Learning Detectors Quickly with Stationary Statistics

Jack Valmadre[1,2], Sridha Sridharan[1] and Simon Lucey[3]

[1]Queensland University of Technology, Brisbane, Australia
[2]CSIRO, Australia
[3]Carnegie Mellon University, Pittsburgh, PA, USA
{j.valmadre, s.sridharan}@qut.edu.au, slucey@cs.cmu.edu

Abstract. Computer vision is increasingly becoming interested in the rapid estimation of object detectors. The canonical strategy of using Hard Negative Mining to train a Support Vector Machine is slow, since the large negative set must be traversed at least once per detector. Recent work has demonstrated that, with an assumption of signal stationarity, Linear Discriminant Analysis is able to learn comparable detectors without ever revisiting the negative set. Even with this insight, the time to learn a detector can still be on the order of minutes. Correlation filters, on the other hand, can produce a detector in under a second. However, this involves the unnatural assumption that the statistics are periodic, and requires the negative set to be re-sampled per detector size. These two methods differ chiefly in the structure which they impose on the covariance matrix of all examples. This paper is a comparative study which develops techniques (i) to assume periodic statistics without needing to revisit the negative set and (ii) to accelerate the estimation of detectors with aperiodic statistics. It is experimentally verified that periodicity is detrimental.

1 Introduction

Historically in computer vision, the time required to train a detector has been considered of minimal consequence because it only needs to be performed once. However, a number of vision algorithms for modern tasks involve learning a multitude of detectors, sometimes even in online settings. Examples include adaptive tracking [1,2], object detection with a large number of classes [3,4], algorithms for discovering discriminable clusters [5,6,7] and exemplar-based methods which train a detector per example [8]. An algorithm which drastically reduces the time and memory in which an effective detector can be trained has a big potential impact on these higher-level tasks.

Detectors are generally trained using machine learning algorithms for classification. One of the immediate and fundamental questions is: how to treat the enormous negative set? Any image which does not contain the object can contribute all of its sub-images, quickly generating myriad negative examples. Support Vector Machines (SVMs) are attractive in this regard, as they seek a

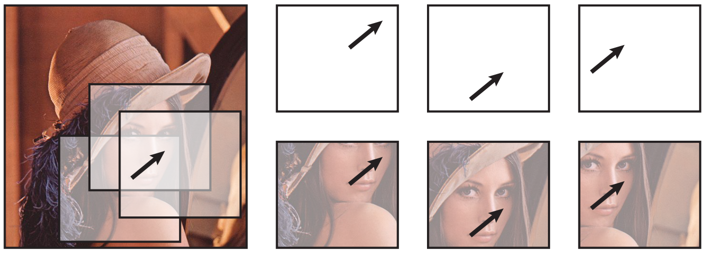

Fig. 1: The set of all translated windows exhibits stationarity since a single pair of pixels (indicated by the arrow) contributes to the statistics of all pairs with the same relative displacement. This results in a covariance matrix with Toeplitz structure, which Hariharan et al. [9] enforce to make estimation of the statistics of all windows feasible.

solution which depends on only a sparse subset of the examples (i.e. the support vectors). Finding this set is, however, no easier than solving the original problem. A popular heuristic is Hard Negative Mining (HNM), which alternates between training a detector and adding possible support vectors to the training set. New examples are found by using the current detector to exhaustively search a large negative set for false positives, making HNM poorly suited to the aforementioned tasks.

An alternative to SVMs is to entertain simple approaches which obtain a detector as the solution to a system of linear equations $\mathbf{w} = \mathbf{S}^{-1}\mathbf{r}$ whose dimension is independent of the number of examples. These include Linear Discriminant Analysis (LDA) and linear least-squares regression, in both of which $\mathbf{S}$ is a covariance matrix. Forming this system, however, tends to be computationally prohibitive without some additional knowledge of the problem. This paper examines two algorithms which make different assumptions regarding the structure of the covariance matrix, each with its own distinct motivation.

The first is the method of Hariharan et al. [9], in which the set of examples is assumed to be stationary, resulting in a Toeplitz covariance matrix

$$S_{ui} = g[i - u] \ . \tag{1}$$

This redundant structure imposes the assumption that the covariance of samples $x[u]$ and $x[i]$ is governed exclusively by their relative position, independent of their absolute position. This is motivated by the observation that any sub-image or “window” of a natural image also belongs to the set of natural images, therefore the statistics of the set of all natural images must be translation invariant (see Figure 1). The redundancy is sufficient to make estimation of the covariance matrix computationally tractable. Adopting this assumption within an LDA framework, comparable detection performance to HNM has been demonstrated [9].

The second method is that of correlation filters [10,11], in which all circular shifts of each example are incorporated into the training set (see Figure 2). This

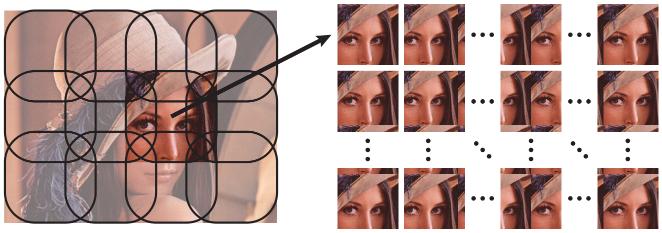

Fig. 2: Henriques et al. approximated the set of all translated windows in an image with all circular shifts (right) of a coarsely-sampled set of windows which cover the image (left). This results in a circulant covariance matrix which can be inverted in closed form. Rounded rectangles illustrate overlap.

manifests in a covariance matrix which is not only Toeplitz, but circulant

$$S_{ui} = h[(i - u) \bmod m] \tag{2}$$

for example signals of length m. The mod operator extends the assumption of stationarity beyond the boundary of the signal, under periodic extension. We refer to this stronger assumption as "periodic stationarity." The set of all circular shifts seems like an unnatural set to want to include, and indeed, the motivation is entirely computational. The discrete Fourier basis constitutes eigenvectors for any circulant matrix, meaning that efficient inversion can be performed using the Fast Fourier Transform (FFT). Thus, while the Toeplitz covariance more closely reflects the nature of the problem, the circulant system can be solved in much less time and memory.

Another critical difference is that the elements of the circulant covariance matrix depend on the signal size m, whereas those of the Toeplitz covariance matrix do not. This is due to the difference between (1) and (2). Therefore the Toeplitz covariance can be used to train detectors of arbitrary size and only needs to be computed once. It can also be elegantly estimated from signals of arbitrary size. In the circulant case, on the other hand, it is necessary to know the size of the examples *a priori*, and then to choose and sample a representative subset of windows of this size as in [12]. To train a detector of a different size, the entire process must be repeated.

The first contribution of this paper is to develop a simple expression for the (circulant) covariance matrix of all circular shifts of a set of windows of arbitrary size with known Toeplitz covariance. This enables correlation filters to be learned from the stationary distribution alone, without the need to ever re-visit the negative set.

The second contribution is to investigate methods for efficiently solving the Toeplitz system, particularly in the case of two-dimensional signals such as images. While Toeplitz matrices are shown to produce higher quality detectors, the raw speed of the non-iterative algorithm for circulant matrix inversion may make it an attractive option despite the degradation of performance. We additionally elucidate and evaluate a heuristic, discovered in previous work, which significantly improves the performance of the circulant regime.

