Fine Segmentation of Head and Torso Using Label Refinement Networks

João Ribeiro Pinto[1,2]
joao.t.pinto@inesctec.pt

Jaime S. Cardoso[1,2]
jaime.cardoso@inesctec.pt

[1]INESC TEC
Porto, Portugal

[2]Faculdade de Engenharia da Universidade do Porto
Porto, Portugal

Abstract

This paper describes the application of an adaptation of the recently proposed Label Refinement Networks (LRN) for fine head and torso segmentation in unconstrained photographs of individuals. The main goal was to obtain segmentations fine and detailed enough to enable high quality replacement of busy backgrounds with plain white environments. The model optionally used VGG-16 pretrained weights, and compared with a U-Net with similar architecture. The results are promising, and successful segmentations are obtained in most cases with LRN-VGG, but further efforts need to be devoted into expanding the training dataset and restructuring the network for larger images, to reach the desired fine segmentations.

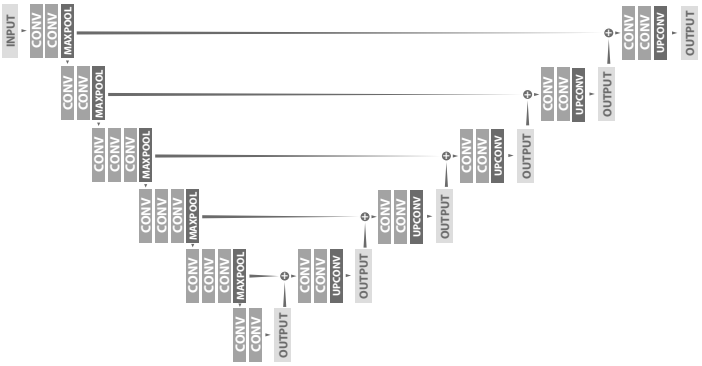

Figure 1: Structure of the implemented Label Refinement Network.

1 Introduction

Semantic segmentation is the process of assigning each pixel of an image a semantic class label. In the case of upper body photograph segmentation, the image pixels are labelled as corresponding to a human or to the background. Detailed segmentation of the head and torso region on upper-body photographs could be very useful for improved face recognition, automatic verification of passport photographs, or artistic photograph manipulation. Several methods have been proposed for this [1, 5, 8], but none present the level of contour fitness that would be required for the target application.

Recently, Islam *et al.* [3, 4] have proposed a network for fine semantic segmentation called Label Refinement Network (LRN). LRN follows the encoder-decoder structure of the U-Net [6], but outputs segmentation results at several resolution levels. This means the model is forced to offer early coarse segmentations that are gradually improved up to the original image resolution. Besides offering superior results in fine segmentation, it is prepared to use VGG-16 [7] weights and has much fewer trainable parameters than other top-performing segmentation models.

This work proposes an adaptation of the Label Refinement Network which is applied to semantic segmentation of upper-body segmentation. The model was trained and evaluated with annotated images from the Labelled Faces in the Wild dataset [2], and compared with a U-Net model with similar architecture, to study the performance improvements brought by the multi-resolution supervision strategy.

2 Methodology

2.1 Label Refinement Network (LRN)

The Label Refinement Network proposed by Islam *et al.* [3, 4] follows an encoder-decoder structure, as illustrated in Fig. 1. This is common in semantic image segmentation networks such as the U-Net, used in this work as reference.

The encoder part of the network mimics the VGG-16 network [7], to allow for easy transfer of weights. It includes thirteen trainable convolutional layers over five levels delimited by pooling layers. The first two levels include two convolutional layers each, with 64 and 128 filters, respectively. The three last levels include three convolutional layers each, with 256, 512, and 512 filters, respectively. Every convolutional layer has filter size of 3×3, padding, and ReLU activation. The max-pooling layers have size 2×2.

The first refinement unit, after the last encoder level, receives the feature maps on a convolutional layer with 64 filters and ReLU activation, followed by one with 1 filter (number of output channels, in this case, is one) with sigmoid activation, thus returning the first coarse label output. The remaining refinement units receive the feature maps from the corresponding encoder levels, through the so-called skip connections, that are concatenated with the previous coarse label output. This will undergo processing over two convolutional layers (identical to those described for the first refinement unit) and an upsampling operation to return the respective coarse labellings.

In the original LRN, the feature maps received by the refinement units are processed by a convolutional layer, with number of filters equal to the number of output channels, before being concatenated with the coarse labellings. However, in this case, since we have only one output channel, this meant the loss of a great amount of information at the start of the refinement units, and the network was unable to learn properly. Hence, we propose the elimination of this convolutional layer in the LRN network, at least when the number of output channels is small.

2.2 Reference: U-Net

A U-Net [6] was used as reference, to adequately assess the performance benefits in fine segmentation of the multi-resolution supervision strategy in LRN. The U-Net is very similar, in its architecture, to the LRN. Its encoder half is completely identical. The decoder part receives feature maps from the corresponding encoder levels through skip connections and performs upsampling and convolution operations over the encoders outputs. The network outputs a single labelling map, of shape $224 \times 224 \times 1$.

2.3 Training

The LRN and U-Net models were trained in two different settings. In the first setting, the parameters of the models were initialised using the Glorot uniform random initialisation method, and trained using the training set prepared for this work. The second setting used VGG-16 pretrained weights. In this case, the encoder half of the networks received pretrained parameters from the VGG-16 model trained on ImageNet, and was frozen. The decoder part was trained until convergence and, after this was achieved, the encoder was unfrozen. Then, training was resumed to improve performance through small encoder parameter adjustments until a new convergence loss plateau was reached.

