Lips landmark detection using CNN

Miroslav Hlaváč[1]

1 Introduction

Current research in the field of Computer Vision mainly focuses on the utilization of Deep Neural Networks (DNN) as a powerful tool for image processing, classification, tracking and regression. There was a boom in neural networks frameworks in recent years and many research groups are now using them for tasks of computer vision, for example Wu et al. (2016), Sun et al. (2013). I present a Convolutional Neural Network (CNN) for lips landmarks detection in this paper.

2 Convolutional Neural Network

I have chosen the Caffe framework for the purpose of this experiment, because it is implemented on the MetaCentrum computing grid with GPU support. The CNN network topology usually includes convolutional layers with RELU activation function, maxpooling layers and fully connected layers at the end. I have used the topology shown in the figure 1.

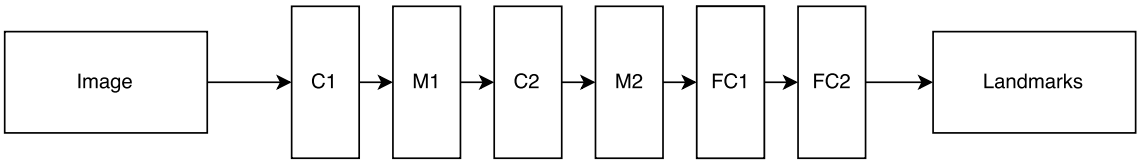

Figure 1: CNN topology

The network consists of two convolutional layers with a RELU activation function, each followed by a maxpooling layer. C1 has kernel size of 3x3 with stride 1 and 32 outputs. C2 has kernel size of 16x16 with stride 1 and 16 outputs. Both maxpooling layers M have kernel size of 2x2 with stride 2. Seventy two landmarks coordinates (x,y) are a direct output from the last fully connected layer.

3 Training data

I have used over 13000 training images of one person under different light conditions with 36 annotated landmarks and 50x76 resolution. The annotations were made by Active Appearance Model (AAM) as in Cootes and Taylor (2001). The shapes with the best fit from the AAM were chosen as training data. Images from the training data set are shown in the figure 2.

[1] student of the PhD program Applied Sciences and Informatics, Department of Cybernetics, Computer Vision specialization, e-mail: mhlavac@kky.zcu.cz

Figure 2: Training data

4 Results

Training of the network was done over 10 000 iterations with a batch size of 32 after which the error converged to 0.74 pixels per point on training set and 0.97 pixels per point on previously unseen data. The comparison between ground truth shape and the output of the network can be seen in the figure 3.

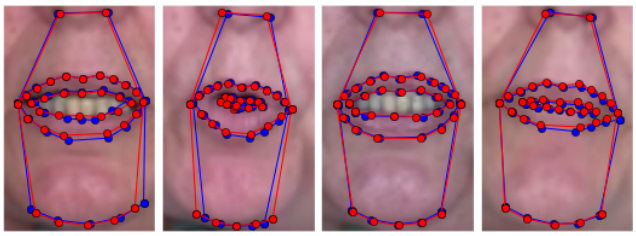

Figure 3: CNN output vs. ground truth comparison.

Blue labels are ground truth. Red labels are output from the network

5 Conclusion

The Convolutional Neural Network proposed in this paper was able to achieve a sub-pixel accuracy in landmarks detection. Most of the errors came from the area of the chin because there are no robust features which would lock the exact position of the landmarks. The network even exceeded the ground truth annotation in some of the pictures.

Acknowledgement

This paper was sponsored by grant project SVK1-2016-023. Access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum, provided under the programme "Projects of Large Research, Development, and Innovations Infrastructures" (CESNET LM2015042), is greatly appreciated.

References

Wu, Y., Hassner, T., Kim, K., Medioni, G., Natarajan, P., 2016. Facial Landmark Detection with Tweaked Convolutional Neural Networks. *arXiv 2016*.

Sun, Y., Wang, X., Tang, X., 2013. Deep Convolutional Network Cascade for Facial Point Detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2013.*.

Cootes, T.F., Taylor, C.J, 2001. Statistical models of appearance for medical image analysis and computer vision. *Proc. SPIE Medical Imaging 2001*.