2 Background

2.1 Linear Discriminant Analysis with Stationarity

Linear Discriminant Analysis (LDA) is a generative approach to binary classification which models both classes with a Gaussian distribution, assuming that the two distributions have the same covariance matrix to ensure that the discriminant is an affine function $f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b$. The optimal template $\mathbf{w}$ is obtained in closed form $\mathbf{w} = \mathbf{S}^{-1}\mathbf{r}$, where $\mathbf{r} = \bar{\mathbf{x}}_+ - \bar{\mathbf{x}}_-$ is the difference between the means of the positive and negative classes and $\mathbf{S}$ is the covariance of all examples. Typically when training a detector, the positive class is "object" and the negative class is "not object."

Let $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathcal{R}^m$ denote n examples, each with m elements. It can be assumed without loss of generality that the examples are zero-mean. The covariance of these examples would typically be estimated

$$\mathbf{S} = \frac{1}{n} \sum_{k=1}^{n} \mathbf{x}_k \mathbf{x}_k^T \quad . \tag{3}$$

This computation requires $\mathcal{O}(nm^2)$ time and $\mathcal{O}(m^2)$ memory. Clearly it is impractical to evaluate this on the set of all windows in a collection of images, where n numbers thousands per image and m is the number of pixels in a window.

Hariharan et al. [9] recognized that the set of natural images exhibits stationarity. This is motivated by observing that all sub-images of a natural image also belong to the set of natural images, therefore the statistics of the set of *all* natural images must be translation invariant (see Figure 1).

Let us first consider each vector $\mathbf{x}_k$ to be a scalar-valued time-series of length m with samples $x_k[0], \dots, x_k[m-1]$, before later generalizing to feature images. If the set of examples is drawn from a stationary distribution, then the covariance matrix possesses a highly redundant Toeplitz structure $S_{ui} = g[i-u]$, which encodes that the correlation of samples $x[u]$ and $x[i]$ depends only on their relative position. For signals of length m, this $m \times m$ matrix is fully specified by the $2m-1$ elements in $g[\delta]$, which is defined for $\delta = -m+1, \dots, m-1$. The symmetry of the covariance matrix further implies $g[\delta] = g[-\delta]$ and therefore m elements are sufficient. The zero-mean assumption in this context is discussed in Appendix B.1.

When considering detection problems in which the negative class is "not object," the covariance and mean are dominated by the negative set. Typically a set of large signals $\phi_1, \dots, \phi_N$ of length $M \geq m$ is available from which every window of length m constitutes a negative example (as in the canonical HNM problem). Under stationarity, the expected covariance is computed per relative displacement δ from all instances of that displacement in the large signals

$$g[\delta] = \frac{1}{N\rho(\delta)} \sum_{k=1}^{N} \sum_{t=a(\delta)}^{b(\delta)-1} \phi_k[t] \cdot \phi_k[t+\delta] \tag{4}$$

where the limits $a(\delta) = \max(0, -\delta)$ and $b(\delta) = M - \max(0, \delta)$ ensure that both t and $t + \delta$ lie within the domain of the signal $\{0, \ldots, M-1\}$. The normalization factor $\rho(\delta) = b(\delta) - a(\delta) = M - |\delta|$ counts the number of occurrences of the displacement δ in each signal (shorter displacements are observed more times). Unlike the true covariance matrix, which is a sum of outer products, a Toeplitz matrix obtained in this fashion is not guaranteed to be positive semidefinite. Given enough data, however, the eigenvalues converge to non-negativity.

If the covariance matrix were computed naively from the full set of $M - m + 1$ overlapping windows contained in each signal, its estimation would take $\mathcal{O}(Mm^2)$ time. Using the above expression, statistics can instead be gathered directly from each large signal in $\mathcal{O}(Mm)$ time. Furthermore, this method only requires $\mathcal{O}(m)$ memory instead of $\mathcal{O}(m^2)$.

Hariharan et al. [9] applied this technique to the problem of computing the covariance of every translated window in a set of larger images, which would otherwise have been intractable. To obtain a detector, they finally instantiated the full covariance matrix and employed a direct method such as Cholesky decomposition, noting that it was necessary to add some small regularization $\lambda\mathbf{I}$. Since the negative examples dominate the statistics, the mean of the negative class is taken to be that of the stationary distribution so that $\mathbf{r} = \bar{\mathbf{x}}_+ - \bar{\mathbf{x}}$.

2.2 Correlation Filters

The algorithm of correlation filters [10] in its unconstrained form [11] is simply linear least-squares regression applied to the set of all circular shifts of every example. While including circular shifts in the training set may seem peculiar, it leads to a system of equations which can be constructed and solved in the Fourier domain.

It is a famous result that LDA is equivalent to linear least-squares regression when the desired outputs of the regression problem take on exactly two distinct values, corresponding to the two classes [13]. Given a set of general vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathcal{R}^m$ with desired outputs $y_1, \ldots, y_n \in \mathcal{R}$, the regularized problem, also known as ridge regression, finds the solution to

$$\min_{\mathbf{w}, b} \quad \frac{1}{2n} \sum_{k=1}^{n} \left(\mathbf{w}^T \mathbf{x}_k + b - y_k\right)^2 + \frac{\lambda}{2} \|\mathbf{w}\|^2 \quad . \tag{5}$$

If the examples are assumed to be zero-mean, then the solution is obtained by taking the bias to be the mean label $b = \bar{y}$, and solving for the template in

$$\min_{\mathbf{w}} \quad \frac{1}{2n} \sum_{k=1}^{n} \left(\mathbf{w}^T \mathbf{x}_k - y_k\right)^2 + \frac{\lambda}{2} \|\mathbf{w}\|^2 \quad . \tag{6}$$

The optimal template is obtained in closed form $\mathbf{w} = (\mathbf{S} + \lambda \mathbf{I})^{-1} \mathbf{r}$ with $\mathbf{S}$ defined as in (3) and the right-hand side given

$$\mathbf{r} = \frac{1}{n} \sum_{k=1}^{n} y_k \mathbf{x}_k \quad . \tag{7}$$

See Appendix B.2 for the derivation. Note that if we choose $y_k = 0$ for negative examples, then they do not appear in the solution beyond their contribution to the covariance and the mean.

Let us again consider each $\mathbf{x}_k$ to be a scalar-valued time-series of length m. Computing the expected covariance of all circular shifts of these examples generates a circulant matrix $S_{ui} = h[(i - u) \bmod m]$. This $m \times m$ matrix is defined by only m unique elements, or $\lceil m/2 \rceil$ accounting for symmetry. These are estimated from data according to

$$h[\delta] = \frac{1}{mn} \sum_{k=1}^{n} \sum_{t=0}^{m-1} x_k[t] \cdot x_k[(t + \delta) \bmod m] \quad . \tag{8}$$

See Appendix B.3 for the derivation. This means that the covariance of two samples $x[u]$ and $x[i]$ is estimated from all pairs of samples which have relative displacement $(i - u) \bmod m$, including some displacements which cross the boundary of the signal and wrap around.

