Context-based additive logistic model for facial keypoint localization

Peng Li
http://www.cs.ucl.ac.uk/staff/P.Li/

Jonathan Warrell
http://www.cs.ucl.ac.uk/people/j.warrell/

Jania Aghajanian
http://www.cs.ucl.ac.uk/staff/j.aghajanian/

Simon J. D. Prince
http://www.cs.ucl.ac.uk/staff/s.prince/

Department of Computer Science,
University College London,
Gower Street, London, United Kingdom

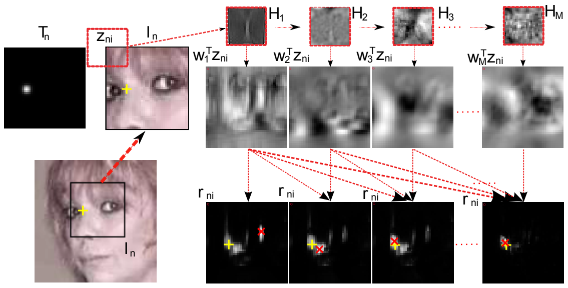

Figure 1: Training image I_n is a region from the original image (bottom left). T_n is the desired output for a keypoint (nose bridge) associated with I_n. I_n is correlated with a set of learned filters $\mathbf{H}_1, \mathbf{H}_2, \cdots, \mathbf{H}_M$ (top row). The concatenation of the filter elements forms $\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_M$ respectively. When applied to the image, the filters give responses $\mathbf{w}_1^T\mathbf{z}_{ni} \ldots \mathbf{w}_1^T\mathbf{z}_{ni}$ respectively (middle row). The bottom row shows the final output of the ALM model r_n as we cumulatively add the filters. Yellow "+" cross denotes ground truth keypoint position. Red "×" is the predicted position (peak output position).

The goal of keypoint localization is to find the coordinates of the keypoints (corners of eyes, mouth, etc.) in a facial image. It is important because face recognition performance degrades if keypoints are not accurately localized. Unfortunately, this remains a difficult problem due to large variations in pose, expression, luminance and occlusion.

One has to search over all the possible locations in the face region to localize a keypoint based on either a generative model [4] or a discriminative model [3]. However, it is difficult to use all of the samples in training, especially the negative samples (non-keypoint class) due to their large number. Recently, the "Average of Synthetic Exact Filters" (ASEF) [2] was proposed the for eye localization. A correlation filter is learned for each training image and these filters are averaged together. One of its advantages is that it uses the whole image in training rather than a subset of patches. Furthermore, ASEF is very fast in both training and testing because correlation can be computed efficiently in the Fourier domain. However, facial keypoints have large variation in their appearance and it is difficult for a single linear filter such as ASEF to capture this complexity. Moreover, it is unclear that the ASEF approach of averaging together individual correlation filters is even the best way to form a single filter.

To this end, we propose a classifier based on **Additive Logistic Models (ALM)** that has a well-defined cost function and can form non-linear combinations of several filters to capture variation of the target keypoint.

Let t_{ni} denote the label at the n'th pixel of the i'th image. The label t_{ni} equals one if the feature is certainly present and zero if certainly absent. Our goal is to find the posterior probability $Pr(t_{ni} = 1|z_{ni})$ of the label based on a data vector $\mathbf{z}_{ni}$ (the concatenated pixel values in a square patch around the point here) extracted from the image at the current pixel. We define a logistic function: $Pr(t_{ni} = 1|\mathbf{z}_{ni}) = \frac{1}{1+\exp(-a_{ni})}$. The activation a_{ni} determines the tendency for the datum $\mathbf{z}_{ni}$ to be considered as belonging to class one. It consists of an additive sum of a constant a_0 and a series of functions f_m each of which acts on an associated linear projection of the datum $\mathbf{w}_m^T\mathbf{z}_{ni}$ and has associated parameters θ_m,

$$a_{ni} = a_0 + \sum_{m=1}^{M} f_m[\mathbf{w}_m^T\mathbf{z}_{ni}, \theta_m] = a_0 + \sum_{m=1}^{M} \alpha_m \arctan[\beta_m \mathbf{w}_m^T\mathbf{z}_{ni} + \gamma_m]. \quad (1)$$

Since the data vector $\mathbf{z}_{ni}$ is a square image patch, the projections $\mathbf{w}_m$ can be interpreted as vectorized filters $\mathbf{H}_m$ (figure 1). Each filter $\mathbf{H}_m$ (or weight vector $\mathbf{w}_m$) captures a characteristic of the facial keypoint and contributes to the overall activation a_{ni} which nonlinearly combines them.

