Available online at www.sciencedirect.com

ScienceDirect

CAAI Transactions on Intelligence Technology 1 (2016) 241–248
http://www.journals.elsevier.com/caai-transactions-on-intelligence-technology/

Original Article

Fourier Locally Linear Soft Constrained MACE for facial landmark localization☆

Wenming Yang*, Xiang Sun, Weihong Deng, Chi Zhang, Qingmin Liao

Graduate School at Shenzhen, Tsinghua University, Tsinghua Campus, The University Town, Shenzhen 518055, China

Available online 20 October 2016

Abstract

This paper proposes a novel nonlinear correlation filter for facial landmark localization. Firstly, we prove that SVM as a classifier can also be used for localization. Then, soft constrained Minimum Average Correlation Energy filter (soft constrained MACE) is proposed, which is more resistent to overfittings to training set than other variants of correlation filter. In order to improve the performance for the multi-mode of the targets, locally linear framework is introduced to our model, which results in Fourier Locally Linear Soft Constraint MACE (FL^2 SC-MACE). Furthermore, we formulate the fast implementation and show that the time consumption in test process is independent of the number of training samples. The merits of our method include accurate localization performance, desiring generalization capability to the variance of objects, fast testing speed and insensitivity to parameter settings. We conduct the cross-set eye localization experiments on challenging FRGC, FERET and BioID datasets. Our method surpasses the state-of-arts especially in pixelwise accuracy.

Keywords: Facial landmark localization; Overfitting; Multimode; FL^2 SC-MACE

1. Introduction

Facial landmark detection has been drawing increasing attentions due to its important role in face recognition, face animation, human computer interaction, expression recognition and 3D face modeling [16,17]. Generally, it can be categorized as face alignment and feature localization. In the task of face alignment, a number of facial landmarks like pupils, nostril and the corner of the mouth are localized interactively. In this vein, perhaps the most popular kind of approaches is Constrained Local Model (CLM [16]), which ultilizes localizer (also referred to as local expert) to detect each landmark separately. The global statistical constraint characterized by the Point Distribution Model (PDM) [3] is then imposed on the localization result for modification. While recent literatures concentrate on how to optimize the global parameters according to the structure of human faces, they simplified the discussion on the local expert by utilizing linear SVM or Adaboost in consideration of its efficiency and effectiveness. In some other circumstances, only the coordinates of crucial feature points are need. Localizing all facial points may decrease the efficiency in practical use. For example, face recognition needs to detect eyes to remove the similarity variances. For these reasons, a precise and efficient localizer purely using local texture is extremely required.

The approaches for feature localization can be divided into two categories: 1) shape based methods and 2) statistics based methods. Shape based methods, which utilize special geometric characteristics of landmarks like pupils, have demonstrated good performance in the situation where the subject show front face and the feature points are visible [4,5]. However they may fail when the subjects show multiple head poses or the features are distorted by some expressions (e.g. closed eyes). Besides, its also hard to transit the method to

☆ This work was supported by the Special Foundation for the Development of Strategic Emerging Industries of Shenzhen under Grant JCY20130402145002441.
* Corresponding author.
E-mail address: yangelwm@163.com (W. Yang).
Peer review under responsibility of Chongqing University of Technology.

http://dx.doi.org/10.1016/j.trit.2016.10.006

other landmarks. On the contrary, statistics based methods learn the classifiers or templates to determine where the target landmark is. As a representative of statistics based method, correlation filter based localizer has gained success recently. Its merits include shift invariance, robustness to noise and speediness of testing. While correlation filter is optimized to generate an explicit peak at the position of the target in the output plane, it suffers the risks of overfitting to the training samples and the incapability to deal with multimode situations. There have been some attempts to overcome the drawback of overfitting. In Ref. [6], the whole correlation output is specified for each training image. In Ref. [7], Maximum Margin Correlation Filter (MMCF) combines Mean Square Error SDF with SVM to make a tradeoff between the localization advantage of synthetic discriminant function (SDF) and the generalization capability of SVM. However, since they are all linear methods, the performances deteriorate when the feature points show multiple patterns. Xie et al. introduced RBF kernel to MACE [8]. Mahalanobis et al. construct a quadratic filter using Rayleigh quotient [9]. Though nonlinear frameworks are applied, they are still prone to overfitting due to hard constraints on the cost function.

In this paper, we propose a novel correlation filter which can be used for individual face landmark detection and as local expert for face alignment algorithms. We first prove why SVM, as a classifier, is also a well performed localizer. Based on the analysis, we propose soft constrained MACE, which minimizes average correlation energy with soft constraint. In order to avoid the fundamental drawback that single correlation filter cant dispose of variant cases, we combine soft constrained MACE with local linear kernel in Fourier domain, and formulate a nonlinear filter named FL^2 SC-MACE. Based on Parseval theorem, in the test phase, all calculations is conducted on spatial domain rather than Fourier domain. Besides, the testing speed is independent of the quantity of training samples. Our method has the merits of superior localization performance, the generalization to the variance of objects, fast testing speed and insensitivity to parameters. Validation experiments for eye localization are conducted on FRGC, BioID and FERET. To verify the generalization performance, we demonstrate cross-set results and show that FL^2 SC-MACE outperforms the state-of-arts by a considerable margin.