Circulant matrices can be inverted efficiently in the Fourier domain because a matrix-vector product amounts to periodic cross-correlation. Let $\star$ denote the periodic cross-correlation operator such that

$$(w \star x)[u] = \sum_{t=0}^{m-1} w[t] \cdot x[(u + t) \bmod m] \tag{9}$$

for $u = 0, \dots, m - 1$, and recall that this is equivalent to element-wise multiplication in the Fourier domain $\mathcal{F}\{\mathbf{w} \star \mathbf{x}\} = \operatorname{conj}(\hat{\mathbf{w}}) \circ \hat{\mathbf{x}}$, where we denote the Fourier transform of a signal $\mathcal{F}\{\mathbf{x}\} = \hat{\mathbf{x}}$. Multiplication by the circulant covariance matrix $\mathbf{z} = \mathbf{S}\mathbf{w}$ computes the cross-correlation $\mathbf{z} = \mathbf{h} \star \mathbf{w}$ or equivalently $\hat{\mathbf{z}} = \operatorname{conj}(\hat{\mathbf{h}}) \circ \hat{\mathbf{w}}$. This enables the re-expression of $\mathbf{w} = (\mathbf{S} + \lambda \mathbf{I})^{-1} \mathbf{r}$ as a diagonal system of equations

$$\hat{\mathbf{w}} = \left[\operatorname{diag}(\operatorname{conj}(\hat{\mathbf{h}})) + \lambda \mathbf{I} \right]^{-1} \hat{\mathbf{r}} \quad . \tag{10}$$

Whereas a general $m \times m$ system of equations requires $\mathcal{O}(m^3)$ time to solve via factorization, the solution to this system is obtained in $\mathcal{O}(m)$ time, with $\mathcal{O}(m \log m)$ additional time required to compute the FFT. Further, while factorization algorithms generally require $\mathcal{O}(m^2)$ space to store the full matrix, this system can be solved in $\mathcal{O}(m)$ space.

The system in (10) can also be constructed in the Fourier domain since

$$h[\delta] = \frac{1}{mn} \sum_{k=1}^{n} (x_k \star x_k)[\delta] \ , \qquad r[u] = \frac{1}{mn} \sum_{k=1}^{n} (y_k \star x_k)[u] \quad . \tag{11}$$

See Appendix B.4 for the derivation. This can be performed in $\mathcal{O}(nm \log m)$ time. Taking the desired response function $y_k[t]$ to be zero everywhere for negative examples and an impulse function (1 at the origin, 0 everywhere else) for positive examples results in the same right-hand side $\mathbf{r} = \bar{\mathbf{x}}_+ - \bar{\mathbf{x}}$ as in LDA.

Henriques et al. [12] proposed correlation filters as an alternative to HNM. They approximated the set of all windows in an image by the set of all circular shifts of a subset of windows which is sufficient to cover the image with significant overlap (see Figure 2). This subset would need to be re-sampled to train a detector of a different size.

2.3 Related Work

A number of other works have used fast correlation in the Fourier domain to accelerate the process of training a detector. Anguita et al. [14] used it to efficiently compute subgradients when training an SVM across all windows in a set of images. However, this is liable to be even slower than HNM, since the negative set must be traversed per gradient descent iteration. Dubout and Fleuret [15] treated images as mini-batches within stochastic descent and used the FFT to efficiently compute the subgradient of the objective function across all windows in an image. While this is undoubtedly more efficient than naively computing inner products, it cannot rival the closed-forms solution of correlation filters.

Rodriguez et al. [16] proposed an objective function which comprises a hinge loss on each un-shifted example plus a least-squares loss on all circular shifts, and showed that it can be solved in a canonical SVM framework. They noted that the loss over circular shifts can be considered a linear transformation of the space in which the margin is measured as in [17]. Our paper provides a method to obtain such a transformation from a large training set, without the need to re-compute it for different sizes. It would also be possible to adopt the Toeplitz matrix in their framework to eliminate periodic effects.

Henriques et al. [12,18] showed that the linear kernel matrix also exhibits block-circulant structure, where the size of the blocks is the number of examples rather than of the number of feature channels. In addition to ridge regression, they used this dual form to consider Support Vector Regression (SVR). We restrict discussion to the primal form in this paper, since we are primarily interested in learning from a large number of examples.

3 Fast Estimation of the Toeplitz Covariance

The previous section established that the circulant system is not only solved but also constructed efficiently in the Fourier domain. In this section, we briefly demonstrate that the FFT can likewise be used to construct the Toeplitz system.

It's clear on inspection of (4) that the elements of the Toeplitz matrix $S_{ui} = g[i - u]$ are computed by a sum of (non-periodic) auto-correlations. Let us introduce ψ_k to denote ϕ_k padded from length M to length $P = M + m - 1$

with zeros. Then g can be obtained via *periodic* auto-correlation

$$g[\delta] = \frac{1}{N\rho(\delta)} \sum_{k=1}^{N} (\psi_k \star \psi_k)[\delta \bmod P] \ , \tag{12}$$

taking only the subset $\delta = -m+1, \dots, m-1$ of each output. See Appendix B.5 for details. Comparing this expression to (11), the unique elements of the Toeplitz matrix $g[\delta]$ can be obtained in almost exactly the same manner as those of the circulant matrix $h[\delta]$.

This can be performed using the FFT in $\mathcal{O}(M \log M)$ time per signal. This implies that the statistics can be gathered for any window size $m \leq M$ without affecting the asymptotic computational complexity. As far as we know, the original authors did not take advantage of this aspect of the problem.

4 From Toeplitz to Circulant

We have now established that the Toeplitz and circulant covariance matrices can be estimated with similar computational effort, and that the circulant system can be solved very efficiently. However, a distinct advantage of adopting the Toeplitz covariance is that, comparing (4) and (8), its elements do not depend on the length m of the template $\mathbf{w}$. This means that the same covariance $g[\delta]$ could be used to learn affine functions $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$ of different sizes.

It also provides a far more elegant way to obtain statistics from a set of larger signals, of which every sub-signal could be considered a negative example. Unlike correlation filters, where it is necessary to sample windows of size m, the stationary covariance can be estimated from the whole signal as in (4).

This section formulates an expression for the elements of the circulant matrix $h[\delta]$ from those of the Toeplitz matrix $g[\delta]$. This is performed in the same way that a circulant matrix is obtained in correlation filters: by incorporating all circular shifts of all signals in some set. The set which we consider is one which possesses stationarity.

Theorem 1. *If a set of length-m signals is stationary with Toeplitz covariance matrix $S_{ij} = g[j-i]$, then the covariance of the set of all circular shifts of these signals is circulant $S_{ij} = h[(j-i) \bmod m]$ with elements*

$$h[\delta] \ = \ (1-\theta)\, g[\delta \bmod m] + \theta\, g[-(-\delta \bmod m)] \tag{13}$$

for $\delta = 0, \dots, m-1$ with $\theta = (\delta \bmod m)/m$.

Proof. See Appendix A.

This is a convex combination of the Toeplitz covariance for the relative displacements of $(\delta \bmod m)$ and $-(-\delta \bmod m)$, with greater weight given to the smaller of the two. The intuition behind this is that, under periodic extension, a given displacement from every position in the signal is more often observed as the

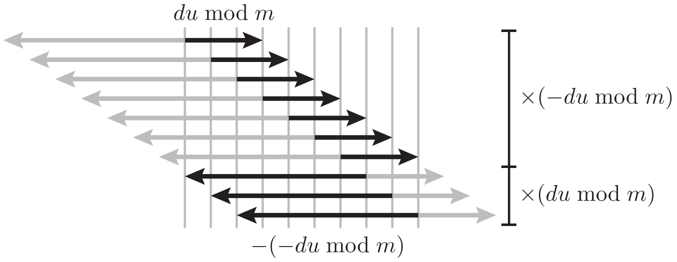

Fig. 3: Under periodic extension, a relative displacement $\delta \neq 0$ from every position in the signal is more often observed as the smaller displacement of the two modulo complements. For example, a small positive and a large negative displacement are both predominantly observed as a small positive displacement.

smaller of it and its modulo complement (see Figure 3). This expression enables correlation filters of arbitrary size to be trained from a stationary distribution, without having to choose explicit negative examples. The result is generalized to two-dimensional vector-valued signals in the following section.

