. One of the best performing pose estimation methods is based on the usage of two neural networks where one computes the heat maps of the body parts and the other computes the part affinity fields between the body parts. However, the convolution filtering with a large kernel filter takes much time in this model. To accelerate the speed in this model, we propose to change the large kernel filters with binarized multi-scale modules. The large receptive field is captured by the multi-scale structure which also prevents the dropdown of the accuracy in the binarized module. The computation cost and number of parameters becomes small which results in increased speed performance.

Keywords: *Deep Learning, Pose Estimation, Binarized Network, Multi-Scale.*

1. Introduction

Pose estimation refers to the technique which estimates the main body parts by some sensors, e.g., camera sensors. The estimation of body parts can be used as body/action recognition, 3D scanning of human body, interactive game interface, controlling systems with body movement, etc. Techniques estimating body parts can be categorized into marker based and marker-less methods. In marker based methods, some markers which the sensor can recognize well are first attached to pre-defined parts of the human body, and then, the sensor senses these markers, thus recognizing the specific body part of the human body. These methods are robust, since the markers can be sensed well, but the attaching of the markers to the body is somewhat inconvenient, and can be done mostly only in indoor environments.

Therefore, nowadays, marker-less methods based on computer vision techniques are preferred [1-4]. These methods can be used for 3D pose estimation. The well-known Kinect device uses several sensors to estimate the depth of the objects to get 3D information of them. Most recently, deep learning has been applied to the pose estimation applications, and get very good results [5],[6]. Particularly, in [6], Zhe Cao et al. proposed a pose estimation method based on deep learning, which takes the part affinity field into account. Normally, other pose estimation methods only compute heat maps which have large values at the positions of certain

Manuscript Received: May. 15, 2018 / Revised: May. 21, 2018 / Accepted: May. 28, 2018
Corresponding Author: petrasuk@gmail.com
Tel: + 82-51-320-1744, Fax: +82-51-327-8955
Department of Computer Engineering, Dongseo, University, Busan, Korea

body parts. In comparison, they proposed also a 2D vector field which indicates the relationship between the body part locations, which they call part affinity field. This method achieves almost the state-of-the-art results in pose estimation.

However, even though they designed the network so that it can work in real-time on hardware systems with high performance GPU, they still use several 7×7 size filters in the convolution, which has a large computation cost. Therefore, there is room left to further accelerate the speed in their model. One of the possible ways is to use binarized modules [7]. However, using binarized modules normally drops down the accuracy in the model. Therefore, special structures for binarized modules have been proposed [8]. In this paper, we adopt the binarized module proposed in [8] to be used as an alternative for the 7×7 size convolution filters. We first give some rationale why this module can be used instead of a large sized filter, then we propose the structure which has a smaller number of parameters than the model in [6]. Experimental results show that the proposed structure estimates almost the same body positions while being faster.

2. Part Affinity Field based Pose Estimation

Figure 1 shows the overall pipeline of the method proposed in [6], where it can be seen that both the heat map and the vector field (part affinity field) are computed by the network. The input is an image, which has to be analyzed, and goes through the first ten layers of the pre-trained VGG-19 network. The output of the tenth layer of the VGG block are feature maps, which are believed to contain compressed characteristics of the input image. The output of the VGG block goes into the first stage of the network which gives intermediate results. These results go into the second stage of the network which refine the initial results. The second stage is repeated several times until results of sufficient accuracy are obtained. Each stage consists of two networks, where the upper network computes the heat maps for the body parts, and the lower network computes the part affinity fields.

