Multi-Channel Correlation Filters

Hamed Kiani Galoogahi
National University of Singapore
Singapore
hkiani@comp.nus.edu.sg

Terence Sim
National University of Singapore
Singapore
tsim@comp.nus.edu.sg

Simon Lucey
CSIRO
Australia
simon.lucey@csiro.au

Abstract

Modern descriptors like HOG and SIFT are now commonly used in vision for pattern detection within image and video. From a signal processing perspective, this detection process can be efficiently posed as a correlation/convolution between a multi-channel image and a multi-channel detector/filter which results in a single-channel response map indicating where the pattern (e.g. object) has occurred. In this paper, we propose a novel framework for learning a multi-channel detector/filter efficiently in the frequency domain, both in terms of training time and memory footprint, which we refer to as a multi-channel correlation filter. To demonstrate the effectiveness of our strategy, we evaluate it across a number of visual detection/localization tasks where we: (i) exhibit superior performance to current state of the art correlation filters, and (ii) superior computational and memory efficiencies compared to state of the art spatial detectors.

1. Introduction

In computer vision it is now rare for tasks like convolution/correlation to be performed on single channel image signals (e.g. 2D array of intensity values). With the advent of advanced descriptors like HOG [5] and SIFT [13] convolution/correlation across multi-channel signals has become the norm rather than the exception in most visual detection tasks. Most of these image descriptors can be viewed as multi-channel images/signals with multiple measurements (such the oriented edge energies) associated with each pixel location. We shall herein refer to all image descriptors as multi-channel images. An example of multi-channel correlation can be seen in Figure 1 where a multi-channel image is convolved/correlated with a multi-channel filter/detector in order to obtain a *single-channel* response. The peak of the response (in white) indicating where the pattern of interest is located.

Like single channel signals, correlation between two multi-channel signals is rarely performed naively in the spatial domain. Instead, the fast Fourier transform (FFT) affords the efficient application of correlating a desired template/filter with a signal. Contrastingly, however, most techniques for estimating a detector for such a purpose (i.e. detection/tracking through convolution) are performed in the spatial domain [5]. It is this dilemma that is at the heart of our paper.

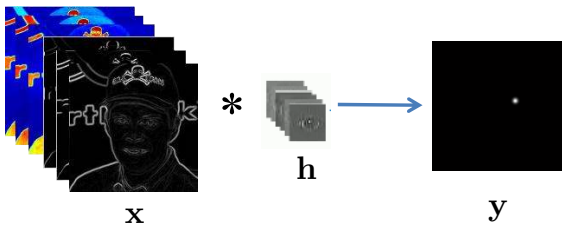


Figure 1. An example of multi-channel correlation/convolution where one has a multi-channel image $\mathbf{x}$ correlated/convolved with a multi-channel filter $\mathbf{h}$ to give a single-channel response $\mathbf{y}$. By posing this objective in the frequency domain, our multi-channel correlation filter approach attempts to give a computational & memory efficient strategy for estimating $\mathbf{h}$ given $\mathbf{x}$ and $\mathbf{y}$.

This has not always been the case. Correlation filters, developed initially in the seminal work of Hester and Casasent [8], are a method for learning a template/filter in the frequency domain that rose to some prominence in the 80s and 90s. Although many variants have been proposed [8, 11, 12], the approach's central tenet is to learn a filter, that when correlated with a set of training signals, gives a desired response (typically a peak at the origin of the object, with all other regions of the correlation response map being suppressed). Like correlation itself, one of the central advantages of the single channel approach is that it attempts to learn the filter in the frequency domain due to the efficiency of correlation/convolution in that domain. Learning multi-channel filters in the frequency domain, however, comes at the high cost of computation and memory usage. In this paper we present an efficient strategy for learning multi-channel signals/filters that has numerous applications throughout vision and learning.

Contributions: In this paper we make the following contributions

- We propose an extension to canonical correlation filter theory that is able to efficiently handle multi-channel signals. Specifically, we show how when posed in the frequency domain the task of multi-channel correlation filter estimation forms a sparse banded linear system. Further, we demonstrate how our system can be solved much more efficiently than spatial domain methods.
- We characterize theoretically and demonstrate empirically how our multi-channel correlation approach affords substantial memory savings when learning on multi-channel signals. Specifically, we demonstrate how our approach does *not* have a memory cost that is linear in the number of samples, allowing for substantial savings when learning detectors across large amounts of data.
- We apply our approach across a myriad of detection and localization tasks including: eye localization, car detection and pedestrian detection. We demonstrate: (i) superior performance to current state of the art single-channel correlation filters, and (ii) superior computational and memory efficiency in comparison to spatial detectors (e.g. linear SVM) with comparable detection performance.