Training was performed using the Adam optimiser, with initial learning rate of 0.0001 for both models, during sufficient epochs to achieve convergence. The loss used was the binary crossentropy. In the case of LRN, the binary crossentropy loss is separately computed for each supervision level, and the optimisation loss is the sum of all individual losses. Horizontal flipping, zooming, and rotation were used for data augmentation.

Figure 2: Examples of results offered by the implemented algorithms.

3 Experimental Settings

The models were trained and evaluated with unconstrained people images from the Labelled Faces in the Wild (LFW) dataset [2]. These were reshaped to 224×224 to match the original VGG-16 input shape, and annotated for semantic head-torso segmentation using the online tool remove.bg[1]. The training set contained a total of 250 images, while the test set was composed of 48 images.

The results were assessed resorting to the Jaccard index and Hausdorff distance. The Jaccard index measures the rate of overlap between the pixel sets of the ground truth L and the prediction P, following the expression:

$$J(L,P) = \frac{L \cap P}{L \cup P}. \quad (1)$$

The Hausdorff distance measures the maximum among the minimum pairwise distances between the true L_c and predicted P_c contours. It follows the expression:

$$d_H(L_c,P_c) = \max\{\sup_{l \in L_c} \inf_{p in P_c} d(l,p), \sup_{p \in P_c} \inf_{l in L_c} d(l,p)\}. \quad (2)$$

Besides Jaccard index and Hausdorff distance, the results were also visually analysed to evaluate and compare the performance of the implemented methods for fine semantic segmentation of head and torso.

4 Results and Discussion

The average Jaccard index and Hausdorff distance results for the images in the test set are presented in Table 1. By either measure, the LRN-VGG model presented the best results, followed by the LRN model. The Jaccard index results are high, but so are the Hausdorff distance results, which may denote the segmentation is mostly correct but the fine contour fitness fails often. Furthermore, the high standard deviation of the results of all methods is worrying, as it shows the segmentation quality is very dependent on the image, and the models may not be as robust as desired.

Some examples of the results in test images are presented in Fig. 2. Despite the high Hausdorff distance results discussed before, the visual analysis reveals the results are, in general, acceptable, especially for the LRN and LRN-VGG models. These two models present very similar results, although LRN-VGG seems slightly better at finding true contours and in difficult parts of the photographs. Between U-Net and U-Net-VGG, the U-Net model offered better results, but both frequently leave out patches of the head-torso regions, unlike LRN and LRN-VGG, that tends towards the false positives.

[1]Remove.bg [Online]. Available on: `https://www.remove.bg/` (visited on June 12th, 2019).

Table 1: Jaccard index and Hausdorff distance results on the test set images for the implemented algorithms.

Method	Jaccard index (%)	Hausdorff distance
LRN	91.67 ± 5.63	32.84 ± 23.05
LRN-VGG	92.73 ± 5.24	26.28 ± 20.07
UNET	89.25 ± 7.19	35.18 ± 19.63
UNET-VGG	87.98 ± 7.53	44.00 ± 16.75

In the offered examples (Fig. 2), we can notice that in the first two rows, LRN and LRN-VGG offered great results, but U-Net and U-Net-VGG failed often in the suit regions. In the fourth row, the difficulty of elaborate hairstyles and dark backgrounds is illustrated, as all models were unsuccessful. In the third and fifth rows, the backgrounds seem to be too busy for any model to perform correctly. Despite the failures in pictures with especially busy backgrounds, the LRN-VGG model was mostly successful in offering fine segmentations of head and torso.

5 Conclusion

In this work, the recently proposed Label Refinement Network was explored for fine segmentation of head-torso in unconstrained images. Trained with 250 images and tested with 48 images of the Labelled Faces in the Wild dataset, the LRN model offered promising results, especially when starting from VGG-16 weights.

In the future, the performance of the model could be improved using a larger training set, and new forms of data augmentation, such as dynamic backgrounds in training images. The network should also be restructured to receive larger images, and thus allow for finer segmentation results. Finally, the introduction of depth map estimation could improve the segmentation of people from the background.

Acknowledgements

This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia within project: "UID/EEA/50014/2019" and within the PhD grant with reference "SFRH/BD/137720/2018".

References

[1] S. Duffner and J.-M. Odobez. Leveraging colour segmentation for upper-body detection. *Pattern Recognition*, 47(6):2222–2230, 2014. doi: https://doi.org/10.1016/j.patcog.2013.12.014.

[2] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts Amherst, 2007.

[3] M. A. Islam, S. Naha, M. Rochan, N. Bruce, and Y. Wang. Label refinement network for coarse-to-fine semantic segmentation. *arXiv*, 2017. 1703.00551.

[4] M. A. Islam, M. Rochan, N. Bruce, and Y. Wang. Gated feedback refinement network for dense image labeling. In *CVPR '17*, 2017.

[5] H. Lu, G. Fang, X. Shao, and X. Li. Segmenting human from photo images based on a coarse-to-fine scheme. *IEEE Transactions on Systems, Man, and Cybernetics*, 42(3):889–899, 2012.

[6] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. *MICCAI 2015*, pages 234–241, 2015.

[7] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv*, 2014. 1409.1556.

[8] R. Tong, D. Xie, and M. Tang. Upper body human detection and segmentation in low contrast video. *IEEE Transactions on Circuits and Systems for Video Technology*, 23(9):1502–1509, 2013.