Hough Networks for Head Pose Estimation and Facial Feature Localization

Gernot Riegler
riegler@icg.tugraz.at
David Ferstl
ferstl@icg.tugraz.at
Matthias Rüther
ruether@icg.tugraz.at
Horst Bischof
bischof@icg.tugraz.at

Institute for Computer Graphics and Vision
Graz University of Technology
Austria

Head pose estimation and facial feature localization are keys to advanced human computer interaction systems and human behavior analysis. Due to their relevance, both tasks have gained a lot of attention in the computer vision community. Recent state-of-the-art methods like [1, 2, 3, 6] report impressive results and are real-time capable. However, those approaches rely on hand-crafted features. In contrast, we try to learn a feature representation from a set of training images. This is done by utilizing Convolutional Neural Networks (CNNs), which have shown to achieve outstanding results on various tasks such as image classification [5].

Instead of segmenting the head in a first step and then regressing the task-dependent parameters, we show in our paper a patch-based approach. Patches are densely extracted from the image along a regular grid and for each patch we perform a joint classification and regression. The classification segments the image patches into foreground and background, whereas the regression casts votes in a Hough space, but only for foreground patches. This is similar to the idea of Hough Forests (HFs) [4]. However, we replace the Random Forest (RF) with a CNN and call it therefore *Hough Network (HN)*.

Assuming that we have a training dataset $\{(x_s, \mathbf{t}_s)\}_{s=1}^{S}$ with S samples, where x_s denotes an image patch, and $\mathbf{t}_s$ encodes the foreground-background information as well as the regression targets, we want to train a CNN that minimizes the following error function

$$E_s(\theta) = \lambda_c E_{s,c} + \lambda_r E_{s,r}, \tag{1}$$

where $E_{s,c}$ and $E_{s,r}$ are the classification and regression error, respectively. The parameters λ_c and λ_r are weighting coefficients of the individual error functions and relate to increased or decreased delta values in the back-propagation algorithm. For classification, we utilize the cross-entropy error that is defined as follows

$$E_{s,c}(\theta) = -\left(t_{s,c} \ln\left(y_{s,c}\right) + \left(1 - t_{s,c}\right) \ln\left(1 - y_{s,c}\right)\right). \tag{2}$$

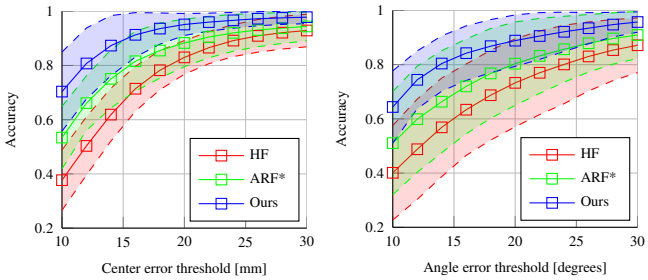
In contrast, for the regression targets we use the L_2 loss that minimizes the Euclidean distance between the target and predicted values:

$$E_{s,r}(\theta) = \frac{1}{2} \|\mathbf{y}_{s,r} - \mathbf{t}_{s,r}\|^2. \tag{3}$$

The objective function in Equation 1 allows that values in the single target vectors can be missing. In such cases we set the gradient values of the involved weights (which only effects connection to the output layer) to zero. We especially utilize this fact, if a patch does not belong to the foreground. In the case of a background patch, we back-propagate only the error values of the class information.

The straight-forward inference process in our HNs would be to densely extract overlapping patches from the image and evaluate the CNN for each patch independently. However, the structure of CNNs allows a more efficient method. We present the whole image as input to the CNN and if the patch stride (distance between two neighboring patch centers) is a multiply of the sum of the pooling widths, then the patches can be separated in the convolution and pooling layers. Only before the fully-connected layers we have to reshape the data to a matrix, where each patch corresponds to a single column. This allows us to perform classification and regression for all patches of an image in a single CNN evaluation.

We evaluated HNs on two challenging computer vision tasks. The first task deals with head pose estimation from consumer depth cameras. Given a depth image, we want to estimate the head center in 3D and its pose in Euler angles. We randomly split the sequences of the Biwi Kinect Headpose Database [3] into a train and a test set. A patch votes for a head center and a pose, if its foreground probability is > 0.99. Using a mean-shift variant [3], we find a single mode in the votes.



(a) (b)

Figure 1: Accuracy for the head center estimation error (a) and the angle error (b) of the HF [3], ARF* [6] and our approach. The curves visualize the fraction of correct estimates over an increasing success threshold. The solid lines represent the mean over five splits, whereas the shaded areas visualize the standard deviation.

Method	F. F. Error.	H. P. Error.
Conditional Random Forests (CRF) [2]	12.0	27.85
Robust Cascaded Pose Regression (RCPR) [1]	5.3	-
Ours	6.3	21.83
Human	4.5	-

Table 1: Performance of HNs compared to CRFs [2], RCPRs [1] and human performance on the LFW dataset as percentage of the inter-ocular distance.

The same approach can be utilized for facial feature localization. We evaluated our approach on the Labeled Faces in the Wild (LFW) dataset [2, 7], which also provides a discrete head pose. This information is incorporated into our HN by extending the error function:

$$E_s(\theta) = \lambda_c E_{s,c} + \lambda_r E_{s,r} - \lambda_h \sum_{i=1}^{5} \mathbf{t}_{s,h}^{(i)} \ln \mathbf{y}_{s,h}^{(i)}. \tag{4}$$

Further details and results can be found in our paper. Our conclusion is that HNs provide a powerful alternative to HFs, because it can learn a rich feature representation. Further, HNs could be adapted to various other tasks, such as human pose estimation and object detection.

[1] Xavier P. Burgos-Artizzu, Pietro Perona, and Piotr Dollár. Robust face landmark estimation under occlusion. In *International Conference on Computer Vision*, pages 1513–1520, 2013.

[2] Matthias Dantone, Juergen Gall, Gabriele Fanelli, and Luc J. Van Gool. Real-time facial feature detection using conditional regression forests. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2578–2585, 2012.

[3] Gabriele Fanelli, Matthias Dantone, Juergen Gall, Andrea Fossati, and Luc J. Van Gool. Random forests for real time 3d face analysis. *International Journal of Computer Vision*, 101(3):437–458, 2013.

[4] Juergen Gall, Angela Yao, Nima Razavi, Luc J. Van Gool, and Victor S. Lempitsky. Hough forests for object detection, tracking, and action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33 (11):2188–2202, 2011.

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1106–1114, 2012.

[6] Samuel Schulter, Christian Leistner, Paul Wohlhart, Peter M. Roth, and Horst Bischof. Alternating regression forests for object detection and pose estimation. In *International Conference on Computer Vision*, pages 417–424, 2013.

[7] Hai Wang, Bong-Nam Kang, and Daijin Kim. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, University of Massachusetts, Amherst, 2007.