Notation: Vectors are always presented in lower-case bold (e.g., $\mathbf{a}$), Matrices are in upper-case bold (e.g., $\mathbf{A}$) and scalars in italicized (e.g. a or A). $\mathbf{a}(i)$ refers to the ith element of the vector $\mathbf{a}$. All M-mode array signals shall be expressed in vectorized form $\mathbf{a}$. M-mode arrays are also known as M-mode matrices, multidimensional matrices, or tensors. We shall be assuming $M = 2$ mode matrix signals (e.g. $2D$ image arrays) in nearly all our discussions throughout this paper. This does not preclude, however, the application of our approach to other $M \neq 2$ signals.

A M-mode convolution operation is represented as the $*$ operator. One can express a M-dimensional discrete circular shift $\Delta\boldsymbol{\tau}$ to a vectorized M-mode matrix $\mathbf{a}$ through the notation $\mathbf{a}[\Delta\boldsymbol{\tau}]$. The matrix $\mathbf{I}$ denotes a $D \times D$ identity matrix and $\mathbf{1}$ denotes a D dimensional vector of ones. $\hat{\mathrm{A}}$ applied to any vector denotes the M-mode Discrete Fourier Transform (DFT) of a vectorized M-mode matrix signal $\mathbf{a}$ such that $\hat{\mathbf{a}} \leftarrow \mathcal{F}(\mathbf{a}) = \sqrt{D}\mathbf{F}\mathbf{a}$. Where $\mathcal{F}()$ is the Fourier transforms operator and $\mathbf{F}$ is the orthonormal $D \times D$ matrix of complex basis vectors for mapping to the Fourier domain for any D dimensional vectorized image/signal. We have chosen to employ a Fourier representation in this paper due to its particularly useful ability to represent circular convolutions as a Hadamard product in the Fourier domain. Additionally, we take advantage of the fact that $\text{diag}(\hat{\mathbf{h}})\hat{\mathbf{a}} = \hat{\mathbf{h}} \circ \hat{\mathbf{a}}$, where $\circ$ represents the Hadamard product, and diag() is an operator that transforms a D dimensional vector into a $D \times D$ dimensional diagonal matrix. The role of filter $\hat{\mathbf{h}}$ or signal $\hat{\mathbf{a}}$ can be interchanged with this property. Any transpose operator ^{T} on a complex vector or matrix in this paper additionally takes the complex conjugate in a similar fashion to the Hermitian adjoint [12]. The operator $\text{conj}(\hat{\mathbf{a}})$ applies the complex conjugate to the complex vector $\hat{\mathbf{a}}$.

2. Related Work

Multi-Channel Detectors: The most notable approach to multi-channel detection in computer vision can be found in the seminal work of Dalal & Triggs [5] where the authors employ a HOG descriptor in conjunction with a linear SVM to learn a detector for pedestrian detection. This same multi-channel detection pipeline has gone on to be employed in a myriad of other detection tasks in vision ranging from facial landmark localization/detection [19] to general object detection [7].

Computational and memory efficiency, however, are issues for Dalal & Triggs style multi-channel detectors. A central advantage of using a linear SVM, over kernel SVMs, for learning a multi-channel detector is the ability to treat that detector as a multi-channel linear filter during evaluation. Instead of inefficiently moving the detector spatially across a multi-channel image, one can take advantage of the fast Fourier transform (FFT) for the efficient application of correlating a desired template/filter with a signal.

During training, however, all learning is done in the spatial domain. This can be a slow and inefficient process. The strategy involves the extraction of positive (aligned) and negative (misaligned) multi-channel image patches of the object/pattern of interest across large amounts of data. From a learning perspective, much of this storage can be viewed as inefficient as it often involves shifted versions of the same multi-channel image. We argue in this paper, that this is a real strength of correlation filters as the objective provides a way for naturally modeling shifted versions of an image without the burden of explicitly storing all the shifted image patches.

Multi-Channel Descriptors: Motivation for working with multi-channel image signals (i.e. descriptors) rather than raw single channel pixel intensities stems from seminal work on the mammalian primary visual cortex (V1) [9]. Here, local object appearance and shape can be well categorised by the distribution of local edge directions, without precise knowledge of their spatial location. It has been noted [10] that V1-inspired descriptors obtain superior photometric and geometric invariance in comparison to raw intensities giving strong motivation for their use in many modern vision applications.

Jarrett et al. [10] showed that many V1-inspired features follow a similar pipeline of filtering an image through a large filter bank, followed by a nonlinear rectification

step, and finally a blurring/histogramming step resulting in a multi-channel signal (where the number of channels was dictated by the size of the filter bank). Canonical features such as HOG and SIFT employ filter banks with strong selectivity to spatial frequency, orientation and scale (e.g. oriented edge filters, Gabor filters, etc.).

