EYE LOCALIZATION BASED ON CORRELATION FILTER BANK

Shiming Ge[1,2], *Rui Yang*[1,2,3], *Hui Wen*[1,2,3], *Shuixian Chen*[1,2], *Limin Sun*[1,2]

[1]State Key Laboratory of Information Security, Institute of Information Engineering, CAS, Beijing
[2]Beijing Key Laboratory of IOT Information Security Technology, IIE, CAS
[3]University of Chinese Academy of Sciences
{geshiming, yangrui, wenhui, chenshuixian, sunlimin}@iie.ac.cn

ABSTRACT

Eye localization is a key step in many face analysis related applications. In this paper, we present a novel eye localization method based on a group of trained filters called correlation filter bank (CFB). We formulate the eye localization problem as an optimization problem with a well-defined cost function based on CFB. The CFB is trained with an EM-like adaptive clustering approach. The trained filter bank includes several discriminative filter templates, each of them suits to a different face condition from the others, thus can provide accurate eye localization ability for variable poses, appearances and illuminations. Simulation comparisons with cascade classifier-based method [1], traditional single correlation filter based methods [2][3] and pictorial structure model based method [4] demonstrates the superiority of the proposed method both in detection ratio and localization accuracy.

Index Terms— Eye Localization, Correlation Filter, Filter Bank, Regression, Adaptive Clustering

1. INTRODUCTION

Face recognition is one of the most important and concentrated problems in computer vision. By providing some basic information of the face, accurate eye localization plays a key role in face recognition and many other face analysis related applications. By contrast with eye detection, eye localization involves a more precise prediction of eye positions. According to the information that is used for model building, the existing eye localization techniques are classified into three categories [4]: characteristics based methods, statistical based methods and hybrid methods.

Characteristics based methods perform eye localization by measuring eye inhere features such as shape, contrast and context. These techniques have limitations under the complex or uncontrolled conditions due to unreliable measuring of characteristics. Statistical based methods learn statistical appearance model from a set of training images to extract useful visual features. Hybrid methods integrate structural information into statistical appearance model to improve eye localization. These methods combine the eye characteristics and the appearance under the same framework. The typical methods include pictorial structure model [4], active shape model (ASM) [5], active appearance model (AAM) [6], and so on. Hybrid methods provide a good mechanism to infer the location of an object by estimate the locations of its parts. They usually localize multiple features simultaneously. In this paper, we focus on statistical based methods.

Statistical based methods aim to find a function to discriminate eye and non-eye classes directly. In this way, the problem of eye localization boils down to a binary classification problem and the trained result is an eye classifier. A well-known classifier is proposed in [1], which uses Haar cascade classifier for face recognition [7] to locate the eyes. When an approximate location of the eye is known, this method could obtain good performance. However, detection errors happen when not having sufficient prior knowledge. Besides, classifier-based methods are set to optimize the classification accuracy rather than localization accuracy, thus the trained classifier may not give the maximal response at the right object location. One way to tackle this problem is to formulate localization task as a regression rather than a classification problem by incorporating the positions of the eyes. In this setting, the training data are given by a set of input images with the corresponding eye positions, and the training goal is to learn a regressor that maps from the input image to the predicted eye position. In [2], Bolme et al. perform regression by constructing a correlation filter that exactly transforms each training image to its correlation image, then simply average all of these exact correlation filters to obtain the final learned filter called average synthetic exact filter (ASEF). ASEF method could achieve good performance on eye localization with very low runtime cost. As an extension, in [3], Bolme et al. proposed a minimum output sum of squared error (MOSSE) filter, which can get better outcome with fewer training samples. In [8], Hefin et al. adopted a similar method to perform eye localization task by first warping the face im-

This work was supported in part by the Strategic Priority Research Program(No. XDA06040101), the Beijing Innovation and Development Program (No. Z131101002813085) and the National Key Technology R&D Program (No. 2012BAH20B03)

age through the candidate response pair and then selecting the one leading to face image with best quality. The crucial step in correlation filter based methods is filter construction. When the filter fits test samples, good outcome could be expected. However, one single correlation filter could hardly fits all the testing face images due to huge appearance variations of face in lighting, expression, pose and so on.

