Body-Boundary-Refined Human Parsing Network based on Fully Convolutional Network and Conditional Random Fields

Hanlin Zhao\textsuperscript{a} and Xin Ma\textsuperscript{b}

School of Control Science and Engineering, Shandong University, Jinan, China

\textsuperscript{a}zhaohanlin@mail.sdu.edu.cn; \textsuperscript{b}maxin@sdu.edu.cn

Abstract. Human parsing is still a big challenge in computer vision task to accurately label every body part in an image. Current methods of semantic segmentation mainly focus on dividing each independent part and ignore the structural priors of human body. However, the position relationships of different parts, especially which around the area of boundary, indicate their associations and also contribute to the semantic segmentation. In our work, we propose a body-boundary-refined part to refine the segmentation result of human part edge by simply utilizing the structure priors around the body boundary. It puts a penalty mechanism on wrong marginal pixels to improve segmentation performance around the area of body boundary. The network achieves competitive performance on the PASCAL-Parts-dataset and especially the area around the body boundary has been refined.

1. Introduction

Human parsing plays an increasingly significant role in many fields of computer vision like human behaviour analysis and human robot handovers. However, the semantic segmentation of human body part is still a challenging task, for every parts of human should be distributed a label with specific and pre-defined semantic information such as head and arms.

Fortunately, human parsing could be regarded as a problem of semantic segmentation. The majority of traditional approaches \cite{1-9} based on the fully convolutional network (FCN) \cite{1} mainly focus on addressing the problem of losing many useful spatial information caused by the operation of pooling or large convolutional stride in the classification network \cite{10-12}. Some networks are favour of using encoder-decoder \cite{3} \cite{5} \cite{8} structure to obtain efficient details from encoder network. And other researches are interested in the area of exploiting the context information \cite{2} \cite{4} \cite{6} \cite{7}. We expect our network could adjust more natural scene with the person images extracted from PASCAL Parts dataset where having more than one multi-scale persons. Therefore, we choose one of the most advanced FCN models called DeepLabv3 \cite{2} as our baseline network with believe that human parsing still can be seen a semantic segmentation aimed to classify each pixel into correct label. However, there exist some differences between semantic segmentation and human parsing. For general semantic segmentation, a several of objects will be distributed everywhere in an image and we can't find obvious interrelationship between different categories. Human body parts have a closely positional relationship, which is driven us to exploit the connection in the pixel level.

Driven by the observation we select the conditional random field (CRF) to process the interaction between each pixel. However, the operation of conditional random field is always used for a post-processing \cite{4} \cite{9} structure followed by a fully convolutional network in most segmentation module because of the difficulties in the CRF training stage. Fortunately, a CRF-RNN structure that combines
the strengths of convolutional neural network and conditional random field to formulate a deep
network which can be trained end-to-end is introduced in this paper [13]. Although it is difficult to
balance the training speed and the internal CRF parameters, we argue that a structural prior of human
body such as CRF is beneficial for our task. Therefore, CRF-RNN is added in the human parsing
network.

But, only structure of CRF not enough obtain all potential position relationship at least around the
body's edge. For example, the potential correlation such that surrounding pixels’ categories of head
marginal pixels are likely to torso pixels or background pixels or head pixels. If prediction results of
surrounding marginal pixels don't align with the structural of the human body, some punishment will
be given to these wrong boundary pixels. To exploit these prior’s relationships, we propose a structure
named body-boundary-refined part to punish the wrong marginal pixels in the final loss function to
refine segmentation result.

2. Related Work

2.1. Semantic segmentation
Semantic segmentation aims to label different objects with relative semantic labels in pixel level.
Recently, FCN [1] has become the most successful basement model in semantic segmentation [2-7]
[9] [14]. It replaces the fully connected layer with convolutional layer to reserve the spatial
information of image to address pixel-level dense prediction and adopt skip connection to solve the
problem of losing spatial information. A universal structure called encoder-decoder module [3] [7] [8]
has been widely used and other successful networks [2] [6] think that it is useful to exploit multi-scale
contextual information in semantic segmentation. Zhao et al. [6] propose a pyramid pooling module
that is applied to harvest different sub-region representations to exploit global and local context
information.

