Maximum Margin Correlation Filter: A New Approach for Localization and Classification

Andres Rodriguez, *Student Member, IEEE*, Vishnu Naresh Boddeti, *Student Member, IEEE*,
B. V. K. Vijaya Kumar, *Fellow, IEEE* and Abhijit Mahalanobis, *Senior Member, IEEE*

***Abstract*—Support vector machine (SVM) classifiers are popular in many computer vision tasks. In most of them, the SVM classifier assumes that the object to be classified is centered in the query image which might not always be valid, e.g., when locating and classifying a particular class of vehicles in a large scene. In this paper we introduce a new classifier called Maximum Margin Correlation Filter (MMCF), which while exhibiting the good generalization capabilities of SVM classifiers is also capable of localizing objects of interest, thereby avoiding the need for image centering as is usually required in SVM classifiers. In other words, MMCF can simultaneously localize and classify objects of interest. We test the efficacy of the proposed classifier on three different tasks: vehicle recognition, eye localization, and face classification. We demonstrate that MMCF outperforms SVM classifiers and also well-known correlation filters.**

I. Introduction

The tasks of object (we use *object* and *target* interchangeably throughout this paper) localization and classification are important in various applications such as automatic target recognition (ATR), biometric recognition, etc. In this paper by *localization* we refer to estimating the location of an object in the scene, by *classification* we refer to determining the class label of a particular object, and by *recognition* we refer to performing both tasks (of localization and classification). Two well-known types of classifiers used for these tasks are support vector machines (SVMs) and correlation filters (CFs).

SVM classifiers [1], [2], [3] (referred to as SVMs throughout this paper) have been investigated for vision tasks such as face localization [4] and pedestrian localization [5]. SVMs are often designed by extracting features from the training images and then using a feature vector to represent an image. When using pixel values as features, the image is lexicographically scanned to form a feature vector. Given N of these training column vectors $\mathbf{x}_i \in \mathbb{R}^d$ and class labels $t_i \in \{-1, 1\}$ $\forall i \in \{1, \cdots, N\}$, the SVM approach (for a 2-class problem) finds the hyperplane that maximizes the smallest L-2 norm distance between the hyperplane and any data sample (also called the margin) by solving

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}\xi_i \\ s.t. \quad & t_i(\mathbf{x}_i^T\mathbf{w} + b) \geq 1 - \xi_i, \end{aligned} \tag{1}$$

where superscript T denotes transpose, $\mathbf{w}$ and b represent the hyperplane ($\mathbf{w}$ denotes the normal to the hyperplane and b is the bias or offset from the origin), $C > 0$ is a trade-off parameter, and the sum of $\xi_i \geq 0$ is a penalty term containing the *slack variables* which offset the effects of outliers. It can be shown [6] that minimizing the squared L-2 norm of $\mathbf{w}$ subject to the above inequality constraints is equivalent to maximizing the margin, and that the solution to Eq. 1 is a linear combination of the training vectors, i.e.,

$$\mathbf{w} = \sum_{i=1}^{N} a_i\mathbf{x}_i = \mathbf{X}\mathbf{a}, \tag{2}$$

where $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N]$, $\mathbf{a} = [a_1, \cdots, a_N]^T$, $\sum_{i=1}^{N} a_i = 0$, $0 \leq a_i \leq C$ $\forall i$ corresponding to class label $t_i = 1$, and $-C \leq a_i \leq 0$ $\forall i$ corresponding to class label $t_i = -1$. The training vectors corresponding to non-zero coefficients a_i are known as the support vectors.

Assuming that training vectors contain pixels values (i.e., images with d pixels lexicographically rearranged into d-dimensional column vectors), one can use the resulting d-dimensional solution vector $\mathbf{w}$ for simultaneous object localization and classification by cross-correlating the 2-D template represented by $\mathbf{w}$, with the query image. Note that the training vectors can represent features other than pixels values, and in that case we cross-correlate the template represented by $\mathbf{w}$ with the features extracted from the query image. Either way, since the template is not optimized to produce sharp correlation peaks (i.e., peaks in the correlation output), the resulting correlation output usually exhibits very broad peaks. Broad peaks result in poor object localization because 1) the top of the peak may be spread over several pixels and 2) in the presence of multiple objects in the scene, the peaks from different objects might overlap, leading to peaks being in wrong locations.

CFs [7] have also been investigated for object recognition. Attractive properties of CFs such as shift-invariance, noise robustness, graceful degradation, and distortion tolerance have been useful in a variety of pattern recognition applications including face localization [8], pedestrian localization [9],

Manuscript received October 18, 2011; revised May 23, 2012; accepted September 12, 2012. This work was supported in part by the U.S. Air Force Research Laboratory.

A. Rodriguez (corresponding author), V. N. Boddeti and B. V. K. Vijaya Kumar are with the Department of Electrical and Computer Engineering at Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213 (phone: 412-268-3026; fax: 412-268-6345; e-mail: {afrodrig, vboddeti, kumar}@ece.cmu.edu).

A. Mahalanobis is with Lockheed Martin Corporation, MP 450, 5600 Sandlake Road, Orlando, FL 32819 (e-mail: abhijit.mahalanobis@lmco.com).

object localization and tracking [10], biometric recognition [11], [12], and vehicle recognition [13]. In this approach, a carefully designed template (loosely called a *filter*) $w(p,q)$ is cross-correlated with the query image $x(p,q)$ to produce the output $g(\tau_x,\tau_y)$. This operation is efficiently carried out in the frequency domain via Fourier transforms (FTs) as follows,

$$\hat{g} = \hat{w} \circ \hat{x}^*, \tag{3}$$

where superscript $*$ denotes complex conjugate, $\circ$ denotes the Hadamard product, and $\hat{g}$, $\hat{x}$ and $\hat{w}$ are the 2-D FTs of the correlation output, the query image and the template, respectively. When the query image is from the true-class (also called authentic or Class-1), $g(\tau_x,\tau_y)$ should exhibit a sharp peak, and when the query image is from the false-class (also called impostor or Class-2) $g(\tau_x,\tau_y)$ should have no such discernible peak. The sharper the peak (i.e., the larger the peak compared to the surrounding values), the greater the probability that the query image is from the true-class, and the location of the peak indicates the location of the target. Thus, CFs offer the ability to simultaneously localize and classify objects of interest. We review some well-known CF designs in Section II-A.

While SVMs are designed to maximize the margin and thus usually offer good generalization (i.e., they usually offer good classification performance for centered images outside the training set), they exhibit poor localization because the peaks resulting from cross-correlation of SVM templates with test images are not sharp. In contrast, CFs can produce sharp peaks and thus offer good localization performance, but they are not explicitly designed to offer good generalization. In this paper we combine the design principles of SVMs and CFs leading to a new classifier called *Maximum Margin Correlation Filter* (MMCF) which has the good generalization capability of SVMs and the good localization capability of CFs. The MMCF template leads to a more distinguishable peak in the correlation outputs than the SVM template. We will show through numerical experiments on different databases that MMCF is able to simultaneously localize and classify objects of interest with improved performance over SVMs and well-known CFs. Figs. 1 and 2 show a comparison of the correlation outputs of SVM and MMCF, respectively, for the same target class and test image. The MMCF output in Fig. 2 is able to localize the desired class tank image in the scene more accurately than the SVM output in Fig. 1. The details of how these outputs are obtained are explained in Section VI.

II. Related Work

There are some previous approaches that attempt to achieve shift-invariant classification. Most of these approaches cross-correlate the template with the query image and are sometimes known as *sliding window* algorithms. We now discuss these approaches and how our approach differs from them.

Scholkopf et al. [14] proposed a method to achieve shift-invariance by training an SVM on centered images, generating shifted images of the support vectors and re-training the SVM. Decoste et al. [15] described different algorithms for training shift-invariant SVMs, and Chapelle et al. [16] proposed algorithms to incorporate shift-invariance in non-linear SVMs. These methods include shift-invariance constraints explicitly as inequalities, i.e.,

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}\sum_{j=1}^{d}\xi_i^j \\ s.t. \quad & t_i(\mathbf{w}^T\mathbf{x}_i^j + b) \geq 1 - \xi_i^j, \end{aligned} \tag{4}$$

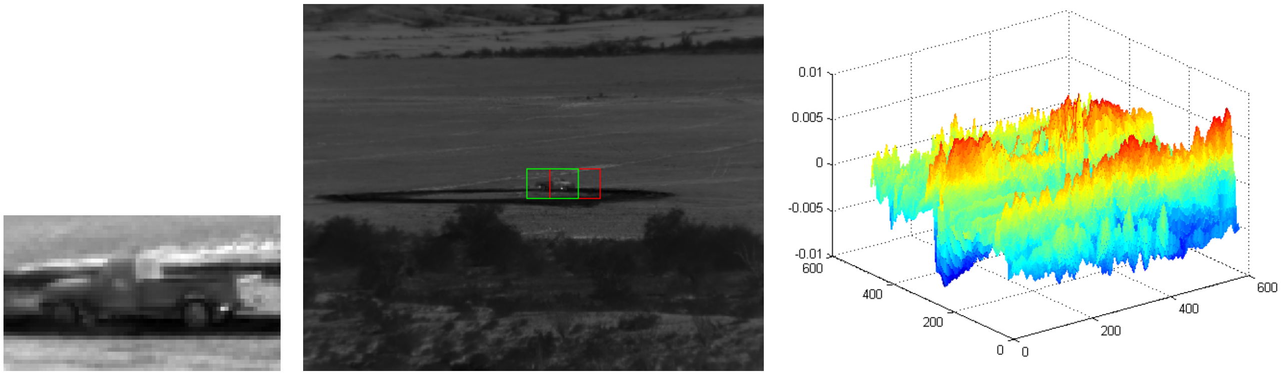
where $\mathbf{x}_i^j$ is the image $\mathbf{x}_i$ shifted by j pixels. For images, j refers to shifts in both the x- and y-directions. In these methods, the number of constraints gets multiplied by the number of shifts making the complexity of these methods prohibitive for large number of shifts. In fact, the approach in [14], [15], [16] make the classifier invariant to just 1 or 2 pixel shifts in the images, and hence precise object localization is still very challenging. In contrast, our proposed method exhibits invariance to arbitrary shifts.