In this paper, we proposed a regression-based method to address eye localization task under diverse conditions. We pose the eye localization problem as an optimization problem and define an energy function with a group of correlation filter called correlation filter bank (CFB). An EM-like approach is exploited to solve the energy minimization problem by adaptively clustering training images and learning correlation filter in an alternative manner. The resulting CFB can provide good ability to locate eye for face images under variable appearances adaptively. We evaluate our method with several eye localization methods, which demonstrates the advantages of our method both in localization performance and robustness.

2. PROPOSED APPROACH

We only discuss the localization for the left eye. The case for the right eye is similar. Suppose the training data include a set of n face images $\{f_1, f_2, \ldots, f_n\}$ and the corresponding labeled positions of eyes $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$. The framework of our approach is shown in Figure 1. In training stage, the goal is to learn a group of discriminative correlation filters called CFB which maps from input face image to eye position. In this stage, an EM-like adaptive clustering method is adopted. The training images are first initially preprocessed and separated into several classes to pre-train the initial filters, then iteratively reclassified and retrained to get the final filter bank. In testing stage, the testing face image is correlated with all the correlation filters in the trained CFB and the final eye position is chosen as the location with peak response from the optimal correlation outputs. More details can be found in 2.2.

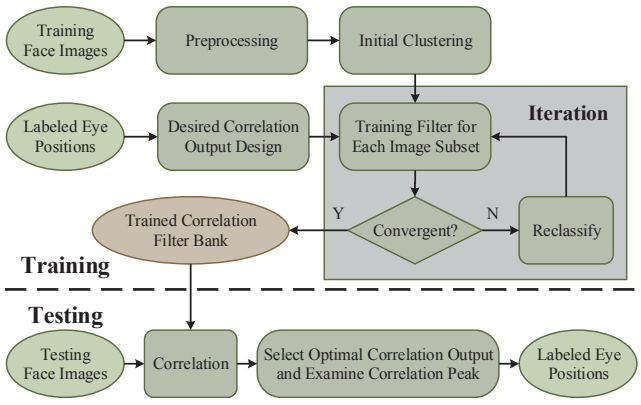

Fig. 1. Block Diagram of CFB based eye localization.

2.1. Correlation filter

Correlation filter have been widely used in pat-tern recognition field [9]. A correlation filter is a spatial-frequency array (equivalently, a template in the image do-main). With correlation filter, 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 output for possible correlation peaks. For computational efficiency, correlation operation is performed in frequency domain, i.e.

$$\hat{\boldsymbol{g}} = \hat{\boldsymbol{f}}\hat{\boldsymbol{h}}^* \tag{1}$$

where $\hat{\boldsymbol{f}}$, $\hat{\boldsymbol{g}}$ and $\hat{\boldsymbol{h}}$ are the 2D Fourier Transform of the query image f, the desired output g and filter template h respectively. The symbol * denotes complex conjugate. The desired correlation output is synthetically generated with a peak at the center of the target and (near) zero value elsewhere. In our case, it is defined with a 2D Gaussian function according to associated position of the target, in our case the left eye,

$$g_i(x, y) = e^{-[(x-x_i)^2+(y-y_i)^2]/\sigma^2} \tag{2}$$

where σ is scale parameter for controlling the sharpness of correlation output. For a single input image f_i and the corresponding output g_i , the exact correlation filter or template (in spatial domain) equivalently is obtained in frequency domain with

$$\hat{\boldsymbol{h}}_i^* = \hat{\boldsymbol{g}}_i / \hat{\boldsymbol{f}}_i = \hat{\boldsymbol{g}}_i \hat{\boldsymbol{f}}_i^* / \hat{\boldsymbol{f}}_i \hat{\boldsymbol{f}}_i^* \tag{3}$$

where the numerator is the correlation between $\hat{\boldsymbol{g}}_i$ and $\hat{\boldsymbol{f}}_i^*$, while the denominator is the energy spectrum of $\hat{\boldsymbol{f}}_i$. A set of exact correlation filters of input data can be used to construct correlation filter for general localization applications. ASEF [2] is constructed by averaging all the exact filters:

$$\hat{\boldsymbol{h}} = \sum_{i=1}^{n} \hat{\boldsymbol{h}}_i \tag{4}$$

In (4), the exact filter can be thought as a weak classifier that only performs perfectly on a particular training image, and then ASEF is a strong classifier consisting of multiple weaker ones. ASEF needs to learn with many training images to perform well. MOSSE filter [3] is proposed to get improved performance with fewer training images. Compared to ASEF filter, MOSSE filter has a more reasonable way of combining all the training images and desired output. Its goal is to find a filter that minimizes the sum of squared error between the actual correlation output and the ideal desired correlation output. It can be presented as an optimization problem as below,

$$\min_{\hat{\boldsymbol{h}}} \sum_{i=1}^{n} \left\| \hat{\boldsymbol{f}}_i \hat{\boldsymbol{h}}^* - \hat{\boldsymbol{g}}_i \right\|^2 \tag{5}$$

Both ASEF filter and MOSSE filter are suitable for localization applications since they design correlation output for mapping from each single training image to its labeled eye position, which makes the final filter flexible and discriminative

for object localization. However, training all the images to get one single correlation filter ignores the huge variance among images, which leads to the trained template work less well in complicate situations.

2.2. Correlation Filter Bank

In order to address the localization errors when using single correlation filter, we train a set of discriminative filter called CFB to adaptively handle localization under different face conditions. The CFB design can be posed as an optimization problem,

$$\min_{\boldsymbol{h}^{(j)}} \sum_{j=1}^{K} \left(\frac{1}{n_j} \sum_{i=1}^{n_j} \left\| f_i^{(j)} \otimes h^{(j)} - g_i^{(j)} \right\|^2 + \lambda \left\| h^{(j)} \right\|^2 \right) \quad (6)$$

where *K* is the number of correlation filters, n_j is the number of training images in the *j*-th subset and $n = \sum_{j=1}^{K} n_j$, $\otimes$ denotes the convolution operation and λ is the regularization parameter. The optimization problem can be solved efficiently in frequency domain where the objective function has the following closed form expression similar with [10],

$$\min_{\boldsymbol{h}^{(j)}} \sum_{j=1}^{K} E\left(\hat{\boldsymbol{h}}^{(j)}, C^{(j)}\right) \quad (7)$$

where total energy in Eq. (7) involves all the energy associated with each training subset. $E\left(\hat{\boldsymbol{h}}^{(j)}, C^{(j)}\right)$ is the energy associate with the *j*-th training subset $C^{(j)}$, and defined as

$$\begin{aligned} E\left(\hat{\boldsymbol{h}}^{(j)}, C^{(j)}\right) = & \frac{1}{n_j} \sum_{i=1}^{n_j} \hat{\boldsymbol{h}}^{(j)\dagger} \hat{\boldsymbol{F}}_i^{(j)\dagger} \hat{\boldsymbol{F}}_i^{(j)} \hat{\boldsymbol{h}}^{(j)} \\ & - \lambda \hat{\boldsymbol{h}}^{(j)\dagger} \hat{\boldsymbol{h}}^{(j)} - \frac{2}{n_j} \sum_{i=1}^{n_j} \hat{\boldsymbol{h}}^{(j)\dagger} \hat{\boldsymbol{F}}_i^{(j)\dagger} \hat{\boldsymbol{g}}_i^{(j)} \end{aligned} \quad (8)$$

where $\hat{\boldsymbol{F}}$ denotes the diagonal matrix whose diagonal entries are the elements of $\hat{\boldsymbol{f}}$, † denotes conjugate transpose, $C^{(j)}$ is the *j*-th cluster of training data. Due to energy independence among all training subsets, the solution has the following closed form expression for the CFB,

$$\begin{aligned} \hat{\boldsymbol{h}}^{(j)} = & \left[\lambda \mathbf{I} + \frac{1}{n_j} \sum_{i=1}^{n_j} \hat{\boldsymbol{F}}_i^{(j)\dagger} \hat{\boldsymbol{F}}_i^{(j)} \right]^{-1} \left[\frac{1}{n_j} \sum_{i=1}^{n_j} \hat{\boldsymbol{F}}_i^{(j)\dagger} \hat{\boldsymbol{g}}_i^{(j)} \right], \\ & j = 1, 2, \ldots, K \end{aligned} \quad (9)$$

where $\mathbf{I}$ is the identity matrix of appropriate dimensions. Solving the optimization problem includes simultaneously clustering the training images into multiple subsets and calculating the corresponding correlation filter for each subset. We use an EM-like method to perform clustering and filter calculating iteratively. All the training data need to be reclassified in clustering step with having the minimal difference between the correlation output and the desired output,

$$j_{best} = \arg\min_{j} \left\| \hat{\boldsymbol{f}} \hat{\boldsymbol{h}}^{(j)} - \hat{\boldsymbol{g}} \right\|^2 \quad (10)$$

The algorithm process is as blow:

Algorithm 1 Training CFB.