Prior Art in Correlation Filters: Bolme et al. [3] recently proposed an extension to traditional correlation filters referred to as Minimum Output Sum of Squared Error (MOSSE) filters. This approach has proven invaluable for many object tracking tasks, outperforming current state of the art methods such as [1, 16]. A strongly related method to MOSSE was also proposed by Bolme et al. [4] for object detection/localization referred to as Average of Synthetic Exact Filters (ASEF) which also reported superior performance to state of the art. A full discussion on other variants of correlation filters such as Optimal Tradeoff Filters (OTF) [15], Unconstrained MACE (UMACE) [17] filters, etc. is outside the scope of this paper. Readers are encouraged to inspect [12] for a full treatment on the topic. Recently, Boddeti et al. [2] introduced vector correlation filter to train multi-channel descriptors in the Fourier domain for car landmark detection and alignment. This approach, however, suffered from huge amount of memory usage and computational complexity, since this approach required to solve a $KD \times KD$ linear system, where K is the number of channels and D is the length of vectorized signals.

3. Correlation Filters

Due to the efficiency of correlation in the frequency domain, correlation filters have canonically been posed in the frequency domain. There is nothing, however, stopping one (other than computational expense) from expressing a correlation filter in the spatial domain. In fact, we argue that viewing a correlation filter in the spatial domain can give: (i) important links to existing spatial methods for learning templates/detectors, and (ii) crucial insights into fundamental problems in current correlation filter methods.

Bolme et. al's [3] MOSSE correlation filter can be expressed in the spatial domain as solving the following ridge regression problem,

$$E(\mathbf{h}) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{D}||\mathbf{y}_i(j) - \mathbf{h}^T\mathbf{x}_i[\Delta\boldsymbol{\tau}_j]||_2^2 + \frac{\lambda}{2}||\mathbf{h}||_2^2 \quad (1)$$

where $\mathbf{y}_i \in \mathbb{R}^D$ is the desired response for the i-th observation $\mathbf{x}_i \in \mathbb{R}^D$ and λ is a regularization term. $\mathbb{C} = [\Delta\boldsymbol{\tau}_1, \ldots, \Delta\boldsymbol{\tau}_D]$ represents the set of all circular shifts for a signal of length D. Bolme et al. advocated the use of a 2D Gaussian of small variance (2-3 pixels) for $\mathbf{y}_i$ centered at the location of the object (typically the centre of the image patch). The solution to this objective becomes,

$$\mathbf{h}^* = \mathbf{H}^{-1}\sum_{i=1}^{N}\sum_{j=1}^{D}\mathbf{y}_i(j)\mathbf{x}_i[\Delta\boldsymbol{\tau}_j] \quad (2)$$

where,

$$\mathbf{H} = \lambda\mathbf{I} + \sum_{i=1}^{N}\sum_{j=1}^{D}\mathbf{x}_i[\Delta\boldsymbol{\tau}_j]\mathbf{x}_i[\Delta\boldsymbol{\tau}_j]^T \quad . \quad (3)$$

Solving a correlation filter in the spatial domain quickly becomes intractable as a function of the signal length D, as the cost of solving Equation 2 becomes $\mathcal{O}(D^3 + ND^2)$.

Efficiency in the Frequency Domain: It is well understood in the signal processing community that circular convolution in the spatial domain can be expressed as a Hadamard product in the frequency domain. This allows one to express the objective in Equation 1 more succinctly and equivalently as,

$$\begin{aligned} E(\hat{\mathbf{h}}) &= \frac{1}{2}\sum_{i=1}^{N}||\hat{\mathbf{y}}_i - \hat{\mathbf{x}}_i \circ \text{conj}(\hat{\mathbf{h}})||_2^2 + \frac{\lambda}{2}||\hat{\mathbf{h}}||_2^2 \quad (4)\\ &= \frac{1}{2}\sum_{i=1}^{N}||\hat{\mathbf{y}}_i - \text{diag}(\hat{\mathbf{x}}_i)^T\hat{\mathbf{h}}||_2^2 + \frac{\lambda}{2}||\hat{\mathbf{h}}||_2^2 \quad . \end{aligned}$$

where $\hat{\mathbf{h}}, \hat{\mathbf{x}}, \hat{\mathbf{y}}$ are the Fourier transforms of $\mathbf{h}, \mathbf{x}, \mathbf{y}$. The complex conjugate of $\hat{\mathbf{h}}$ is employed to ensure the operation is correlation not convolution. The equivalence between Equations 1 and 4 also borrows heavily upon another well known property from signal processing namely, Parseval's theorem which states that

$$\mathbf{x}_i^T\mathbf{x}_j = D^{-1}\hat{\mathbf{x}}_i^T\hat{\mathbf{x}}_j \quad \forall i, j, \quad \text{where } \mathbf{x} \in \mathbb{R}^D \quad . \quad (5)$$