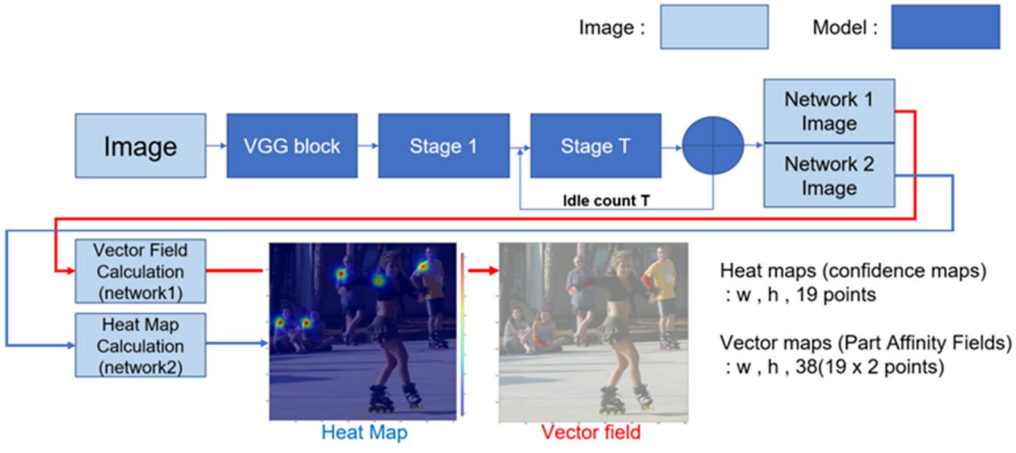

Figure 1. The pipeline of the Part Affinity Field based pose estimation method

Figure 2 shows the detailed structure of stage 1 and stage 2, where the upper part computes the set of confidence maps $S = (S_1, S_2, \ldots S_J)$, where J denotes the total number of body parts. The lower part computes the set of part affinity fields denoted by $L = (L_1, L_2, \ldots L_C)$, where C is the number of part

affinity fields. The final confidence maps and affinity fields are parsed together and results in the keypoints for the people in the image.

The total network is implemented in Caffe library, and has an inference speed of about 3~5 fps when run on a PC with NVIDIA GTX 1080Ti GPU. The bottlenecks of the computation are the 7×7 convolution blocks in the second stage. The large size of the kernel is due to the large receptive field of body parts in the image. A 7×7 convolution requires 7×7 floating point multiplications and 7×7 floating point additions for one pixel of a single input channel. As there are 19 confidence maps and part affinity fields for a single input image, the computation is very large. Therefore, we propose a network structure which reduces the computation cost while retaining a certain level of accuracy.

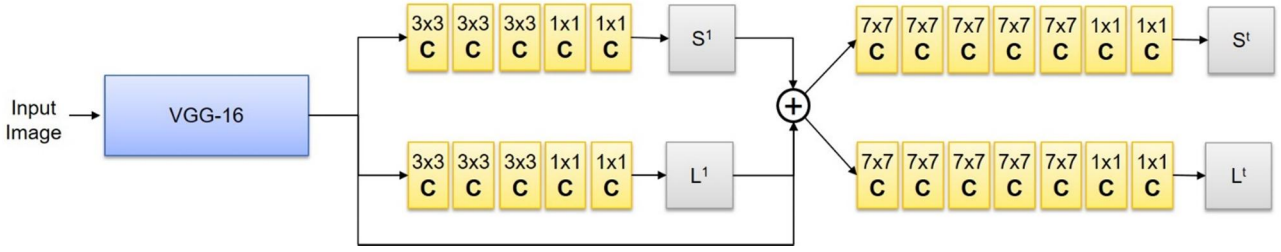

Figure 2. The detailed network structure of the Part Affinity Field based pose estimation method [6]

3. Proposed Structure

Most of the computation is done in the floating point 7×7 convolution. Therefore, we substitute the 7×7 convolution with a lighter computation module shown in Fig. 3. The module we use is the multi-scale binary structure module which has been proposed in [8] to be used in the Hourglass model for face alignment.

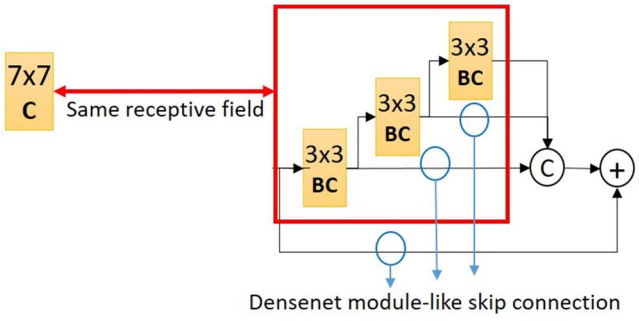

Figure 3. Comparison between the 7×7 convolution filter and the multi-scale binary module