Input:

Training face images $\{f_1, f_2, \ldots, f_n\}$;
Labeled eye positions $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$
The number of filters in CFB K;
The maximal iteration time T;

Output:

Trained CFB $\{\hat{\boldsymbol{h}}^{(j)}\}$;

Training:

1: Initialize K image subsets with k-means;
2: E step: calculate $\{\hat{\boldsymbol{h}}^{(j)}\}$ with Eq. (9) according to the current clustered subsets;
3: M step: reclassified all the training data with Eq. (10);
4: Check the converge condition: finish training and output the final CFB if the maximal iteration reaches, go back to step 2 otherwise;

At testing step, the input face image first correlates with all the filters in trained CFB to get the correlation outputs with $g^{(j)} = IFFT(\hat{f}\hat{h}^{(j)})$, where $IFFT(\cdot)$ is the Inverse Fast Fourier Transform operation. The optimal correlation output $g^{(o)}$ is selected as the one with the strongest peak-to-sidelobe ratio (PSR),

$$g^{(o)} = \max_{g^{(j)}} PSR(g^{(j)}) \quad (11)$$

The PSR of a correlation output g is calculated with,

$$PSR(g) = (g_{max} - \mu_{sl})/\sigma_{sl} \quad (12)$$

where g_{max} denotes the peak value in g, μ_{sl} and σ_{sl} are the mean and standard deviation of the sidelobe respectively. Then, the eye position is selected as the position with the maximal peak value in the optimal correlation output.

3. EXPERIMENT

In order to verify the efficiency of our proposed eye localization methods, we simulated the experiments on the BioID dataset which contains 1471 images. For these experiments, the dataset was randomly partitioned into two sets. One set was used for training with 1000 images and another for testing with 471 images. The parameters were set as $K = 5$ and $T = 20$ in all experiments. The left eye normalized distance is calculated as follows:

$$D_l = \|P_l - L_l\| / \|L_l - L_r\| \quad (13)$$

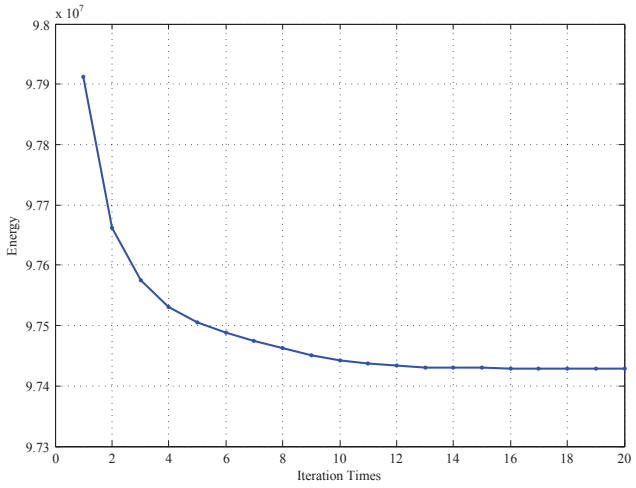

Fig. 2. Energy cost reduces against iteration times.

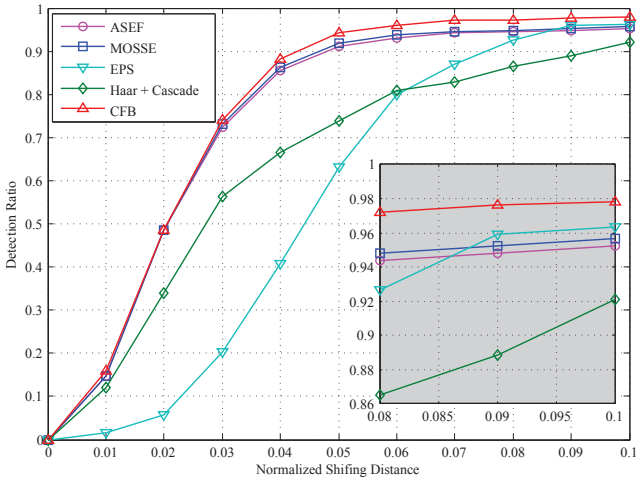

Fig. 4. Detection ratio under different D_l threshold.