The solution to Equation 4 becomes

$$\begin{aligned} \hat{\mathbf{h}}^* &= [\text{diag}(\hat{\mathbf{s}}_{xx}) + \lambda\mathbf{I}]^{-1}\sum_{i=1}^{N}\text{diag}(\hat{\mathbf{x}}_i)\hat{\mathbf{y}}_i \quad (6)\\ &= \hat{\mathbf{s}}_{xy} \circ^{-1} (\hat{\mathbf{s}}_{xx} + \lambda\mathbf{1}) \end{aligned}$$

where $\circ^{-1}$ denotes element-wise division, and

$$\hat{\mathbf{s}}_{xx} = \sum_{i=1}^{N}\hat{\mathbf{x}}_i \circ \text{conj}(\hat{\mathbf{x}}_i) \quad \& \quad \hat{\mathbf{s}}_{xy} = \sum_{i=1}^{N}\hat{\mathbf{y}}_i \circ \text{conj}(\hat{\mathbf{x}}_i) \quad (7)$$

are the average auto-spectral and cross-spectral energies respectively of the training observations. The solution for $\hat{\mathbf{h}}$ in Equations 1 and 4 are identical (other than that one is posed in the spatial domain, and the other is in the frequency domain). The power of this method lies in its computational efficiency. In the frequency domain a solution to $\hat{\mathbf{h}}$ can be

found with a cost of $\mathcal{O}(ND \log D)$. The primary cost is associated with the DFT on the ensemble of training signals $\{\mathbf{x}_i\}_{i=1}^N$ and desired responses $\{\mathbf{y}_i\}_{i=1}^N$.

Memory Efficiency: Inspecting Equation 7 one can see an additional advantage of correlation filters when posed in the frequency domain. Specifically, memory efficiency. One does not need to store the training examples in memory before learning. As Equation 7 suggests one needs to simply store a summation of the auto-spectral $\hat{\mathbf{s}}_{xx}$ and cross-spectral $\hat{\mathbf{s}}_{xy}$ energies. This is a powerful result not often discussed in correlation filter literature as unlike other spatial strategies for learning detectors (e.g. linear SVM) whose memory usage grows as a function of the number of training examples $\mathcal{O}(ND)$, correlation filters have fixed memory overheads $\mathcal{O}(D)$ irrespective of the number of training examples.

4. Our Approach

Inspired by single-channel correlation filters we shall explore a multi-channel strategy for learning a correlation filter. We can express the multi-channel objective in the spatial domain as

$$E(\mathbf{h}) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{D}||\mathbf{y}_i(j) - \sum_{k=1}^{K}\mathbf{h}^{(k)T}\mathbf{x}_i^{(k)}[\Delta\boldsymbol{\tau}_j]||_2^2 + \frac{\lambda}{2}\sum_{k=1}^{K}||\mathbf{h}^{(k)}||_2^2 \quad (8)$$

where $\mathbf{x}^{(k)}$ and $\mathbf{h}^{(k)}$ refers to the kth channel of the vectorized image and filter respectively where K represents the number of filters. As with a canonical filter the desired response is single channel $\mathbf{y} = [\mathbf{y}(1), \ldots, \mathbf{y}(D)]^T$ even though both the filter and the signal are multi-channel. Solving this multi-channel form in the spatial domain is even more intractable than the single channel form with a cost of $\mathcal{O}(D^3K^3 + ND^2K^2)$ since we now have to solve a $KD \times KD$ linear system.

Fourier Efficiency: Inspired by the efficiencies of posing single channel correlation filters in the Fourier domain we can express Equation 8 equivalently and more succintly

$$E(\hat{\mathbf{h}}) = \frac{1}{2}\sum_{i=1}^{N}||\hat{\mathbf{y}}_i - \sum_{k=1}^{K}\text{diag}(\hat{\mathbf{x}}_i^{(k)})^T\hat{\mathbf{h}}^{(k)}||_2^2 + \frac{\lambda}{2}\sum_{k=1}^{K}||\hat{\mathbf{h}}^{(k)}||_2^2 \quad (9)$$

where $\hat{\mathbf{h}} = [\hat{\mathbf{h}}^{(1)T}, \ldots, \hat{\mathbf{h}}^{(K)T}]^T$ is a KD dimensional supervector of the Fourier transforms of each channel. This can be simplified further,

$$E(\hat{\mathbf{h}}) = \frac{1}{2}\sum_{i=1}^{N}||\hat{\mathbf{y}}_i - \hat{\mathbf{X}}_i\hat{\mathbf{h}}||_2^2 + \frac{\lambda}{2}||\hat{\mathbf{h}}||_2^2 \ . \quad (10)$$

where $\hat{\mathbf{X}}_i = [\text{diag}(\hat{\mathbf{x}}_i^{(1)})^T, \ldots, \text{diag}(\hat{\mathbf{x}}_i^{(K)})^T]$. At first glance the cost of solving this linear system looks no different to the spatial domain as one still has to solve a $KD \times KD$ linear system:

$$\hat{\mathbf{h}}^* = (\lambda\mathbf{I} + \sum_{i=1}^{N}\hat{\mathbf{X}}_i^T\hat{\mathbf{X}}_i)^{-1}\sum_{i=1}^{N}\hat{\mathbf{X}}_i^T\hat{\mathbf{y}}_i \quad (11)$$