We give some rationale, why this substitution will work. First, as is known from the result of the VGG network, a combination of three 3×3 convolution filters has the same receptive field as a single 7×7 convolution filter. Therefore, the three consecutive 3×3 convolution filters look at the same size of region as the 7×7 convolution filter. Second, the skip connection helps the training of the weight filters, because only the residual has to be learned by the filters as in the ResNet module. In fact, the skip connection at every scale is similar to the DenseNet or UNet structure, which helps to pass the information of previous filtering results forward. This makes the preserve some accuracy even though the number of parameters is reduced.
The substitution of the 7×7 floating point convolution filter to the floating point multi-scale structure already

drops down the computation cost to a large extent. With the proposed structure, the number of multiplication drops down from 49 to 27 (three filters with 3×3 multiplication) and the number of addition drops from 48 to 26. This gives a speed acceleration from 3~5 fps to 7~10 fps, depending on the hardware. Furthermore, when binarizing the floating point calculation with the XNOR net model, the computation becomes very fast.

Figure 4 shows the proposed structure, where the 7×7 floating point convolution filters are substituted with the binarized multi-scale modules. As the stage T is computed iteratively, the computational cost is reduced much with this structure. Furthermore, the number of channels after the first stage are normally about 32~64, so such kind of computation reduction is done 32~64 times. Figure 4 shows the proposed structure applying the substituted module.