Dalal and Triggs [5] proposed cross-correlating the 2-D template (represented by the vector $\mathbf{w}$) with training images, adding the false positives as false-class training images, and retraining the template. We observe that this retraining method greatly improves the performance of both SVM and MMCF.

Shivaswamy et al. [17] recently showed that the *type* of margin (e.g., L-2 norm margin) maximized is important while designing maximum margin classifiers. Ashraf et al. [18] maximized a non-Euclidean margin for their task of applying Gabor filters in a lower dimensional feature space. We mention their work since we also optimize a non-Euclidean margin in the MMCF formulation. We, however, are motivated by peak sharpness criterion which improves object localization performance, while Ashraf et al.'s work is motivated by reduction in computational complexity of designing classifiers on potentially infinite dimensional features extracted from infinite number of Gabor filters.

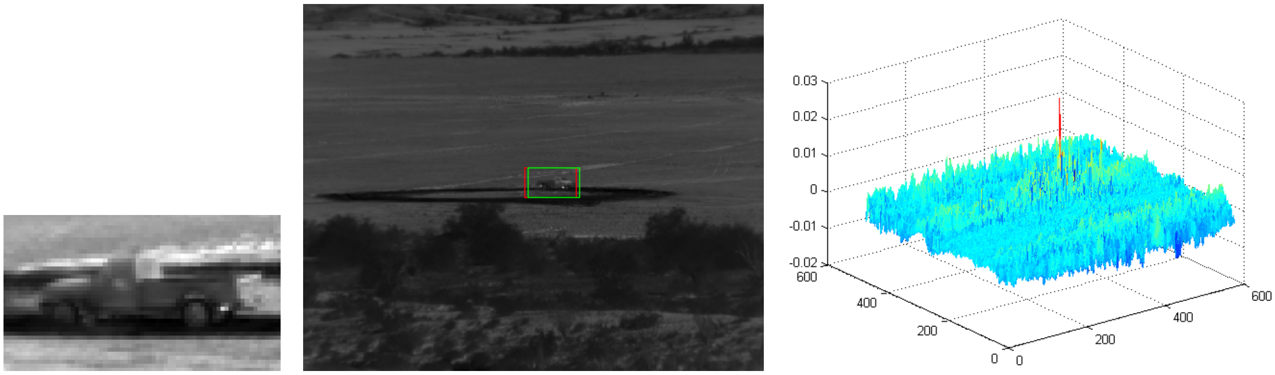
Thornton et al. [19] proposed what they called SVM Correlation Filter, but their work is very different from that proposed in this paper. They simply treat shifted versions of the true-class training images as virtual false-class training images, which does not scale well with the number of training images, i.e., the number of false-class training images gets multiplied by the number of shifts making the complexity of these methods prohibitive for large number of shifts. Moreover, they do not deal with actual false-class training images as well as their shifted versions which, if included, could further increase the complexity of the problem making the optimization problem intractable. In contrast, our proposed method exhibits invariance to arbitrary shifts without using shifted images as false-class training images.

Kumar et al. [20] proposed a CF design that provides some rotation response control, and Kerekes et al. [21] proposed a CF design that provides some scale response control. These designs, however, sacrifice some recognition performance by ignoring some of the circular harmonics and Mellin radial harmonics in order to achieve some rotation and scale control, respectively. In [22] Lowe proposed the popular scale-invariant feature transform (SIFT) features which achieves scale invariance. However, the SIFT approach is not aimed at localization.



(a) Pickup truck (b) Target localization (c) Correlation Output

Figure 1. The SVM template response (in (c)) to the test image (512×640 pixels) (in (b)). The SVM template is designed (using retraining [5]) to produce a positive value for the pickup truck image (70×40 pixels) (in (a)) and negative values for background. This vehicle appears in the test image (see the green box in the test image) but the correlation output does not show any noticeable sharp peak. The red box shows the window corresponding to the highest correlation value.



(a) Pickup truck (b) Target localization (c) Correlation Output

Figure 2. The MMCF response (in (c)) to the test image (in (b)). The MMCF is designed (using retraining [5]) to produce a large value for the pickup truck image (in (a)) and small values for background. The green box shows the ground truth target window, and the red box show the window at the highest correlation value. There is considerable overlap between the windows. In comparison to Fig. 1 there is distinguishable sharp peak leading to a more accurate object localization.

A. *Correlation Filters*

We will briefly introduce the best known CFs. A detailed review of CFs can be found elsewhere [7]. Because CFs can be used for classification, we loosely use *classifier* and *filter* interchangeably. Many CF designs can be interpreted as optimizing a distance metric between an ideal desired correlation output for an input image and the correlation output of the training images with the filter template, i.e.,

$$\mathbf{w} = \underset{\mathbf{w}}{\operatorname{argmin}} \sum_{i=1}^{N} \|\mathbf{w} \otimes \mathbf{x}_i - \mathbf{g}_i\|_2^2, \tag{5}$$

where the $\otimes$ symbol denotes the implied 2-D cross-correlation operation of the 2-D input image and the template represented by their vectors versions $\mathbf{x}_i$ and $\mathbf{w}$, respectively, and $\mathbf{g}_i$ is the vector versions of the desired correlation output. Fig. 3 shows the desired correlation output to an input image correlated with the filter template. We now discuss the two main kinds of CF designs, namely unconstrained CFs (optimizing Eq. 5 for different forms of $\mathbf{g}_i$) and constrained CFs (optimizing Eq. 5 along with additional constraints on the correlation value at the object's location).

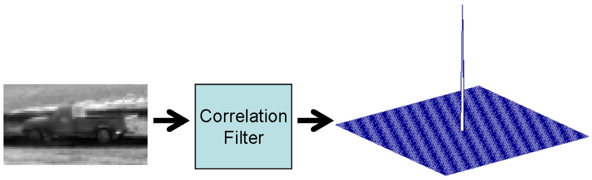


Figure 3. The desired output (right) to an input image (left) correlated with the filter template. CFs can be designed to output a sharp peak when the input is the desired target.

Four popular unconstrained CFs are the Unconstrained Minimum Average Correlation Energy (UMACE) filter, the Minimum Output Sum of Squared Error (MOSSE) filter, the Average of Synthetic Exact Filter (ASEF), and the Unconstrained Optimal Trade-off Synthetic Discriminant Function (UOTSDF) filter. The UMACE [23] filter is designed using Eq. 5 with $\mathbf{g}_i = [0, \cdots, 0, 1, 0, \cdots, 0]^T$ with a one at the target location and zeros everywhere else, meaning a sharp (delta function-like) peak is desired at the target's location. This filter, however, overfits to the training images and does not produce good correlation peaks in response to images outside the training set. The MOSSE [10] filter is also designed using

Eq. 5 with $g_i(x) = \exp((x-\mu)^2/(2\sigma^2))$ meaning a Gaussian function-like correlation shape centered at the target's location is desired. Using a Gaussian function-like shape instead of a delta function-like shape is one approach that can improve performance for images outside the training set. The ASEF [8] filter also uses $g_i(x) = \exp((x-\mu)^2/(2\sigma^2))$. The difference between ASEF and MOSSE is that ASEF computes one filter template $\mathbf{w}_i$ per image and then averages them, i.e.,

$$\mathbf{w} = \frac{1}{N}\sum_{i=1}^{N} \underset{\mathbf{w}_i}{\text{argmin}} \|\mathbf{w}_i \otimes \mathbf{x}_i - \mathbf{g}_i\|_2^2. \tag{6}$$

The UOTSDF [24] filter is another approach to improve the performance for images outside the training set. The filter is designed with $\mathbf{g}_i = [0, \cdots, 0, 1, 0, \cdots, 0]^T$ and has an additional penalty term $\gamma||\mathbf{w}||_2^2$ in the objective function, i.e.,

$$\mathbf{w} = \underset{\mathbf{w}}{\text{argmin}} \left(\sum_{i=1}^{N} \|\mathbf{w} \otimes \mathbf{x}_i - \mathbf{g}_i\|_2^2 + \gamma||\mathbf{w}||_2^2 \right), \tag{7}$$

where $\gamma \geq 0$. It can be shown [25] that this penalty term represents the output noise variance when the input is corrupted by additive white noise. Including $\gamma||\mathbf{w}||_2^2$ makes the filter more robust to noise which can improve classification performance. For convenience, we re-formulate Eq. 7 as follows,

$$\mathbf{w} = \underset{\mathbf{w}}{\text{argmin}} \left((1-\lambda) \sum_{i=1}^{N} \|\mathbf{w} \otimes \mathbf{x}_i - \mathbf{g}_i\|_2^2 + \lambda||\mathbf{w}||_2^2 \right), \tag{8}$$

where $\lambda = \frac{\gamma}{1+\gamma}$ is $0 \leq \lambda \leq 1$ to avoid having the variable γ with no upper limit.

Constrained CFs constrain the peak correlation output to be equal to a certain value. Two popular constrained CFs are the Minimum Average Correlation Energy (MACE) [23] filter, and the Optimal Trade-off Synthetic Discriminant Function (OTSDF) [26] filter. These filters are designed using Eqs. 5 and 8, respectively, with the additional constraint of $\mathbf{w}^T\mathbf{x}_i = c_i$, where $c_i = 1$ when $\mathbf{x}_i$ represents a true-class training image and $c_i = 0$ when $\mathbf{x}_i$ represents a false-class training image. These constraints ensure a desired output at the target's location for training images.