Fortunately, $\hat{\mathbf{X}}$ is sparse banded and inspecting Equation 10 one can see that the *j*th element of each correlation response $\hat{\mathbf{y}}_i(j)$ is dependent only on the K values of $\mathcal{V}(\hat{\mathbf{h}}(j))$ and $\mathcal{V}(\hat{\mathbf{x}}(j))$, where $\mathcal{V}$ is a concatenation operator that returns a $K \times 1$ vector when applied on the *j*th element of a K-channel vectors $\{\mathbf{a}^{(k)}\}_{k=1}^K$, i.e. $\mathcal{V}(\mathbf{a}(j)) = [\text{conj}(\mathbf{a}^{(1)}(j)), ..., \text{conj}(\mathbf{a}^{(K)}(j))]^T$. Therefore, we can equivalently express Equation 10 through a simple variable re-ordering as:

$$E(\mathcal{V}(\hat{\mathbf{h}}(j))) = \frac{1}{2}\sum_{i=1}^{N}||\hat{\mathbf{y}}_i(j) - \mathcal{V}(\hat{\mathbf{x}}_i(j))^T\mathcal{V}(\hat{\mathbf{h}}(j))||_2^2 + \frac{\lambda}{2}||\mathcal{V}(\hat{\mathbf{h}}(j))||_2^2, \quad for \quad j = 1, ..., D. \quad (12)$$

Therefore, an efficient solution of Equation 10 can be found by solving D independent $K \times K$ linear systems using Equation 12 as:

$$\mathcal{V}(\hat{\mathbf{h}}(j))^* = \hat{\mathbf{H}}^{-1}\sum_{i=1}^{N}\mathcal{V}(\hat{\mathbf{x}}_i(j))\hat{\mathbf{y}}_i(j) \quad (13)$$

where,

$$\hat{\mathbf{H}} = \lambda\mathbf{I} + \sum_{i=1}^{N}\mathcal{V}(\hat{\mathbf{x}}_i(j))\mathcal{V}(\hat{\mathbf{x}}_i(j))^T \quad (14)$$

This results in a substantially smaller computational cost of $\mathcal{O}(DK^3 + NDK^2)$ than solving this objective in the spatial domain $\mathcal{O}(D^3K^3 + ND^2K^2)$.

Memory Efficiency: As outlined in Section 3 an additional strength of single channel correlation filters are their memory efficiency. Specifically, one does not need to hold all the training examples in memory. Instead, they need to just

compute the auto-spectral $\hat{\mathbf{s}}_{xx}$ and cross-spectral $\hat{\mathbf{s}}_{xy}$ energies respectively of the training observations (see Equation 7). The memory saving become sizable as the number of training examples increase as the memory overhead remains constant $\mathcal{O}(D)$ instead of $\mathcal{O}(ND)$ if one was to employ a spatial objective. A similar strategy can be taken advantage of in our multi-channel correlation form. For multi-channel correlation filters this saving becomes even more dramatic as the memory overhead remains constant $\mathcal{O}(K^2D)$ as opposed to $\mathcal{O}(NDK)$. This property stems from the sparse banded structure of multi-channel correlation filters such that the problem can be posed as D independent $K \times K$ linear systems.

5. Experiments

We evaluated our method across a number of challenging localization and detection tasks: facial landmark localization, car detection, and pedestrian detection. For all our experiments we used the same parametric form for the desired correlation response, which we defined as a 2D Gaussian function with a spatial variance of two pixels whose the peak is centered at the location of the target of interest (facial landmarks, cars, pedestrians, etc.). Across all our experiments we used the same multi-channel image representation, specifically HOG [5]. All correlation filters, both single-channel and multi-channel, employed in this paper used a 2D cosine window (as suggested by Bolme et al. [3]) to reduce boundary effects.

5.1. Facial Landmark Localization

We evaluated our method for facial landmark localization on the Labeled Faces in the Wild (LFW) database[1], including 13,233 face images stemming from 5749 subjects. The images were captured *in the wild* with challenging variations in illumination, pose, quality, age ,gender, race, expression, occlusion and makeup. For each image, there are ground truth annotations for 10 facial landmarks as well as the bounding box of the face. We used the bounding box to crop a 128×128 face image enclosing all the landmarks. We then employed a 10-fold cross validation procedure to compute evaluation results across folds. 10% of images were approximately used for testing, with the remaining 90% being used for learning/training the detectors. The folds were constructed carefully to have no subjects in common.

