A New Hand Segmentation Method Based on Fully Convolutional Network

Shiyu Zhao[1], Wankou Yang[1], Yangang Wang[1,*]

[1]Southeast University, Central building, Nanjing P.R. China, 210096
E-mail: charlottezsy@foxmail.com, wkyang@seu.edu.cn, ygwangthu@gmail.com

Abstract: Hand segmentation has several important applications such as human-machine interaction, person behaviors identification and etc. However, traditional hand segmentation methods cannot be widely used due to the complexities of hand motion and environment. With the development of deep learning, convolutional neural networks are demonstrated as powerful in many vision tasks. In this paper, we present a hand segmentation method based on fully convolutional networks (FCNs). We transfer the FCN-8s architecture of VGG 16-layer net (VGG16) into a hand segmentation network. Through fine-tuning the version of VGG16 model in ILSVRC-2014 competition, we obtain a professional hand segmentation model. Experiments show that our method achieves a 91.0% mean IU on our hand dataset and gives a great performance on hand segmentation.

Key Words: Hand detection, Hand Segmentation, Fully Convolutional Networks

1 INTRODUCTION

Hand detection and segmentation is very important in the area of computer vision. It has many applications, such as hand gesture estimation, driver behaviors monitoring, human-computer interaction, and etc. However, estimating the hand segmentation mask is very challenge because of not only the flexibilities of the hand motion, but also the lighting, shadows and occlusions between hand interactions. Although there are many works target to detecting the hand and estimating the semantic hand mask year by year, it is still an interesting and open problem for researchers to solve.

Traditional hand segmentation methods rely on the skin color [2, 3, 5], feature extraction [4, 6] and etc. The per-pixel segmentation accuracy and performance of these methods are not easily to meet the high-quality requirements. In recent years, deep neural networks (DNNs), especially convolutional neural networks (CNNs), has achieved great success. They have been widely used and are all very successful in many computer vision tasks, such as object recognition, human face tracking, human pose estimation and etc. Lots of pioneer works have applied CNNs for hand detection, including hand key-point detection [8], hand skeleton detection [9], bounding box detection [10, 11]. However, the per-pixel hand segmentation results of these typical convolutional neural networks still need lots of efforts.

Among the majority of different convolutional neural networks, fully convolutional networks (FCN) [1], as one important type of convolutional neural networks, can be trained end-to-end and pixels-to-pixels. It has been widely used by transferring common classification networks to the segmentation task and is suitable for the task of semantic segmentation. A large number of experiments have demonstrated that FCN is efficient in per-pixel tasks and can exceed the state-of-the-art method in the task of semantic segmentation.

In this paper, we build a new hand segmentation dataset and present a new method for hand semantic segmentation on the newly built hand dataset. We integrate the FCN-8s net architecture with a well-designed hand segmentation network. In the training stage, we fine-tune the VGG16 model from the one in ILSVRC-2014 competition [12], and adapt the trained model to be a hand segmentation model by training on our new hand segmentation dataset. The experiments have demonstrated that the proposed method can obtain higher per-pixel segmentation accuracy, which can achieve a 91.0% mean IU.

2 METHODS

2.1 Fully Convolutional Networks

Fully convolutional networks are confirmed to be able to define a new architecture for semantic segmentation that consists coarse, semantic and local, appearance information to improve the prediction.

In this study, the VGG 16-layer net [7] is chosen as the basic architecture of our method. To build the FCN-8s net of VGG16 (See Fig 1), add a 1×1 convolution layer on top of the pool4 layer to produce the pool4 predictions (FCN-32s net) and add a $2\times$ upsampling layer on top of conv7 (convolutionalized fc7) to produce the conv7 predictions at stride 32 simultaneously. Then fuse the two predictions together and sum them as a fuse-output (FCN-16s net).

*Corresponding author: Yangang Wang

This work is supported by National Nature Science Foundation under Grant 61703203, Natural Science Foundation of Jiangsu Province under Grant BK20170812.

978-1-5386-1243-9/18/$31.00 ©2018 IEEE

Similarly, add another 1×1 convolution layer on top of the pool3 layer to produce the pool3 predictions. Then fuse the predictions together with the $2\times$ upsampling of predictions of the fuse-output.

The fusion improvements can be reflected in both the mean IU metric during training and validation and the visible output results [1]. It is obvious that the performance of the FCN-16s net must be better than the performance of the FCN-32s net. And the FCN-8s net absolutely performs best of the three nets.