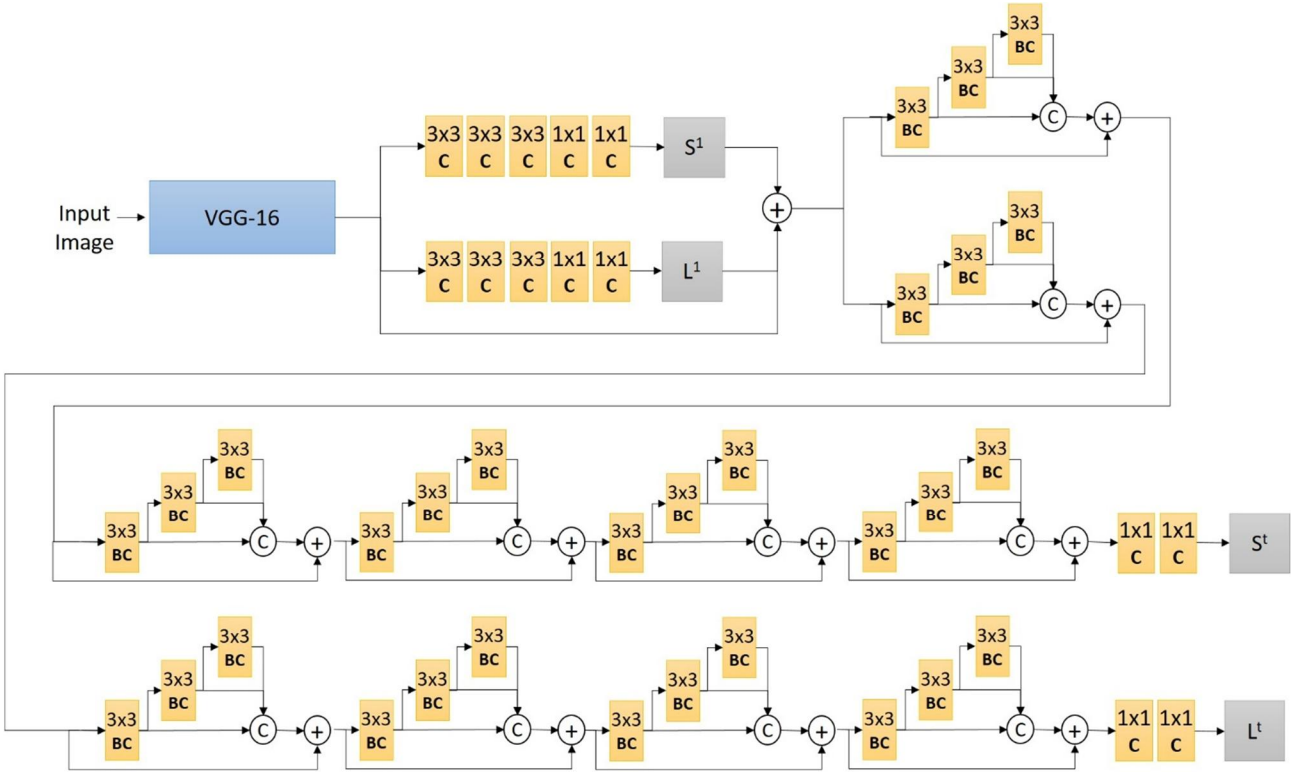

Figure 4. Proposed structure for Part Affinity based pose estimation with multi-scale binary module

4. Experimental Results

For the training we used the Coco Dataset which includes all the annotations as Json files. We modified the tensorflow version of the model proposed in [6], and run it on Linux OS with a GTX 1080 GPU on image of size 368×654. In the testing time, we compared the results of the model in [6], the mere binarized version of it, and the proposed structure. Figure 5 shows the decrease of the loss function as the training is proceeded with the proposed structure.

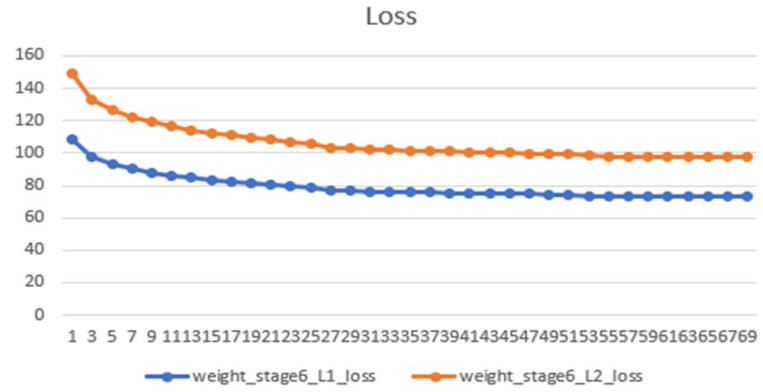

Figure 5. Showing the loss values for the training epochs with the proposed structure

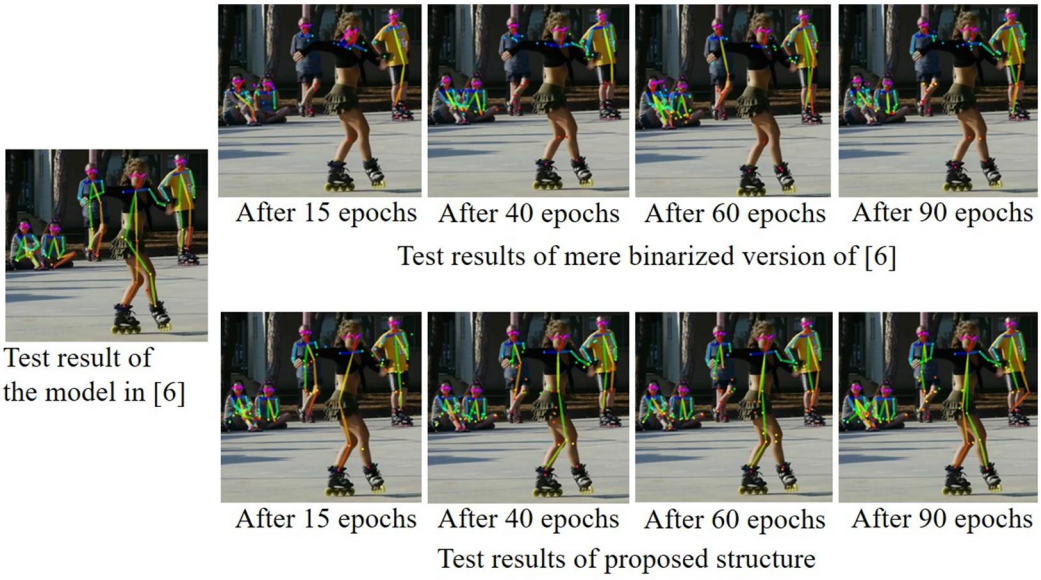

Figure 6. Comparing the test results at different training epochs of the mere binarized version and the proposed structure

Figure 6 shows the testing results with several different parameters obtained from different training epochs. As can be seen in the upper row of Fig. 6, the test results with the mere binarized code cannot estimate the part affinity fields well even after 100 epochs of training. However, with the proposed structure, the part affinity field is obtained almost the same as in the model in [6], while the number of parameters have been reduced largely.

