Face Parsing for Mobile AR Applications

Conference Paper · October 2018

DOI: 10.1109/ISMAR-Adjunct.2018.00119

5 authors, including:

Xavier Naturel
Fitting Box
18 PUBLICATIONS **172** CITATIONS
SEE PROFILE

Yongzhe Yan
Mines Saint-Etienne
5 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Anthony Berthelier
Institut National des Sciences Appliquées de Lyon
4 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Face Parsing for Mobile AR Applications

Yongzhe Yan*
Université Clermont-Auvergne
Wisimage

Benjamin Bout
Université Clermont-Auvergne
Wisimage

Anthony Berthelier
Université Clermont-Auvergne
Wisimage

Xavier Naturel
Wisimage

Thierry Chateau†
Université Clermont-Auvergne

ABSTRACT

Face parsing is a segmentation task over facial components, which is very important for a lot of facial augmented reality applications. We present a demonstration of face parsing for mobile platforms such as iPhone and Android. We design an efficient fully convolutional neural network (CNN) in an hourglass form that is adapted to live face parsing. The CNN is implemented on the iPhone with the CoreML framework. In order to visualize the segmentation results, we superpose a mask with false colors so that the user can have an instant AR experience.

Keywords: Face Parsing, Mobile AR, Semantic Segmentation, Video Segmentation, Computer Vision

1 INTRODUCTION

The goal of face parsing is to classify every pixel of an image into a category of facial components. Detecting different facial components is of great interest for a lot of augmented reality (AR) applications such as facial image beautification and facial image editing. For example, given the lip area, we can apply virtual lipstick-wearing effect on it by colorizing the region with proper colors. In this paper, we focus on designing an efficient face parsing methods by neural networks since lots of these applications are aimed at the mobile platforms such as iOS and Android.

Deep convolutional networks have been proved to be the leading methods for lots of computer vision tasks especially semantic segmentation [1, 8]. Nonetheless, these methods are either time-consuming or over-sized due to the excessive amount of parameters and computation. Most of the semantic segmentation methods are designed for general complex scenes or street scenes. However, face parsing is a quite different task for the following reasons.

- Face parsing is usually done based on an ROI given by preliminary face detection compared to the general semantic segmentation which is generally performed on the entire image.
- Sharp boundaries are demanded for facial AR applications in order to render better visual effects.
- Facial components have more deformable variance but less position and size variance compared to semantic segmentation.

In this paper, we present a real-time AR face parsing demonstration on iPhone with an efficient deep convolutional neural network. Unlike the precedent face parsing methods, we consider the adaption of neural networks on video in order to provide a fluid and temporally consistent rendering. The users are able to visualize the segmentation results by a mask which indicates different facial regions. A rendering result is shown in figure 1. More AR applications can be realized based on our results.

*e-mail: yongzhe.yan@etu.uca.fr
†e-mail: thierry.chateau@uca.fr

Figure 1: The visual results of our method on iPhone. Our method is robust to extreme expressions and poses

2 RELATED WORK

A commonly shared consensus in deep semantic segmentation area is that there exists two kinds of mainstream methods [3], the spatial pyramid pooling (SPP) module based structures and encoder-decoder structures [1, 8]. Encoder-decoder structures adopt progressive up-sampling with skip connection to reconstruct the object boundary as sharp as possible. Skip connections play an important role in the network structure so that the CNN can transfer the low-level detailed information to the output layers.

To ensure the best user experience in AR applications, short inference time and low latency are required. Some researchers proposed to use optical flow and reuse the feature map of the past frames if the static scene persists. Another way to accelerate the inference time is to use network acceleration techniques like quantification or pruning. Recent works such as Mobile-net [9] and Shuffle-net proposed to use depth-wise separable convolution to reduce computational complexity of CNN with almost the same performance.

CNN based methods [6] brought a huge advance to face parsing compared to exemplar based methods [10]. Liu et al. [6] proposed to use Conditional Random Field (CRF) to provide sharp facial component contours. Sharing the same motivation, another work [5] proposed to use spatially variant recurrent unit which enables the regional information propagation.

3 MOBILE FACE PARSING DEMO DESCRIPTION

In this part, we present the design of our segmentation network, how we adapt it to the video as well as some implementation details.

3.1 Mobile Hourglass Network

We follow the design of the hourglass model in human pose estimation [7]. Our network structure is presented in Figure 2. The Hourglass model is a symmetric encoder-decoder fully convolutional network with a depth of 4, which means the encoder downsamples the input for 4 times and the decoder upsamples the feature map for 4 times to reconstruct the output. Each yellow block represents a network block which enables the CNN to learn the information flow at each stage and skip connection.