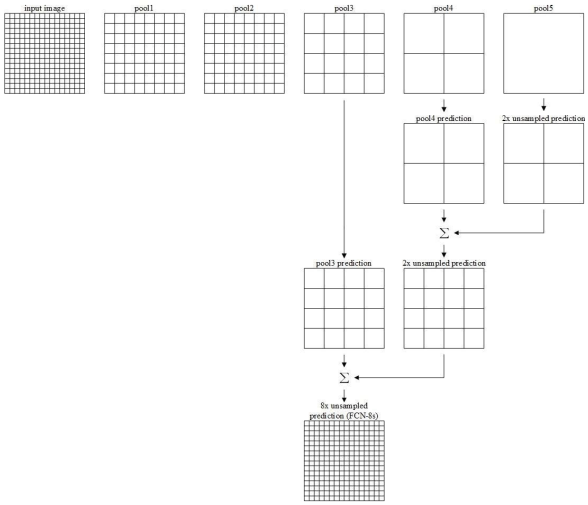


Fig 1. The FCN-8s net can learn to consist coarse high-layer information together with fine low-layer information. For convenience, only the pooling layers in VGG16 are shown in this figure.

2.2 Hand Dataset

For it is hard to find a published hand segmentation dataset which can perfectly meet our training and validation requirements, we labeled a hand segmentation dataset by ourselves.

For this study, we collected the hand images from two sources. We first collected the public images with only single hand and rick background from the internet. However, the collected hand images cannot cover all hand gestures. Then, we asked several volunteers to perform the lacked typical hand gestures and capture the hand images.

With the collected real hand images, the hand skeleton (key points) and hand mask are both labeled. In total, we have collected more than 10000 images for our neural network training.

2.3 Hand Detection

Fully convolutional networks transform fully connected layers into convolutional layers so that they can enable classification network to export a heatmap. The FCN-8s net detect the probabilities of every pixel of belong to each class and then compare these probabilities one by one to get the largest one. This pixel is identified as the corresponding class.

The FCN-8s net has been trained on several famous datasets already, including PASCAL VOC 2011, NYUDv2, SIFT Flow and SDBB. And most of trained models has been published. However, nearly all of these datasets have more than 20 classes. It means that we cannot use them directly, We need to change the layers directly related to the score layer.

For our aim is to make the FCN-8s net do semantic segmentation only for hands and ignore other objects, we first adapt the number of output into two, which refers to the hands and the background. Fig 2 shows the architecture of our networks for hand segmentation transferred from VGG16.

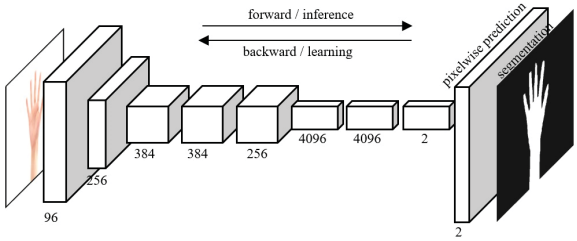


Fig 2. Fully convolutional networks is efficient on per-pixel tasks and can be transferred into a hand segmentation detector.

3 EXPERIMENTS

3.1 Dataset

For it is hard to find a published hand segmentation dataset which can perfectly meet our requirements, we test our method on a hand-mask dataset labeled by ourselves, which is described in Section 2.2.

3.2 Optimization

The model is trained by SGD together with momentum. We choose momentum of 0.99 and fixed learning rates of 10^{-9} for FCN-VGG16. We use weight decay of 5×10^{-4} and fixed the learning rates for biases. Dropout ratio is set 0.5 as used in the original VGG16 net.

We tried to zero-initialize the score layer and find that random initialization is able to reach the same performance and speed of convergence.

3.3 Fine-tuning

We decide to fine-tune the parameters of all layers by back-propagation through the whole VGG16. We need to notice that the VGG net should be trained in stages while initializing from the full 16-layer versions for achieving better performances. If we want to get a better model, we should fine-tune the FCN-32s net first. And then use the FCN-32s model to upgrade the FCN-16s net. Finally, use the FCN-16s model to upgrade the parameters of FCN-8s net.

For convenience, in this experiment, we use the improved version of VGG16 model in ILSVRC-2014 competition to

directly fine-tune our FCN-8s net. For the model has been trained perfectly by a large amount of data which makes it powerful in extracting high layer features of images. We transfer all the parameters of the same layers between the original VGG16 and our transferred networks and just learn the parameters of those different layers which are adapted by us.

3.4 Class balancing

For there are only two classes in our dataset, hand and background to detect, it is obvious that the percentage of background area in all images is far away larger than 50% and it means that the labels are definitely unbalanced.