2. Backgrounds and soft constrained MACE

2.1. Correlation filter

In the past decades, different correlation filters that emphasize different goals have been proposed. A filter is synthesized based on the training samples, which is then cross-correlated with the test image. There is supposed to be a distinct peak at the ground truth location of the target (a blunt peak would affect the accuracy of the localization), and flat responses elsewhere on the correlation plane. In real application, the calculation is always conducted on spatial frequency domain:

$$Y(u,v) = H(u,v)X^*(u,v) \tag{1}$$

where $H(u,v)$ and $X(u,v)$ is the DFT of the synthesized template and the test image respectively. $(\cdot)^*$ denotes conjugate operator. The framework is much like sliding window, in which each subregion of test image is matched with a template. However there are several differences between correlation operator and sliding window: 1) each subregion could be normalized in the framework of sliding window, which could not be achieved by correlation; 2) because of the lack of normalization, in the testing process peak-to-sidelobe ratio (PSR) [24] rather than simply the output is measured.

A. Minimum Average Correlation Energy

MACE is the most widely used and researched kind of correlation filter. The earliest version of MACE is proposed in [15], which requires the inner product of template and the true class samples to strictly reach a fixed value:

$$\begin{aligned} &\min_{\tilde{\mathbf{h}}} \tilde{\mathbf{h}}^H \mathbf{S} \tilde{\mathbf{h}} \\ &s.t. \quad \tilde{\mathbf{h}}^H \tilde{\mathbf{x}}_i = u_i, \quad i = 1,2,\ldots l \end{aligned} \tag{2}$$

where $\tilde{\mathbf{h}}$ is the template to be calculated and u_i is the desired inner product value (e.g. 1 for positive samples and 0 for negative ones). $(\cdot)^H$ is conjugate transposition and $\tilde{\mathbf{x}}_i$ is the DFT of the training sample re-ordered into a column vector. $\mathbf{S} = \sum_i \mathbf{X}_i^* \mathbf{X}_i$ is the average energy diagonal matrix, where $\mathbf{X}_i$ is a $N^2 \times N^2$ diagonal matrix with the elements of $\tilde{\mathbf{x}}$ ordered into its diagonal. By minimizing $\tilde{\mathbf{h}}^H \mathbf{S} \tilde{\mathbf{h}} = \sum_i \left|\left| X_i^* \tilde{\mathbf{h}} \right|\right|_2^2$, the energy of the correlation output between the template and all the training samples is suppressed overall. It can be proved that if only one training sample is involved, the optimized correlation result will be a delta function $\mathbf{X}_i^* \tilde{\mathbf{h}} = c(1,0,\ldots 0)^T$. When there are more than one training samples, it is overdetermined for all the correlation results to be delta functions. Observations show that the hard constraint as the form of (2) often leads to the sensitivity to distortions and the overfitting to the training samples, which affects the generalization capability of the filter.

Aiming at solving the drawbacks brought by hard constraint, unconstrained MACE is proposed [2]:

$$\max_{\tilde{\mathbf{h}}} \frac{\tilde{\mathbf{h}}^H \tilde{\mathbf{m}} \tilde{\mathbf{m}}^H \tilde{\mathbf{h}}}{\tilde{\mathbf{h}}^H \mathbf{S} \tilde{\mathbf{h}}} \tag{3}$$

where $\tilde{\mathbf{m}} = 1/N \sum \tilde{\mathbf{x}}_i$ is the mean of the DFTs of training samples. Unconstrained MACE shows better tolerance to noises than constrained filters. By removing the hard constraint, it permits the correlation planes to adjust to whatever value that best optimizes the criterion. However it still doesn't exclude the risk of overfitting. Assuming the response of one training sample $\tilde{\mathbf{h}}^H \tilde{\mathbf{x}}_i$ is extremely high, it would cause

large value of $\tilde{\mathbf{h}}^H\tilde{\mathbf{m}}\tilde{\mathbf{m}}^H\tilde{\mathbf{h}}$, which results in the overlook of the influence of other training samples.

B. Why is SVM a good localizer?

SVM is an optimal classifier in the sense of maximum margin. According to Parseval theorem, the objective function of SVM converted to the Fourier domain can be written as:

$$\min_{\tilde{\mathbf{h}}} \tilde{\mathbf{h}}^H\tilde{\mathbf{h}} + C\sum_i \left[1 - y_i(\tilde{\mathbf{h}}^H\tilde{\mathbf{x}}_i + b)\right]_+ \tag{4}$$

where C is the penalty parameter. $[\cdot]_+ = \max(\cdot, 0)$ is referred to as hinge loss function. While $\tilde{\mathbf{h}}^H\tilde{\mathbf{h}}$ is optimized to maximize the margin (minimize $\tilde{\mathbf{h}}^H\tilde{\mathbf{h}}$ is equivalent to maximize $(\tilde{\mathbf{h}}^H\tilde{\mathbf{h}})^{-1}$, which is proportional to the margin), the penalty item seeks to punish the outliers. Here we re-interpret the objective function of SVM in the sense of localization rather than classification, and reveal how SVM achieves good localization performance.

The localization task involves two key capabilities: 1) when the object is not the target, the response should stay low to keep from false detecting, and 2) when a target appears, the localizer should not only detect it, but localize it accurately, which requires the output plane to be a sharp peak rather than a blunt one. As $\mathbf{S}$ is a diagonal positive definite matrix defined in (2), $\tilde{\mathbf{h}}^H\mathbf{S}\tilde{\mathbf{h}}$ satisfies the axiom of norm and thus is referred to as $\mathbf{S}$-norm. The equivalence of $\mathbf{S}$-norm and l_2-norm can be written as:

$$s_1\tilde{\mathbf{h}}^H\tilde{\mathbf{h}} \le \tilde{\mathbf{h}}^H\mathbf{S}\tilde{\mathbf{h}} \le s_2\tilde{\mathbf{h}}^H\tilde{\mathbf{h}} \tag{5}$$

where s_1 and s_2 are constant and correspond to the minimum and maximum eigenvalues of $\mathbf{S}$. SVM suppresses the energy of the correlation output by minimizing the l_2 regularizer. Meanwhile, as $\tilde{\mathbf{h}}^H\tilde{\mathbf{x}}$ is the origin of the correlation output, $[1 - y_i(\tilde{\mathbf{h}}^H\tilde{\mathbf{x}}_i + b)]_+$ is actually the soft constraint on the correlation peak value.

