Advanced correlation filters for facial landmark localization

Vitomir Štruc[1], **Jerneja Žganec Gros**[2], **Nikola Pavešić**[1]

[1]*Faculty of Electrical Engineering, University of Ljubljana*
Tržaška 25, SI-1000 Ljubljana, Slovenia
Alpineon Ltd., Ulica Iga Grudna 15, SI-1000 Ljubjana, Slovenia
E-mail: vitomir.struc@fe.uni-lj.si, jerneja.gros@alpineon.si, nikola.pavesic@fe.uni-lj.si

Abstract

The paper develops a novel technique for facial landmark localization based on advanced correlation filters. Specifically, it introduces a new class of advanced correlation filters, named Principal Directions of Synthetic Exact Filters or PSEFs for short, and applies them to the problem of eye localization. To improve upon the basic performance of the PSEF filter for eye localization two types of constraints (i.e., soft and hard constraints) that affect the outcome of the localization procedure are also proposed and incorporated into the procedure. The effectiveness of the developed localization technique is demonstrated on more than 40000 facial images pooled from the FERET and LWF databases. The results of our experiments suggest that the PSEF filters produce significantly better localization results than, for example, the Haar-cascade object detector, while ensuring a more than 10-fold improvement in the processing time.

1 Introduction

In recent years we have witnessed an increased interest in so-called advanced correlation filters, which have proven extremely successful in solving complex tasks related to pattern recognition in computer vision, e.g., face or palmprint recognition, object detection, tracking, etc. The interest in these types of filters is fueled not only by their efficiency, but also by some of their properties, such as mathematical simplicity, computational efficiency and robustness to distortions [1].

In general, advanced correlation filters bear a resemblance to templates and correlation-based template matching techniques, where patterns of interest in images are searched for by cross-correlating the input image with one or more example templates and examining the resulting correlation plane for large values - also known as correlation peaks. With properly designed templates, these correlation peaks can be exploited to determine the presence and/or location of patterns of interest in the given input image [1]. Early template matching techniques relied on rather primitive templates, computed, for example, through simple averaging of the available training images. Contemporary methods, on the other hand, use correlation templates (also referred to as *advanced correlation filters*) that are constructed by optimizing specific performance criteria [1], [2]. Examples of existing correlation filters can be found in [3], [4], [5] or [6].

In this paper we focus on a class of correlation filters called Principal directions of Synthetic Exact Filters (PSEFs) that we have originally introduced in [2]. These filters generalize upon the recently proposed class of advanced correlation filters called Average of Synthetic Exact Filters (ASEF) [6]. Based on these filters and a number of localization constraints we develop a facial landmark localization procedure and demonstrate its effectiveness in comparison with ASEF filters and the established Haar cascade classifier proposed in [7].

2 Preliminaries

ASEF filters represent a recently proposed class of advanced correlation filters that have already proven successful in various computer vision problems [6]. Similar to other correlation filters, a pattern of interest in an image is detected with an ASEF filter by cross-correlating the input image with the given ASEF filter and examining the resulting correlation plane for possible correlation peaks. However, ASEF filters differ from most existing correlation filters in the way they are constructed.

Unlike the majority of correlation filters, which define only a single correlation value per training image, ASEF filters predefine the entire correlation plane for each available training image. As stated by Bolme et al. [6], this correlation plane commonly features a high peak centered at the pattern of interest and (near) zeros at all other image locations (second image in Fig. 1) [2]. Such a synthetic correlation output results in a synthetic exact filter (SEF) (third image in Fig. 1) that can be used to locate the pattern of interest in its corresponding training image.

Obviously, SEF filters do not exhibit broad generalization capabilities, instead they produce distinct peaks only for those images that were used for their construction. To overcome this shortcoming Bolme et al. [6] com-

This work has been supported in parts by the national research program P2-0250(C) Metrology and Biometric Systems, the postdoctoral project BAMBI (ARRS ID Z2-4214), by the European Union, European Regional Fund, in scope of the framework of the Operational Programme for Strengthening Regional Development Potentials for the Period 2007-2013, contract No. 3211-10-000467 (KC Class) and the BioID project financed by the European Union from the European Social Fund, contract No. PP11/2010-(1/2009). The authors additionally appreciate the support of COST Actions IC 1106 and IC1103.