It is well known that when training a fully convolutional network, weighting and sampling the loss of the network can be quite helpful in balancing classes. But we choose to drop class balancing for class balancing is unnecessary to our method by actual experiments.

3.5 Implementation

All the hand segmentation models of our method are trained and tested by Caffe on a server with the GPU of a single NVIDIA Tesla K40c, which has a compute capability of 3.5 shown in the official website of NVIDIA.

Our hand segmentation dataset includes 11703 images in total. There are many hand images taken continuously by the same volunteers one time are numbered consecutively, which means that the background, light and some other elements of these images are nearly the same. To guarantee the performance of the training and validation, we disturbed all the images, randomly chose 10700 images for training and used the other 1003 images for validation.

Here we choose to set the number of iterations into 100,000 and do a segmentation tests every 4,000 iterations. We also save a snapshot of our training every 4,000 iterations. To quicken the speed of training, we resize all images and ground truth of our hand dataset into 50×50 pixel. It cost us about 5 hours for training and validation and about 180ms per iteration.

There are many images taken by our mobile phones to test the performance of our model in the dataset. For the size of these images are too large for Caffe to compute, resizing all these images into the one tenth of their original sizes is completed before training. It takes about 1.8s to test an image in size of about 300×400 pixel and save the result as JPG format.

At the same time, we set the normalization of loss from false to true so that it is more convenient and more intuitive to observe the variation tendency of loss during the whole training period. The loss reduces from the original 0.693 to the final 0.098 after training.

4 RESULTS

Metrics: We test our method on our hand segmentation dataset. The dataset consists of 11703 images in different sizes and their corresponding masks.

In this study, we choose four metrics from common semantic segmentation which are able to evaluate and represent the variations on pixel accuracy and region intersection over union (IU).

Let $m_i = \sum_j n_{ij}$ be the number of pixels of class i. Let n_{ij} be the number of pixels which are predicted to belong to class j but actually belong to class i.

We compute:

- Pixel accuracy:

  $$\sum_i n_{ii} / \sum_i m_i$$

- Mean accuracy:

  $$\frac{1}{2}\sum_i n_{ii} / m_i$$

- Mean IU:

  $$\frac{1}{2}\sum_i n_{ii} / (m_i + \sum_j n_{ij} - n_{ii})$$

- Frequency weighted IU:

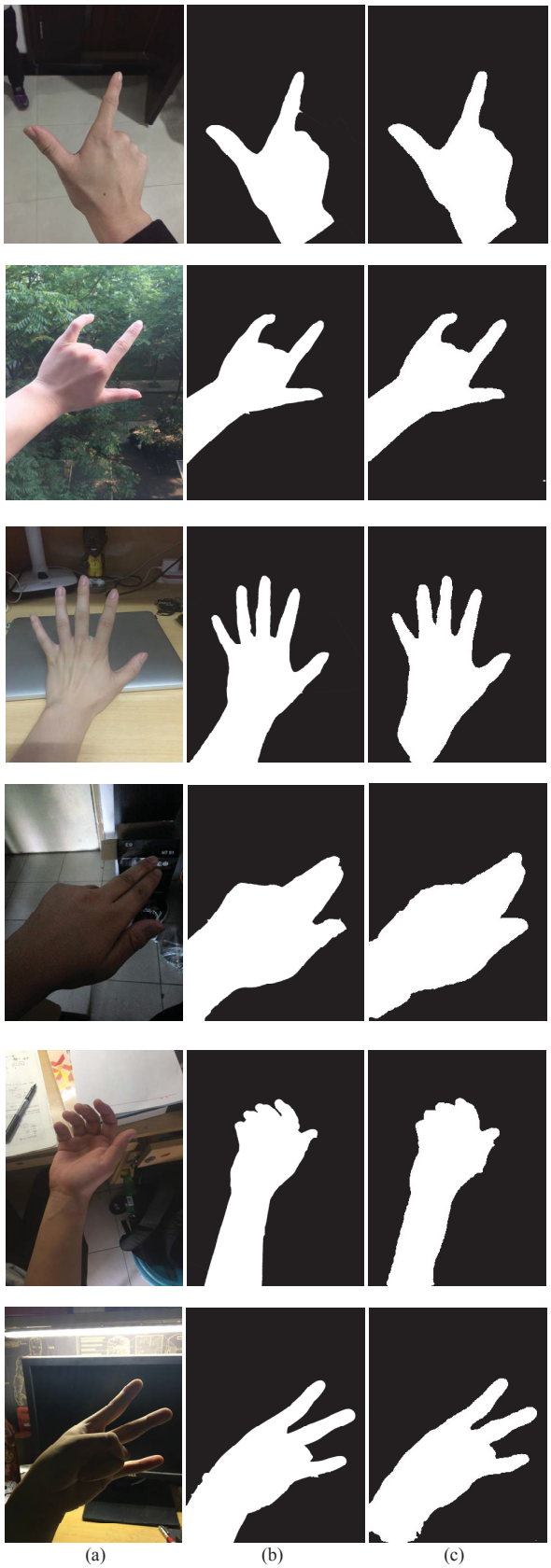
  $$(\sum_k m_k)^{-1}\sum_i m_i n_{ii} / (m_i + \sum_j n_{ij} - n_{ii})$$