C. The formulation of soft constrained MACE

Savvides et al. indicate that the correlation template works to whiten the amplitude spectrum and cancel the phase of the image, which yields a flat spectrum plane [24]. As the energy decays with the increase of the frequency, the correlation template is expected to amplify the energy of the high frequency region. Though (5) gives an upper limit of the average correlation energy, SVM doesn't emphasize the characteristics of objects' spectrums. Soft constrained MACE which minimizes the average correlation energy is written as:

$$\min_{\tilde{\mathbf{h}}} \tilde{\mathbf{h}}^H\mathbf{S}\tilde{\mathbf{h}} + C\sum_i \left[1 - y_i(\tilde{\mathbf{h}}^H\tilde{\mathbf{x}}_i + b)\right]_+ \tag{6}$$

In order to minimize the objective function, if a diagonal entry of $\mathbf{S}$ is large, the corresponding entry of the template $\tilde{\mathbf{h}}$ is supposed to be small. On the other hand, the entries of the template with respect to the small energy are expected to be large enough to guarantee a high value of $\tilde{\mathbf{h}}^H\tilde{\mathbf{x}}_i$. Note that though both using soft constraint, soft constrained MACE is different from (MMCF [12]). While MMCF is a tradeoff between MSE-SDF and SVM, soft constrained MACE seeks the melioration of SVM and MACE simultaneously.

For further analysis, the dual problem writes as:

$$\begin{aligned} &\max_{0\le\alpha_i\le C} \sum_i \alpha_i - \frac{1}{2}\sum_i\sum_j \alpha_i\alpha_j y_i y_j \tilde{\mathbf{x}}_i\mathbf{S}^{-1}\tilde{\mathbf{x}}_j \\ &s.t. \sum_i \alpha_i y_i = 0 \end{aligned} \tag{7}$$

The frequency components of the training samples are equilibrized by $\mathbf{S}^{-1/2}$. This could be regarded as a pre-whitening process in the Fourier domain. Fig. 1 shows the typical correlation output plane of soft constrained MACE (top) and SVM (bottom). The sharpness of the formal one means it is more suitable for localization task. In addition, because soft constrained MACE calculates the maximum margin hyperplane for the weighted DFTs, only the samples that approximate to the decision hyperplane contribute to the synthesis of the correlation template, which eliminates the risks of overfitting to the training set. While all variants of MACE aim to minimize average correlation energy and maximize the peak value, soft constrained MACE is less likely to overfit samples.

In summary, there exits close relationships between the classification by SVM and the localization by correlation filter. In addition to the sense of maximum margin classification, the objective function of SVM can also be interpreted as improving the correlation peak value and suppressing the

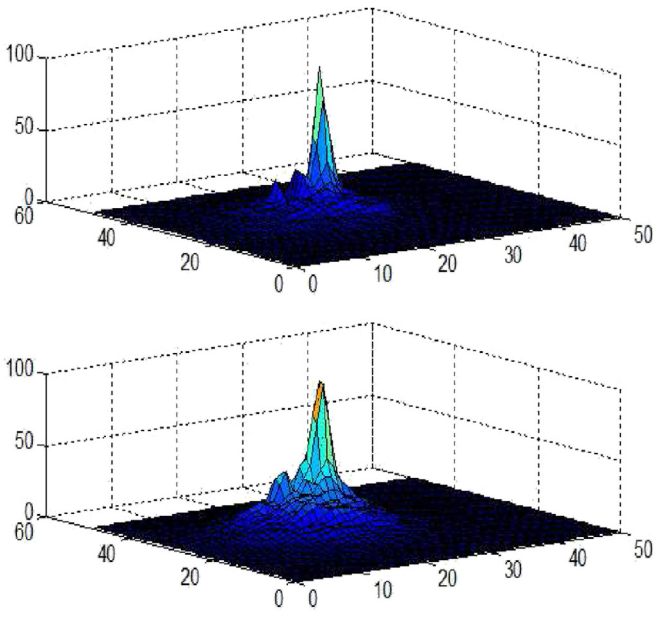

Fig. 1. The typical correlation output plane of soft constrained MACE (top) and SVM (bottom). Both results are multiplied by a Gaussian function. Though SVM suppresses the energy of the correlation output plane as well, soft constrained MACE show more obvious effect, which results in a sharper peak.

energy of the whole correlation output. In comparison, the objective function of MACE can be interpreted as calculating the optimal decision hyperplane for the weighted DFT of samples. Based on the analysis, we propose soft constrained MACE, which is formulated to overcome the overfitting.

3. Fourier local soft constrained MACE

The previous discussion illustrates training the correlation filter is equivalent to participating the weighted DFTs in frequency domain. Therefore, though soft constrained MACE is less likely to overfit data, it doesn't settle the problem that the optimization of linear correlation filter is overdetermined and ill-posed when the data are characterized of multiple patterns. Recently some researchers proposed local based classifiers whose essential idea is that a query sample is judged by a hyperplane which is estimated only by the samples that adjoin to it [27,32]. The performances are comparable to traditional kernel methods. However, they suffer from huge quantity of time cost during test. In this section, we aim to resolve the overfitting efficiently by means of Fourier local linear approach.