5 Multi-Channel, Two-Dimensional Signals

This section generalizes the results thus far from time-series to feature images, or from single-channel, one-dimensional signals to multi-channel, two-dimensional signals. We denote elements samples of a feature image $x^p[u, v]$ for channel p at position (u, v).

5.1 Toeplitz Covariance Matrix

For the covariance matrix of two-dimensional signals of size $m \times \ell$ with c channels, we replace $u \leftarrow (u, v, p)$ and $i \leftarrow (i, j, q)$, using (u, v) and (i, j) to denote 2D positions and p and q denote feature indices. Stationarity of such signals is expressed in the constraint

$$S_{(u,v,p),(i,j,q)} = g_{pq}[i - u, j - v] \ , \tag{14}$$

where the four-dimensional array $g_{pq}[du, dv]$ contains $c^2(2m-1)(2\ell-1)$ unique elements. We refer to the structure of this covariance matrix as "block two-level Toeplitz." The symmetry of $\mathbf{S}$ gives the further redundancy $g_{pq}[du, dv] = g_{qp}[-du, -dv]$. Note that there is no assumption of stationarity across channel indices p and q. This matrix can also be efficiently estimated in the Fourier domain with appropriate zero-padding as in Section 3 (see Appendix B.6). The multi-channel stationary mean is constant per-channel $\bar{x}^p[u, v] = \mu_p$.

For a particular vectorization of the feature image, this $m\ell c \times m\ell c$ matrix is an $m \times m$ Toeplitz matrix of $\ell \times \ell$ Toeplitz matrices of $c \times c$ blocks. However, we prefer to remain agnostic to the order of vectorization and use joint indices

(u, v, p). What matters is that $\mathbf{S}$ is a linear operator which maps $\mathcal{R}^{m\ell c} \to \mathcal{R}^{m\ell c}$ such that $\mathbf{z} = \mathbf{S}\mathbf{x}$ implies

$$z^p[u, v] = \sum_{i=0}^{m-1} \sum_{j=0}^{\ell-1} \sum_{q=1}^{c} g_{pq}[i-u, j-v] \cdot x^q[i, j] \ . \tag{15}$$

5.2 Circulant Covariance Matrix

Variously known as Vector Correlation Filters or Multi-Channel Correlation Filters, the periodic case is similar to the pure stationary case, with elements of the covariance matrix defined

$$S_{(u,v,p),(i,j,q)} = h_{pq}[(i-u) \bmod m, (j-v) \bmod \ell] \ . \tag{16}$$

The only difference is the introduction of the modulo operators. The four-dimensional array $h_{pq}[du, dv]$ contains $c^2 m\ell$ unique elements, with the symmetry of the matrix yielding the further redundancy $h_{pq}[du, dv] = h_{qp}[-du \bmod m, -dv \bmod \ell]$.

Rather than being diagonalized by the 1D Fourier transform, this "block two-level circulant" matrix is block-diagonalized by applying the 2D Fourier transform to each channel independently, a fact which a slew of recent vision papers have taken advantage of [19,20,12,21]. After transforming each channel, the problem decomposes into a $c \times c$ complex linear system of equations per sample. See Appendix B.7 for the form of these equations.

Introducing $d = m\ell$ to denote the number of pixels, the time required to compute necessary transforms and then solve these systems of equations is $\mathcal{O}(c^2 d \log d + c^3 d)$. Once the system has been constructed and each block factorized, subsequent solutions can be be obtained in $\mathcal{O}(c^2 d + cd \log d)$ time for back-substitution and inverse transforms. The memory required is the same as to store $h_{pq}[du, dv]$. In contrast, to solve this system using factorization would take $\mathcal{O}(c^3 d^3)$ time and $\mathcal{O}(c^2 d^2)$ memory, with subsequent solutions obtained in $\mathcal{O}(c^2 d^2)$ time. For even modest template sizes, this makes an enormous difference. Furthermore, transforms of each channel and inversions of each block can be performed in parallel.

5.3 From Toeplitz to Circulant

The case for 2D signals is more involved since displacements can wrap around horizontal and/or vertical boundaries. Elements of the circulant matrix are given

$$\begin{aligned} h_{pq}[du, dv] = \ & (1-\alpha)(1-\beta)\, g_{pq}[du \bmod m, dv \bmod \ell] \\ & + (1-\alpha)\beta\, g_{pq}[du \bmod m, -(-dv \bmod \ell)] \\ & + \alpha(1-\beta)\, g_{pq}[-(-du \bmod m), dv \bmod \ell] \\ & + \alpha\beta\, g_{pq}[-(-du \bmod m), -(-dv \bmod \ell)] \end{aligned} \tag{17}$$

with $\alpha = (du \bmod m)/m$, $\beta = (dv \bmod \ell)/\ell$. The derivation follows the same technique as the one-dimensional case.

6 Solving Toeplitz Systems

Unfortunately, Toeplitz matrices are not diagonalized by the Fourier transform as circulant matrices are. There is, however, an extensive and varied body of literature surrounding the solution of Toeplitz systems, and we briefly review some key results.

6.1 Direct Methods

Recall that a general $m \times m$ system of equations can be factorized in $\mathcal{O}(m^3)$ time with subsequent solutions obtained in $\mathcal{O}(m^2)$ time. Levinson recursion [22,23] allows Toeplitz systems to instead be factorized in $\mathcal{O}(m^2)$ time, with the Gohberg-Semencul formula [24] enabling solutions to then be obtained in $\mathcal{O}(m \log m)$ time. This is entirely without inflicting the $\mathcal{O}(m^2)$ memory requirement of instantiating the explicit matrix or its inverse. There also exist "superfast" or "asymptotic" algorithms [25,26] which solve a system in $\mathcal{O}(m \log^2 m)$ time without factorization, although the hidden coefficients can be large. Levinson recursion has been generalized to solve $mc \times mc$ block Toeplitz systems, comprising an $m \times m$ Toeplitz structure of arbitrary $c \times c$ blocks, in an algorithm that takes $\mathcal{O}(c^3 m^2)$ time [27]. This is useful for multi-channel, *one*-dimensional signals.

Unfortunately, in the extension to two-level Toeplitz matrices, which are our primary interest in vision, algorithms based on Levinson recursion cannot do better than to treat one level as a general matrix [28,29]. For $m \times \ell$ images with c feature channels, this only enables inversion of the Toeplitz covariance matrix in $\mathcal{O}(c^3 \min(m^2 \ell^3, m^3 \ell^2))$ time. A handful of obscure exceptions have been identified [30,29], although they do not seem pertinent to us.

6.2 Iterative Methods

While the Fourier transform cannot be used directly to invert a Toeplitz matrix, it does enable fast evaluation of matrix-vector products. This is achieved by extending an $m \times m$ Toeplitz matrix to form a $(2m - 1) \times (2m - 1)$ circulant matrix. This *does* extend to block two-level Toeplitz matrices, as $\mathbf{z} = \mathbf{S}\mathbf{x}$ gives

$$z^p[u, v] = \sum_{q=1}^{c} \sum_{i=0}^{m-1} \sum_{j=0}^{\ell-1} g_{pq}[i - u, j - v]\, x^q[i, j] = \sum_{q=1}^{c} (g_{pq} \star \tilde{x}^q)[u, v] \qquad (18)$$

where $\tilde{x}^q[u, v]$ denotes a zero-padded version of $x^q[u, v]$. For images with d pixels and c channels, this allows multiplication to be performed in $\mathcal{O}(c^2 d \log d)$ time, all without instantiating the full matrix.

