Machine Learning and Computer Vision - Object Detection with a Part-Based Model

Olivier Jais-Nielsen

Belief propagation

If $p \in P$ represents a part other than the nose n, the message sent from p to the nose is:

$$\log\left(m_{p\to n}(x,y)\right) = \max_{u,v}\left(\log\left(\Phi_p(u,v)\right) - \frac{\left(u-x-\mu_p^x\right)^2}{2\sigma_p^{x\,2}} - \frac{\left(v-y-\mu_p^y\right)^2}{2\sigma_p^{y\,2}}\right)$$

The belief at the nose is:

$$B_n(x,y) = \log\left(\Phi_n(x,y)\right) + \sum_{p\in P}\log\left(m_{p\to n}(x,y)\right)$$

The message received from the nose is:

$$\log\left(m_{n\to p}(x,y)\right) = \max_{u,v}\left(\log\left(\Phi_n(u,v)\right) - \frac{\left(u-x-\mu_p^x\right)^2}{2\sigma_p^{x\,2}} - \frac{\left(v-y-\mu_p^y\right)^2}{2\sigma_p^{y\,2}} + \sum_{p'\in P-\{p\}}\log\left(m_{p'\to n}(u,v)\right)\right)$$

And the belief at a part is:

$$B_p(x,y) = \log\left(\Phi_n(x,y)\right) + \log\left(m_{n\to p}(x,y)\right)$$

The file “get_configurationDP.m” has been modified to compute all the Beliefs and the messages received by the parts from the root, and to plot them. The file “get_configurationFL.m” implements the exact same functionality with no dynamic programming for the maximization.

Here are two outputs on the same input data (for computational time issues, the test image was reduced and therefore was not well detected by the algorithm, hence the lack of obvious meaning in the different messages).

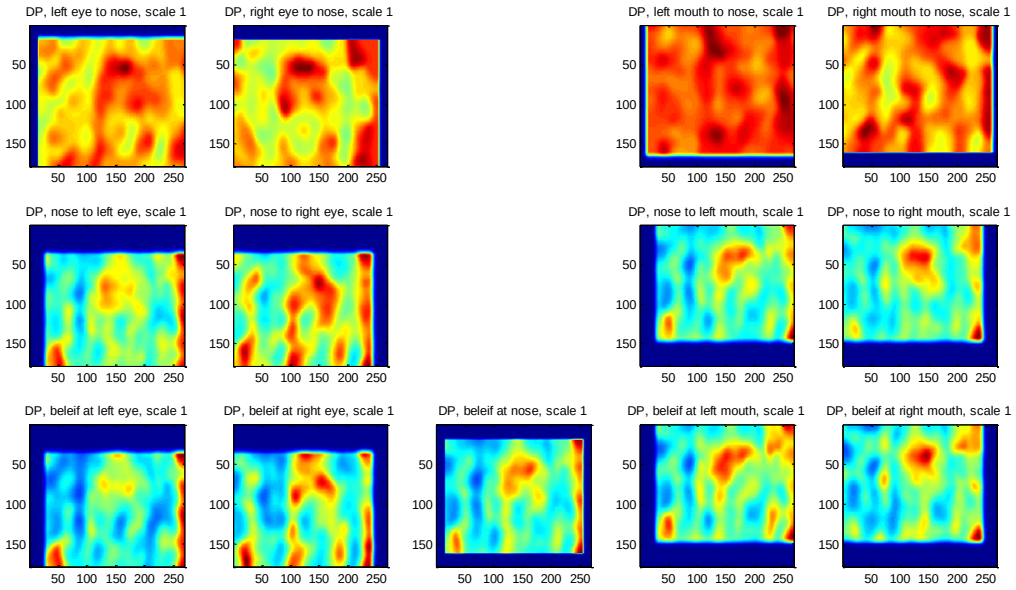

Figure 1 – Messages computed using dynamic programming optimization

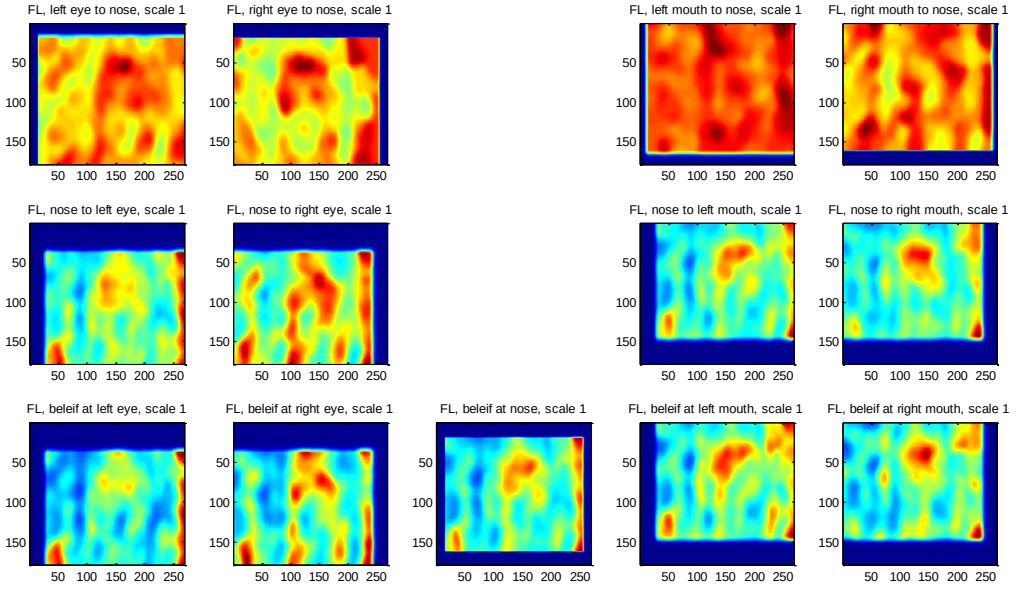

Figure 2 – Messages computed using direct Matlab implementation

Influence of the training set size

Here are the results of the detection using SIFT features after training the algorithm on image sets of different sizes.