All the cropped images were first pre-processed using Gamma correction and Difference of Gaussian (DoG) filtering to compensate for the large variations in illumination. Multi-channel HOG descriptors were computed using 9 orientation bins normalized by cell and block sizes of 6×6 and 3×3, respectively. Localization occured by correlating each landmark detector across the cropped face image where the peak response location was used as the predicted landmark location. The facial landmark localization was evaluated using normalized distance between the desired location and the predicted coordinate of the landmarks:

$$d = \frac{\|\mathbf{p}_i - \mathbf{m}_i\|_2}{\|\mathbf{m}_l - \mathbf{m}_r\|_2} \tag{15}$$

where $\mathbf{m}_r$ and $\mathbf{m}_l$ respectively indicate the ground truth of the right and left eye, and $\mathbf{m}_i$ and $\mathbf{p}_i$ are respectively the true and predicted locations of the landmark of interest. A localization with $d < \tau$ was considered successful where τ is a threshold defined as a fraction of the inter-ocular distance (the denominator of the above equation).

Results and Analysis: Inspecting Figure 2 one can see the superiority of our multi-channel approach compared to state of the art single-channel correlation filter methods MOSSE and ASEF. Further, we compare our performance to leading non-correlation filter methods: specifically Everingham et al. [6] and Valstar et al. [18] which also show the superiority of our approach. Some visual examples of the output from our approach employed for facial landmark localization can be seen in Figure 3. It should be noted that this approach to landmark localization employs no shape prior, relying instead solely on the landmark detectors making a fair comparison with more recent methods in facial landmark localization such as Zhu and Ramanan [19] difficult.

5.2. Car Detection

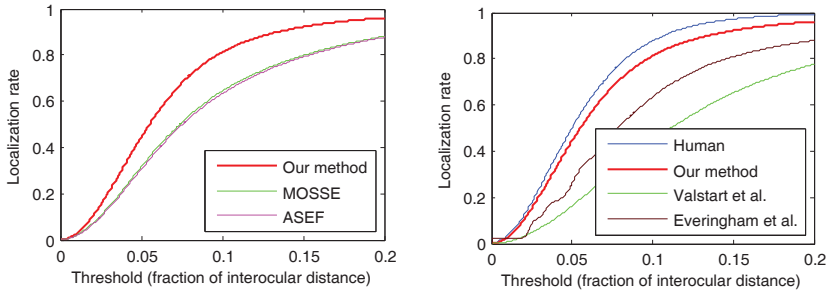
The objective of this experiment is to evaluate our proposed multi-channel correlation filter (MCCF) strategy for car localization in street scene images. We selected 1000 images from the MIT StreetScene [2] database, each image contains one car taken from an approximate left-half-frontal view. All the selected images were first cropped to a size of 360×360, and then power normalized to have zero-mean and unit norm. Our MCCF was trained and evaluated in the same manner to the previous experiment using 100×180 car patches cropped from training images (excluding street scenes). The peak of the Gaussian desired responses was located at the center of the car patches. We selected the peak of the correlation output as the predicted location of a car in street scene of the testing images. Figure 5.2 depicts our localization performance in comparison to leading single-channel correlation filters MOSSE and ASEF where we obtain superior performance across all thresholds. Visual examples of our car detection results can be seen in Figure 5.

5.3. Pedestrian Detection

We evaluated our method for pedestrian detection using Daimler pedestrian dataset [14] containing five disjoint im-

[1] http://vis-www.cs.umass.edu/lfw

[2] http://cbcl.mit.edu/software-datasets/streetscenes



(a) (b)

Figure 2. The performance of facial features localization: localization rate versus threshold (best viewed in color).

Figure 3. Visualizing facial features localization, first and second rows show successful localizations, and the third row show wrong localizations.

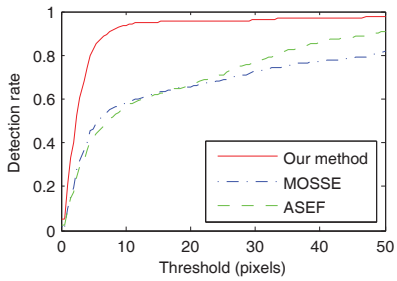


Figure 4. Car detection rate as a function of threshold (pixels).

ages sets, three for training and two for testing. Each set consists of 4800 pedestrian and 5000 non-pedestrian images of size 36×18. The oriented gradient channels were computed using 5 orientation bins with cell and block sizes of 3×3. Our MCCF was trained using all the negative and positive training samples with their corresponding desired responses. Given a test image, we first correlate it with the trained MCCF and then measure the Peak-to-Sidelobe Ratio (PSR)[3] of the output with a threshold for detection. This threshold was chosen through a cross-validation process.