Figure 1: Construction of a synthetic exact filter (SEF): normalized input image multiplied with a cosine window (left), the synthetic correlation output plane (middle), the synthetic exact filter corresponding to the training image on the left (right).

puted a new filter by averaging all of the synthetic exact filters corresponding to a specific pattern of interest. Through the averaging operation the authors ensured better generalization capabilities of the ASEF filters when compared to the SEFs and avoided the over-fitting problem that affects many existing correlation filters

Consider a set of n training images $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ and n corresponding image locations of the pattern of interest. The first step towards computing the ASEF filter for a pattern of interest is the construction of the desired correlation outputs $\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n$ for all n training images:

$$\mathbf{y}_i(x, y) = e^{-\frac{(x-x_i)^2+(y-y_i)^2}{\sigma^2}}, \text{ for } i = 1, 2, ..., n, \quad (1)$$

where σ denotes the standard deviation of the Gaussian-shaped correlation output and (x_i, y_i) represents the coordinate pair corresponding to the location of the pattern of interest in the i-th training image.

Once the correlation outputs have been determined, SEFs are calculated for all n pairs $(\mathbf{x}_i, \mathbf{y}_i)$ as follows:

$$H_i^* = \frac{Y_i \odot X_i^*}{X_i \odot X_i^* + \epsilon}, \text{ for } i = 1, 2, ..., n, \quad (2)$$

where, $X_i = \mathcal{F}(\mathbf{x}_i)$ and $Y_i = \mathcal{F}(\mathbf{y}_i)$ denote the Fourier transforms of the i-th training image and its corresponding synthetic correlation output, $H_i = \mathcal{F}(\mathbf{h}_i)$ stands for the Fourier transform of the i-th SEF filter $\mathbf{h}_i$, ϵ denotes a small constant that prevents divisions by zero, $\odot$ stands for the Schur product and * for the conjugate operator.

In the final step, all n SEFs are simply averaged to produce an ASEF filter (see second image of Fig. 2 for a visual example) that can be used to locate the pattern of interest in a given input image. Here, the ASEF filter in the frequency domain is defined as [6]:

$$H^* = \frac{1}{n}\sum_{i=1}^{n} H_i^*, \quad (3)$$

To apply the ASEF filters for localization of a pattern of interest in an input image, the input image in first cross-correlated with the appropriate ASEF filter and the correlation output is then examined for its maximum. The location of the maximum is simply declared the location of the pattern of interest. In the frequency domain this can be defined as follows:

$$Y = X_t \odot H^*, \quad (4)$$

where Y denotes the correlation output in the frequency domain, $X_t = \mathcal{F}(\mathbf{x}_t)$ denotes the Fourier transform of a test image $\mathbf{x}_t$, H stands for the ASEF filter in the frequency domain and $\odot$ again represents the Schur product. The procedure is also illustrated in Fig. 2.

Figure 2: Visualization of the facial landmark localization procedure (from left to right): the input image, the ASEF filter (with shifted quadrants), the correlation output, the input image with the detected correlation maximum.

3 PSEF filters

The filter construction procedure described in the previous section ensures high generalization capabilities of the ASEF filters through an averaging procedure applied on the SEF filters. However, it implicitly presumes that the SEF filters represent a random variable drawn from a unimodal symmetric distribution and, thus, that their distribution is adequately described by their sample mean.

To derive our PSEF filters we will make a similar assumption and assume that the SEF filters are drawn from a multi-variate Gaussian distribution. Under this assumption, we are able to extend the concept of ASEF filters to a more general form The basic reasoning for our generalization stems from the fact that the first eigenvector of the correlation matrix of some sample data corresponds to the data's mean (or average), while the remaining eigenvectors encode the variance of the sample data in directions orthogonal to the data's average. By using more than only the first eigenvector of the SEF correlation matrix for the localization procedure, we should be able to further improve upon the localization performance of the original ASEF filters [2].