where P_l is the predicted eye location estimating by the algorithm, L_l and L_r are the real labeled location of the left and right eye. The operating threshold of $D_l < 0.10$ is considered as the criteria for a successful localization. Localization accuracy is evaluated by the mean of the normalized distance, which represents how precise a localization algorithm is. Localization robustness is evaluated by the standard deviation of the normalized distance. Detection ratio is the percentage of successful localizations.

Before training, a preprocessing step was conducted similar with ASEF method [2]. All the face images are resized to 64×64 pixels, and normalized by first taking the log operation and then normalizing the pixel values to have a mean of 0.0 and energy of 1.0. To reduce the frequency effect in Fast Fourier Transform (FFT) operation, a cosine window is applied to the image. In the end, a random similarity transform is operated to enlarge the database and improve the robustness. The similarity transform contains rotates by up to $\pm 15°$, scales by up to 1.0 ± 0.1 and shifts by up to ± 4 pixels. Each training image is randomly perturbed 16 times which results in 16,000 training examples.

The initialization of training method was performed as follows: an 11×15 region centered at the eye position is cropped from each training images to conduct k-means clustering. Besides, to improve the training performance, some examples are removed out due to huge deviation measuring with L2 norm. Figure 2 shows that the total energy defined in equations (7) and (8) is reducing when iteration grows. The resulting CFB includes a group of filters, which is shown in Figure 3. The training process adaptively divides the training set into several clusters according to similar appearance and yields several discriminative templates.

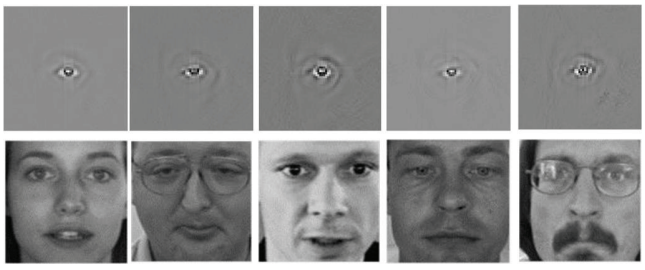

Fig. 3. The trained CFB and their corresponding examples.

Figure 4 gives the experimental comparisons of our proposed CFB method with other methods including Haar classifier based method [1], ASEF filter method [2], MOSSE filter method [3] and enhanced pictorial structure method (EPS) [4]. The grey part in the figure shows the details when normalized distance threshold D_l is above 0.08. We evaluate the detection ratio (percent of successful localization) under different normalized distance threshold. The evaluation examines the robustness of eye localization methods. It shows that our proposed method has the highest detection ratio under the same D_l threshold.

Table 1 gives the results in detail. The detection ratio is evaluated when normalized distance threshold $D_l < 0.10$. Mean and standard deviation represents the average error of normalized shifting distance D_l in (13). It can be seen that

Table 1. Performance comparisons of five methods

Method	Detection Ratio	Mean	Standard Deviation
Haar+Cascade[1]	92.10%	0.0865	0.3617
ASEF[2]	95.27%	0.0631	0.2008
MOSSE[3]	95.70%	0.0532	0.1750
EPS[4]	96.34%	0.0494	0.1349
CFB	**97.85%**	**0.0395**	**0.1304**

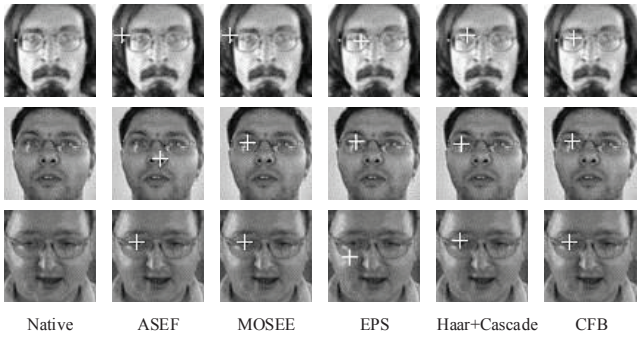

Fig. 5. Comparison of the localization results on BioID.