Comparison with Linear SVM: In this experiment we chose to compare our MCCF directly with a spatial detector learned using a linear SVM (as originally performed by Dalal and Triggs [5]). The linear SVM was trained in almost exactly the same fashion as our MCCF so as to keep the comparison as fair as possible. Inspecting Figure 6 (a) one can see our MCCF obtains similar detection results to linear SVM in terms of detection performance as a function of different false positive rates. This result is not that surprising as the linear SVM objective is quite similar to the MCCF objective (which can be interpreted as a ridge regression when posed in the spatial domain). It is well understood that the linear SVM objective enjoys better tolerance to outliers than ridge regression, but in practice we have found that advantage to be only marginal when learning multi-channel detectors.

[3]Peak-to-Sidelobe Ratio (PSR) is a common metric used in correlation filter literature for detection/verification tasks. It is the ratio of the peak response to the local surrounding response, more details on this measure can be found in [12].

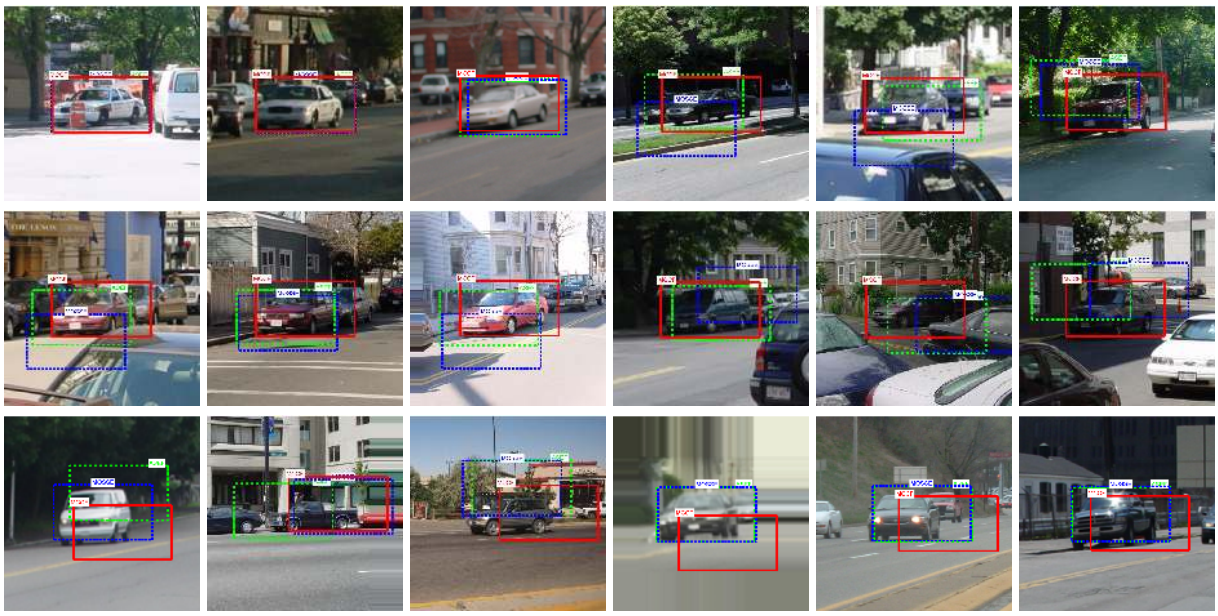


Figure 5. Car detection results. The first and second rows: true detections, and the third row: wrong detections. The red, blue and green boxes represents detection by our method, MOSSE and ASEF, respectively.

	250	500	1000	2000	4000	8000	16000	24000
MCCF	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
SVM	6.17	12.35	24.68	49.36	98.87	197.44	395.88	592.32

Table 1. Comparing minimum required memory (MB) of our method with SVM as a function of number of training images.

Inspecting Figure 6 (b) one can see detection performance as a function of number of training data. It is interesting to note that our MCCF objective can achieve good detection performance with substantially smaller amounts of training data when compared to linear SVM. This superior performance can be attributed to how correlation filters implicitly use synthetic circular shifted versions of images within the learning process without having to explicitly create the images. As a result our MCCF objective can do "more with less" by achieving good detection performance with substantially less training data.

