Face Tracking and Head Pose Estimation using Convolutional Neural Networks

Stylianos Asteriadis[*] KostasKarpouzis[†] Stefanos Kollias[‡]
Image, Video, Multimedia Lab, National Technical University of Athens, Greece

1 Introduction

In applications where face orientation is necessary, but in unpretending environments in terms of lighting, equipment, resolution, etc, employing local tracking techniques would usually fail to give accurate results, regarding the issue of head pose estimation. However, in a similar manner, holistic techniques require the face to be well aligned with the training data. This pre-assumes correct and accurate face tracking, which is also a challenging issue. Here, we propose a face tracker, adjusted to each person's face chrominance values, and learnt online. Based on the face bounding box, Convolutional Neural Networks (CNNs) are employed, in order to calculate face orientation. CNNs are ideal for cases where a lot of distortions exist in the data, and the proposed architecture only utilizes subsets of classifiers, excluding those corresponding to rotation angles far from the current.

2 Face Tracking

Face is initially detected using the detector of Viola-Jones [Viola and Jones 2001]. Subsequently, a sample area C_{skin} of face is used for each person; we used the saturation values of this area, and the saturation values of face pixels C_{fp} in subsequent frames are expected to be within certain limits (defined by threshold T)with regards to the mean saturation value s_M of C_{skin}. The binary image of these pixels C_{fp} is formed, followed by binary opening. The threshold T is automatically selected for each user, at the detection step, according to the hypothesis below: it is expected that, at the first frame, the amount of pixels with saturation values close to the mean of C_{skin} is close to the amount of pixels that account for the real face region. Thus, the threshold that gives amount of pixels closer to the expected face size, is the one kept.

To reduce the number of candidate facial pixels, the rules defined in [Jure Kovac and Solina 2003] are used, in order for skin clusters in RGB colorspace to be built and the resulting binary map C_{fp} is combined using logical AND with C_{sp}. Finally, the proposed method uses connected component labelling and chooses the largest component as the final face region.

3 Convolutional Neural Networks

In a generic image recognition problem, an image is used as input in the first layer of a CNN [LeCun et al. 1990] and it is convolved with a series of filters in the second layer ($C1$), resulting in feature maps. In the third layer ($S2$), the feature maps are subsampled. Further such levels may follow. This architecture guarantees limited number of free parameters, automatic extraction of complex and spatial features and robustness to distortions. The architecture employed in the proposed scheme, is a 6-layer convolutional neural network, with the layers being in the order $C1$ (6 maps), $S2$, $C3$ (6 maps), $C4$ (80 maps), $F5$ (10 neurons), $F6$ (output - 2 neurons), which is a 2-element vector. The convolutional layers use 7 × 7 kernels, while the sub-sampling layer uses a downscale factor of 2. Inputs are 32 × 32 normalized face imagettes.

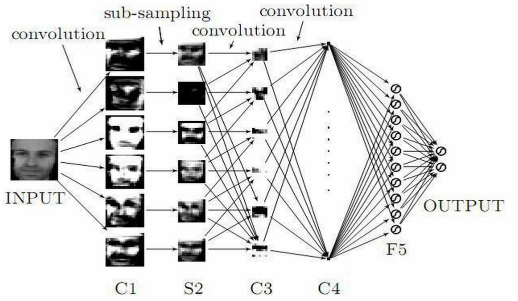

Figure 1: *Employed CNN architecture.*

3.1 Training Procedure and Results

For training, we created a pose space consisting of classes centered at pitch angles $\{-60^o, 0, 60^o\}$ and yaw angles $\{-90^o, -45^o, 0, 45^o, 90^o\}$. We trained one CNN for each combination of neighboring classes, resulting in a total of 38 classifiers. Using *yaw* and *pitch* information from the previous frame, all n networks considered include the class C_c whose center is closer to the the *yaw* and *pitch* values of the previous frame. In this way, only a subset of CNNs are used at each frame, constituting the system faster and more reliable, as the possibility of erroneous classification is reduced. The final estimate of the yaw (or pitch) angle is done by employing regression models including elements of the differences of outputs, as well as the centre of class C_c.

The method has been tested on the Boston University Face dataset: and errors regarding *yaw* and *pitch* were $5.6°$ and $4.7°$, respectively.

Acknowledgements

This work was funded by ICT-Project 'Siren', (under contract FP7-ICT-258453)

References

JURE KOVAC, P. P., AND SOLINA, F. 2003. Human skin colour clustering for face detection. In *IEEE International Conference on Computer as a Tool*, vol. 2.

LECUN, Y., BOSER, B., DENKER, J. S., HENDERSON, D., HOWARD, R. E., HUBBARD, W., AND JACKEL, L. D. 1990. Handwritten digit recognition with a back-propagation network. In *Advances in Neural Information Processing Systems*, Morgan Kaufmann, 396–404.

VIOLA, P. A., AND JONES, M. J. 2001. Rapid object detection using a boosted cascade of simple features. In *Conference on Computer Vision and Pattern Recognition*, vol. 1, 511–518.

[*] e-mail: stiast@image.ntua.gr
[†] e-mail: kkarpou@image.ntua.gr
[‡] e-mail: stefanos@cs.ntua.gr