Table 1 gives the performance of our method on the hand segmentation dataset. We achieve a 96.1% mean accuracy and a 91.0% mean IU. While the pixel accuracy is 96.1% and the frequency weighted IU is 93.6. And our final loss of the network is 0.098 after normalization.

Table1. The performance of our hand segmentation method.

Metric	Result
pixel acc.	96.6
mean acc.	96.1
mean IU	91.0
f.w. IU	93.6

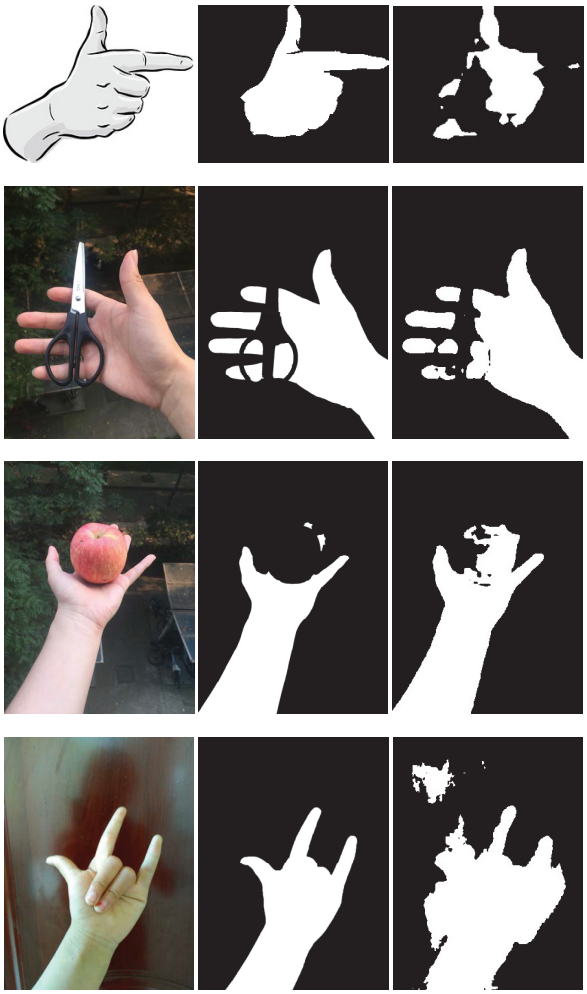
Fig 3 shows the results with good performance of our method. From these images, it is easy to see that our method can reach great performance in different intensity of light and in different kinds of background. It is clear that our results are very close to the ground truth labeled artificially. Meanwhile, there is still room for improvement that the edges of these results are not accurate enough (See Fig 3, Row 4 and Row 5).



(a) (b) (c)

Fig 3. Results of our method. Column (a) shows the input images. Column (b) shows the ground truth. Column (c) shows the results of our method.

And Fig 4 shows the results with bad performance of our method. For some images in which hands are not in the color of skin, such as black and white images, inverse color images and paintings of hands, it is hard for our model to detect hands from the background (See Fig 4, Row 1). When hands have complex interaction with other goods, the results are not that good in those interaction area (See Fig 4, Row 2). And we also find it hard for our model to tell hands from the background in some particular cases. For example, when hands have interaction with an apple or when the background is brown wood, results are unsatisfied for some part of the background or goods are identified as hands (See Fig 4, Row 3 and Row 4). These problems need to be addressed in our future study.