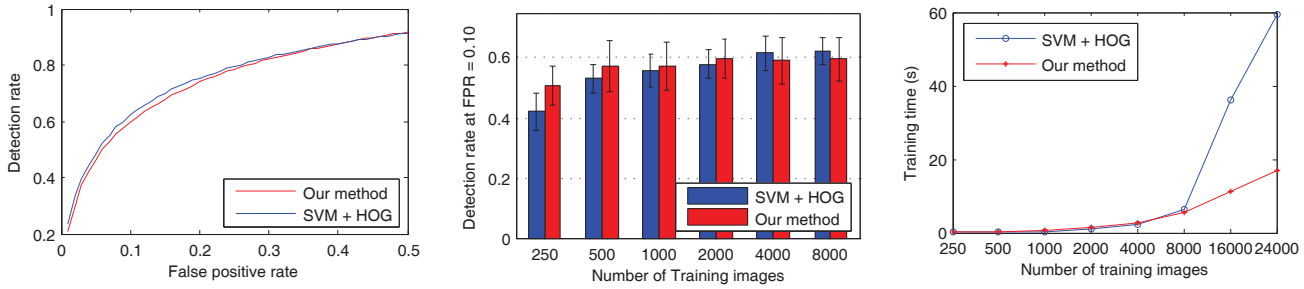
Computation and Memory Efficiency: Figure 6(c) depicts one of the major advantages of MCCF, and that is its superior scalability with respect to training set size. One can see how training time starts to increase dramatically for linear SVM[4] where as our training time only increases modestly as a function of training set size. The central advantage of our proposed approach here is that the solving of the multi-channel linear system in the frequency domain is independent to the number of images. Therefore the only component of MCCF that is dependent on training set size is the actual FFT on the training images which should only have the moderate computational cost $\mathcal{O}(ND \log D)$ as the training set size N increases.

Finally, inspecting Table 1 one can see the superior nature of our MCCF approach in comparison to linear SVM with respect to memory usage. As discussed in Section 4 our proposed MCCF approach has a modest fixed memory requirement independent of the training set size, whereas the amount of memory used by the linear SVM approach is a linear function of the number of training examples.

6. Conclusion

In this paper, we propose a novel extension to correlation filter theory which allows for the employment of multi-channel signals with the efficient use of memory and computations. We demonstrate the advantages of our new approach across a variety of detection and localization tasks.

[4]We employed the efficient and widely used LibLinear linear SVM package `http://www.csie.ntu.edu.tw/~cjlin/liblinear` in all our experiments.



(a) (b) (c)

Figure 6. Comparing our method with SVM + HOG (a) ROC curve of detection rate as a function of false positive rate (8000 training images), (b) pedestrian detection rate at FPR = 0.10 versus number of training images, and (c) training time versus the number of training images.

Figure 7. Some samples of (top) true detection of pedestrian (true positive), (middle) false detection of non-pedestrian (false negative), and (bottom) false detection of pedestrian (false positive).

References

[1] B. Babenko, M. H. Yang, and S. Belongie. Visual tracking with online multiple instance learning. In *CVPR*, 2009.

[2] V. N. Boddeti, T. Kanade, and B. V. K. V. Kumar. Correlation filters for object alignment. In *CVPR*, 2013.

[3] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui. Visual object tracking using adaptive correlation filters. In *CVPR*, 2010.

[4] D. S. Bolme, B. A. Draper, and J. R. Beveridge. Average of synthetic exact filters. In *CVPR*, 2009.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[6] M. Everingham, J. Sivic, and A. Zisserman. "hello! my name is... buffy"–automatic naming of characters in tv video. In *BMVC*, 2003.

[7] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 32(9):1627–1645, 2010.

[8] C. F. Hester and D. Casasent. Multivariant technique for multiclass pattern recognition. *Appl. Opt.*, 19(11):1758–1761, 1980.

[9] D. Hubel and T. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160(1):106, 1962.

[10] K. Jarrett, K. Kavukcuoglu, M. A. Ranzato, and Y. LeCun. What is the best multi-stage architecture for object recognition? *ICCV*, pages 2146–2153, 2009.

[11] B. V. K. V. Kumar. Minimum-variance synthetic discriminant functions. *J. Opt. Soc. Am. A*, 3(10):1579–1584, 1986.

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

[13] D. Lowe. Object recognition from local scale-invariant features. *ICCV*, pages 1150–1157, 1999.

[14] S. Munder and D. M. Gavrila. An experimental study on pedestrian classification. *PAMI*, 28(11):1863–1868, 2006.

[15] P. Refregier. Optimal trade-off filters for noise robustness, sharpness of the correlation peak, and horner efficiency. *Optics Letters*, 16:829–832, 1991.

[16] D. Ross, J. Lim, R. Lin, and M. Yang. Incremental learning for robust visual tracking. *IJCV*, 77(1):125–141, 2008.

[17] M. Savvides and B. V. K. V. Kumar. Efficient design of advanced correlation filters for robust distortion-tolerant face recognition. In *AVSS*, pages 45–52, 2003.

[18] M. Valstar, B. Martinez, X. Binefa, and M. Pantic. Facial point detection using boosted regression and graph models. In *CVPR*, 2010.

[19] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, 2012.