III. Maximum Margin Correlation Filters

The Maximum Margin Correlation Filter (MMCF) classifier combines the design principles of SVMs and CFs. In addition to maximizing the smallest Euclidean distance between the hyperplane and data points (i.e., the margin), we also want to minimize the mean square error $\|\mathbf{w} \otimes \mathbf{x}_i - \mathbf{g}_i\|_2^2$ (see Eq. 5) where we choose $\mathbf{g}_i$ to be a delta function-like in order to have a sharp peak in the correlation output at the target's center location (hereinafter referred to as the target's location) to improve the localization capability of SVMs. For this purpose we write the MMCF multi-objective function as follows,

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \left(\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N} \xi_i, \sum_{i=1}^{N} \|\mathbf{w} \otimes \mathbf{x}_i - \mathbf{g}_i\|_2^2 \right) \\ s.t. \quad & t_i(\mathbf{w}^T\mathbf{x}_i + b) \geq c_i - \xi_i, \end{aligned} \tag{9}$$

where $c_i = 1$ for true-class images and $c_i = 0$ or other small value ε for false-class images. That is, for true-class images, we expect a value near 1 and for false-class we expect a value that is close to 0. This allows us to detect the true targets and ignore everything else. We refer to $\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N} \xi_i$ as the margin criterion and $\sum_{i=1}^{N} \|\mathbf{w} \otimes \mathbf{x}_i - \mathbf{g}_i\|_2^2$ as the localization criterion. The smaller the value of $\mathbf{w}^T\mathbf{w}$, the large the margin [6]; a large margin usually results in better generalization and classification performance. The smaller the value of $\sum_{i=1}^{N} \|\mathbf{w}\otimes\mathbf{x}_i - \mathbf{g}_i\|_2^2$, the sharper the correlation peak (assuming $\mathbf{g}_i$ is a delta function-like); a sharp peak usually results in better localization performance. We set the desired CF output to $\mathbf{g}_i = [0, \cdots, 0, \mathbf{w}^T\mathbf{x}_i, 0, \cdots, 0]^T$ where we use $\mathbf{w}^T\mathbf{x}_i$ as the cross-correlation value of $\mathbf{w}$ and $\mathbf{x}_i$ at the target's location. In other words, we desire a peak centered at the target's location and zeros everywhere else for better object localization.

We express the multi-objective function shown in Eq. 9 in the frequency domain in order to take advantage of the well-known property that cross-correlation in the spatial domain is equivalent to multiplication in the frequency domain. The localization criterion can be expressed in the frequency domain as follows (using *Parseval's Theorem* [27]),

$$\begin{aligned} \sum_{i=1}^{N} \|\mathbf{w} \otimes \mathbf{x}_i - \mathbf{g}_i\|_2^2 &= \frac{1}{d}\sum_{i=1}^{N} \|\hat{\mathbf{X}}_i^*\hat{\mathbf{w}} - \hat{\mathbf{g}}_i\|_2^2 \\ &\propto \sum_{i=1}^{N} \|\hat{\mathbf{X}}_i^*\hat{\mathbf{w}} - \hat{\mathbf{g}}_i\|_2^2, \end{aligned} \tag{10}$$

where d is the dimensionality of $\mathbf{x}_i$, $\hat{\mathbf{X}}_i$ is a diagonal matrix whose diagonal entries are the elements of $\hat{\mathbf{x}}_i$, and $\hat{\mathbf{x}}_i$, $\hat{\mathbf{w}}$, and $\hat{\mathbf{g}}_i$ are the vector representations of the 2-D discrete Fourier transforms (DFTs) of the 2-D images represented by the vectors $\mathbf{x}_i$, $\mathbf{w}$, and $\mathbf{g}_i$, respectively. We ignore the scalar $\frac{1}{d}$ because it does not affect the value of $\hat{\mathbf{w}}$ that minimizes the localization criterion. The vector $\hat{\mathbf{g}}_i$ can be expressed as

$$\hat{\mathbf{g}}_i = \mathbf{1}\mathbf{x}_i^T\mathbf{w} = \frac{1}{d}\mathbf{1}\hat{\mathbf{x}}_i^\dagger\hat{\mathbf{w}}, \tag{11}$$

where superscript $\dagger$ denotes conjugate transpose, and $\mathbf{1} = [1 \cdots 1]^T$ with d ones in it. The vector $\hat{\mathbf{x}}_i$ can be expressed as $\hat{\mathbf{x}}_i = \hat{\mathbf{X}}_i\mathbf{1}$. We expand Eq. 10 to get the following expression,

$$\begin{aligned} \sum_{i=1}^{N} \|\hat{\mathbf{X}}_i^*\hat{\mathbf{w}} - \hat{\mathbf{g}}_i\|_2^2 &= \sum_{i=1}^{N} \left(\hat{\mathbf{w}}^\dagger\hat{\mathbf{X}}_i\hat{\mathbf{X}}_i^*\hat{\mathbf{w}} - 2\hat{\mathbf{w}}^\dagger\hat{\mathbf{X}}_i\hat{\mathbf{g}}_i + \hat{\mathbf{g}}_i^\dagger\hat{\mathbf{g}}_i \right) \\ &= \sum_{i=1}^{N} \left(\hat{\mathbf{w}}^\dagger\hat{\mathbf{X}}_i\hat{\mathbf{X}}_i^*\hat{\mathbf{w}} - \frac{2}{d}\hat{\mathbf{w}}^\dagger\hat{\mathbf{X}}_i\mathbf{1}\hat{\mathbf{x}}_i^\dagger\hat{\mathbf{w}} + \frac{1}{d^2}\hat{\mathbf{w}}^\dagger\hat{\mathbf{x}}_i\mathbf{1}^\dagger\mathbf{1}\hat{\mathbf{x}}_i^\dagger\hat{\mathbf{w}} \right) \\ &= \sum_{i=1}^{N} \left(\hat{\mathbf{w}}^\dagger\hat{\mathbf{X}}_i\hat{\mathbf{X}}_i^*\hat{\mathbf{w}} - \frac{2}{d}\hat{\mathbf{w}}^\dagger\hat{\mathbf{x}}_i\hat{\mathbf{x}}_i^\dagger\hat{\mathbf{w}} + \frac{1}{d}\hat{\mathbf{w}}^\dagger\hat{\mathbf{x}}_i\hat{\mathbf{x}}_i^\dagger\hat{\mathbf{w}} \right) \\ &= \hat{\mathbf{w}}^\dagger \left(\sum_{i=1}^{N} \hat{\mathbf{X}}_i\hat{\mathbf{X}}_i^* - \frac{1}{d}\sum_{i=1}^{N} \hat{\mathbf{x}}_i\hat{\mathbf{x}}_i^\dagger \right) \hat{\mathbf{w}} = \hat{\mathbf{w}}^\dagger\hat{\mathbf{Z}}\hat{\mathbf{w}}, \end{aligned} \tag{12}$$

where

$$\hat{\mathbf{Z}} = \sum_{i=1}^{N} \hat{\mathbf{X}}_i\hat{\mathbf{X}}_i^* - \frac{1}{d}\sum_{i=1}^{N} \hat{\mathbf{x}}_i\hat{\mathbf{x}}_i^\dagger = \hat{\mathbf{D}} - \hat{\mathbf{Y}}\hat{\mathbf{Y}}^\dagger, \tag{13}$$

where $\hat{\mathbf{D}} = \sum_{i=1}^{N} \hat{\mathbf{X}}_i \hat{\mathbf{X}}_i^*$ is a diagonal matrix and $\hat{\mathbf{Y}} = \frac{1}{\sqrt{d}}[\hat{\mathbf{x}}_1, \cdots, \hat{\mathbf{x}}_N]$. Note from Eq. 12 that $\hat{\mathbf{w}}^\dagger \hat{\mathbf{Z}} \hat{\mathbf{w}}$ is non-negative for any $\hat{\mathbf{w}}$, and thus $\hat{\mathbf{Z}}$ is positive semidefinite.

We can formulate the SVM in the frequency domain making use of the fact that inner products are only scaled by $\frac{1}{d}$ [27]. Thus, the SVM frequency domain formulation is as follows,

$$\begin{aligned} \min_{\hat{\mathbf{w}}, b'} \quad & \hat{\mathbf{w}}^\dagger \hat{\mathbf{w}} + C \sum_{i=1}^{N} \xi_i \\ s.t. \quad & t_i(\hat{\mathbf{w}}^\dagger \hat{\mathbf{x}}_i + b') \geq 1 - \xi_i, \end{aligned} \tag{14}$$

where $b' = b \times d$, and the other $\frac{1}{d}$ scalars are ignored because everything gets scaled appropriately.

We can now express the MMCF multi-criteria shown in Eq. 9 in the frequency domain as follows,

$$\begin{aligned} \min_{\hat{\mathbf{w}}, b'} \quad & \left(\hat{\mathbf{w}}^\dagger \hat{\mathbf{w}} + C \sum_{i=1}^{N} \xi_i, \hat{\mathbf{w}}^\dagger \hat{\mathbf{Z}} \hat{\mathbf{w}} \right) \\ s.t. \quad & t_i(\hat{\mathbf{w}}^\dagger \hat{\mathbf{x}}_i + b') \geq c_i - \xi_i, \end{aligned} \tag{15}$$

where $\hat{\mathbf{w}}^\dagger \hat{\mathbf{w}} + C \sum_{i=1}^{N} \xi_i$ is the margin criterion and $\hat{\mathbf{w}}^\dagger \hat{\mathbf{Z}} \hat{\mathbf{w}}$ is the localization criterion.