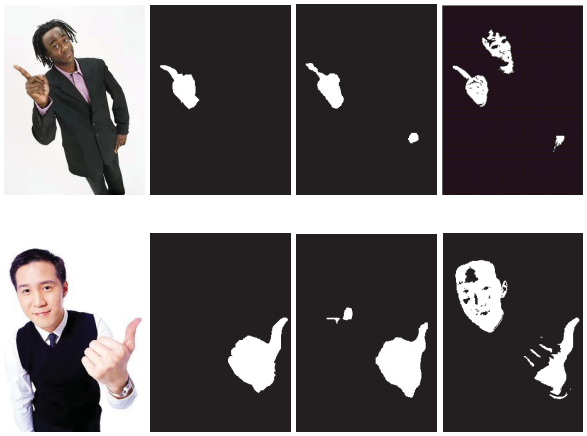


(a) (b) (c)

Fig 4. Results of our method. Column (a) shows the input images. Column (b) shows the ground truth. Column (c) shows the results of our method.

Fig 5 shows the comparation with the results of traditional methods called adaptive skin color model for hand segmentation [13] which based on skin color detection. The adaptive skin color model changes the RGB color space into YCbCr color space to construct a clustered region of skin color for the person. It is clear that this method isn't able to tell faces from hands for they have the same color while our method can achieve more accurate results. It shows that our

model is able to learn the high layer feature of hands through training.



(a) (b) (c) (d)

Fig 5. Column (a) shows the input images. Column (b) shows the ground truth. Column (c) shows the results of our method. Column (d) shows the results of adaptive skin color model.

5 CONCLUSION

Fully convolutional networks are a large set of models, including classification convnets. A large number of research shows that extending these convnets to semantic segmentation and improving its architecture can powerfully improve the state-of-the-art.

In this paper, we adapt a new hand detection architecture from FCN-VGG16. We have achieved good performance on mean IU semantic segmentation metric in hand segmentation. And a large number of experiments on natural hand images can affirm that our method is able to perform well on natural hand images. But there are still a lot of progressive space.

The next step is using the improved version of VGG16 model to fine-tune the FCN-32s net and then use the model it trained to upgrade the FCN-16s net and the FCN-8s net successively. This is assumed to be able to achieve a better performance on hand segmentation. And expend our method to other convolutional networks, such as the AlexNet, GoogLeNet and so on, has been taken into consideration.

And in view of the bad performance mentioned in Section 4, many solutions are taking into consideration. Due to the fact that some kind of background and goods may have similar high layer features with hands, which makes convolutional neural networks hard to tell hands accurately, we are still looking for methods to solve this problem. For those images in which hands are not in the color of human skin color, we can consider to enrich our hand segmentation dataset with black and white images and inverse color images.

6 ACKNOWLEGEMENT

This work is supported by National Nature Science Foundation under Grant 61703203, Natural Science Foundation of Jiangsu Province under Grant BK20170812. We thank all the volunteers for their contributions to taking a large number of hand images and labeling the key points and the masks of these images for our hand segmentation dataset.

REFERENCES

[1] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2015:3431-3440.

[2] Jones M J, Rehg J M. Statistical color models with application to skin detection[M]. Kluwer Academic Publishers, 2002.

[3] Zhu X, Yang J, Waibel A. Segmenting Hands of Arbitrary Color[M]// Segmenting hands of arbitrary color. 2000:446-453.

[4] Kolsch M, Turk M. Robust hand detection[C]// IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings. IEEE, 2004:614-619.

[5] Kurata T, Okuma T, Kourogi M, et al. The Hand Mouse: GMM Hand-Color Classication and Mean Shift Tracking[C]// IEEE ICCV Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems. IEEE Computer Society, 2001:119.

[6] Wu Y, Huang T S. View-independent recognition of hand postures[C]// Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on. IEEE, 2000:88-94 vol.2.

[7] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. Computer Science, 2014.

[8] Cao Z, Simon T, Wei S E, et al. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields[J]. Computer Vision and Pattern Recognition, 2017.

[9] Simon T, Joo H, Matthews I, et al. Hand Keypoint Detection in Single Images using Multiview Bootstrapping[J]. Computer Vision and Pattern Recognition, 2017.

[10] Le T H N, Zhu C, Zheng Y, et al. Robust hand detection in Vehicles[C]// International Conference on Pattern Recognition. IEEE, 2017:573-578.

[11] Das N, Ohn-Bar E, Trivedi M M. On Performance Evaluation of Driver Hand Detection Algorithms: Challenges, Dataset, and Metrics[C]// IEEE, International Conference on Intelligent Transportation Systems. IEEE, 2015:2953-2958.

[12] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. Computer Science, 2014.

[13] Dawod A Y, Abdullah J, Alam M J. Adaptive skin color model for hand segmentation[C]// International Conference on Computer Applications and Industrial Electronics. 2011:486-489.