An UNet-based Head Shoulder Segmentation Network

Hong-Xia Xie[1] , Chih-Yang Lin[2]*, Hua Zheng[1], and Pei-Yu Lin[3]
[1]College of Photonic and Electronic Engineering, Fujian Normal University, Fuzhou, China
[2]Dept. of Electrical Engineering, Yuan-Ze University, Taoyuan, Taiwan
[3]Dept. of Information Communication, and Innovation Center for Big Data and Digital Convergence, Yuan-Ze University, Taiwan

ABSTRACT

Within the rapidly developing field of computer vision, pedestrian detection is a fundamental and challenging task for both industry and academia. However, object segmentation information can help the network to capture the attention of the model during training. In this paper, we propose a head-shoulder segmentation network based on modified U-Net network. The architecture consists of a contracting path to capture information from a lower layer and a symmetric expanding path to enable precise localization. The proposed model aims to effectively segment the head-shoulder portion of pedestrian without a huge annotated training sample. Segmentation of a random image takes less than a second on NVIDIA GTX 1070. This paper will show the mean IOU and some segmentation results to prove effectiveness of this model.

INTRODUCTION

Currently, detecting people on the street, in supermarkets, airports and railway is a difficult problem that remains to be solved. Variations in pose, lighting conditions and partial occlusions, etc., can result in the head-shoulder area being occluded. This area is not as flexible as the rest of the human body, and an omega-like shape has been proven to be a salient feature of the head-shoulder region. Thus, the pedestrian detection task can be transformed into omega shape detection.

In recent decades, researchers have proposed many head-shoulder detection methods. The main differences between these methods lie in the feature extraction and classifier design. Li et al. [9]-[8] propose an effective head-shoulder detection method based on boosting local HOG features. Experimental results from[8] indicate HOG feature performs better than Hair feature [9] and SIFT descriptor. Another edge-based feature similar to HOG, named Oriented Integration of Gradients(OIG), is introduced to describe subparts of human head-shoulder in [5]. Julio et al. [7]-[3] propose a graph-based segmentation model to estimate the head-shoulder contour. As can be seen from the literature review, most of the work uses the handcrafted features.

The success of ImageNet [2] and the Pascal Visual Object Classes(VOC) Challenge [4] showed that object classification is a very promising research area that has led to many great advancements. In general, the best results have been achieved through deep Learning methods. The excellent performance and flexibility of convolutional neural networks (CNNs) has led to the VOC's use for pedestrian detection. [13] analyzes the main reasons account for the insufficient accuracy of Faster RCNN for pedestrian detection and proposes an RPN followed by boosted forests pipeline. [11] aggregates extra features into CNN-based pedestrian detection framework. [1]utilize the heatmap of semantic scene parsing, in which detectors benefit from the semantic information within a large receptive field.

However, less research has been conducted in the realm of head shoulder CNN-based methods for pedestrian detection because of a lack of large annotated training samples.

In this article, we try to tackle the problem of head-shoulder semantic segmentation and propose a deep learning based end-to-end learning framework.

THE ALGORITHMS

The model we chose is a scaled down version of an image segmentation architecture called U-Net. U-Net is an encoder-decoder type network architecture for image segmentation. The name of the architecture comes from its unique shape, where the feature maps from down-sampling block are fed into the up-sampling block. We modify some details in U-Net to achieve better segmentation performance.

Semantic segmentation faces an imbalance between semantics and location: global information resolves "what" while local information resolves "where" [10]. Unlike the coarse skip connection used in [10], the up-sampling part of U-Net has a large number of feature channels, which allow the network to propagate context information to higher resolutions [12]. To enhance learning rate and overall accuracy, we used batch normalization [6] after each convolutional layer and relu layer. By normalizing the data in each mini-batch, the "internal covariate shift" problem is largely avoided

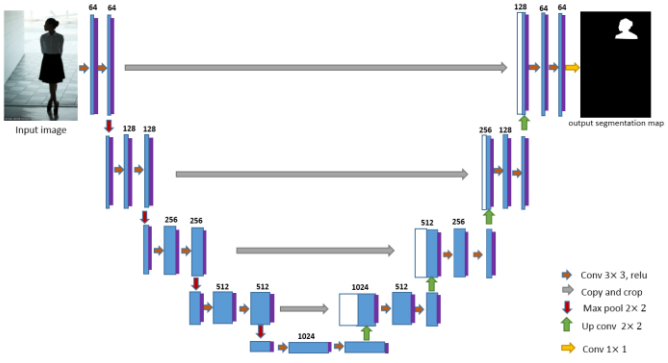

Fig.1 An modified U-NET architecture. Each blue box represents the feature map in each convolutional layer and purple box following is the batch normalization layer. The number of channels we used is indicated on top of the blue box. Since we didn't adopt the overlap-tile strategy used in [12], the output image should be the same size with the input image.