2.2. Human parsing
Recently, variant of FCN also has been successfully used in human parsing. Human Part Discovery
[15] designs a like-FCN structure to obtain different human parts labels in two granularity levels.
However, it misses out the structural priors of the human body. Look into Person (LIP) [16] not only
contributes significantly to make a new benchmark, but also adopts a self-supervised structure-
sensitive learning to give a self-supervise for the net. Qi et al. [17] performed a holistic and instance-
level human parsing using conditional random field to allocate every part into every object with given
detections and category-level segmentation module. HAZN [18] designs an efficient strategy of first
zooming into objects and then zooming into parts to get a great performance. Other works like Chen e
al. [19] introduce an attention strategy which allocates different weights according to different scales
of objects into network to improve the performance in the small objects. In our work, the proposed
body-boundary-refined part can utilize the structural of the human body to refine the body boundary in
some extent.

2.3. Conditional Random Field
Currently, many works combine fully convolutional network with conditional random fields to
improve the performance of semantic segmentation [4] [17] [20] via incorporating surrounding
information with pairwise energy. More researches [13] [20] attempt to add CRFs into fully
convolutional network to train end-to-end. Thanks to the great contribute of densecrf [21], it's more
convenient to integrate fully connected CRFs with deep convolutional neural network in semantic
segmentation via Mean-field approximate inference and efficient message passing. In the human
parsing, the human body has many potential structural information could be used, such features could
well solve by CRFs, so we incorporate the densecrf with deeplab to refine the prediction result.

3. Proposed Approach
Body-boundary-refined network consists of three components as shown in Figure 1. A variant fully convolutional network named DeeplabV3 [2] and a fully connected CRFs model named CRFasRNN [13] and our proposed model named body-boundary-refined part to refine the boundary of human body.

3.1. Deeplabv3 Network
For the task of semantic segmentation, there are two challenges in exploiting convolutional neural network. One problem is reduced feature resolution and another is multiple scales of objects in the image. We choose Deeplabv3 as baseline network to solve the case of having multi-scale persons.

3.2. Fully Connected Conditional Random Fields
Conditional random field, which can establish relationships between mutual independence pixels, is a probability graph model with Markov property. In our work, five labels such as head, torso, arms, legs, and background should be given in an image. Although this task is similar with semantic segmentation, the special structural relationship is unique in the human segmentation. Thus, human body image should contain so many useful priors that we can’t ignore. Therefore, CRFs is followed by the deeplabv3 to represent structural priors of human body.

3.3. Body-Boundary-Refined Part
When we only use deeplabv3 to segment human semantic body part, some terrible results appear that these connected pixels between head part and torso part have disappeared. Therefore, we realise that it is significant to encode the structural priors of human body part especially around boundary into network. Motivated by this intuition, we design a body-boundary-refined part which includes penalty weight $\beta$ and category score ($\text{score}_{\text{category}}$, as in equation (2)) to improve the importance of human body boundary. We put the $L_{\text{boundary}}$ (as in equation (1)) as a small part of total loss to punish the terrible boundary results of prediction. In addition, in order to conveniently compute the punishment, we have fixed the penalty weight $\beta$ as 0.01 in our work.

$$L_{\text{boundary}} = \beta \cdot \text{score}_{\text{category}}$$ (1)

$$\text{score}_{\text{category}} = \sum_{i=1}^{N} \text{Boundary}(I_{\text{score}}, I_{\text{pred}})$$ (2)

The category score (as in equation (2)) represents that how much punishment network would put into the total loss. Fortunately, these four categories have some fixed position relationship especially around the human body boundary. For example, if boundary pixels of head have been predicted by network, obviously we can infer categories of surrounding pixels. In another words, the eight neighborhood of boundary pixels in prediction represent the degree of good or bad about structural information. That is, for example, if there is a boundary pixel connected with head and torso, then
head pixels and torso pixels will be found in the eight neighborhoods of this pixel, thus the eight neighborhood of boundary pixels in prediction will represent the structural information about human body. Therefore, some norms \(I_{\text{norm}}\) as in equation (3) are defined to present the structural priors of human body. Let \(L = \{l_0, l_1, l_2, l_3, l_4\}\) be the labels of background, head, torso, arms, and legs and \(X = \{x_1, x_2, \ldots, x_N\}\) be the boundary pixels. Let \(H = \{\eta_0, \eta_1, \eta_2, \eta_3, \eta_4, \eta_5, \eta_6\}\) be the categories of every pixels in the eight neighborhood and \(\eta_0\) be the boundary pixels in the central position of eight neighborhood. Where \(P(\eta = l_i \mid X = l_j)\) means the probability of appearing background or head or torso pixels is larger than other situations around the head marginal pixels. Similarly, \(P(\eta = l_i \mid X = l_j)\) means the probability of appearing background or torso or arms pixels is larger than other situations around the arms marginal pixels and \(P(\eta = l_i \mid X = l_j)\) means the probability of appearing background or torso or legs pixels is larger than other situations around the legs marginal pixels.