The existence of a fast multiplication routine suggests iterative first-order methods. In fact, a number of past works have proposed to solve Toeplitz systems using the Preconditioned Conjugate Gradient (PCG) method. The convergence rate of this algorithm depends on the condition number of the matrix and how tightly clustered its eigenvalues are [31]. PCG considers the equivalent problem

$\mathbf{MSw} = \mathbf{Mr}$, where the preconditioner $\mathbf{M}$ must be full rank and $\mathbf{MS}$ has more desirable spectral properties than $\mathbf{S}$ alone. Most works have centered around the choice of preconditioner, with Chan and Ng [32] in particular arguing that an effective preconditioner renders the number of iterations a small constant, yielding the solution to an $m \times m$ Toeplitz system in $\mathcal{O}(m \log m)$ time.

The ideal choice is $\mathbf{M} = \mathbf{S}^{-1}$, however to obtain this matrix is to solve the original problem. Inverse circulant matrices make attractive preconditioners because they are easily computed and circulant matrices are in some sense "close" to Toeplitz matrices. Strang [33] originally proposed a circulant matrix which used only the inner diagonals of the Toeplitz matrix and was shown to guarantee superlinear convergence for a large class of problems [34]. Chan [35] instead considered the nearest circulant matrix and observed empirically that it was more effective at reducing the condition number and producing a clustered spectrum. Two-level circulant preconditioners have previously been explored for block Toeplitz [36] and two-level Toeplitz systems [32], but to our knowledge not for block two-level Toeplitz systems. Serra Capizzano and Tyrtyshnikov [37] presented the theoretical result that multi-level circulant preconditioners are not guaranteed superlinear convergence for multi-level Toeplitz matrices by the same mechanism as one-level, noting that fast convergence is still possible in practice.

Somewhat surprisingly, the circulant covariance matrix which we obtained in Section 4 is in fact the nearest (block multi-level) circulant matrix, analogous to the preconditioner in [35]. Therefore we can optionally employ the circulant solution (i.e. learn a correlation filter) as a preconditioner within conjugate gradient. This preconditioner results in significantly faster convergence (see Figure 6).

To summarize, this leaves us with several options to learn a detector. Firstly, we can choose to solve either the Toeplitz or the circulant system. If we choose to solve the circulant system, it is done in closed form. If we instead decide to solve the Toeplitz system, then we can either solve it directly by Cholesky decomposition or iteratively using conjugate gradient, with or without the circulant inverse as a preconditioner.

6.3 An Effective Heuristic

The performance of the detector learned using circulant covariance can be greatly increased with a simple heuristic, which is to train a larger detector than desired and then crop it down to size *after* training. One extra feature pixel on all sides was found to be sufficient. This was discovered in the code of [12], although as far as we are aware, it has not previously been discussed. Figure 4 shows that nearly identical performance to the Toeplitz method is achieved.

While we do not have a theoretical analysis of the cropping heuristic, it at least makes intuitive sense. The most highly correlated feature pixels are those which are adjacent. A circulant matrix considers two pixels on opposite edges to be adjacent. The probable discontinuity between these elements in the mean positive image is likely to be something which the detector learns about the positive set. That one feature pixel is sufficient suggests that the correlation of samples decays rapidly with increasing distance.

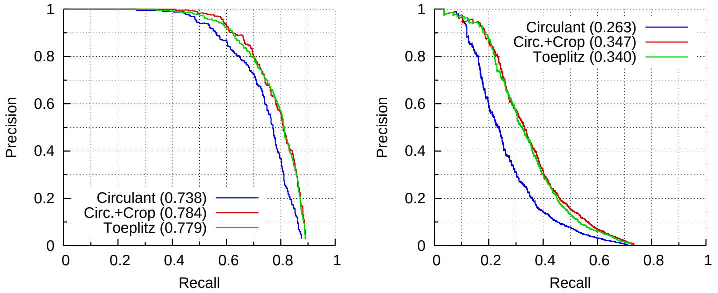

Fig. 4: Precision versus recall for INRIA (left) and Caltech (right) pedestrian detection datasets. Average precision is shown in parentheses.

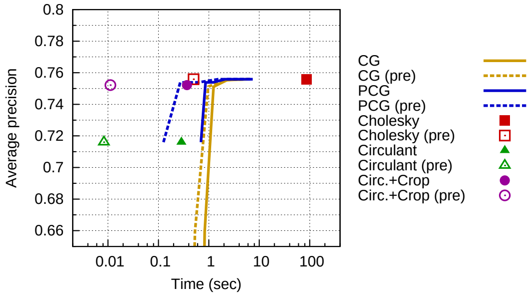

Fig. 5: Average precision versus training time for pedestrian detection on the INRIA dataset. Iterative methods trace a path, closed-form methods mark a single point. Timing shown for full computation and to obtain subsequent detectors (after *pre*-computation). The template was 12 × 33 features with 31 channels.

7 Empirical Study

Experiments were conducted on HOG images [38] using the 31-channel implementation of [39]. The stationary statistics were estimated once from four million random images in ImageNet [3] to illustrate that all techniques can draw on a huge number of negative examples. Regularization $\lambda \mathbf{I}$ was added with $\lambda = 10^{-2}$. Further practical details are found in Appendix C.

7.1 Detection Performance

The detectors learned under Toeplitz and circulant assumptions were compared for the task of pedestrian detection (see Figure 4). Toeplitz was found to consistently outperform circulant. Surprisingly, learning with a circulant matrix and using the extend-and-crop heuristic rivals the performance of the Toeplitz method. The detectors were evaluated on the ETHZ Shape Classes dataset [40], although the results were found to be noisy and less conclusive due to its insufficient size (see Appendix D).

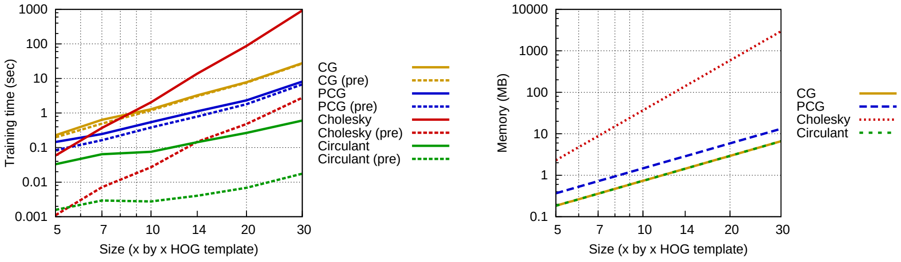

Fig. 6: Time and memory complexity of the different algorithms as the size of the template increases. These specific results are for 31-channel HOG images. Training time is empirical and memory is theoretical.

7.2 Time and Memory

Figure 5 plots the average performance of each detector against its training time. Figure 6 shows how the time and memory demands of the different algorithms grow with template size. We present times with and without pre-computable factorizations and transforms included (note that these must be performed per detector size). Algorithms were implemented in Go, making use of FFTW and LAPACK where appropriate.

Cholesky factorization is fast for compact templates. However, as the template size grows, it becomes relatively slow unless the factorization can be precomputed. The memory required to store such a factorization also grows rapidly with the template size, soon reaching gigabytes. This makes it impractical to cache and load factorizations for several detector sizes, and may simply be infeasible or restrictive in some scenarios. For the problem of pedestrian detection, conjugate gradient offers a speed increase of nearly two orders of magnitude over computing the factorization. The direct circulant method is several times faster again, however this requires one to either accept diminished performance or employ the cropping heuristic, the behaviour of which is not yet well understood.