EXPERIMENTS

A. Experiments setup

1) Datasets preparation

The dataset contains 672 human head-shoulder images collected from the PASCAL VOC2012 dataset. Since we focus on human head-shoulder segmentation, the images we selected contain at least one person, and span several scenes, different illumination conditions, and multi-scale size. Due to a lack of public head-shoulder datasets, we used the Photoshop to manually label the ground truth for all the images. Data augmentation is an important step since a limited number of training samples are available. In the case of our application, we mainly focus on shift and rotation variance as well as scale variations. Ultimately, we generated 4032 training samples.

2) Training

To ensure the best performance, epoch is set to be 20, batch size to 4, and momentum to 0.2. The initial learning rate is 0.001, when the number of epochs reaches half of the total epochs; then, the learning rate is adjusted to 0.0001.

Loss function: Softmax function is a common way to provide a probabilistic similarity score between a predicted result and ground truth value. In this head-shoulder segmentation task, the modified UNet divides the pixels into head-shoulder and background areas; thus the softmax is simplified correspondingly.

3) Evaluation metrics

Here we choose an average IOU("intersection over union") to evaluate the proposed method. We ultimately achieve a 91% accuracy rate.

B. Experimental results

Figure2 shows some testing results.

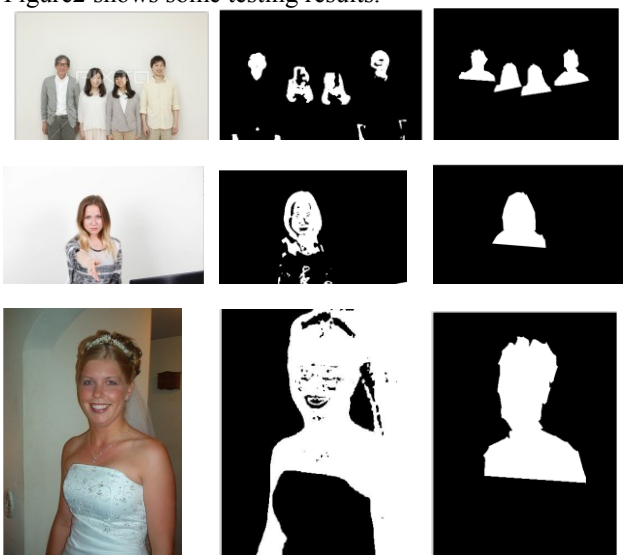

Fig.2. (a) input image (b) segmentation result (c) ground truth

CONCLUSION AND FUTURE WORKS

In this paper, we proposed a new deep learning method for human head-shoulder segmentation, i.e. modified U-Net network. It is an end-to-end fully convolutional network with U-shape connections and Plus connection. Under limited training images, experiments results show that the proposed network performed well in background noise suppression and edge detection. What's more, it can accurately recognize the omega shape that indicates the head-shoulder region of pedestrians.

In the future, we intend to augment our training dataset and integrate multi-scale technique into our architecture to further enhance performance.

REFERENCES

[1] A. Daniel Costea and S. Nedevschi, "Semantic channels for fast pedestrian detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2360-2368, 2016.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 248-255, 2009.

[3] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 4, pp. 743-761, 2012.

[4] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303-338, 2010.

[5] F. He, Y. Li, S. Wang, and X. Ding, "A novel hierarchical framework for human head-shoulder detection," *Image and Signal Processing (CISP), 2011 4th International Congress on*, pp. 1485-1489, 2011.

[6] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *International Conference on Machine Learning*, pp. 448-456, 2015.

[7] J. C. Jacques, C. R. Jung, and S. R. Müsse, "Head-shoulder human contour estimation in still images," *Image Processing (ICIP), 2014 IEEE International Conference on*, pp. 278-282, 2014.

[8] M. Li, Z. Zhang, K. Huang, and T. Tan, "Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection," *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pp. 1-4, 2008.

[9] M. Li, Z. Zhang, K. Huang, and T. Tan, "Rapid and robust human detection and tracking based on omega-shape features," *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pp. 2545-2548, 2009.

[10] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431-3440, 2015.

[11] J. Mao, T. Xiao, Y. Jiang, and Z. Cao, "What Can Help Pedestrian Detection?," *arXiv preprint arXiv:1705.02757*, no., 2017.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234-241, 2015.

[13] L. Zhang, L. Lin, X. Liang, and K. He, "Is faster r-cnn doing well for pedestrian detection?," *European Conference on Computer Vision*, pp. 443-457, 2016.