A principled approach to remove false alarms by modelling the context of a face detector

Cosmin Atanasoaei[1][2]
cosmin.atanasoaei@idiap.ch
Christopher McCool[2]
christopher.mccool@idiap.ch
Sébastien Marcel[2]
marcel@idiap.ch

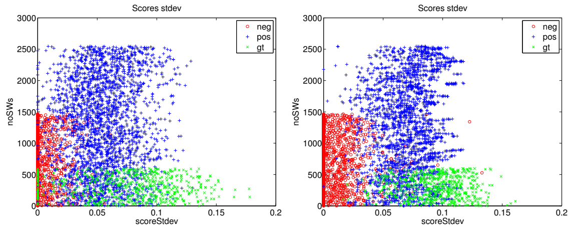
[1] École Polytechnique Fédérale de Lausanne (EPFL)
Lausanne, CH
[2] Idiap Research Institute
Martigny, CH

Face detection [1, 6] is the task of classifying a sub-window as being a face or not. There are many ways to obtain sub-windows from an image, with the sliding window approach being the most well known. This can result in multiple detections and false alarms. A merging and pruning heuristic algorithm is then typically used to output the final detections [3]. Recent work has been done to overcome the limitations of the sliding window approach by using a branch-and-bound technique to evaluate all possible sub-windows in an efficient way [2]. A different approach was recently proposed in [4] and [5] where they show that the score distribution is significantly different around a true object location than around a false alarm location.

We propose a model to enhance a given face classifier, by discriminating false detections (sub-windows) from true detections using the contextual information. Our approach follows the work of [4, 5], but we propose a more discriminative approach and we extract a larger variety of features. We investigate the detection distribution around some sub-window (which we call the *context*) from which we compute features from every possible axis combination (location and scale). The main advantages of our method is that it can be initialized with any sub-window collection and it poses no restriction regarding the object classifier to run on top of.

To build the context of a target sub-window $T_{sw} = (x, y, s)$, we *sample* in the 3D space of location (x, y) and scale (s) to collect detections. Then the *context of* T_{sw} consists of collection of 4D points $C(T_{sw}) = \{(x_i, y_i, s_i, ms_i)_{i=1,..}\}$, where ms is the classifier score. We propose two strategies for context sampling: full and axis. The *full* strategy consists of sampling by varying the location and scale at the same time, while the *axis* strategy the sampling is done just along one axis at a time.

The *feature vectors* are defined by their attribute and the axis combination (x, y and s) used to obtain the attribute. We use 5 attributes that capture the *global information* (counts), the *geometry of the detection distribution* (hits) and the *detection confidence* (score) obtained from the face classifier.



(a) (b)

Figure 1: Cloud of points of score standard deviation for 3 axis (x, y, scale) using axis sampling (a) and full sampling (b).

It is possible then to use these features to separate the two classes of context (Fig. 1). We found experimentally that: significantly more detections are generated around a true detection than around a false alarm and false alarms present contexts with smaller score variation and less spread along each axis.

Next we build a logistic linear model for each feature:

$$M(x, w) = \frac{1}{1 + exp\left(-w_0 + \sum_{i=1}^{d} x_i w_i\right)}. \tag{1}$$

Training the model is done by minimizing the negative of the likelihood of the model output being generated from the input data with additional L_1 and L_2 norm regularization terms. Then the function to optimize becomes:

$$E(w, \lambda_1, \lambda_2) = \frac{\sum l(w, x^+)}{N^+} + \beta \frac{\sum l(w, x^-)}{N^-} + \lambda_1 \underbrace{\sum_{i=1}^{n} |w_i|}_{L_1} + \lambda_2 \underbrace{\sum_{i=1}^{n} |w_i|^2}_{L_2}, \tag{2}$$

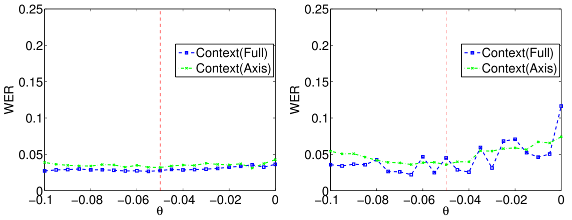
where $l(w, x) = -y\ log(M(x, w)) - (1 - y)\ log(1 - M(x, w))$ is the negative log likelihood of the sample x using the model weights w and having the label y. The log likelihoods are averaged separately over the N^+ positive samples and the N^- negative samples respectively because of the unbalanced nature of the training samples.

In the case of object detection (in particular face detection) it is preferred to have higher false alarms than to miss objects. This implies that $\beta < 1$. The λ_1 and λ_2 and τ (threshold) are chosen to minimize the Weighted Error Rate (WER) defined as:

$$WER(\beta, \tau) = \frac{\beta \times FAR + FRR}{\beta + 1}, \tag{3}$$

where FAR is the False Acceptance Rate and FRR is the False Rejection Rate.

Next we combine the feature classifiers using the same logistic linear model. The inputs to the combined classifier are the normalized outputs of the context feature classifiers $x_{k,l} = M_{k,l}(x, w) - \tau_{k,l}$, where k indicates the attribute type $(k = 1..n, n = 5)$ and l corresponds to the axis combination $(l = 1..m, m = 7)$ and $\tau_{k,l}$ is the optimum threshold value of the $M_{k,l}$ model.



(a) XM2VTS-WER (b) MIT+CMU-WER

Figure 2: Context-based model's weighted error rate (WER) for the test sets of XM2VTS (a) and MIT+CMU (b) databases, with blue for full sampling and with green for axis sampling.

Experimental results have shown that the context-based classifier distinguishes reasonable well false alarms and true detections (see Fig. 2) for multiple threshold values of the MCT face classifier [1]. This affects face detection by reducing exponentially the number of false alarms with a relative small drop in TAR.

[1] B. Froba and A. Ernst. Face detection with the modified census transform. *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 91–96, 2004.

[2] Christoph H. Lampert, Matthew B. Blaschko, and Thomas Hofmann. Beyond sliding windows: Object localization by efficient subwindow search. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008.

[3] Yann Rodriguez. *Face Detection and Verification using Local Binary Patterns*. PhD thesis, Ecole Polytechnique Fédérale de Lausanne, 2006.

[4] Hiromasa Takatsuka, Masayuki Tanaka, and Masatoshi Okutomi. Spatial merging for face detection. In *SICE-ICASE International Joint Conference*, pages 5587–5592, 2006.

[5] Hiromasa Takatsuka, Masayuki Tanaka, and Masatoshi Okutomi. Distribution-based face detection using calibrated boosted cascade classifier. In *ICIAP '07: Proceedings of the 14th International Conference on Image Analysis and Processing*, pages 351–356, 2007.

[6] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Comput. Surv.*, 35(4):399–458, 2003. ISSN 0360-0300.