8 Conclusion

Toeplitz and circulant covariance matrices have both previously been employed, within simplistic classifiers, to avoid Hard Negative Mining when learning from a large negative set. This paper has elucidated commonalities between these two techniques and proposed improvements to each. Compared to existing methods which use Toeplitz structure, identical detectors are obtained in orders of magnitude less time and memory. Circulant methods were shown to offer a further order of magnitude increase in speed for a small degradation of performance. Compared to existing methods which use circulant structure, the negative set does not need to be revisited per detector size. These are exciting developments for higher-level vision algorithms which involve learning linear templates.

Acknowledgement. This research was supported by an Australian Research Council (ARC) Discovery Grant DP140100793.

References

1. Kalal, Z., Matas, J., Mikolajczyk, K.: P-N learning: Bootstrapping binary classifiers by structural constraints. In: CVPR, IEEE (2010) 49–56
2. Bolme, D.S., Beveridge, J.R., Draper, B.A., Lui, Y.M.: Visual object tracking using adaptive correlation filters. In: CVPR, IEEE (2010) 2544–2550
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: CVPR, IEEE (2009) 248–255
4. Dean, T., Ruzon, M., Segal, M., Shlens, J., Vijayanarasimhan, S., Yagnik, J.: Fast, accurate detection of 100,000 object classes on a single machine. In: CVPR. (2013)
5. Bourdev, L., Malik, J.: Poselets: Body part detectors trained using 3D human pose annotations. In: ICCV, IEEE (2009) 1365–1372
6. Singh, S., Gupta, A., Efros, A.A.: Unsupervised discovery of mid-level discriminative patches. In: ECCV. (2012)
7. Doersch, C., Singh, S., Gupta, A., Sivic, J., Efros, A.A.: What makes Paris look like Paris? In: ACM Transactions on Graphics. Volume 31. (2012)
8. Malisiewicz, T., Gupta, A., Efros, A.A.: Ensemble of exemplar-SVMs for object detection and beyond. In: ICCV. (2011) 89–96
9. Hariharan, B., Malik, J., Ramanan, D.: Discriminative decorrelation for clustering and classification. In: ECCV, Springer (2012)
10. Mahalanobis, A., Vijaya Kumar, B.V.K., Casasent, D.: Minimum average correlation energy filters. Applied Optics **26** (1987) 3633–40
11. Mahalanobis, A., Vijaya Kumar, B.V.K., Song, S., Sims, S.R.F., Epperson, J.F.: Unconstrained correlation filters. Applied Optics **33** (1994) 3751–9
12. Henriques, J.F., Carreira, J., Caseiro, R., Batista, J.: Beyond hard negative mining: Efficient detector learning via block-circulant decomposition. In: ICCV, IEEE (2013)
13. Fukunaga, K.: Introduction to Statistical Pattern Recognition. 2nd edn. Academic Press (1990)
14. Anguita, D., Boni, A., Pace, S.: Fast training of support vector machines for regression. In: International Joint Conference on Neural Networks. Volume 5., IEEE (2000) 210–214
15. Dubout, C., Fleuret, F.: Accelerated training of linear object detectors. In: CVPR Workshop on Structured Prediction, IEEE (2013) 572–577
16. Rodriguez, A., Boddeti, V.N., Vijaya Kumar, B.V.K., Mahalanobis, A.: Maximum Margin Correlation Filter: A new approach for localization and classification. Transactions on Image Processing **22** (2013) 631–43
17. Ashraf, A.B., Lucey, S., Chen, T.: Reinterpreting the application of Gabor filters as a manipulation of the margin in linear support vector machines. PAMI **32** (2010) 1335–41
18. Henriques, J.F., Caseiro, R., Martins, P., Batista, J.: High-speed tracking with kernelized correlation filters. Technical report, University of Coimbra (2014)
19. Boddeti, V.N., Kanade, T., Vijaya Kumar, B.V.K.: Correlation filters for object alignment. CVPR (2013) 2291–2298
20. Bristow, H., Eriksson, A., Lucey, S.: Fast convolutional sparse coding. In: CVPR, IEEE (2013) 391–398

21. Kiani Galoogahi, H., Sim, T., Lucey, S.: Multi-channel correlation filters. In: ICCV, IEEE (2013)
22. Levinson, N.: The Wiener RMS error criterion in filter design and prediction. Journal of Mathematics and Physics (1947)
23. Trench, W.F.: An algorithm for the inversion of finite Toeplitz matrices. Journal of the Society for Industrial and Applied Mathematics **12** (1964) 515–522
24. Gohberg, I., Semencul, A.: On the inversion of finite Toeplitz matrices and their continuous analogs. Mat. Issled. **2** (1972) 201–233
25. Brent, R.P., Gustavson, F.G., Yun, D.Y.Y.: Fast solution of Toeplitz systems of equations and computation of Padé approximants. Journal of Algorithms **295** (1980) 259–295
26. Ammar, G.S., Gragg, W.B.: Superfast solution of real positive definite Toeplitz systems. SIAM Journal on Matrix Analysis and Applications **9** (1988) 61–76
27. Akaike, H.: Block Toeplitz matrix inversion. SIAM Journal on Applied Mathematics **24** (1973) 234–241
28. Wax, M., Kailath, T.: Efficient inversion of Toeplitz-block Toeplitz matrix. IEEE Transactions on Acoustics, Speech, and Signal Processing **31** (1983) 1218–1221
29. Yagle, A.E.: A fast algorithm for Toeplitz-block-Toeplitz linear systems. In: ICASSP. Volume 3., IEEE (2001) 1929–1932
30. Turnes, C.K., Balcan, D., Romberg, J.: Image deconvolution via superfast inversion of a class of two-level Toeplitz matrices. In: ICIP, IEEE (2012) 3073–3076
31. Nocedal, J., Wright, S.J.: Numerical Optimization. 2nd edn. Springer (2006)
32. Chan, R.H., Ng, M.K.: Conjugate gradient methods for Toeplitz systems. SIAM Review **38** (1996) 427–482
33. Strang, G.: A proposal for Toeplitz matrix calculations. Studies in Applied Mathematics (1986)
34. Chan, R.H.: Circulant preconditioners for Hermitian Toeplitz systems. SIAM Journal on Matrix Analysis and Applications **10** (1989) 542–550
35. Chan, T.F.: An optimal circulant preconditioner for Toeplitz systems. SIAM Journal on Scientific and Statistical Computing **9** (1988) 766–771
36. Chan, T.F., Olkin, J.A.: Circulant preconditioners for Toeplitz-block matrices. Numerical Algorithms **6** (1994) 89–101
37. Serra Capizzano, S., Tyrtyshnikov, E.E.: Any circulant-like preconditioner for multilevel matrices is not superlinear. SIAM Journal on Matrix Analysis and Applications **21** (2000) 431–439
38. Dalal, N., Triggs, B.: Histograms of Oriented Gradients for Human Detection. In: CVPR. Volume 1., IEEE (2005) 886–893
39. Felzenszwalb, P.F., Girshick, R.B., McAllester, D.A., Ramanan, D.: Object detection with discriminatively trained part based models. PAMI **32** (2010) 1627–1645
40. Ferrari, V., Tuytelaars, T., Van Gool, L.: Object detection by contour segment networks. In: ECCV 2006. (2006)