\[
I_{\text{norm}} = \begin{cases} 
    & P(\eta = l_i \mid X = l_j) \mid X = l_j \\
    & P(\eta = l_i \mid X = l_j) \mid l_j \in H \\
    & P(\eta = l_i \mid X = l_j) \mid l_j = l_i 
\end{cases} \quad (3)
\]

After acquiring the pixels of boundary in the prediction results, the norm is performed to decide whether giving punishment to the total loss. As in equation (4), where \(\text{Boundary}(I_{\text{norm}}, I_{\text{pre}})\) means that if the boundary pixel in the prediction result is satisfied with the norm, then zero punishment will give to the loss. However, the punishment should put to the loss when the prediction result breaks the norms.

\[
\text{Boundary}(I_{\text{norm}}, I_{\text{pre}}) = \begin{cases} 
    0, & I_{\text{norm}} = I_{\text{pre}} \\
    1, & I_{\text{norm}} \neq I_{\text{pre}}
\end{cases} \quad (4)
\]

4. Experiments
We evaluate the performance of our network on the PASCAL Parts dataset and some images from Freiburg Sitting People dataset [15] to visualize the segmentation effect of our network.

4.1. Experimental Settings
PASCAL Parts dataset: In our experiment, we extract 3589 images which contain human body by key annotation word of person from original PASCAL Parts dataset and fuse 24 original human parts classes into 4 classes to evaluate our network. Then, we randomly divide dataset into 80% training and 20% testing.

Metrics: In our experiment, pixel accuracy and intersection over union are used to evaluate performance of our network. The intersection over union (IOU) is also used to balance quantitative gap from background and foreground pixels. Confusion matrix is used to compute the mean IOU and IOU of every classes.

Training: We use the pre-trained model from Resnet101 and the scale of input images is fixed as 300x300 for training. Two types of augmentations to our training data is mirroring and cropping. Polynomial decay is adopted in the setting of learning rate. The initial learning rate is 7e-3 and the end learning rate is 1e-4. After training Deeplabv3 for 80 epochs we add the CRF and body-boundary-refined part into our end-to-end network by the trained deeplabv3 to increase the training speed.

4.2. Results and Comparisons
In our task, only four labels are divided into human body parts and there exist few same works with us. Therefore, we compare FCN-8s [1] and Deeplabv3 [2] with our network in the same experiment environment to show the improvement of our network. Table 1 show the mean intersection over union (IOU) and pixel accuracy of FCN-8s, Deeplabv3 and Deeplabv3+CRF in the person parts dataset.
extracted from PASCAL Parts dataset by us with four parts labels and background label. We first train FCN-8s [1] with VGG16 [10] extracting feature network in the same dataset getting bad mean IOU result. Then when using Deeplabv3 [2] with ResNet101 [11] extracting feature network, we obtain good mean IOU result. However, many details are lost especially in the small area like hand and foot. According to the third column results of Figure 2, human body parts are refined by adding CRF into Deeplabv3 and the mean-IOU is improved about 1.5%. Compared with previous model, although our network obtains a little improvement in the mean-IOU ranging from 58.72% to 60.64%, the segmenting effect on the body boundary like the area of arms or head or legs is greatly improved as the fourth column shown, especially in the joints of the two parts. The refined inference image and a little improvement in mean IOU have confirmed the effectiveness of our body-boundary-refined part.