Again consider a set of n training images $\mathbf{x}_1, ..., \mathbf{x}_n$, for which we have already computed n corresponding SEFs for some pattern of interest $\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_n$, (where $\mathbf{h}_i = \mathcal{F}^{-1}(H_i)$ stands for the i-th SEF filter defined in the spatial domain). Assume also that the SEFs reside in a d-dimensional space and that they are arranged into a column-data matrix $\boldsymbol{\zeta} \in \mathbb{R}^{d \times n}$. Instead of simply averaging the SEFs to produce an ASEF filter, we compute the sample correlation matrix $\boldsymbol{\Sigma}$ of the SEFs: $\boldsymbol{\Sigma} = \boldsymbol{\zeta}\boldsymbol{\zeta}^T \in \mathbb{R}^{d \times d}$, and adopt its leading eigenvectors as our PSEF filters, i.e.:

$$\boldsymbol{\Sigma}\mathbf{f}_j = \lambda_j \mathbf{f}_j, \text{where} j = 1, 2, ..., \min(d, n) \quad (5)$$

and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_j \cdots \geq \lambda_{\min(d,n)}$.

One problem arising as a consequence of such a construction procedure is the sign ambiguity of the PSEF filters $\mathbf{f}_j$. Since the computed filters can be multiplied by -1 and still represent valid eigenvectors of $\boldsymbol{\Sigma}$, we have to alleviate this sign ambiguity. In the experimental section we will try to solve the sign ambiguity of our filters through preliminary experiments on some training data.

3.1 Utilizing linearity

The landmark localization procedure using PSEF filters is identical the one presented in Section 2, except for the fact that we have more than a single filter at our disposal and, hence, obtain more than one correlation output:

$$Y_j = X_t \odot F_j^*, \text{ for } j \in \{1, 2, ..., \min(d, n)\}, \quad (6)$$

Figure 3: Comparison of the visual appearance of an ASEF filter (left) and the combined PSEF filter (right).

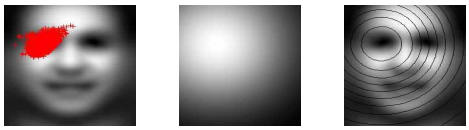


Figure 4: Illustration of the soft constraint concept.

where $X_t = \mathcal{F}(\mathbf{x}_t)$ denotes the Fourier transform of a given test image $\mathbf{x}_t$, F_j denotes the Fourier transform of the j-th PSEF filter $\mathbf{f}_j$ and Y_j refers to the j-th correlation output in the Fourier domain.

To determine the location of our pattern of interest in the given input image, we need to examine all correlation outputs Y_j for maxima and combine all obtained information. A straight forward way of doing this is to examine only the linear combination of all correlation outputs for its maximum and use the location of the detected maximum as the location of our pattern of interest. Thus, we have to examine the following combined correlation output: $\mathbf{y}_c = \sum_{i=1}^{k} w_i \mathbf{y}_i$, where $\mathbf{y}_i$ denotes the correlation output (in the spatial domain) of the i-th PSEF filter, w_i denotes the weighting coefficient of the i-th correlation output, $\mathbf{y}_c$ denotes the combined correlation output, and k stands for the number of PSEF filters used ($1 \leq k \leq \min(d, n)$). From the above descriptions we can deduce that if $k = 1$ the combined correlation output is identical to the correlation output of the ASEF filter. On the other hand, if $k > 1$ we add additional information to the combined correlation output by including additional PSEF filters into the localization procedure. The presented procedure requires one filtering operation for each PSEF filter used. However, the computation can be speeded up by exploiting the linearity of the combination procedure. Instead of combining the correlation outputs, we simply combine all employed PSEF filters into one single filter with enhanced localization capabilities, i.e.:

$$\mathbf{y}_c = \sum_{i=1}^{k} w_i(\mathbf{f}_i \otimes \mathbf{x}_t) = (\sum_{i=1}^{k} w_i \mathbf{f}_i) \otimes \mathbf{x}_t = \mathbf{f}_c \otimes \mathbf{x}_t, \quad (7)$$

where $\mathbf{f}_c = \sum_{i=1}^{k} w_i \mathbf{f}_i$, and $\sum_{i=1}^{k} w_i = 1$. In the presented equations $\mathbf{f}_c$ stands for the combined PSEF filter and $\otimes$ denotes the convolution operator. Note that the localization procedure with the combined PSEF filter has exactly, the same computational complexity as the procedure relying on ASEF filters regardless of the number of PSEF filters used. For our experiments the weights of the individual PSEF filters were selected as: $w_i = \frac{\lambda_i}{\sum_{i=1}^{k} \lambda_i}$.

An example of the visual appearance of the combined PSEF filter obtained with the presented weighting procedure (after the sign ambiguity has been eliminated - see Section 4) is shown on the right hand side of Fig. 3.