The additive logistic model can be learned by minimizing the negative log binomial posterior probability over patches $\mathbf{z}_{ni}$ centered at N_K positions of N training images: $\arg\min_{a_0,\mathbf{w}_{1\ldots M},\theta_{1\ldots M}} \sum_{n=1}^{N}\sum_{i=1}^{N_K} -\log[Pr(t_{ni}|\mathbf{z}_{ni})]$. We take a boosting-style sequential approach to optimize this cost function in which we add the functions $f_{1\ldots M}$ one at a time. At each stage we optimize the log likelihood criterion using a gradient descent method.

Keypoints	Left Eye	Right Eye	Nose Bridge	Mouth Tip
ASEF	0.069(0.003)	0.062(0.002)	0.065(0.002)	0.128(0.004)
AdaBoost	0.086(0.003)	0.086(0.003)	0.065(0.002)	0.160(0.005)
Bayesian	0.106(0.004)	0.117(0.005)	0.268(0.007)	0.265(0.006)
ALM	0.056(0.002)	**0.058**(**0.002**)	0.057(0.002)	0.092(0.003)
CALM [3]	**0.054**(**0.002**)	0.059(0.002)	**0.050**(**0.001**)	**0.081**(**0.003**)
CALM [5]	0.057(0.002)	0.059(0.002)	0.051(0.001)	**0.081**(**0.003**)

Table 1: Result of localization error (mean and standard error) in terms of normalized Euclidean distance on UCL data subset.

The appearances of the facial keypoints vary considerably due to factors like expression, lighting and pose which we term *contexts*. If we have information about the context, we can adapt the model accordingly. To this end, we propose a **Context–based Additive Logistic Model (CALM)** which modifies its responses based on the context. In this paper, we illustrate the properties of the proposed algorithm by taking face pose as the context, which was estimated using the approach of [1]. For each training image I_n, the context $c_n \in \{1 \ldots J\}$ is discrete. Equation 1 now depends on the context j so that

$$a_{ni} = \sum_{j=1}^{J} Pr(c_n = j)[a_{0j} + \sum_{m=1}^{M} f_{mj}(\mathbf{w}_{mj}^T\mathbf{z}_{ni}, \theta_{mj})] \quad (2)$$

where we incorporate uncertainty in the context estimate using the posterior $Pr(c_n = j)$ returned by the pose classifier. Note that the context not only determines the function parameters θ_{mj}, but also the data projections $\mathbf{w}_m^T\mathbf{z}_{ni}$: we measure the same aspects of the image, but interpret these measurements differently depending on the situation.

Although we model the appearance of facial keypoints independently using a discriminative model, the keypoint localization can be further improved by exploring the relations of the keypoint positions in a generative model. In our case, we have access to robust pose information, and we exploit this by building a **Context–based Pictorial Structure (CPS)** model: we learn a separate pictorial structure model for each of the pose clusters. In the localization stage, we estimate the MAP context of the test image and use the corresponding tree to help infer the keypoint positions.

We evaluate our proposed method on a subset of the UCL database [1] where the images were captured from on-line dating websites and contains significant variation in pose, lighting and expression. Evaluation of the keypoint detection algorithms is based on the distance from the predicted position to the manually labeled position, normalized by the inter–ocular distance. Experimental results show that our ALM outperforms three the-state-of-art algorithms: ASEF [2], AdaBoost [4] and the Bayesian approach of [4] (Table 1). By exploiting the context information and relation among keypoints, our CALM model and CPS model can further improve the localization performance than ALM model.

[1] J. Aghajanian and S. J. D. Prince. Face pose estimation in uncontrolled environments. In *BMVC*, 2009.

[2] D. S. Bolme, B. A. Draper, and J. R. Beveridge. Average of synthetic exact filters. In *CVPR*, pages 2105–2112, 2009.

[3] D. Cristinacce and T. Cootes. Facial feature detection using AdaBoost with shape constraints. In *BMVC*, pages 231–240, 2003.

[4] M. Everingham and A. Zisserman. Regression and classification approaches to eye localization in face images. In *FGR*, pages 441–448, 2006.