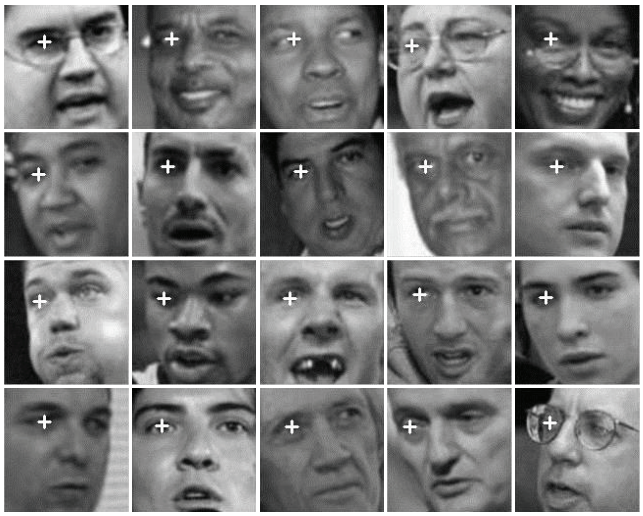

Fig. 6. Examples of the localization results on LFW.

our method outperforms the others not only in detection ratio, but also in localization accuracy. Figure 5 gives some visual examples of the localization results. It shows CFB have advantages in these complex conditions comparing with other four methods.

Besides the experiments on BioID database, to check robustness of our algorithm, we run extra experiments on LFW database, which contains more pose changes. The CFB is trained with the images from LFW database. The results of the experiments are quite promising, which shows that our algorithm can handle the eye localization with pose changes. Figure 6 gives some samples of the results.

4. CONCLUSION

This paper proposed an eye localization approach with adaptively constructed correlation filter bank. The eye localization problem is posed as an optimization problem with a well-defined energy function. The solution involves adaptive clustering of face images and correlation filter construction in a unified framework. We exploit an EM-like method to get the final correlation filter bank. The trained CFB contains a group of discriminative correlation filters that could handle eye localization task under variable face appearance. Experimental comparisons with other methods show the superiority of the proposed method both in detection ratio and localization accuracy. Next step we want to incorporate the structure information into our CFB method, exploit it in more challenge circumstances and extend the applications to other fields, such as car alignment [10], biometric security [11], and so on.

5. REFERENCES

[1] M. C. Santana, J. L. Navarro, O. D. Suarez, and et.al, "Multiple face detection at different resolutions for perceptual user interfaces," *Lecture Note on Computer Science, Pattern Recognition and Image Analysis*, pp. 445–452, 2005.

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

[3] D. S. Bolme, J. R. Beveridge, and B. A. Draper, "Visual object tracking using adaptive correlation filtersc," *CVPR*, pp. 2544–2550, 2010.

[4] X. Tan, F. Song, Z. Zhou, and S. Chen, "Enhanced pictorial structures for preciseeye localization under uncontrolled conditions," *CVPR*, pp. 1621–1628, 2009.

[5] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," *CVPR*, pp. 2887–2894, 2012.

[6] I. Matthews and S. Baker, "Active appearance models revisited," *IJCV*, vol. 60, pp. 135–164, 2004.

[7] P. Viola and M. J. Jones, "Robust real-time face detection," *IJCV*, vol. 57, pp. 137–154, 2004.

[8] B. Heflin, W. Scheirerc, and T. E. Boult, "For your eyes only," *IEEE Workshop on the Applications of Computer Vision*, pp. 193–200, 2012.

[9] B. V. Kumar, A. Mahalanobis, and R. Juday, "Correlation pattern recognition," Cambridge University Press, 2005.

[10] V. N. Boddeti, T. Kanade, and B. V. Kumar, "Correlation filters for object alignment," *CVPR*, pp. 2291–2298, 2013.

[11] V. N. Boddeti and B. V. Kumar, "A framework for binding and retrieving class-specific information to and from image patterns using correlation filters," *PAMI*, vol. 35, pp. 2064–2077, 2013.