3.2 Incorporating localization constraints

To improve upon the basic performance of the PSEF filters we incorporate two constraints into the the facial landmark localization procedure.

The first, which we will refer to as our *soft constraint* in the remainder, represents a weighting function that is multiplied with the correlation output to give more weight to more probable landmark locations. The weighting function can be considered as sort of a prior model and is estimated by analyzing the locations of the landmark of interest on some training data and fitting a Gaussian distribution (with a diagonal covariance matrix) to these locations. The procedure is illustrated in Fig. 4. Here the first image depicts the average of our training set of 15520 face images after the face detection step with superimposed coordinates of the left eye from all images in the training set. The second image shows the estimated weighting function and the third image presents isohypses of the estimated Gaussian weighting function superimposed over the average face.

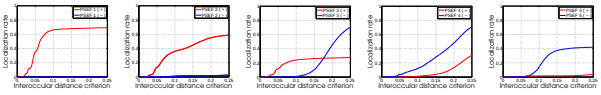
The second constraint incorporated into the landmark localization procedure, referred to as our *hard constraint* in the remainder, is to limit the search space for the facial landmark of interest. When using this heuristic, we look for the left eye only in the upper left quadrant of the image and, similarly, we search for the right eye only in the upper right quadrant of the image.

4 Experiments and results

To assess the landmark localization procedure relying on PSEF filters we make use of two face databases, namely, the FERET database [8] and the Labeled Faces in the Wild (LFW) database [9]. We extract the facial regions from all images of the two databases using the Haar cascade classifier proposed by Viola and Jones [7]. After detecting the facial regions in all images, we select 640 images from the LFW database and manually label the locations of the left and right eye. Next, we produce 40 variations of the facial region of each of the 640 LFW images by randomly shifting the location of the facial regions by up to ± 5 pixels, rotating them by up to $\pm 15°$, scaling them by up to 1.0 ± 0.15 and mirroring them around the y axes. Through these transformations, we augment the initial set of 640 images to a set of 25600 images (of size 128×128 pixels) and employ them for training of the ASEF and PSEF filters.

For testing purposes we apply the same random transforms to 3815 images from the FERET database. Here, we produce 12 modifications of each facial region, which results in 45780 facial images that can be used for our assessment. Prior to subjecting the face images to the proposed localization procedure, all face images are subjected to a log transform and normalized to zero mean and unit variance. In the last step the images are weighted with a cosine window to reduce the frequency effects of the edges encountered when applying the Fourier transform [6]. To measure the effectiveness of the localization procedure we adopt the following criterion [10]:

$$\eta = \frac{\max(\|l_{le} - r_{le}\|, \|l_{re} - r_{re}\|)}{\|r_{le} - r_{re}\|}, \quad (8)$$



(a) PSEF 1 (b) PSEF 2 (c) PSEF 3 (d) PSEF 4 (e) PSEF 5

Figure 5: Results of preliminary experiments aimed at alleviating the sign ambiguity of the computed PSEFs.

where l_{le} and l_{re} denotes the location of the left and right eye found by the assessed procedure, r_{le} and r_{re} denote the reference location of the left and right eye, respectively, and the expression $\|r_{le} - r_{re}\|$ represents the reference interoccular (L_2) distance. For our assessment we examine the correct localization rate for different operating points, i.e., $\eta < \Delta \in \{0.10, 0.15, 0.20, 0.25\}$. We use the soft constraint in all of our experiments with correlation filters, and state explicitly when we also adopt the hard constraint.

The goal of our first series of experiments is to alleviate the sing ambiguity of the computed PSEF filters. To this end, we compute 5 PSEF filters (corresponding to the 5 largest, non-zero eigenvalues of Eq. 5), derive two filters from each of the 5 PSEF filters by multiplying them with $+1$ and -1, and normalizing the result to zero mean and unit variance. With the 5 computed filter pairs, we conduct localization experiments with the 45780 face images of the FERET database and plot the results in form of graphs as shown in Fig. 5. We select a threshold of $\Delta = 0.25$ as the relevant operating point of our localization procedure and based on this value determine the appropriate sign of each of the five PSEF filters. Note here that more (or less) filters than 5 could be used for our experiments, the presented results, however, are enough to show the feasibility of our approach.

If we take a look at the presented results in Fig. 5, we can see that in our case the best localization results are obtained with the first two filters being multiplied with $+1$ and the remaining filters being multiplied with -1. Furthermore, we can notice, that the best localization performance is obtained with the first PSEF filter, which in fact corresponds to an ASEF filter, while the remaining filters perform worse.