Local linear-SVM (LL-SVM) [25] deems the decision template as a function of the sample being judged. The sample is first projected on a local basis. Then a decision function is obtained by using stochastic gradient descent. Instead of using local coding scheme, Zhang et al. [30] coded samples by the singular vectors of the matrix composed by the vectorized training data: $\mathbf{M} = (\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_n) = \mathbf{U}\mathbf{\Sigma}\mathbf{V}$, and proved the model is equivalent to a finite kernel which can be written as:

$$\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = <\mathbf{c}_i\mathbf{x}_i^T, \mathbf{c}_j\mathbf{x}_j^T> \tag{8}$$

where $<\cdot,\cdot>$ is the Frobenius inner product and $\mathbf{c}_i = \mathbf{U}_p^T\mathbf{x}_i$ is the coefficient of $\mathbf{x}_i$ projecting on the principle left singular vectors. If the DFTs of samples are chosen to be the feature vectors: $\tilde{\mathbf{M}} = (\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots \tilde{\mathbf{x}}_n) = \tilde{\mathbf{U}}\tilde{\Sigma}\tilde{V}$, it can be proved via the theory of SVD that $\tilde{\mathbf{U}} = \mathbf{F}\mathbf{U}$, where $\mathbf{F}$ is the Fourier transform matrix. Therefore we have:

$$\tilde{\mathbf{c}}_i = \tilde{\mathbf{U}}_p^T\tilde{\mathbf{x}}_i = \mathbf{c}_i \tag{9}$$

Note that (8) can be simplified as:

$$\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{c}_i^T\mathbf{c}_j \sum_m x_{mi}x_{mj} = \mathbf{c}_i^T\mathbf{c}_j\mathbf{x}_i^T\mathbf{x}_j \tag{10}$$

where x_{mi} is the m-th element of the i-th sample. Use Parseval theorem:

$$\mathbf{K}(\tilde{\mathbf{x}}_i\ ,\ \tilde{\mathbf{x}}_j) = \tilde{\mathbf{c}}_i^H\tilde{\mathbf{c}}_j\tilde{\mathbf{x}}_i^H\tilde{\mathbf{x}}_j = \mathbf{c}_i^T\mathbf{c}_j\mathbf{x}_i^T\mathbf{x}_j = K(\mathbf{x}_i\ ,\ \mathbf{x}_j) \tag{11}$$

That is, the kernel value preserves when the manipulations are conducted in Fourier domain. Now consider implementing the local linear scheme to soft constrained MACE. All the DFTs of the samples are weighted by $\mathbf{S}^{-1/2}$. The FLSC kernel then has a form of:

$$\mathbf{K}_{FLSC}(\mathbf{x}_i, \mathbf{x}_j) = \widehat{\mathbf{c}}_i^H\widehat{\mathbf{c}}_j\widehat{\mathbf{x}}_i^H\widehat{\mathbf{x}}_j = \widehat{\mathbf{c}}_i^H\widehat{\mathbf{c}}_j\tilde{\mathbf{x}}_i^H\mathbf{S}^{-1}\tilde{\mathbf{x}}_j \tag{12}$$

where $\widehat{\mathbf{x}}_i$ is the IDFT of $\mathbf{S}^{-1/2}\tilde{\mathbf{x}}_i$ and $\widehat{\mathbf{c}}_i$ is the SVD codes with respect to $\widehat{\mathbf{x}}_i$. An interpretation of FLSC kernel is that as the samples distribute in several different manifolds, the inner product value can judge whether two samples belong to the same one. In compensation, $\widehat{\mathbf{c}}_i^H\widehat{\mathbf{c}}_j$ provides local informations about whether they are also resemblant within the subspace. Therefore only when two samples are from the same manifold and share the similar coding information can FLSC-kernel get large value.

In testing process, the decision function gives the form as:

$$\mathbf{H}_{FLSC}(\mathbf{x}) = \sum_i y_i\alpha_i K_{LS}(\mathbf{x}\ ,\ \mathbf{x}_i) + b \tag{13}$$

where y_i and α_i are the label and weight of $\mathbf{x}_i$. Like the case of kernel SVM, as the number of training samples increases, the computation complexity grows linearly with it. To alleviate the computation complexity, noting in (12) that:

$$\widehat{\mathbf{c}}_i^T\widehat{\mathbf{c}}_j = \widehat{\mathbf{x}}_i^T\left(\frac{\widehat{\mathbf{u}}_1}{\widehat{\sigma}_1}, \frac{\widehat{\mathbf{u}}_2}{\widehat{\sigma}_2}, .., \frac{\widehat{\mathbf{u}}_{n_0}}{\widehat{\sigma}_{n_0}}\right)\left(\frac{\widehat{\mathbf{u}}_1}{\widehat{\sigma}_1}, \frac{\widehat{\mathbf{u}}_2}{\widehat{\sigma}_2}, .., \frac{\widehat{\mathbf{u}}_{n_0}}{\widehat{\sigma}_{n_0}}\right)^T\widehat{\mathbf{x}}_j \tag{14}$$

$$= \left(\mathbf{S}^{-\frac{1}{2}}\tilde{\mathbf{x}}_i\right)^H\sum_{n=0}^{n_0}\frac{\breve{\mathbf{u}}_n\mathbf{u}_n^T}{\widehat{\sigma}_n^2}\widehat{\mathbf{x}}_j \tag{15}$$

$$= \mathbf{x}_i^T\sum_{n=0}^{n_0}\frac{\overline{\mathbf{u}}_n\mathbf{u}_n^T}{\widehat{\sigma}_n^2}\widehat{\mathbf{x}}_j \tag{16}$$

where $\widehat{\mathbf{u}}_i$ are the singular vectors trained by the DFTs of the samples, $\breve{\mathbf{u}}_j$ is the DFT of $\widehat{\mathbf{u}}_j$ and $\overline{\mathbf{u}}_j$ is the IDFT of $\mathbf{S}^{-1/2}\breve{\mathbf{u}}_j$, the decision function of local soft constrained MACE is a quadratic form of $\mathbf{x}$:

$$\mathbf{H}_{FLSC}(\mathbf{x}) = \mathbf{x}^T\sum_i\sum_n\frac{\widehat{\mathbf{u}}_n\mathbf{u}_n^T}{\sigma_n^2}\widehat{\mathbf{x}}_i\ddot{\mathbf{x}}_i^T\mathbf{x} + b \tag{17}$$

$$= \mathbf{x}^T\widehat{\mathbf{U}}U\widehat{X}\ddot{\mathbf{X}}\mathbf{x} + b \tag{18}$$

$$= \mathbf{x}^T\mathbf{A}\mathbf{x} + b \tag{19}$$

where $\ddot{\mathbf{x}}_i$ is the DFT of $\alpha_i\mathbf{S}^{-1}\tilde{\mathbf{x}}_i$. The bold capital letters denote the vectors ordered into the correspond matrixes (e.g. $\widehat{\mathbf{U}}$ is constructed by the singular vectors divided by the corresponding singular value $\widehat{\mathbf{u}}_i/\sigma_i$), and $\mathbf{A} = \widehat{\mathbf{U}}U\widehat{X}\ddot{\mathbf{X}}$ is the quadratic matrix. To further accelerate the decision process, considering the localization tasks to which bias b is ignorable and

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = \mathbf{x}^T\mathbf{A}_e\mathbf{x} \tag{20}$$

where the symmetric matrix $\mathbf{A}_e = \mathbf{A} + \mathbf{A}^T/2$ is the even part of $\mathbf{A}$, the decision function can be reformulated as:

$$\mathbf{H}_{FLSC}(\mathbf{x}) = \mathbf{x}^T\mathbf{A}_e\mathbf{x} = \mathbf{x}^T\mathbf{Z}\mathbf{\Lambda}\mathbf{Z}^T\mathbf{x} = \sum_{i=1}^{m_0}\lambda_i\left(\mathbf{z}_i^T\mathbf{x}\right)^2 \tag{21}$$

where $\mathbf{Z}$ is an orthogonal matrix and $\mathbf{\Lambda}$ is a diagonal matrix whose elements are all real. $\mathbf{z}_i$ are the eigenvectors and λ_i are the corresponding eigenvalues sorted in descend order with respect to its absolute value. Since the rank of $\mathbf{A}$ is no more than that of $\mathbf{U}$, m_0 *is equal to two times of the number of singular vectors used for coding*. Experimentally, we found the selection of the number of singular vector doesn't affect much to the performance, which guarantees the fast implementation of FL^2 SC-MACE. By further eliminating the eigenvectors $\mathbf{z}_i$ with small eigenvalue energy, the computation can be achieved more efficiently without deteriorating the performance.

4. Experiment

We conduct a cross-set eye localization experiment based on BioID, FERET and FRGC databases. This experiment is suitable for verifying the performance of our method because there are many cases when the subjects wear glasses and close eyes, which are different patterns from the standard eye model. We choose cross-set experiment because 1) many state-of-the-art methods have demonstrated satisfying results of intra-class experiment but not exactly in the case of cross-set experiment, and 2) cross-set experiment is more analogous to the real practice, in which the collect environment of training set may not be the same as that of the testing set. Though current researches also attempt to localize other landmarks of faces, we focus on eye localizing experiment because almost all the landmark detection methods use the special structure of the face, which is not the main point of this paper. Nevertheless, our method is nicely compatible for many landmark detection approaches, which also use a localizer like SVM to give a evaluation of candidate pixels.

BioID is a challenging dataset for eye localization task with image resolution fixed at 384 × 286. The samples are collected under a large variety of illumination, expression and pose. The FERET database was constructed ranging from 1993 to 1997 over 15 sessions for automatic face recognition experiments. It consists 994 subjects photoed at different angles with the size of 512 × 768 pixels. We choose regular frontal images and alternative frontal images (referred to as fa and fb) for our experiment. FRGC aims to provide sufficient data for researchers who are engaged in face recognition. It comprises six experiments. Among 50000 available images, some are collected in still controlled lighting condition and some are not. In all datasets, there are a number of images when the subjects either close their eyes or wear glasses.

Normalized error is used to assess our outcome. This measurement was firstly proposed by Jesorsky [19] and is a reasonable criterion with respect to the real application of eye localizer. In this metric, the worse localized eye is measured:

$$e_n = \frac{\max(e_l, e_r)}{d} \tag{22}$$

where e_l and e_r are the Euclidian distances between the ground truth and the estimated location with respect to left and right eyes respectively, and d is the ground truth distance between the two eyes.

A. Procedures

We use OpenCV to crop faces from all three datasets, and resize them to 100 × 100. The images which are not detected are discarded. Since the cropped face image fits the anthropometric ratio, we will localize eyes in a rough region that is large enough to cover all kinds of extreme cases. Rationally, the negative samples are randomly selected in the same region. Also, they should keep away from the authentic positions at least about 0.1 times of d. Each image provides one positive feature and 20 negative features for two eyes respectively. All the feature vectors have the dimension of 21 × 21. Each of them is normalized by taking a $log(1 + v)$, deducting its mean and normalizing the energy. The training samples are firstly convoluted by $\mathbf{S}^{-1/2}$. SVD is then implemented on the filtered sample features. SVM packets like Libsvm [14] can be used to train the weighting coefficients α_i, with the kernel as a form of (12). While $\mathbf{A}$ is obtained following (19), the correlation templates $\mathbf{z}_i$ and eigenvalues λ_i can be calculated. The position with the highest peak-to-sidelobe ratio is regarded as eye location.