A Proof of Theorem 1

Proof. Recall that assuming stationarity of the set of examples from which the statistics are estimated results in a covariance matrix with Toeplitz structure

$$S_{ui} = \frac{1}{n} \sum_{k=1}^{n} x_k[u] \cdot x_k[i] = g[i-u] \ . \tag{19}$$

To obtain a circulant covariance matrix, instead consider the statistics of the augmented set containing all circular shifts $t = 0, \dots, m-1$ of every example

$$S_{ui} = \frac{1}{mn} \sum_{k=1}^{n} \sum_{t=0}^{m-1} x_k[(t+u) \bmod m] \cdot x_k[(t+i) \bmod m] \ . \tag{20}$$

This is shown to be circulant by replacing $t \leftarrow t - u$

$$S_{ui} = \frac{1}{mn} \sum_{k=1}^{n} \sum_{t=0}^{m-1} x_k[t] \cdot x_k[(t+i-u) \bmod m] = h[(i-u) \bmod m] \tag{21}$$

since $(t+\delta) \bmod m = (t + (\delta \bmod m)) \bmod m$. To obtain $h[\delta]$ from $g[\delta]$, we split the summation based on whether $(t+\delta) \bmod m < t$. Thus the inner sum in the above expression becomes

$$\begin{aligned} \sum_{t=0}^{m-1} x_k[t] \cdot x_k[(t+\delta) \bmod m] = & \sum_{t=0}^{(-\delta \bmod m)-1} x[t] \cdot x[t + (\delta \bmod m)] \\ & + \sum_{t=(-\delta \bmod m)}^{m-1} x[t] \cdot x[t - (-\delta \bmod m)] \ . \end{aligned} \tag{22}$$

Combining (19), (21) and (22) yields the final formula. □

B Derivations

B.1 Zero-Mean Assumption in LDA

If the examples are not zero-mean, then the covariance matrix is computed

$$\tilde{\mathbf{S}} = \frac{1}{n} \sum_{k=1}^{n} (\mathbf{x}_k - \bar{\mathbf{x}})(\mathbf{x}_k - \bar{\mathbf{x}})^T \ . \tag{23}$$

Rather than subtract the mean from every example, this "centering" can be performed using the identity $\tilde{\mathbf{S}} = \mathbf{S} - \bar{\mathbf{x}}\bar{\mathbf{x}}^T$.

Under stationarity, the mean is a constant image $\bar{\mathbf{x}} = \mu \mathbf{1}$, where μ is the mean sample value and $\mathbf{1}$ is vector of all ones. Therefore the difference between the two matrices $\mathbf{S} - \tilde{\mathbf{S}} = \mu^2 \mathbf{1}\mathbf{1}^T$ is a uniformly-valued matrix, and they have identical structure.

B.2 Ridge Regression Formulation

Differentiating the objective function in (5) with respect to b yields

$$\frac{1}{n}\sum_{k=1}^{n}\left(\mathbf{w}^T\mathbf{x}_k + b - y_k\right) = 0 \tag{24}$$

$$b = \bar{y} - \mathbf{w}^T\bar{\mathbf{x}} \ . \tag{25}$$

We will not enforce the zero-mean assumption in this derivation. Making this substitution, the weights can be found by solving

$$\min_{\mathbf{w}} \quad \frac{1}{2n}\sum_{k=1}^{n}\left[\mathbf{w}^T(\mathbf{x}_k - \bar{\mathbf{x}}) - (y_k - \bar{y})\right]^2 + \frac{\lambda}{2}\left\|\mathbf{w}\right\|^2 \ . \tag{26}$$

This is equivalent to

$$\min_{\mathbf{w}} \quad \tfrac{1}{2}\mathbf{w}^T(\tilde{\mathbf{S}} + \lambda\mathbf{I})\mathbf{w} - \mathbf{w}^T\tilde{\mathbf{r}} \tag{27}$$

and hence also to the linear equation

$$(\tilde{\mathbf{S}} + \lambda\mathbf{I})\mathbf{w} = \tilde{\mathbf{r}} \tag{28}$$

where $\tilde{\mathbf{S}}$ is defined in Appendix B.1 and the right-hand side is given

$$\tilde{\mathbf{r}} = \frac{1}{n}\sum_{k=1}^{n}(y_k - \bar{y})(\mathbf{x}_k - \bar{\mathbf{x}}) \tag{29}$$

or any of the following equivalent expressions

$$\tilde{\mathbf{r}} \ = \ \frac{1}{n}\sum_{k=1}^{n} y_k(\mathbf{x}_k - \bar{\mathbf{x}}) \ = \ \frac{1}{n}\sum_{k=1}^{n}(y_k - \bar{y})\mathbf{x}_k \ = \ \frac{1}{n}\sum_{k=1}^{n} y_k\mathbf{x}_k - \bar{y}\bar{\mathbf{x}} \ . \tag{30}$$

Returning to the zero-mean assumption, $\tilde{\mathbf{S}} = \mathbf{S}$ and $\tilde{\mathbf{r}} = \mathbf{r}$.

B.3 Circulant Covariance of Circularly Shifted Examples

Let $\mathbf{x}_{kt}$ denote example $\mathbf{x}_k$ shifted by t pixels such that its elements are given

$$x_{kt}[i] = x_k[(t+i) \bmod m] \ . \tag{31}$$

The expected covariance of samples u and i computed from every shift $t = 0, \ldots, m-1$ of every example $k = 1, \ldots, n$ is

$$\begin{aligned}
S_{ui} &= \frac{1}{mn}\sum_{k=1}^{n}\sum_{t=0}^{m-1} x_{kt}[u] \cdot x_{kt}[i] && (32)\\
&= \frac{1}{mn}\sum_{k=1}^{n}\sum_{t=0}^{m-1} x_k[(t+u) \bmod m] \cdot x_k[(t+i) \bmod m] && (33)\\
&= \frac{1}{mn}\sum_{k=1}^{n}\sum_{t=0}^{m-1} x_k[t] \cdot x_k[(t+i-u) \bmod m] && (34)\\
&= h[(i-u) \bmod m] && (35)
\end{aligned}$$

and therefore the matrix is circulant with elements

$$h[\delta] = \frac{1}{mn} \sum_{k=1}^{n} \sum_{t=0}^{m-1} x_k[t] \cdot x_k[t + \delta \bmod m] \ . \tag{36}$$

B.4 Forming Circulant System in Fourier Domain

It follows from (8) that

$$h[\delta] = \frac{1}{mn} \sum_{k=1}^{n} \sum_{t=0}^{m-1} (x_k \star x_k)[\delta] \tag{37}$$

or equivalently

$$\hat{\mathbf{h}} = \frac{1}{mn} \sum_{k=1}^{n} \operatorname{conj}(\hat{\mathbf{x}}_k) \circ \hat{\mathbf{x}}_k \ . \tag{38}$$

Considering (7) computed on the set of circular shifts $\mathbf{x}_{kt}$ as defined in Appendix B.3 for all $t = 0, \ldots, m-1$ yields

$$\begin{aligned} r[u] &= \frac{1}{mn} \sum_{k=1}^{n} \sum_{t=0}^{m-1} y_k[t] \, x_{kt}[u] & (39) \\ &= \frac{1}{mn} \sum_{k=1}^{n} \sum_{t=0}^{m-1} y_k[t] \, x_k[(u+t) \bmod m] & (40) \\ &= \frac{1}{mn} \sum_{k=1}^{n} (y_k \star x_k)[u] & (41) \end{aligned}$$

or equivalently

$$\hat{\mathbf{r}} = \frac{1}{mn} \sum_{k=1}^{n} \operatorname{conj}(\hat{\mathbf{y}}_k) \circ \hat{\mathbf{x}}_k \ . \tag{42}$$