Our second series of experiments comprises two types of tests. The first type does not rely on the hard constraint while the second type does. The results for the first type of experiments are shown on the left side of Fig. 6, while the results of the second type of experiments are shown on the right side of Fig. 6. Some numerical results for different values of Δ are also summarized in Table 1. Note that the proposed PSEF filters outperform both tested alternatives to eye localization, namely, ASEF filters as well as the Haar cascade classifier.

In the third series of experiments we measured the execution times needed for the localization procedure. The best average time, computed by conducting the (left and right eye) localization procedure 10 times on all test images, was 46.3 ms for the Haar classifier (25.1 ms with the hard constraint) and 1.00 ms for the correlation filters (1.01 ms with the hard constraint).

As a final note let us say that the ASEF filters require only a few minutes to be trained, since the rely only on a simple averaging operation. The PSEF filters require a few hours for their training, as this involves the computation of a large correlation matrix and its decomposition. Finally, the Haar classifier is known to have training times in the order of days or weeks.

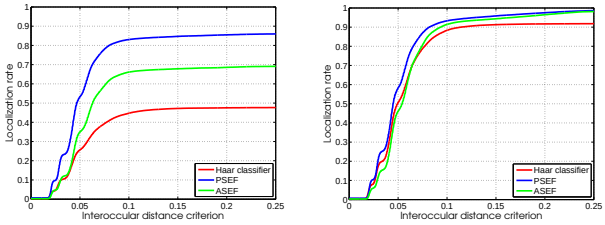


Figure 6: Comparison of different localization techniques with (right) and without (left) hard constraint.

Table 1: Localization rates (in %) at different values of the localization criterion.

η	Without hard constraint			With hard constraint		
	Haar	ASEF	PSEF	Haar	ASEF	PSEF
0.10	44.7	66.1	83.0	88.3	91.4	93.3
0.15	47.2	67.8	84.7	91.3	94.4	95.8
0.20	47.5	68.6	85.5	91.7	96.5	97.5
0.25	47.7	69.1	86.0	91.8	98.1	98.6

5 Conclusion

We have presented a new class of correlation filters called Principal directions of Synthetic Exact Filters and applied them to the task of eye localization. We have shown that the filters outperform the recently proposed ASEF filters and the established Haar cascade classifier at this task.

References

[1] B.V.K.V. Kumar, A. Mahalanobis, A. Takessian: Optimal tradeoff circular harmonic function correlation filter methods providing controlled in-plane rotation response. IEEE Trans. on Image Proc., vol. 9, no. 6, 1025–1034, 2000.

[2] V. Štruc, J. Žganec-Gros, N. Pavešić: Principal Directions of Synthetic Exact Filters for Robust Real-Time Eye Localization, In: Proc. of BioID, pp. 180-192, 2011.

[3] R.A. Kerekes, B.V.K.V. Kumar: Correlation filters with controlled scale response. IEEE Transactions on Image Processing, vol. 15, no. 7, 1794–1802, 2006.

[4] C.F. Hester, D. Casasent: Mulitvariant technique for multiclass pat. rec. App. Opt., vol. 19, no. 11, 1758–1761, 1980.

[5] A. Mahalanobis, B.V.K.V. Kumar, D. Casasent: Minimum average correlation energy filters. Applied Optics, vol. 26, no. 17, 3633–3640, 1987.

[6] D.S. Bolme, B.A. Draper, J.R. Beveridge: Average of synthetic exact filters. In: CVPR'09, pp. 2105–2112, 2009.

[7] P. Viola, M.J. Jones: Robust real-time face detection. Int. J. of Comp. Vis., vol. 57, 137–154, 2004.

[8] P.J. Phillips, H. Moon, S.A. Rizvi, P.J. Rauss: The FERET evaluation methodology for face-recognition algorithms. IEEE TPAMI, vol. 22, no. 10, 1090–1104, 2000.

[9] G.B. Huang, M. Ramesh, T. Berg, E. Learned-Miller: Labeled Faces in the Wild. Technical Report 07-49, 2007.

[10] O. Jesorsky, K.J. Kirchberg, R.W. Frischholz: Robust face detection using the Hausdorff distance. AVBPA'01, Springer LCNS-2091, pp. 90–95, 2001.