We drop several convolutional and max-pooling layers at the beginning of the network and augment the size of the output map to 256, which is the same size as the input image. This will make the boundaries of facial objects sharper but increase the computational complexity as well. In order to accelerate the inference, we replace

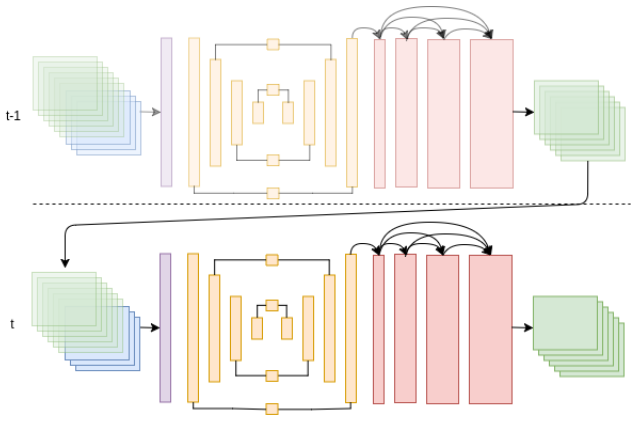

Figure 2: Overview of our method for video face parsing. Blue channels: RGB facial image, Green channels: Mask predictions, More transparency signifies earlier predictions in time dimension. The output predictions of $t-1$ is reinjected to t for robustness. Purple block: convolutional layer + batch norm layer. Yellow blocks: The Hourglass model composed of Mobile-net Blocks. Red blocks: Dense [4] Blocks

all of the ResNet blocks in the hourglass network by MobileNet [9] inverted bottleneck block. Expansion factor t is an important hyper parameter which indicates how many times the number of feature channels are expanded in its blocks. We chose the number of features f as 32 and the expansion factor t as 6 by finding the best compromise between the speed and segmentation quality.

3.2 Video Adaption

Inspired by this blog [2], we implemented two strategies to adapt our model to video face parsing.

For inference, we take the mask of the frame $t-1$ as input to the frame t to stabilize the segmentation. Due to the lack of video dataset, we apply a randomly transformed mask as a fake previous mask during the training. According to our experiments, this will eliminate the random segmentation noise which is present without taking the previous frame mask as input.

We add four dense layers [4] with a growth rate of 8 at the end of the network for more robust rendering.

4 Experiments

We train our model on the Helen dataset [10] which contains 2330 images manually labeled in 11 classes including background, skin, hair nose, left/right eyebrows, left/right eyes, upper/bottom lips as well as inner-mouth. The models are trained on all of the labels except the hair because the hair annotations are not precise. The images are cropped with a margin of 30%-70% of bounding box size according to the facial landmark annotations.

We use the RMSprop as optimizer and softmax cross-entropy function as loss function. We apply an initial learning rate of 0.0005 with decay of 0.1 for each 40 training epochs until total 190 epochs are finished. We use ONNX as an intermediate format to transform our Pytorch model to CoreML model, which is optimized for iOS devices. We adopt the Vision framework for face detection that is anterior to face parsing .

We compare our methods with several well-known segmentation networks [1, 8]. The results are measured in F1-score in Table1. The number of parameters are also listed aside to provide more information about the model size, which is critic for mobile platform.

We measure the runtime of different MobileNet block settings by changing the number of channels f and expansion factor t. We also provide a profiling time analysis on the face detection, array transformation and colorization in Table2.

Table 1: Quantitative evaluation on Helen dataset

Model	Overall F-score	Num. of parameters
SegNet [1]	92.90	29.45M
Unet [8]	93.72	13.40M
Mobile-Hourglass-f16-t3	93.00	0.09M
Mobile-Hourglass-f16-t6	93.08	0.12M
Mobile-Hourglass-f32-t3	93.18	0.17M
Mobile-Hourglass-f32-t6	93.55	0.27M

Table 2: Run-time (in ms) profiling on iPhone X

Model	FD	MI	Colorization	Total
Unet [8]	7	203	54	269
Mobile-Hourglass-f16-t3	8	77	18	106
Mobile-Hourglass-f16-t6	7	75	18	103
Mobile-Hourglass-f32-t3	7	76	18	104
Mobile-Hourglass-f32-t6	7	78	18	106
Dense Mobile-Hourglass-f32-t6	7	75	18	110

FD: Face Detection MI: Model Inference

5 Conclusion

We present a real-time encoder-decoder video face parsing mobile AR demonstration. Our method will draw interest because it is simple and interesting for everybody to test. More importantly,

- Face parsing is crucial for a lot of facial editing applications for example virtual make-up, hair dying, skin analysis, face morphing, reenactment etc.
- Our method is not only limited to facial AR applications but also interesting for fine-grained segmentation based AR applications.

References

[1] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[2] V. Bazarevsky and A. Tkachenka. Mobile real-time video segmentation. `https://ai.googleblog.com/2018/03/mobile-real-time-video-segmentation.html`, 2018.

[3] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *arXiv preprint arXiv:1802.02611*, 2018.

[4] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, 2017.

[5] S. Liu, J. Shi, J. Liang, and M.-H. Yang. Face parsing via recurrent propagation. *arXiv preprint arXiv:1708.01936*, 2017.

[6] S. Liu, J. Yang, C. Huang, and M.-H. Yang. Multi-objective convolutional learning for face labeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3451–3459, 2015.

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

[8] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

[9] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, 2018.

[10] B. M. Smith, L. Zhang, J. Brandt, Z. Lin, and J. Yang. Exemplar-based face parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3484–3491, 2013.