B.5 Fast Estimation of the Toeplitz Covariance

Recall that the purpose of $a(\delta)$ and $b(\delta)$ in (4) was to limit the summation over t to the range of values such that both t and $t+\delta$ are in $\{0, \ldots, M-1\}$. Defining the zero-padded signal ψ

$$\psi_k[t] = \begin{cases} \phi_k[t], & 0 \le t < M \\ 0, & M \le t < P \ , \end{cases} \tag{43}$$

it is no longer necessary to limit the range. If $|\delta| \le m-1$ and either $t + \delta < 0$ or $t + \delta > M - 1$, then $\psi_k[(t+\delta) \bmod P] = 0$. This enables the expression to be written as a modulo-P summation over $t = 0, \ldots, P-1$

$$g[\delta] = \frac{1}{N\rho(\delta)} \sum_{k=1}^{N} \sum_{t=0}^{P-1} \psi_k[t] \cdot \psi_k[(t+\delta) \bmod P] \ , \tag{44}$$

which is thus *periodic* cross-correlation as in (12).

B.6 Multi-Channel Two-Dimensional Toeplitz Covariance

The elements of the multi-channel two-dimensional Toeplitz covariance matrix can be estimated from N larger signals $\phi^p[u,v]$ of size $M \times L$ with c channels. Individual elements are found from all instances of the relative displacement (du, dv) in the image using

$$g_{pq}[du,dv] = \frac{1}{N\rho_M(du)\rho_L(dv)} \sum_{k=1}^{N} \sum_{u=a_M(du)}^{b_M(du)-1} \sum_{v=a_L(dv)}^{b_L(dv)-1} \phi_k^p[u,v] \cdot \phi_k^q[u+du, v+dv]. \tag{45}$$

The region $(u,v) \in$

$$\{a_M(du), \ldots, b_M(du)-1\} \times \{a_L(dv), \ldots, b_L(dv)-1\} \tag{46}$$

ensures that both (u,v) and $(u+du, v+dv)$ are within the domain of the image $\{0,\ldots,M-1\} \times \{0,\ldots,L-1\}$. The normalization factor $\rho_M(du)\rho_L(dv)$ measures the size of this region. These functions are defined

$$a_M(du) = \max(0, -du) \qquad b_M(du) = M - \max(0, du) \tag{47}$$

$$\rho_M(du) = b_M(du) - a_M(du) = M - |du| \ . \tag{48}$$

As was the case for scalar-valued time-series, the Toeplitz covariance matrix can be efficiently computed in the Fourier domain. Let ψ_k denote the signal ϕ_k padded with zeros to size $P \times Q$ with $P = M + m - 1$ and $Q = L + \ell - 1$. The covariance can be computed as a sum of cross-correlations between channel pairs

$$g_{pq}[du,dv] = \frac{1}{N\rho_M(du)\rho_L(dv)} \sum_{k=1}^{N} (\psi_k^p \star \psi_k^q)[du \bmod P, du \bmod Q]. \tag{49}$$

This takes $\mathcal{O}(c^2 D \log D)$ time per image, where $D = ML$ is the number of pixels.

B.7 Multi-Channel Circulant Inverse

Examining the product $\mathbf{z} = \mathbf{S}\mathbf{w}$ element-wise reveals

$$z^p[u,v] = \sum_{i=0}^{m-1} \sum_{j=0}^{\ell-1} \sum_{q=1}^{c} h_{pq}[(i-u) \bmod m, (j-v) \bmod \ell] \cdot w^q[i,j] \ , \tag{50}$$

which is a sum over circular cross-correlations

$$z^p[u,v] = \sum_{q=1}^{c} (h_{pq} \star w^q)[u] \ . \tag{51}$$

Let $\hat{\mathbf{h}}_{pq} = \mathcal{F}\{\mathbf{h}_{pq}\}$ and $\hat{\mathbf{w}}^p = \mathcal{F}\{\mathbf{w}^p\}$ denote the 2D Fourier transform of these $m \times \ell$ signals. Then the matrix-vector product is equivalent to an independent $c \times c$ complex system of equations per pixel (u,v) in the Fourier domain

$$\hat{\mathbf{z}}[u,v] = \hat{\mathbf{H}}_{uv}\, \hat{\mathbf{w}}[u,v] \tag{52}$$

where $\hat{\mathbf{x}}[u, v] = (\hat{x}^1[u, v], \ldots, \hat{x}^c[u, v])$ samples the Fourier transform of all channels at frequency (u, v), and the matrix $\hat{\mathbf{H}}_{uv}$ is constructed from the Fourier transform of all cross-channel covariance pairs (p, q) according to

$$\hat{\mathbf{H}}_{uv} = \left(\hat{h}^*_{pq}[u, v] \right)_{pq} . \tag{53}$$

Therefore to compute $\mathbf{w} = \mathbf{S}^{-1}\mathbf{r}$, it is sufficient to compute

$$\hat{\mathbf{w}}[u, v] = \hat{\mathbf{H}}_{uv}^{-1} \, \hat{\mathbf{r}}[u, v] \tag{54}$$

for every position (u, v) and then take the inverse transform $\mathbf{w}^p = \mathcal{F}^{-1}\{\hat{\mathbf{w}}^p\}$ of each channel.

C Practical Details

Our INRIA examples were 12 × 33 feature images, each extracted from a region of 62 × 146 pixels centered on a bounding box of 43 × 128 pixels. For the Caltech dataset, each example was a 9 × 19 feature image extracted from a region of 50 × 90 pixels, corresponding to a bounding box of 26 × 64 pixels. The HOG descriptor of [39] was slightly modified to eliminate minor boundary effects. The results presented here all used a cell size of four pixels.

Figure 4 was produced using the code of [12], with a few minor changes, to facilitate comparison. However, the average precision cited in Figures 5 and 6 was measured using our own implementation. For these experiments, practical details were as follows. Multi-scale search was performed in geometric steps of 1.07 (roughly 10 scales per octave). Detections were selected greedily by score, with each detection suppressing all candidates with which it shared an intersection-over-union of more than 30%. Candidates which were not a maximum in their local four-connected neighborhood were not considered. Detections were deemed to be true positives if they have more than 50% intersection-over-union with a ground-truth box. Each ground truth label can only match to one detection and this is performed greedily, with the highest scoring detection taking the rectangle with which it overlaps the most.

D Further Empirical Results

These results were not included in the main text because they are noisy and less conclusive, due to the insufficient size of the dataset.

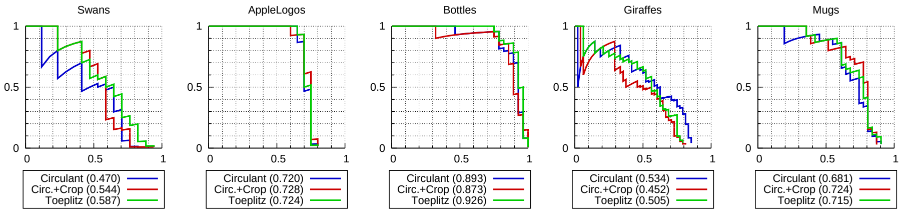

Fig. 7: Precision versus recall for ETHZ Shape Classes dataset. Average precision shown in parentheses.