B. Result

In the first experiment, we test the performance changes with respect to the number of singular vectors $\widehat{\mathbf{u}}_i$ to code. We partition FRGC into two groups. The training group consists 2000 images and the remaining are assigned to testing group. Fig. 3 shows the localization performance as a function of the

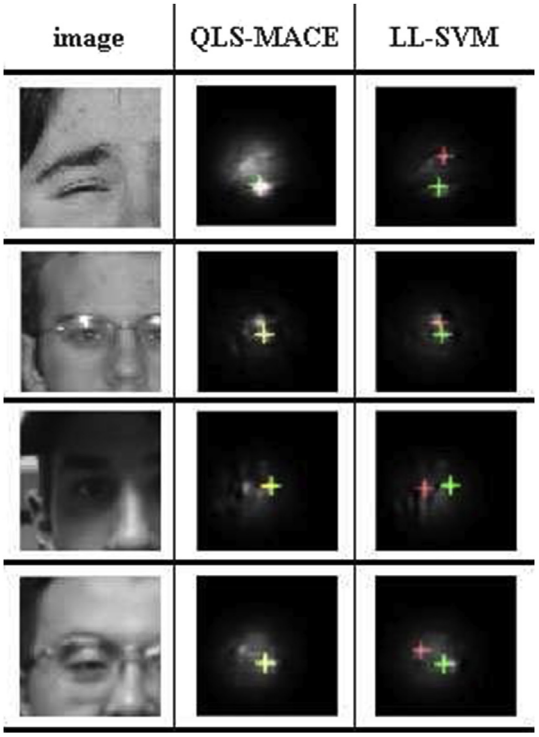

Fig. 2. The green crosses denote the ground truth positions while the red ones denote the predict positions.

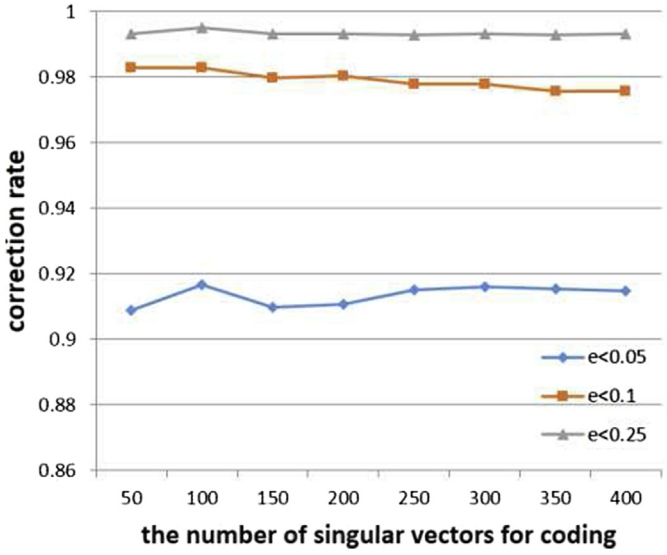

Fig. 3. The performance as a function of the number of basis singular vectors.

number of singular vectors $\hat{\mathbf{u}}_i$. As can be seen, the performance is insensitive to the parameter.

Fig. 4 gives the cross-set outcome of FL^2 SC-MACE as well as other eye localization approaches. 2000 positive samples and 20000 negative samples outside the testing database are used for training. As can be seen, QLS-MACE shows overall advantages beyond others, especially in accurate localization performance ($e < 0.05$). Fig. 2 demonstrates some examples of complex situations (i.e. closing eyes, wearing glasses or dark lighting conditions).

Since a number of eye locating methods are tested on BioID with the same normalization error measurement, it's facilitative to compare our methods with the preceding results. In this experiment we still guarantee that all training samples are from other datasets than BioID. Prior knowledge of eye position is also used to improve performance. Following MAP criterion, the posterior probability writes as:

$$P((x,y)|\mathbf{I}) \propto P(\mathbf{I}|(x,y))P((x,y)) \tag{23}$$

where (x,y) indicates the event that (x,y) is the ground truth location of the eye and $\mathbf{I}$ is the given pixel values. While the output of SVM can be deemed as $P(\mathbf{I}(x,y))$ via binomial log-likelihood [36], we model prior probability $P((x,y))$ by a simple unimodal normal distribution whose mean and variance are those of training set.

Table 1 shows the results of our approach and others. We provide results with or without prior knowledge. If the authors didn't provide a complete result, we will make an approximation based on the performance curve or simulate their methods. Note that although some papers demonstrate a good localization rate, i.e. [31], which reaches 89.6%, when $e = 0.05$ and 99.1% when $e = 0.25$, we exclude those results from Table 1 because the training and validation process are executed on BioID as well. As can be seen from the results presented, FL^2 SC-MACE demonstrates a desired result in both accurate and coarse localization.

Tables 2 and 3 show the localization results of several facial landmarks, i.e., eye corner, mouth corner and nose on LFPW [21] and HELEN database [22]. Since these two da-

Table 1
The comparison of cross-set eye localization performance. If the author doesn't give result, we will estimate it according to the given curve or simulate the algorithm ourselves (underlined figures).

Method	$e \leq 0.05$	$e \leq 0.1$	$e \leq 0.25$
Asteriadis et al. [26]	73.5%	81.7%	97.4%
Kroon et al. [28]	65.0%	87.0%	98.8%
Türkan et al. [18]	20.0%	73.7%	**99.6%**
Niu et al. [23]	76.0%	93.0%	97.0%
Cristinacce et al. [20]	57.2%	**96.0%**	97.0%
Timm and Barth [35]	82.5%	93.4%	98.0%
Valenti and Gevers [29]	**86.1%**	91.7%	97.0%
Bolme et al. [33]	78.9%	91.3%	98.1%
Tan et al. [34]	65.9%	95.0%	99.2%
Kumar et al. [12]	78.2%	85.2%	96.6%
Ours	**87.3%**	94.2%	**99.6%**
Ours with prior knowledge	**90.2%**	**98.2%**	**99.9%**

The bold is the best performance in the corresponding correction level.