Refregier [28] showed that two quadratic criteria are optimized (i.e., the solution yields the best for one criterion for a fixed value of the other) by minimizing a weighted sum of the two criteria (see also [29]), i.e.,

$$\begin{aligned} \min_{\hat{\mathbf{w}}, b'} \quad & \hat{\mathbf{w}}^\dagger \hat{\mathbf{w}} + C \sum_{i=1}^{N} \xi_i + \delta \hat{\mathbf{w}}^\dagger \hat{\mathbf{Z}} \hat{\mathbf{w}} \\ s.t. \quad & t_i(\hat{\mathbf{w}}^\dagger \hat{\mathbf{x}}_i + b') \geq c_i - \xi_i, \end{aligned} \tag{16}$$

where $\delta \geq 0$. For convenience in our experiments, we reformulate Eq. 16 as follows,

$$\begin{aligned} \min_{\hat{\mathbf{w}}, b'} \quad & \lambda \hat{\mathbf{w}}^\dagger \hat{\mathbf{w}} + C' \sum_{i=1}^{N} \xi_i + (1 - \lambda) \hat{\mathbf{w}}^\dagger \hat{\mathbf{Z}} \hat{\mathbf{w}} \\ s.t. \quad & t_i(\hat{\mathbf{w}}^\dagger \hat{\mathbf{x}}_i + b') \geq c_i - \xi_i, \end{aligned} \tag{17}$$

where $\lambda = \frac{1}{1+\delta}$ is $0 < \lambda \leq 1$ to avoid having a variable δ with no upper limit and $C' = \lambda C$. Subsuming one quadratic term into the other quadratic term we rewrite Eq. 17 as follows,

$$\begin{aligned} \min_{\hat{\mathbf{w}}, b'} \quad & \hat{\mathbf{w}}^\dagger \hat{\mathbf{S}} \hat{\mathbf{w}} + C' \sum_{i=1}^{N} \xi_i \\ s.t. \quad & t_i(\hat{\mathbf{w}}^\dagger \hat{\mathbf{x}}_i + b') \geq c_i - \xi_i, \end{aligned} \tag{18}$$

where $\hat{\mathbf{S}} = \lambda \mathbf{I} + (1 - \lambda) \hat{\mathbf{Z}}$. Here λ is the parameter which trades-off margin (i.e., L-2 norm margin maximization between the centered true-class and false-class training images) and object localization. Setting $\lambda = 1$ will ignore the localization criterion and result in the conventional SVM classifier for centered images. Therefore the SVM objective function is a special case of this more general MMCF objective function with $\lambda = 1$. Smaller values of λ improve object localization by having sharper peaks in the correlation output.

Since $0 < \lambda \leq 1$, $\hat{\mathbf{S}}$ is a positive definite matrix, and we can transform the data such that $\tilde{\mathbf{w}} = \hat{\mathbf{S}}^{\frac{1}{2}} \hat{\mathbf{w}}$ and $\tilde{\mathbf{x}}_i = \hat{\mathbf{S}}^{-\frac{1}{2}} \hat{\mathbf{x}}_i$ and rewrite the MMCF criterion as follows,

$$\begin{aligned} \min_{\tilde{\mathbf{w}}, b'} \quad & \tilde{\mathbf{w}}^\dagger \tilde{\mathbf{w}} + C' \sum_{i=1}^{N} \xi_i \\ s.t. \quad & t_i(\tilde{\mathbf{w}}^\dagger \tilde{\mathbf{x}}_i + b') \geq c_i - \xi_i. \end{aligned} \tag{19}$$

This means that we can implement the MMCF design using a standard SVM solver by using transformed images to find $\tilde{\mathbf{w}}$.

IV. Implementation

A. Sequential minimal optimization

The dual of Eq. 19 can be shown [6] to be

$$\begin{aligned} \min_{\mathbf{a}} \quad & \mathbf{a}^T \mathbf{T} \tilde{\mathbf{X}}^\dagger \tilde{\mathbf{X}} \mathbf{T} \mathbf{a} + \mathbf{c}^T \mathbf{T} \mathbf{a} \\ s.t. \quad & \mathbf{0} \leq \mathbf{a} \leq \mathbf{1} C', \ \mathbf{a}^T \mathbf{t} = 0, \end{aligned} \tag{20}$$

where $\tilde{\mathbf{X}} = [\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_N]$, $\mathbf{t} = [t_1, \ldots, t_N]^T$, $\mathbf{T}$ is a diagonal matrix with $\mathbf{t}$ along the diagonal, and $\mathbf{c} = [c_1 \ldots, c_N]^T$. Sequential minimal optimization (SMO) [30] is used to find $\mathbf{a}$. Instead of simultaneously solving for the entire $\mathbf{a} = [a_1, \cdots, a_N]^T$ vector, SMO recursively solves for different $(a_i, a_j)_{i \neq j}$ pairs and can be implemented efficiently. Once we solve for $\mathbf{a}$, the MMCF filter $\hat{\mathbf{w}}$ can be computed as

$$\hat{\mathbf{w}} = \hat{\mathbf{S}}^{-\frac{1}{2}} \tilde{\mathbf{w}} = \hat{\mathbf{S}}^{-\frac{1}{2}} \tilde{\mathbf{X}} \mathbf{a}. \tag{21}$$

B. Sum of correlation energies

Computing $\hat{\mathbf{S}}^{-\frac{1}{2}}$ is computationally expensive especially when d is large. Therefore to reduce the computational complexity we ignore $\hat{\mathbf{Y}}\hat{\mathbf{Y}}^\dagger$ and approximate $\hat{\mathbf{S}}$ by a diagonal matrix, i.e.,

$$\hat{\mathbf{S}} = \lambda \mathbf{I} + (1 - \lambda) \hat{\mathbf{Z}} \approx \lambda \mathbf{I} + (1 - \lambda) \hat{\mathbf{D}}, \tag{22}$$

thus avoiding the inversion of a non-diagonal matrix. This simplifies the localization criterion shown in Eq. 15 to $\hat{\mathbf{w}}^\dagger \hat{\mathbf{D}} \hat{\mathbf{w}}$. The value

$$\hat{\mathbf{w}}^\dagger \hat{\mathbf{D}} \hat{\mathbf{w}} = \hat{\mathbf{w}}^\dagger \left(\sum_{i=1}^{N} \hat{\mathbf{X}}_i \hat{\mathbf{X}}_i^* \right) \hat{\mathbf{w}} = \sum_{i=1}^{N} \left| \hat{\mathbf{w}}^\dagger \hat{\mathbf{X}}_i \right|^2 \tag{23}$$

is a measure of the sum of the energies of the correlation outputs. This localization criterion ignores the single pixel at the target's location. However, the energy contribution of the value at the target's location to the energy of the entire correlation output is negligible, and thus our approximation does not adversely affect the filter. Empirically, in our experiments we observed that the overall loss in filter performance is negligible when using this approximation.

C. Limitations

MMCFs (as well as SVMs, and CFs) are sensitivity to different scales (e.g., as the sensor moves closer/farther to the target), target occlusions, and challenging lighting conditions. To deal with scale changes, Dalal and Triggs [5] use a pyramid approach; they compute a template and compare different scaled versions of the template to the test image. To deal with occlusions, Rodriguez and Kumar [13] combine the CF output with a tracker to estimate the target's location even when it

Table I
COMPUTATIONAL COMPLEXITY BIG O AND MEASURED (SEC.)

	Training one template		Testing one image	
Template	Big O	time	Big O	time
MMCF	$\min(N^3, N^2 d) + N d \log d$	0.89	$d_s \log d_s$	0.20
SVM	$\min(N^3, N^2 d)$	0.48	$d_s \log d_s$	0.20
OTSDF	$N^3 + N d \log d$	0.61	$d_s \log d_s$	0.20
ASEF	$N d \log d$	0.41	$d_s \log d_s$	0.20
MOSSE	$N d \log d$	0.38	$d_s \log d_s$	0.20
UOTSDF	$N d \log d$	0.35	$d_s \log d_s$	0.20

is temporarily occluded (e.g., going under a bridge). To deal with different lighting conditions, Kumar and Hassebrook [31] present different performance measures, including the peak-to-correlation-energy (PCE) to measure a peak value relative to the surrounding values. In our dataset these techniques are not required but may be useful with datasets that have these challenges.

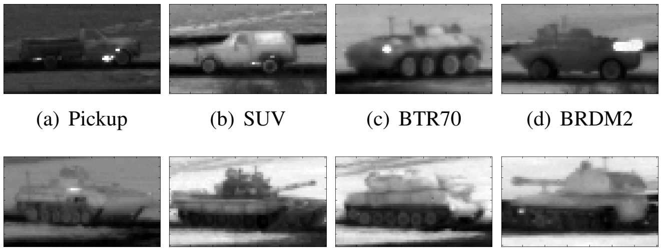
V. COMPUTATIONAL COMPLEXITY

The theoretical and measured complexity of each method is shown in Table I. These comparisons were done using MATLAB on a 2.91 GHz, 3.25 GB RAM Dual Core Windows XP desktop. To measure the training and testing time, we used $N = 100$ training images of dimension $d = 40 \times 70 = 2800$, and 200 testing images of dimension $d_s = 512 \times 640 = 327680$ and report the average time (in seconds) per image. MMCF, OTSDF, UOTSDF, ASEF, and MOSSE are designed by transforming the images into the frequency domain thereby requiring N FFTs of size d, i.e., $O(N d \log d)$. In addition to the Fourier transforms of the images, MMCF and SVM solve the quadratic optimization problem using SMO which has a computational complexity of $O(\min(N^3, N^2 d))$, and OTSDF requires a matrix inversion of complexity $O(N^3)$. The computation required to test any of these filters on a given query image is exactly the same, i.e., it involves the cross-correlation of the query image with the template which is computed efficiently in the frequency domain which has a computational complexity of $O(d_s \log d_s)$.

Using non-linear classifiers such as Quadratic CFs [32] and Kernel SVMs [6] may be useful to handle data non-linearities. However, these classifiers require $O(N d_s \log d_s)$ operations so the testing time grows linearly with the number of training images, making their usage impractical in some scenarios. Extending linear MMCF to kernel MMCF is not straightforward and is a topic of future research.

VI. NUMERICAL EXPERIMENTS

To demonstrate the efficacy of the MMCF approach, we consider three different computer vision tasks: vehicle recognition in large scenes, eye localization in face images, and face classification in centered images. For each of the tasks we consider six different classifiers, SVM, ASEF, MOSSE, UOTSDF, OTSDF, and MMCF. Since MACE and UMACE are special cases of OTSDF and UOTSDF, respectively, we will not explicitly consider these classifiers. In addition to MMCF, we selected these other classifiers because they have been shown to outperform other classifiers [33], [8], [10]. We determine target location by cross-correlating the 2-D template (obtained from the $\mathbf{w}$ vector described earlier) with the query image and determining the location of the largest value in the resulting correlation output. In addition, we applied the retraining technique described in Dalal and Triggs [5], i.e., we iteratively apply the template to the training frames (training images plus background) and include the false positives as false-class training images and recompute the template. Empirically we observed that for ASEF, MOSSE, and UOTSDF the performance degrades when we include false-class training images, therefore retraining was done only for SVM, OTSDF and MMCF. The goal of these experiments is to compare the performance of the different template designs for simultaneous localization and classification. We conducted several tests with various parameters (e.g., the λ shown in Eq. 17 or the σ used in ASEF and MOSSE–see Section II-A) for these filters and present the best results for each filter type.