In the test time, the inference time of the model in [6] was about 0.19 seconds in average, while with the proposed structure the inference time reduces to 0.09 seconds in average showing almost twice as fast an inference performance. However, in the experiments we used binarized values in floating point format, therefore, not fully utilizing the advantages of binarization. Thus, the reduce in the computation time is mainly due to the reduce in the number of parameters with the proposed structure, but not the full usage of binarized operation of the floating point values. This is because we could not make a function module which can implement a true binarized operation. However, we showed that with the binarized module similar accuracy can be obtained. Therefore, if we further make a function which computes the binarized operations

with fast binary shifting and XNOR operations, we will get the same accuracy while having a much faster speed performance.

5. Conclusion

We proposed a structure which accelerates the part affinity field based pose estimation algorithm while preserving similar accuracy. Experimental results verified this fact. As a future work, we think to adopt operations which make full usage of the binarized weighting values such as XNOR operation and shifting operation. Furthermore, when the pose estimation is done with a video, it seems to be redundant to computer the part affinity field from the sketch for every new input image. Therefore, initializations with previous results can further accelerate the estimation.

Acknowledgement

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT and Future Planning (No. NRF-2016R1D1A3B03931875)

References

[1] H.M. Kwon, V. Kumaran, and S. Gupta, “Real-time Tracking and Identification for Multi-Camera Surveillance System,” The Journal of the Institute of Internet, Broadcasting and Communication(JIIBC), Vol. 10, No. 1, pp. 16-22, Feb. 2018.
DOI: https://doi.org/10.7236/IJIBC.2018.10.1.3.

[2] J. Deutscher and I. Reid. “Articulated body motion capture by stochastic search,” International Journal of Computer Vision, Vol. 61, No. 2, pp.185–205, 2005.
DOI: https://doi.org/10.1023/B:VISI.0000043757.18370.9c

[3] Y. Du, Y. Wong, Y. Liu, F. Han, Y. Gui, Z. Wang, M. Kankanhalli, and W. Geng, “Marker-less 3d human motion capture with monocular image sequence and height-maps,” In European Conference on Computer Vision, pp. 20–36, 2016.
DOI: https://doi.org/10.1007/978-3-319-46493-0_2

[4] J. Gall, B. Rosenhahn, T. Brox, and H.-P. Seidel, “Optimization and filtering for human motion capture,” International Journal of Computer Vision, Vol. 87, No. 1, pp.75–92, 2010.
DOI: https://doi.org/10.1007/s11263-008-0173-1

[5] M. Trumble, A. Gilbert, A. Hilton, and J. Collomosse, “Deep convolutional networks for marker-less human pose estimation from multiple views,” In Proceedings of CVMP 2016. The 13th European Conference on Visual Media Production, 2016.
DOI: https://doi.org/10.1145/2998559.2998565

[6] Z. Cao, T. Simon, S-E Wei, Y. Sheikh, “Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields,” in Proc. Computer Vision and Pattern Recognition, pp. 7291-7299, July 21-26, 2017.

[7] M. Rastegari V. Ordonez, J. Redmon, and A. Farhadi, "XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks," In European Conference on Computer Vision, pp. 525-542, Sep. 2016.
DOI: https://doi.org/10.1007/978-3-319-46493-0_32

[8] A. Bulat and G. Tzimiropoulos, “Binarized Convolutional Landmark Localizers for Human Pose Estimation and Face Alignment with Limited Resources,” Proc. International Conference on Computer Vision, March 2017.
DOI: https://doi.org/10.1109/ICCV.2017.400