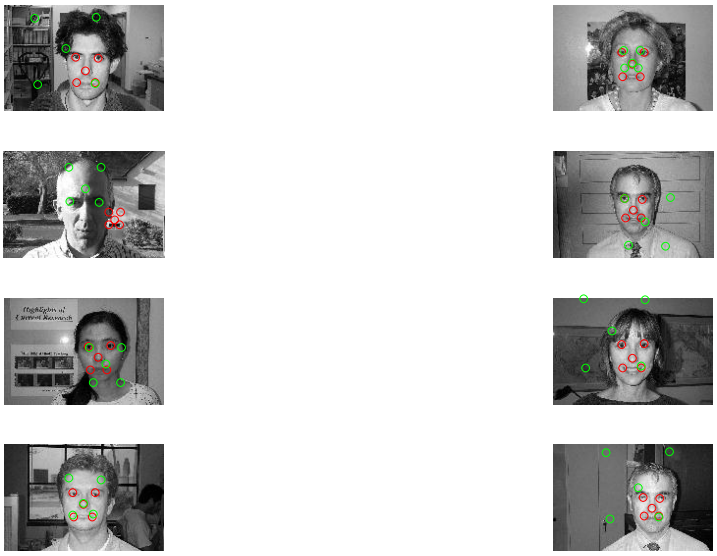

Figure 3 – SIFT based detection after training on 20 images

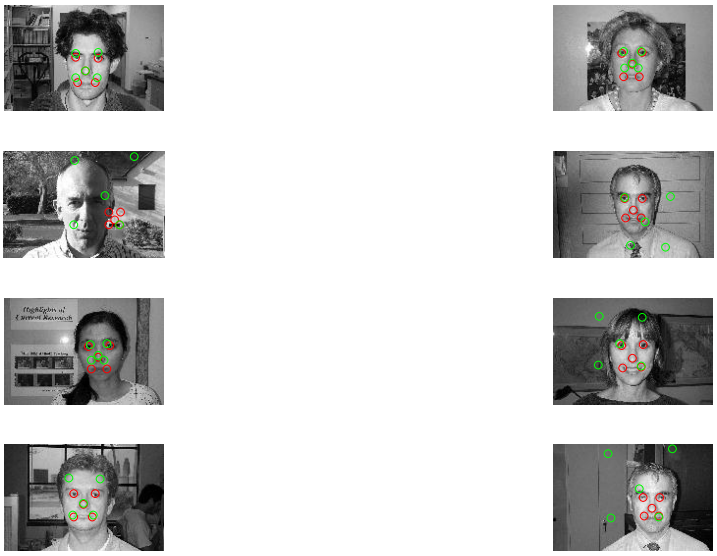

Figure 4 – SIFT based detection after training on 50 images

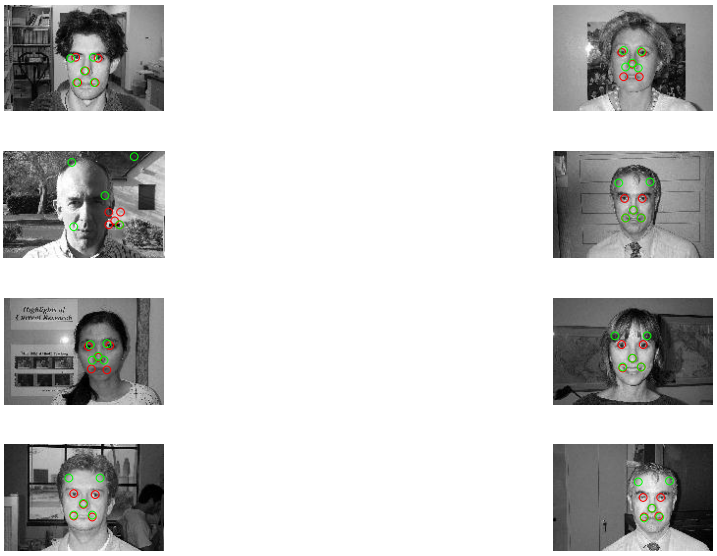

Figure 5 – SIFT based detection after training on 100 images

As expected, the images that presented very stable and successful results (2nd and 7th images) for a small training set remain that way: the best result is near-perfect and the second best is realistic.

The majority of the images had a near perfect best result and a second best result not successful (1st, 4th, 5th, 6th and 8th images). The second best result becomes realistic when the training set grows.

Finally, images with an identifiable difficulty (the lighting in the 3rd image) are not detected whatever the size of the training set is, most likely as long as no image with similar pattern are present in the training set.

Influence of the features

Here are the results of the detection using Filterbank features after training the algorithm on image sets of different sizes.

Figure 6 – Filterbank based detection after training on 20 images

Figure 7 – Filterbank based detection after training on 100 images

It is striking to see that in most cases, no matter the training set size, the filterbank based detector performs much worse than the SIFT based detector. One might explain it by the very nature of SIFT features: the key points are extrema of a difference of Gaussians pyramid, which can easily catch blobs such as the eyes. The descriptors are based on gradient histograms, which can discriminate corner areas such as the mouth extremities. Finally, it seems that SIFT global robustness is an advantage here, even

though the SIFT transform might seem more adapted for the recognition of specific object (i. e. and not classes of objects).

One exception to the previous observations, and not the least, concerns the third image which was not successfully detected with the SIFT based method: it has a rather acceptable best result when trained with 100 images.