![Figure 2. Qualitative results on the PASCAL dataset and the Freiburg sitting people dataset.](image)

**Table 1.** Performance comparison in terms of mean IOU and pixel accuracy on the validations.

| Method                  | Head | Torso | Arms | Legs | Bkg | mIOU | PA     |
|-------------------------|------|-------|------|------|-----|------|-------|
| FCN-8s [1]              | 41.31| 17.98 | 3.61 | 0.02 | 87.87 | 30.16 | 86.92 |
| Deeplabv3 [2]          | 71.21| 54.65 | 37.98| 32.90| 89.30 | 57.21 | 88.17 |
| Deeplabv3+CRFas RNN    | 75.54| 54.26 | 42.99| 30.96| 89.89 | 58.72 | 89.18 |
| Our network             | **75.64** | **58.55** | **46.42** | **32.19** | **90.40** | **60.64** | **89.54** |

5. Conclusion
In this paper, we present a novel body-boundary-refined part based on semantic segmentation and fully connected CRF to improve the performance of human parsing. Our approach applies Deeplabv3 as our baseline network and fuses denseCRFs into Deeplabv3. To exploit more structural priors of human body part particularly in the area of boundary, we introduce a new body-boundary-refined part for human parsing task, including 3,589 images extracted from PASCAL Parts dataset with pixel-wise annotations on four semantic part labels. Experimental results on the dataset demonstrate that our proposed approach refine boundary of human, and achieves a good performance for the human parsing.

Acknowledgments
This work was supported in part by the Chinese National Key Research and Development Plan (2018YFB1305803), Chinese National Natural Science Foundation (61673245), Chinese National Programs for High Technology Research and Development (2015AA042307).

References
[1] Long, Jonathan and Shelhamer, Evan and Darrell, Trevor. Fully convolutional networks for semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015:3431–3440.
[2] Chen, L. C., Papandreou, G., Schroff, F., Adam, H. Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587, 2017.

[3] Badrinarayanan V, Kendall A, Cipolla R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation [J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 39(12): 2481-2495.

[4] Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence, 40(4): 834-848, 2018.

[5] Lin, G., Milan, A., Shen, C., Reid, I. D. RefineNet: Multi-path Refinement Networks for High-Resolution Semantic Segmentation. CVPR. Vol. 1. No. 2, 2017.

[6] Zhao, Hengshuang and Shi, Jianping and Qi, Xiaojuan and Wang, Xiaogang and Jia, Jiaya. Pyramid scene parsing network. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2017:2881--2890.

[7] Chen L C, Zhu Y, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]. Proceedings of the European Conference on Computer Vision (ECCV). 2018: 801-818.

[8] Ronneberger, Olaf and Fischer, Philipp and Brox, Thomas. U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention, 2015:234--241.

[9] Chen L C, Papandreou G, Kokkinos I, et al. Semantic image segmentation with deep convolutional nets and fully connected crfs[J]. arXiv preprint arXiv:1412.7062, 2014.

[10] K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition. International Conference on Learning Representations, 2015.

[11] He, Kaiming and Zhang, Xiangyu and Ren, Shaoqing and Sun, Jian. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016:770--778.

[12] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.

[13] Zheng S, Jayasumana S, Romera-Paredes B, et al. Conditional random fields as recurrent neural networks[C]. Proceedings of the IEEE international conference on computer vision. 2015: 1529-1537.

[14] Yu, Fisher, and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122, 2015.

[15] Oliveira, G. L., Valada, A., Bollen, C., Burgard, W., Brox, T. Deep learning for human part discovery in images. IEEE International Conference, 2016:1634-1641.

[16] Gong, Ke and Liang, Xiaodan and Zhang, Dongyu and Shen, Xiaohui and Lin, Liang. Look into Person: Self-Supervised Structure-Sensitive Learning and a New Benchmark for Human Parsing. CVPR, 2017.

[17] Li, Qizhu and Arnab, Anurag and Torr, Philip HS. Holistic, Instance-level Human Parsing. arXiv preprint arXiv:1709.03612, 2017.

[18] Xia, Fangting and Wang, Peng and Chen, Liang-Chieh and Yuille, Alan L. Zoom better to see clearer: Human and object parsing with hierarchical auto-zoom net. European Conference on Computer Vision, 2016:648--663.

[19] Chen, Liang-Chieh and Yang, Yi and Wang, Jiang and Xu, Wei and Yuille, Alan L. Attention to scale: Scale-aware semantic image segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016:3640--3649.

[20] Teichmann, Marvin TT and Cipolla, Roberto. Convolutional CRFs for Semantic Segmentation. arXiv preprint arXiv:1805.04777, 2018.

[21] Krähenbühl, Philipp and Koltun, Vladlen. Efficient inference in fully connected crfs with gaussian edge potentials. Advances in neural information processing systems, 2011:109--117.