Table 2
The percentage that the localization error of the landmark is less than 10% of the eye-to-eye distance on LFPW database.

Method	Left eye corner	Left mouth corner	Nose
ASEF	81.64%	71.35%	58.39%
LL-SVM	84.53%	76.74%	67.07%
HOG + SVM	82.94%	73.25%	59.83%
Adaboost	83.24%	74.38%	65.22%
Ours	85.07%	78.28%	68.78%

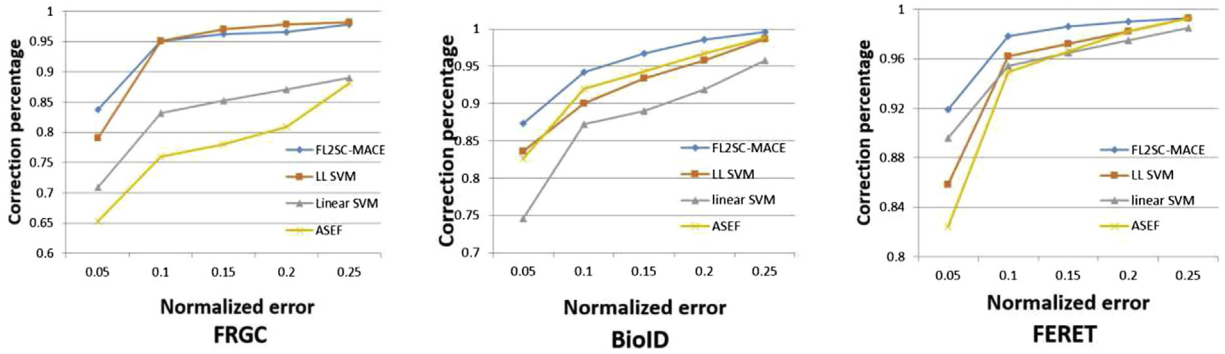

Fig. 4. Cross-set performances in FRGC FERET and BioID databases. QSL-MACE demonstrates overall advantages in both accurate and coarse localization.

Table 3
The percentage that the localization error of the landmark is less than 10% of the eye-to-eye distance on HELEN database.

Method	Left eye corner	Left mouth corner	Nose
ASEF	79.32%	64.21%	51.28%
LL-SVM	81.79%	69.90%	56.89%
HOG + SVM	78.83%	63.49%	53.84%
Adaboost	80.04%	65.72%	55.21%
Ours	82.67%	69.33%	57.33%

Table 4
The comparison of running time for detecting one eye. The number of training samples (including both positive and negative ones) accounts to 40000. All algorithm are implemented by C on Intel I7 CPU.

Method	Running times (ms)
ESP	1131
LL-SVM	853
RBF-SVM	632
Correlation filter	0.5
LS-MACE	7.5

tabases consist of unconstrained faces from Internet, it is not necessary to perform cross-set experiment for validating the generalization capability. Therefore, we combine the training sets of the databases as the generic training set. Note that we didn't compare our method with face alignment algorithms because 1) FL^2 SC-MACE detects feature points only relying on local appearance, while face alignment algorithms utilize the global face shape to rectify the results, and 2) our approach is compatible with many face alignments [16,17] by locally detecting candidate feature points. As can be seen, our algorithm outperforms the commonly used localizers. It means FL^2 SC-MACE is still effective in complex situations such as multiple face poses, expressions, disturbance and so on.

In Table 4 we compare the running speed of FL^2 SC-MACE with that of other typical methods. We conduct algorithms by C on Intel I7 CPU among which SVMs are implemented by LibSVM [14]. All localizers are trained by 2000 images, which lead to roughly 4000 SVs for kernel SVM. As can be seen, the fast testing speed of FL^2 SC-MACE enables real-time applications.

5. Conclusion

This letter proposes a correlation filter referred to as FL^2 SC-MACE. Soft constraint is imposed on the cost function of MACE to overcome the overfitting. Fourier local linear model is further introduced so that it could deal with multimode situations. We conducted cross-set eye localization experiments to demonstrate the superior performance and efficiency of our method. In future work we will seek an accurate and fast face alignment algorithm based on FL^2 SC-MACE.

Acknowledgment

This work was supported in part by the Natural Science Foundation of China under Grant No.61471216 and in part by the Special Foundation for the Development of Strategic Emerging Industries of Shenzhen under Grant No.JCYJ20150831192224146 and No.JCYJ20150601165744635.