(a) Pickup (b) SUV (c) BTR70 (d) BRDM2

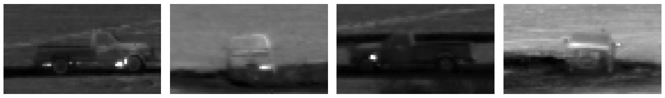
(e) BMP2 (f) T72 (g) ZSU23-4 (h) 2S3

Figure 4. Example of the different classes of military vehicles.

A. Vehicle Recognition

We investigate vehicle recognition (i.e., classification and localization) using a set of infrared images (frames from videos) where the vehicle's class-label and location are unknown. We use the recently approved for public release *ATR Algorithm Development Image Database* [34] produced by the Military Sensing Information Analysis Center. The database contains infrared videos of 512×640 pixels/frame of eight military vehicles (one vehicle in each video), shown in Fig. 4, taken at multiple ranges during day and night time at 30 frames/sec. Note in Fig. 4 that some of the military vehicles are very similar, making the classification task challenging. These vehicles were driven at about 5 m/sec. making a full circle of diameter of about 100 m., therefore exhibiting $360°$ of azimuth rotation. Each video is 60 sec. long (i.e., 1800 frames), allowing the vehicle to complete at least one full circle. We used videos from each vehicle collected during day time at a range of 1000 m. and compared our results to the ground truth data provided in the database.

We conducted two sets of experiments. In Exp. 1 we divided each video into four segments and trained one different MMCF per vehicle (we use all 8 vehicles) per segment (32 MMCFs in total). Each video segment contains one vehicle exhibiting approximately $90°$ azimuth range (i.e., $-45°$ to $45°$, $45°$ to $135°$, $135°$ to $225°$, and $225°$ to $315°$). Fig. 5 shows examples of the vehicle *Pickup* at approximately $0°$, $90°$, $180°$, and $270°$, so that four segments cover $360°$. From each segment,



(a) 0° (b) 90° (c) 180° (d) 270°

Figure 5. Images of vehicle *Pickup* at $\sim 0°$, $90°$, $180°$, and $270°$ views.

Table II
RECOGNITION PERFORMANCE (%) FOR EXPS. 1 AND 2

Exp.	MMCF	SVM	OTSDF	ASEF	MOSSE	UOTSDF
1	97.4	86.7	94.6	67.3	81.3	66.9
2	74.3	56.7	37.9	34.4	32.8	26.0

we selected 15 true-class images (manually cropped from the corresponding frames of width $w = 70$ and height $h = 40$ pixels) per filter and 100 non-overlapping background images as false-class images for training, and 100 test frames (images plus background). We verified that none of the testing frames were used in training. Note that the low quality frame and the general background (see Fig. 1b) makes the recognition task challenging. For comparison, we similarly trained 32 SVMs, OTSDFs, ASEFs, MOSSEs, and UOTSDFs classifiers and correlated them with the scene.

It is important to note that we did not include shifted images of true-class vehicles as false-class images as done in some recognition approaches [32], [14], [15]. Including every possible shift would have required an additional $15 \times \{(2 \times 70 - 1) \times (2 \times 40 - 1) - 1\} = 1647000$ false-class images per filter. One of the strengths of the MMCF approach is avoiding the need to use these shifted images during training without sacrificing the ability to accurately localize the vehicle.

To investigate the ability of the various approaches to localize and classify targets, we did *not* use any tracker in these experiments but assumed that the vehicle can be anywhere in each frame, and we treated each frame independently from other frames. Including a tracker may improve localization performance but is omitted from our experiments in order to analyze the performance of the unaided MMCF.

We declare a correct recognition when the correct template produces the maximum response to a given frame (i.e., correct classification) *and* produces the peak within a specified window centered at the correct location (i.e., correct localization). This means that it is considered an error 1) when the largest correlation peak is close to the target's ground truth location but is from the incorrect class, or 2) when the largest correlation peak is from the correct class but the peak's location is not near the target's ground truth location. In these experiments, we defined the window as follows,

$$window = \left(\frac{|P_x - \hat{P}_x|}{w} \leq D\right) \cap \left(\frac{|P_y - \hat{P}_y|}{h} \leq D\right), \quad (24)$$

where P_x and P_y are the ground truth location coordinates, $\hat{P}_x$ and $\hat{P}_y$ are correlation peak location coordinates, and $0 \leq D \leq 1$ is the normalized distance. Recall that in our experiments, width $w = 70$ and height $h = 40$. $D = 0$ requires that the correlation peak location be the same as the ground truth location. $D = 0.5$ requires that the peak location be within 35 and 20 pixels of the ground truth location in the $x-$ and $y-$directions, respectively. $D = 1$ requires that the peak location be within 70 and 40 pixels of the ground truth location in the $x-$ and $y-$directions, respectively.

Exp. 2 is much more challenging because we required one filter per vehicle to correctly classify the vehicle in the presence of 360° azimuth variation. For each vehicle, we selected 50 true-class images and 100 non-overlapping background images as false-class images for training, and 1000 test frames. The rest of the set up is the same as Exp. 1.

Table II shows the average recognition percentages of the filters in these two experiments using $D = 0.5$. As expected, the recognition rates in Exp. 1 when using four filters per vehicle are higher than in Exp. 2 when using only one filter per vehicle. We observe that the unconstrained CFs UOTSDF, ASEF, and MOSSE perform poorly. This is because the design formulations for these unconstrained filters use only true-class training images; including false-class training images decreases their performance. Therefore, these filters lack the advantage of the other filter's usage of false-class training images, and hence, also of retraining.

We performed retraining [5] on SVM, MMCF, and OTSDF. Retraining greatly helps the performance of SVM and MMCF. Note that Table II shows the results *after* retraining. In Exp. 1 retraining improved SVM from 51.3% recognition rate (not shown in the table) to 86.7% and MMCF from 75.4% to 97.4%, and in Exp. 2 retraining improved SVM from 21.8% to 56.7% and MMCF from 52.5% to 74.3%. We observed (not shown in the table) that OTSDF reaches a point of saturation and adding more training images actually decreases its performance (Table II shows OTSDF at its best performance). This is because, as mentioned earlier, OTSDF uses hard constraints, i.e., it must satisfy $\mathbf{w}^T\mathbf{x}_i = c_i$, resulting in overfitting to the training data.

In Exp. 2, MMCF outperformed the next best classifier SVM by 31%. This is due to the localization criterion $\hat{\mathbf{w}}^\dagger\hat{\mathbf{D}}\hat{\mathbf{w}}$ discussed earlier that MMCF has in addition to the margin criterion $\hat{\mathbf{w}}^\dagger\hat{\mathbf{w}}$. The localization criterion results in sharp correlation peaks. Fig. 2 shows an example of a test frame showing a target-sized rectangle around the ground truth data and around the corresponding location of the correlation peak. It can be seen that the MMCF correlation peak is well defined and leads to more accurate localization than the SVM template result shown in Fig. 1.

We also compare performance as a function of the normalized distance D in Fig. 6 for Exps. 1 and 2. In both sets of experiments, MMCF outperforms all the other classifiers for all values of D. We observe that the improvement when $D > 0.3$ is insignificant over $D = 0.3$. This means when the target is correctly recognized, the classifier usually puts the target's location within 21 and 12 pixels in the x- and y-direction, respectively, of the ground truth location.

We also investigated the performance of MMCF as a function of λ for a given set of training images. Earlier, we discussed that MMCF is equivalent to SVM when $\lambda = 1$. We now show the benefits of MMCF by testing the MMCF classifier over a variety of λ values using the same training

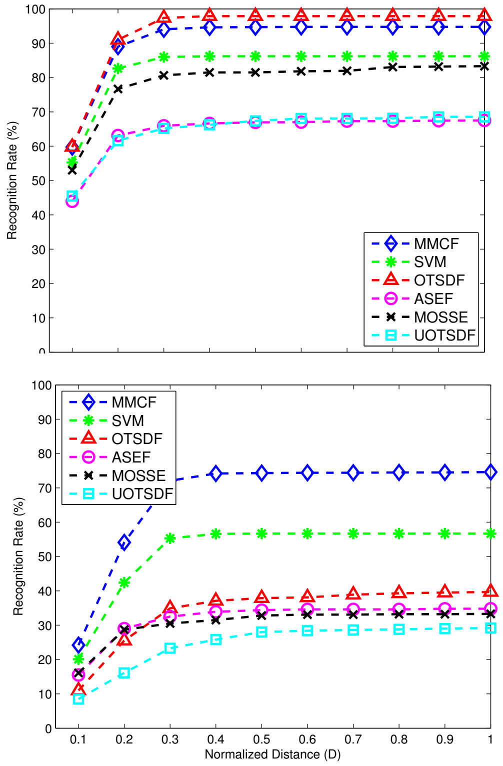


Figure 6. Recognition rate as a function of normalized localization error D for different classifiers (top: Exp. 1, bottom: Exp. 2). In every case, MMCF outperforms all the other classifiers.

images that we obtained for SVM in Exp. 2 (after retraining). To be clear, before retraining, all the filters have the same training images, but as we retrain, each filter selects a different set of false-class training images. Therefore, MMCF usually has a different set of false-class training images for each λ value after retraining. However, to investigate the effect of changing λ for a fixed set of training images, we designed the filters using the same set of training images obtained for SVM after retraining. We call these set of experiments *MMCF* $\lambda = 1$. For comparison, we repeat the experiment using the set of training images obtained after retraining for $\lambda = 0.67$ (since that value gave the best MMCF results) and call these set of experiments *MMCF* $\lambda = 0.67$. In addition, we investigate the effect of changing λ (see Eq. 8) on OTSDF and UOTSDF before retraining (since retraining decreases their performance). The results are shown in Fig. 7. Note from Fig. 7 that even when using the training images obtained for SVM after retraining, MMCF can outperform SVM, i.e., SVM recognition rate is 56.7% whereas MMCF recognition rate is 68.7% when $\lambda = 0.92$.

All the previous experiments focus on recognition (i.e., classification *and* localization) rate. Localization rate only tests a filter against its targeted class, therefore it does not test for classification. To be clear, suppose that the test image contains Target 3. To test classification, the image is cross-correlated with all eight templates. If Template 7 produces the best correlation peak value then that image is incorrectly classified as Target 7. To test localization, the image is cross-correlated with the correct template, i.e., Template 3, and if the correlation output (only from Template 3) produces the best peak near the correct location, then this image is correctly localized. Because this image is correctly localized but incorrectly classified, the image is deemed to have been incorrectly recognized. Therefore, recognition requires both correct classification and localization and this is what we report. Localization rates are always better (or at worse, they are equal) than the recognition rates. Fig. 8 shows the localization rates.

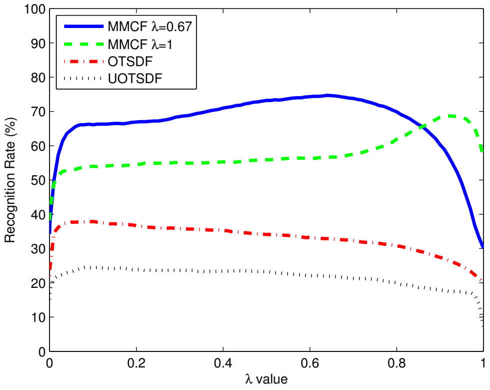


Figure 7. The recognition performance of the classifiers as a function of λ. In every case, MMCF outperforms all the other classifiers.

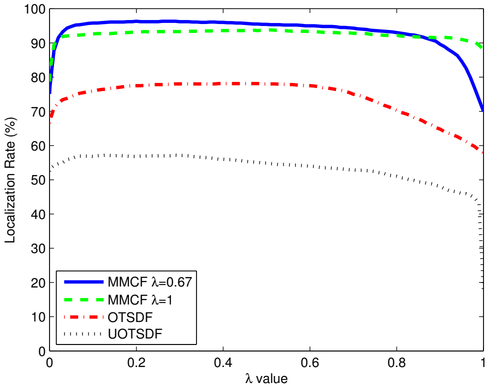


Figure 8. The localization performance of the classifiers as a function of λ. In every case, MMCF outperforms all the other classifiers.

Figs. 7 and 8 show that as the value of λ increases from 0 to 1, the performance of MMCF first increases and then decreases as it approaches SVM (i.e., $\lambda = 1$). Therefore in these set of experiments, we can always find a λ value for which MMCF outperforms SVM.

We next investigated the relation between the number of support vectors that MMCF, SVM, and OTSDF have. The MMCF, SVM, and OTSDF templates can all be formulated as weighted sums of training images (or transformed images) as in Eqs. 2 and 21 (see [28] for OTSDF's formulation). The vectors $\mathbf{a}$ used in MMCF and SVM have many zero elements while OTSDF does not have any zero elements. We referred to the training vectors corresponding to non-zero elements of $\mathbf{a}$ as support vectors. Fig. 9 compares the average percentage of supports vectors to the number of training images over the

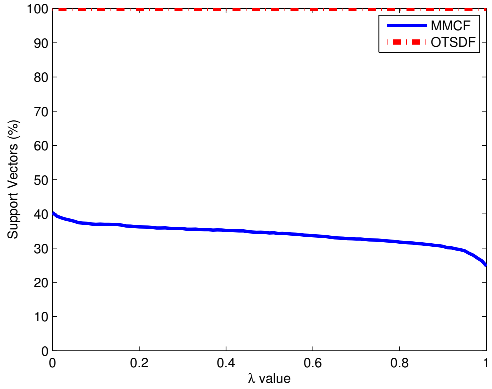


Figure 9. The percent of support vectors (i.e., support vectors over number of training images) for MMCF and OTSDF as a function of λ.

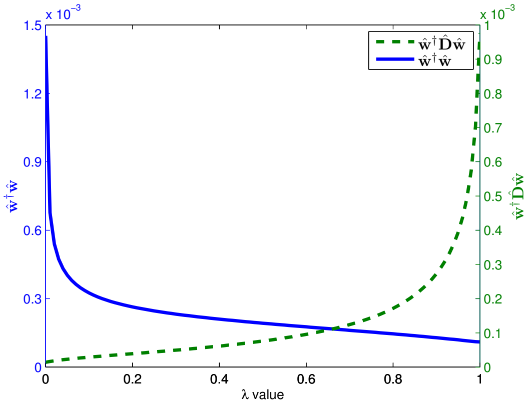


Figure 10. The margin criterion $\hat{\mathbf{w}}^\dagger\hat{\mathbf{w}}$ and localization criterion $\hat{\mathbf{w}}^\dagger\hat{\mathbf{D}}\hat{\mathbf{w}}$ as a function of λ for the MMCF. Note that as the MMCF classifiers approach SVM ($\lambda = 1$), the margin criterion value decreases (i.e., gets better), and the localization criterion value increases.

eight (one per vehicle) templates. We see from this figure that SVM (i.e., MMCF with $\lambda = 1$) has the fewest number of support vectors. As the value of λ decreases, the number of support vectors that MMCF has slightly increases. OTSDF uses all training vectors. This trend was observed for all SVMs, MMCFs, and OTSDFs in all our experiments. We note that using all training images as support vectors (as is the case in OTSDF), actually decreases performances because the classifier overfits to the training images. MMCF uses more support vectors than SVM to improve localization (and therefore recognition) while maintaining a much smaller number of support vectors than OTSDF which avoids overfitting.

In addition, we investigated how the margin criterion $\hat{\mathbf{w}}^\dagger\hat{\mathbf{w}}$ and the localization criterion $\hat{\mathbf{w}}^\dagger\hat{\mathbf{D}}\hat{\mathbf{w}}$ vary as a function of λ. When $\lambda = 0$ we minimize only the localization criterion $\hat{\mathbf{w}}^\dagger\hat{\mathbf{D}}\hat{\mathbf{w}}$, and when $\lambda = 1$ we minimize only the margin criterion $\hat{\mathbf{w}}^\dagger\hat{\mathbf{w}}$. For MMCF, we show the average (over the eight filters–one per vehicle) values of criteria $\hat{\mathbf{w}}^\dagger\hat{\mathbf{w}}$ and $\hat{\mathbf{w}}^\dagger\hat{\mathbf{D}}\hat{\mathbf{w}}$ as function of λ in Fig. 10. We do not show the values for OTSDF, UOTSDF, ASEF, and MOSSE because they are much higher than MMCF (i.e., they are off this chart). As expected, as $\lambda \to 1$, the MMCF margin criterion $\hat{\mathbf{w}}^\dagger\hat{\mathbf{w}}$ decreases and the localization criterion $\hat{\mathbf{w}}^\dagger\hat{\mathbf{D}}\hat{\mathbf{w}}$ increases. We observe that a slight increase in one criterion can significantly improve the other criterion. For example, increasing the value of $\hat{\mathbf{w}}^\dagger\hat{\mathbf{w}}$ from 1.1×10^{-4} at $\lambda = 1$ to 1.3×10^{-4} at $\lambda = 0.9$, decreases the average correlation energy $\hat{\mathbf{w}}^\dagger\hat{\mathbf{D}}\hat{\mathbf{w}}$ from 9.8×10^{-4} to 2.7×10^{-4}. MMCF optimally trades-off between these two criteria resulting in a filter with improved recognition.

We also investigated the effects of additive white Gaussian noise (AWGN). We used the same setup as Exp. 2 and applied AWGN to the test images. Fig. 11 shows the performance loss (%) as a function of signal-to-noise-ratio (SNR) in dB. SNR is defined as $10\log_{10}$(Signal Power/AWGN variance). From this figure we see that MMCF is comparable to SVM in performance loss. Even at -20dB, the MMCF has only 16% performance loss. ASEF, MOSSE, and UOTSDF have a smaller performance loss but their recognition rates were much smaller to begin with.

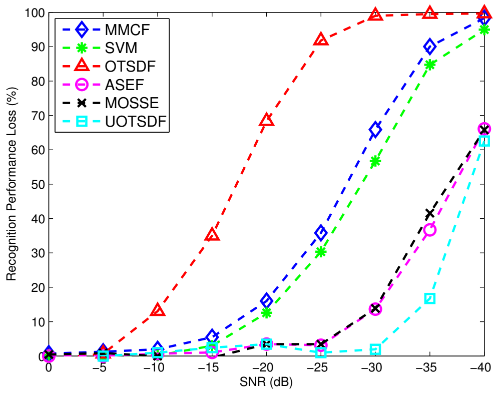


Figure 11. Recognition performance loss as a function of decreasing SNR.