References

[2] A. Mahalanobis, B.V.K.V. Kumar, S. Song, et al., Appl. Opt. 33 (17) (1994) 3751–3759.
[3] T.F. Cootes, C.J. Taylor, D.H. Cooper, et al., Comput. Vis. Image Underst. 61 (1) (1995) 38–59.
[4] R. Valenti, T. Gevers, Accurate eye center location and tracking using isophote curvature, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2008, pp. 1–8.
[5] Y. Ren, S. Wang, B. Hou, et al., IEEE Trans. Image Process. 23 (1) (2014) 226–239.
[6] D.S. Bolme, B. Draper, J.R. Beveridge, Average of synthetic exact filters, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2009, pp. 2105–2112.
[7] A. Rodriguez, V.N. Boddeti, B.V.K.V. Kumar, et al., IEEE Trans. Image Process. 22 (2) (2013) 631–643.
[8] C. Xie, M. Savvides, B.V.K. VijayaKumar, Kernel correlation filter based redundant class-dependence feature analysis (KCFA) on FRGC2. 0 data, in: Analysis and Modelling of Faces and Gestures, Springer Berlin Heidelberg, 2005, pp. 32–43.
[9] A. Mahalanobis, R.R. Muise, S.R. Stanfill, Appl. Opt. 43 (27) (2004) 5198–5205.
[12] A. Rodriguez, V.N. Boddeti, B.V.K.V. Kumar, et al., Maximum margin correlation filter: a new approach for localization and classification, in: IEEE Transactions on Image Processing, 2013, pp. 631–643.
[14] C.C. Chang, C.J. Lin, ACM Trans. Intell. Syst. Technol. (2011).
[15] Abhijit Mahalanobis, B.V.K. Kumar, David Casasent, Appl. Opt. 26 (1987) 3633–3640.
[16] Jason M. Saragih, Simon Lucey, Jeffrey F. Cohn, Face alignment through subspace constrained mean-shifts, in: International Conference on Computer Vision – ICCV, 2009, pp. 1034–1041.
[17] Xiaowei Zhao, Shiguang Shan, Xiujuan Chai, Xilin Chen, Cascaded Shape Space Pruning for Robust Facial Landmark Detection, ICCV, 2013.
[18] M. Türkan, M. Pardas, A.E. Cetin, Human eye localization using edge projections, in: VISAPP, 2007, pp. 410–415.
[19] Oliver Jesorsky, Klaus J. Kirchberg, Robert W. Frischholz, Robust Face Detection Using the Hausdorffdistance, Audio-and Video-based Biometric Person Authentication, 2001.
[20] D. Cristinacce, T. Cootes, Scott, A multistage approach to facial feature detection, in: Proceedings of the 15th BMVC, 2004, pp. 277–286.
[21] P.N. Belhumeur, D.W. Jacobs, D.J. Kriegman, N. Kumar, Localizing parts of faces using a consensus of exemplars, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), IEEE, Colorado Springs, USA, 2011, pp. 545–552.
[22] V. Le, J. Brandt, Z. Lin, Interactive facial feature localization, in: Proc. IEEE Eur. Conf. Comput. Vis. (ECCV), Firenze, Italy, 2012, pp. 679–692.
[23] Zhiheng Niu, et al., 2d Cascaded Adaboost for Eye Localization, ICPR, 2006.
[24] M. Savvides, B.V.K.V. Kumar, P.K. Khosla, "Corefaces"-robust Shift Invariant PCA Based Correlation Filter for Illumination Tolerant Face Recognition, CVPR, 2004.
[25] Lubor Ladicky, Philip Torr, Locally Linear Support Vector Machines, ICML, 2011.
[26] Stylianos Asteriadis, et al., An eye detection algorithm using pixel to edge information, in: Int. Symp. on Control, Commun. and Sign. Proc, 2006.
[27] Hao Zhang, et al., Discriminative Nearest Neighbor Classification for Visual Category Recognition, CVPR, 2006.
[28] Bart Kroon, Alan Hanjalic, Sander MP. Maas, Eye localization for face matching: is it always useful and under what conditions?, in: Proceedings of the 2008 International Conference on Content-based Image and Video Retrieval, ACM, 2008.

[29] R. Valenti, T. Gevers, Accurate Eye Center Location and Tracking Using Isophote Curvature, CVPR, 2008.
[30] Zhang Ziming, et al., Learning Anchor Planes for Classification, NIPS, 2011.
[31] Fei Yang, et al., Eye Localization through Multiscale Sparse Dictionaries, Automatic Face & Gesture Recognition and Workshop, 2011.
[32] M. Gönen, E. Alpaydin, Localized Multiple Kernel Learning, ICML, 2008.
[33] David S. Bolme, Bruce A. Draper, J. Ross Beveridge, Average of Synthetic Exact Filters, CVPR, 2009.
[34] Xiaoyang Tan, et al., Enhanced Pictorial Structures for Precise Eye Localization under Incontrolled Conditions, CVPR, 2009.
[35] Fabian Timm, Erhardt Barth, Accurate Eye Centre Localisation by Means of Gradients, VISAPP, 2011.
[36] J. Platt, Adv. Large Margin Classif. 10 (3) (1999) 61–74.

Wenming Yang received the B.S. and M.S. in material science from Harbin University of Science and Technology, Harbin, China in 2000 and 2003, respectively. In 2006, he received the Ph.D in electronic engineering from Zhejiang University, Hangzhou, China. From 2007 to 2009, He was a Post doctoral Fellow in Department of Electronic Engineering, Tsinghua University. Since 2013, He has been associate professor in Department of Electronic Engineering at Graduate School at Shenzhen, Tsinghua University. His research interests include image processing and pattern recognition, computer vision, biometrics, video surveillance, image super-resolution and non-destructive testing based on artificial vision.

Xiang Sun received the B.S. of Communication Engineering in 2013, Beijing University of Posts and Telecommunications, Beijing, China. In 2016, he received M.S. of electronic engineering from Tsinghua University, Beijing, China. He is now working at Big Data Inventation Centre of Creditease, Beijing, China. His research interests include face recognition and automatic credit to business.

Qingmin Liao born in 1963. He received the B.S. degree in radio technology from the University of Electronic Science and Technology of China, Chengdu, China, in 1984, and the M.S. and Ph.D. degrees in signal processing and telecommunications from the University of Rennes 1, Rennes, France, in 1990 and 1994, respectively. Since 1995, he has been joining with Tsinghua university, Bei-jing, China. He became Professor in the Department of Electronic Engineering of Tsinghua University, in 2002. From 2001 to 2003, he served as the Invited Professor with a tri-year contract at the University of Caen, France. Since 2010, he has been Director of the Division of Information Science and Technology in the Graduate School at Shenzhen, Tsinghua University, Shenzhen, China. His research interests include image/video processing, transmission and analysis; biometrics; and their applications to teledetection, medicine, industry, and sports. He has published over 90 papers internationally.