B. Eye Localization

Accurate localization of the eyes in face images is an important component of face, ocular, and iris recognition. In this experiment we consider the task of accurately determining the location of the left and the right eye given a bounding box around a face obtained from a face detector. Since a good face detector makes eye localization overly simple, following the experimental setup outlined in [8], we make the problem harder by introducing errors in face localization. We first center the faces obtained using the *OpenCV* face detector [35] to produce 128×128 images with the eyes centered at (32.0, 40.0) and (96.0, 40.0) and then apply a random similarity transform with translation of up to ± 4 pixels, scale factor of up to 1.0 ± 0.1 and rotations of up to $\pm{}^{\pi}/_{16}$ radians. We used the FERET [36] database for this task which has about 3400 images of 1204 people. We randomly partitioned the database with 512 images used for training, 675 for parameter selection by cross-validation, and the rest for testing. The different template design approaches are compared by evaluating the normalized distance defined as follows,

$$D = \frac{\|P - \hat{P}\|}{\|P_l - P_r\|}, \tag{25}$$

where P is the ground truth location, $\hat{P}$ is the predicted location, and P_l and P_r are the ground truth locations of the left

Table III
EYE LOCALIZATION PERFORMANCE (%)

Eye	MMCF	SVM	OTSDF	ASEF	MOSSE	UOTSDF
L	95.1	87.8	81.2	91.2	94.1	94.6
R	93.6	89.4	78.5	90.6	92.9	93.2

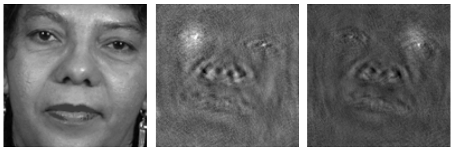

(a) Face (b) Left (c) Right

Figure 12. Example showing the output of the left and right MMCF eye detector on a sample face image (a).

and the right eye, respectively. The localization performance results averaged over 5 different runs with random partitions for training and testing and random similarity transforms are shown in Table III. Fig. 12 shows an example of the response of the left and the right MMCF eye detector on a sample face image. While our results for ASEF are consistent with those reported by Bolme et al. [8] for the same task, our results for UOTSDF and OTSDF are better than those reported. This is because Bolme et al. used the full face images to train ASEF but only a window around each eye to train UOTSDF and OTSDF, and we used the full face images to train UOTSDF and OTSDF. It must be noted that the performance of SVM and MMCF improves significantly after at least one round of retraining (we do 3 rounds of retraining in our experiments), while the performance of OTSDF degrades due to overfitting. As is consistent with the other experiments, SVM uses the fewest support vectors (200) followed by MMCF (600) followed by OTSDF (1024) among 1024 initial training images.

C. Face Classification

We consider face classification on 64×64 images where the bounding boxes for the faces have been pre-determined by running a face detector like the Viola-Jones face detector [37]. We use the Multi-PIE database [38] which is an extension of the CMU PIE database [39] with images that have been captured over multiple sessions (maximum of 5 sessions). It has a total of 337 subjects with different face poses, expressions and illumination variations. We present results using frontal images of neutral expressions with different illuminations (see Fig. 13) of which there are over 23000 images. We conducted eight experiments using these images. In each face classification experiment, we select a set of true-class training images for each subject, and used the true-class training images from all the impostor subjects as false-class training images. The descriptions of the eight experiments are:

- **Exp. 1:** We selected 1 true-class training image per subject (frontal illumination) from session 1 and tested on images from all the other sessions (i.e, excluding images from session 1).
- **Exp. 2:** We selected 2 true-class training images per subject (one with illumination from the right and one with illumination from the left) from session 1 and tested on images from all the other sessions (i.e., excluding images from session 1).
- **Exp. 3:** We selected 3 true-class training images per subject (one image with frontal illumination, one image with illumination from the left and one image with illumination from the right) from session 1 and tested on images from all the other sessions (i.e., excluding images from session 1).
- **Exp. 4:** Similar to Exp. 1 except that we test on all the images from session 1 only excluding the training images.
- **Exp. 5:** Similar to Exp. 2 except that we test on all the images from session 1 only excluding the training images.
- **Exp. 6:** Similar to Exp. 3 except that we test on all the images from session 1 only excluding the training images.
- **Exp. 7:** Similar to Exp. 3 except that we test on images with varying degrees of occlusion.
- **Exp. 8:** Similar to Exp. 6 except that we test on images with varying degrees of occlusion.

Table IV
MULTI-PIE DATABASE RANK-1 CLASSIFICATION ACCURACY (%)

Exp.	MMCF	SVM	OTSDF	ASEF	MOSSE	UOTSDF
1	58.3	17.0	50.2	26.5	24.8	32.0
2	71.9	21.6	56.1	33.9	57.6	50.3
3	73.5	24.7	55.2	34.1	64.3	50.8
4	97.7	37.0	98.3	53.9	51.6	97.7
5	99.9	47.3	99.9	58.3	88.7	99.9
6	99.9	50.2	99.9	61.0	92.2	99.9

In each experiment, each test image is cross-correlated with all 337 templates (each template is designed to positively classify one subject). A correct classification means that the highest correlation value was produced by the correct template. Table IV presents the classification accuracy (%) for the first six experiments using different classifiers. Exps. 1 to 3 are more challenging than Exps. 4 to 6 since the testing session is different from the training session, while in Exps. 4 to 6 the testing and training session are the same. We observe that MMCF outperforms the other classifiers (especially OTSDF) in Exps. 1 to 3 due to its better generalization capability while OTSDF performs better in Exp. 4 since the test images are from the same session as the training images. We can also see that MMCF exhibits higher classification accuracies than SVMs. We conjecture that this is mainly due to the implicit increase in the number of training images when minimizing the localization criterion used by CFs. Minimizing the localization criterion is equivalent to making the correlation output approximate a delta function. This is the same as requiring the inner products of the template with centered true-class training images to be 1 and the inner products with the shifted training images to be 0, effectively using centered training images as well as all shifted versions in designing the CF. In contrast, SVMs only constrain the inner product of the template and the centered training images and does not constrain the other values of the correlation plane.

In addition to the above mentioned experiments, we further compare the robustness of various template designs to occlusions. Towards this purpose, in Exps. 7 and 8 we perform face recognition on the same test set as in Exps. 3 and 6,

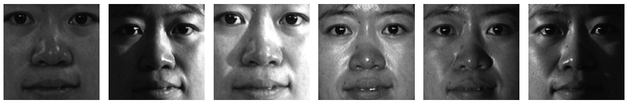
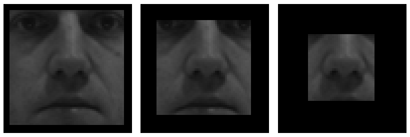


Figure 13. Sample Multi-PIE database images with illumination variations.



(a) 25% (b) 50% (c) 75%

Figure 14. Test images with 25%, 50%, and 75% occlusion.

respectively, with varying degrees of occlusions (results for the other experimental settings follow the same general trend). In Fig. 14 we show an example of the simple occlusion pattern that we performed experiments on and in Fig. 15 we present the percentage loss in Rank-1 identification accuracy, as a function of the percent of missing pixels in the image, in comparison to the Rank-1 identification accuracy with no occlusions/missing pixels. In the more challenging setting of Exp. 7 MMCF is more robust to small amounts of occlusion compared to the other template designs while in the relatively easier Exp. 8, UOTSDF and OTSDF show more robustness across all degrees of occlusions. A more thorough analysis of robustness to different kinds of occlusions for some of the CF template designs considered here can be found in [40].

VII. Conclusions

Conventional object recognition approaches based on SVMs do not explicitly take into account object localization criterion while earlier CF designs were not explicitly designed to provide good generalization. In this work, we introduced the Maximum Margin Correlation Filter (MMCF), which is an extension of SVMs and CFs. It combines the generalization capability of SVMs and the localization capability of CFs. We evaluated this classifier on three distinct tasks (vehicle recognition, eye localization, and face classification) and demonstrated that MMCF outperforms well-known CF designs and SVMs for object recognition.

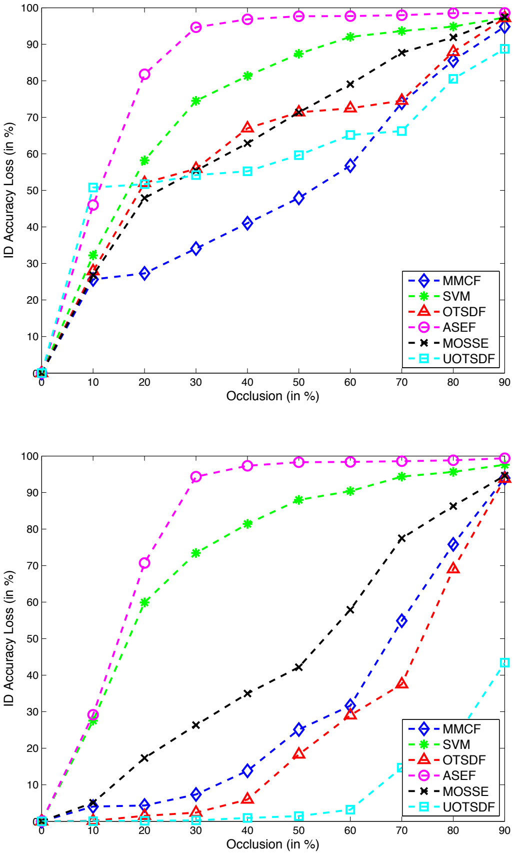


Figure 15. ID Accuracy loss for the task of face recognition (top: Exp. 7, bottom: Exp. 8) as a function of the percent of occlusion.

References

[1] B. Boser, I. Guyon, and V. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc. Fifth Annual Workshop on Computational Learning Theory*, pp. 144–152, 1992.

[2] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[3] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer, 1995.

[4] E. Osuna, R. Freund, and F. Girosit, "Training support vector machines: an application to face detection," in *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 130–136, 1997.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 886–893, 2005.

[6] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[7] B. V. K. Vijaya Kumar, A. Mahalanobis, and R. D. Juday, *Correlation Pattern Recognition*. Cambridge Univ. Press, 2005.

[8] D. Bolme, B. Draper, and J. Beveridge, "Average of synthetic exact filters," in *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 2105–2112, 2009.

[9] D. Bolme, Y. Lui, B. Draper, and J. Beveridge, "Simple real-time human detection using a single correlation filter," in *IEEE Int'l Workshop on Performance Evaluation of Tracking and Surveillance*, pp. 1–8, 2010.

[10] D. Bolme, J. Beveridge, B. Draper, and Y. Lui, "Visual object tracking using adaptive correlation filters," in *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 2544–2550, 2010.

[11] M. Savvides, J. Heo, J. Thornton, P. Hennings, C. Xie, K. Venkataramani, R. Kerekes, M. Beattie, and B. V. K. Vijaya Kumar, "Biometric identification using advanced correlation filter methods," in *Springer-Verlag Lecture Notes in Computer Science: Ambient Intelligence*, 2005.

[12] M. Savvides and B. V. K. Vijaya Kumar, "Efficient design of advanced correlation filters for robust distortion-tolerant face recognition," in *IEEE Conf. Advanced Video and Signal Based Surveillance*, pp. 45–52, 2003.

[13] A. Rodriguez and B. V. K. Vijaya Kumar, "Automatic target recognition of multiple targets from two classes with varying velocities using correlation filters," in *IEEE Int'l Conf. on Image Processing*, pp. 2781–2784, 2010.

[14] B. Scholkopf, C. Burges, and V. Vapnik, "Incorporating invariances in support vector learning machines," *Artificial Neural Networks ICANN*, vol. 1112, pp. 47–52, 1996.

[15] D. Decoste and B. Scholkopf, "Training invariant support vector machines," *Machine Learning*, vol. 46, no. 1, pp. 161–190, 2002.

[16] O. Chapelle and B. Scholkopf, "Incorporating invariances in non-linear support vector machines," *Advances in Neural Information Processing Systems*, vol. 1, no. 1, pp. 609–616, 2002.

[17] P. Shivaswamy and T. Jebara, "Relative margin machines," *Advances in Neural Information Processing Systems*, vol. 21, no. 21, 2008.

[18] A. B. Ashraf, S. Lucey, and T. Chen, "Re-interpreting the application of gabor filters as a manipulation of the margin in linear support vector machines," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 32, no. 7, pp. 1335–1341, 2010.

[19] J. Thornton, M. Savvides, and B. V. K. Vijaya Kumar, "Linear shift-invariant maximum margin svm correlation filter," in *Proc. Intelligent*

Sensors, Sensor Networks and Information Processing Conf., pp. 183–188, 2005.

[20] B. V. K. Vijaya Kumar, A. Mahalanobis, and A. Takessian, “Optimal tradeoff circular harmonic function correlation filter methods providing controlled in-plane rotation response,” *IEEE Trans. Image Processing*, vol. 9, no. 6, pp. 1025–1034, 2000.

[21] R. Kerekes and B. V. K. Vijaya Kumar, “Correlation filters with controlled scale response,” *IEEE Trans. Image Processing*, vol. 15, no. 7, pp. 1794–1802, 2006.

[22] D. Lowe, “Object recognition from local scale-invariant features,” in *Proc. Seventh IEEE Int’l Conf. on Computer Vision*, vol. 2, pp. 1150–1157, 1999.

[23] A. Mahalanobis, B. V. K. Vijaya Kumar, and D. Casasent, “Minimum average correlation energy filters,” *Applied Optics*, vol. 26, no. 5, pp. 3633–3640, 1987.

[24] A. Mahalanobis, B. V. K. Vijaya Kumar, S. Song, S. Sims, and J. Epperson, “Unconstrained correlation filters,” *Applied Optics*, vol. 33, no. 17, pp. 3751–3759, 1994.

[25] B. V. K. Vijaya Kumar, “Minimum-variance synthetic discriminant functions,” *J. Opt. Soc. Am. A*, vol. 3, no. 10, pp. 1579–1584, 1986.

[26] B. V. K. Vijaya Kumar, D. W. Carlson, and A. Mahalanobis, “Optimal trade-off synthetic discriminant function filters for arbitrary devices,” *Optics Letters*, vol. 19, no. 19, pp. 1556–1558, 1994.

[27] A. V. Oppenheim, A. S. Willsky, and S. Hamid, *Signals and Systems*. Prentice Hall, 1997.

[28] P. Réfrégier, “Filter design for optical pattern recognition: multicriteria optimization approach,” *Opt. Let.*, vol. 15, no. 15, pp. 854–856, 1990.

[29] S. P. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[30] J. Platt, “Sequential minimal optimization: A fast algorithm for training support vector machines,” *Advances in Kernel Methods–Support Vector Learning*, vol. 208, no. 14, pp. 98–112, 1998.

[31] B. V. K. Vijaya Kumar and L. Hassebrook, “Performance measures for correlation filters,” *Applied Optics*, vol. 29, no. 20, pp. 2997–3006, 1990.

[32] A. Mahalanobis, R. Muise, and S. R. Stanfill, “Quadratic correlation filter design methodology for target detection and surveillance applications,” *Applied Optics*, vol. 43, no. 27, pp. 5198–5205, 2004.

[33] R. Kerekes and B. V. K. Vijaya Kumar, “Selecting a composite correlation filter design: a survey and comparative study,” *Opt. Eng.*, vol. 47, no. 6, pp. 1–18, 2008.

[34] Military Sensing Information Analysis Center, “ATR algorithm development image database.” www.sensiac.org, 2011.

[35] G. Bradski, “The OpenCV Library,” *Dr. Dobb’s Journal of Software Tools*, 2000.

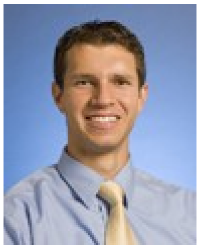
[36] P. Phillips, H. Moon, S. Rizvi, and P. Rauss, “The FERET evaluation methodology for face-recognition algorithms,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1090–1104, 2000.

[37] P. Viola and M. J. Jones, “Robust real-time face detection,” *Int’l Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[38] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, “Multi-pie,” *Image and Vision Computing*, vol. 28, no. 5, pp. 807–813, 2010.

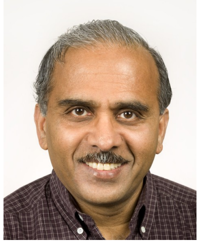
[39] T. Sim, S. Baker, and M. Bsat, “The CMU pose, illumination, and expression (PIE) database,” in *Proc. Fifth IEEE Conf. on Automatic Face and Gesture Recognition*, pp. 46–51, 2002.

[40] M. Savvides, B. V. K. Vijaya Kumar, and P. Khosla, “Robust shift-invariant biometric identification from partial face images,” in *Proc. of SPIE*, vol. 5404, p. 124, 2004.

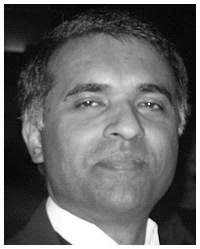


Andres Rodriguez is a Ph.D. candidate in Electrical and Computer Engineering at Carnegie Mellon University and a research scientist at the Air Force Research Laboratory Sensors Directorate. His research interests are in Pattern Recognition and Machine Learning. He received a BS and MS degree in Electrical Engineering from Brigham Young University in 2006 and 2008, respectively, while working at BYU’s MAGICC laboratory. He was awarded the Carnegie Institute of Technology Dean’s Tuition Fellowship in 2008 and the Frank J. Marshall Graduate Fellowship in 2009. He has published 10 technical papers and received the best student paper award at the SPIE ATR conference in 2010.

Vishnu Naresh Boddeti received a BTech degree in Electrical Engineering from the Indian Institute of Technology, Madras in 2007. He is currently in the Electrical and Computer Engineering program at Carnegie Mellon University where he received the MSc degree and is a candidate for the Ph.D degree. His research interests are in Computer Vision, Pattern Recognition and Machine Learning.



B. V. K. Vijaya Kumar (Prof. Kumar) is a Professor of Electrical and Computer Engineering and the Interim Dean of College of Engineering at Carnegie Mellon University (CMU). Professor Kumar’s research interests include Pattern Recognition, Biometrics and Coding and Signal Processing for Data Storage Systems. His publications include the book entitled Correlation Pattern Recognition, fifteen book chapters and more than 500 technical papers. He served as a Pattern Recognition Topical Editor for Applied Optics and as an Associate Editor for IEEE Transactions on Information Forensics and Security. Professor Kumar serves on many biometrics and data storage conference program committees and was a co-general chair of the 2004 Optical Data Storage conference, a co-chair of the 2008, 2009 and 2010 SPIE conferences on Biometric Technology for Human Identification and is a co-chair of the 2012 Biometrics: Theory, Applications and Systems (BTAS) conference. Professor Kumar is a Fellow of IEEE, a Fellow of SPIE, a Fellow of Optical Society of America and a Fellow of the International Association of Pattern Recognition. He serves on the IEEE Biometric Council and was a former member of IEEE Signal processing Society’s Technical Committee on Information Forensics and Security. Prof. Kumar received the 2003 Eta Kappa Nu award for Excellence in Teaching in the ECE Department at CMU and the 2009 Carnegie Institute of Technology’s Outstanding Faculty Research Award (jointly with Prof. Marios Savvides).



Abhijit Mahalanobis is a Senior Fellow of the Lockheed Martin Corporation, in the Applied Research division of the company. His primary research areas are in Optical information processing, Computational Sensing and Imaging, and Video/Image processing for information exploitation and ATR. He has over 140 journal and conference publications in this area. He also holds three patents, co-authored a book on pattern recognition, contributed several book chapters, and edited special issues of several journals. Abhijit completed his B.S. degree with Honors at the University of California, Santa Barbara in 1984. He then joined the Carnegie Mellon University and received the MS and Ph.D. degrees in 1985 and 1987, respectively. Prior to joining Lockheed Martin, Abhijit worked at Raytheon in Tucson, and was a faculty at the University of Arizona and the University of Maryland. Abhijit was elected a Fellow of SPIE in 1997, and a Fellow of OSA 2004 for his work on optical pattern recognition and ATR. He served as an associate editor for Applied Optics from 2004-2009. He was as an associate editor for the journal of the Pattern Recognition Society from 1994-2003. He is a current member of the OSA Board of Meetings, and served on OSA’s Science and Engineering council in the capacity of Pattern Recognition Chair from 2001-2004. He also serves on the organizing committees for the SPIE conferences, and OSA’s annual and topical meetings. Abhijit received the Hughes Business unit Patent Award in 1998. He was recognized as the Innovator of the Year by the State of Arizona in 1999, and was elected to the Raytheon Honors program for distinguished technical contribution and leadership. At Lockheed Martin, he was elected to the rank of Distinguished Member of Technical Staff in 2000, and twice received the Lockheed Martin Technical Excellence award, the Author of the Year award in 2001, and the Inventor of the Year in 2005 for designing novel target recognition systems. In 2005, he received the prestigious Lockheed Martin NOVA award, the Corporation’s highest honor. Most recently, Abhijit was recognized as the 2006 Scientist of the Year by Science Spectrum Magazine.