Correlation Filters with Limited Boundaries

Hamed Kiani[1], Terence Sim[2] and Simon Lucey[3]
[1]PAVIS, Istituto Italiano di Tecnologia, [2]SoC, National University of Singapore, [3]RI, Carnegie Mellon University

Interest in correlation filters has been reignited in the vision world through the recent work of Bolme et al. [5] on Minimum Output Sum of Squared Error (MOSSE) correlation filters for object detection and tracking. Bolme et al.'s work was able to circumvent some of the classical problems with correlation filters and performed well in tracking under changes in rotation, scale, lighting and partial occlusion. MOSSE correlation filter [1] can be expressed in the spatial domain as solving a ridge regression problem,

$$E(\mathbf{h}) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{D}||\mathbf{y}_i(j) - \mathbf{h}^\top \mathbf{x}_i[\Delta\tau_j]||_2^2 + \frac{\lambda}{2}||\mathbf{h}||_2^2 \tag{1}$$

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. $\mathbb{C} = [\Delta\tau_1, \dots, \Delta\tau_D]$ represents the set of all circular shifts for a signal of length D.

The major problem with the objective in Equation 1, however, is that the shifted image patches $\mathbf{x}[\Delta\tau]$ at all values of $\Delta\tau \in \mathbb{C}$, except where $\Delta\tau = \mathbf{0}$ (no shift), are exploited to estimate a discriminative template from an unbalanced set of "real-world" and "synthetic" examples (Figure 1(c)). In signal-processing, one often refers to this as the *boundary effect*. These synthetic examples are created through the application of a circular shift on the real-world examples, $\mathbf{x}[\Delta\tau]$, and are supposed to be representative of those examples at different translational shifts. We use the term synthetic, as all these shifted examples are plagued by circular boundary effects and are not truly representative of the shifted example (see Figure 1(c)). As a result, the training set used for learning the template is extremely unbalanced with one real-world example for every $D-1$ synthetic examples (where D is the dimensionality of the examples). These boundary effects can dramatically affect the resulting performance of the estimated template [18].

we proposed to circumvent this problem by allowing the training signal $\mathbf{x} \in \mathbb{R}^T$ to be a larger size than the filter $\mathbf{h} \in \mathbb{R}^D$ such that $T > D$. Through the use of a $D \times T$ masking matrix $\mathbf{P}$, Equation 1 can be expressed as:

$$E(\mathbf{h}) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{T}||\mathbf{y}_i(j) - \mathbf{h}^\top \mathbf{P}\mathbf{x}_i[\Delta\tau_j]||_2^2 + \frac{\lambda}{2}||\mathbf{h}||_2^2 \tag{2}$$

The masking matrix $\mathbf{P}$ of ones and zeros encapsulates what part of the signal should be active/inactive. The central benefit of this augmentation in Equation 2 is the dramatic increase in the proportion of examples unaffected by boundary effects ($\frac{T-D+1}{T}$ instead of $\frac{1}{D}$ in canonical correlation filters). From this insight it becomes clear that if one chooses $T >> D$ then boundary effects become greatly diminished (Figure 1(d)). The computational cost $\mathcal{O}(D^3 + NTD)$ of solving this objective is only slightly larger than the cost of Equation 1, as the role of $\mathbf{P}$ in practice can be accomplished efficiently through a lookup table. A major contribution of this paper is to solve this objective function efficiently in terms of computational cost.

A problem arises, however, when one attempts to solve the objective in 2 in the same Fourier domain for computational efficiency. Equation 2 can be expressed in the Fourier domain as:

$$E(\mathbf{h}) = \frac{1}{2}\sum_{i=1}^{N}||\hat{\mathbf{y}}_i - \text{diag}(\hat{\mathbf{x}}_i)^\top \sqrt{D}\mathbf{F}\mathbf{P}^\top \mathbf{h}||_2^2 + \frac{\lambda}{2}||\mathbf{h}||_2^2 \tag{3}$$

Unfortunately, since we are enforcing a spatial constraint $\mathbf{P}^\top$ on $\mathbf{h}$ the efficiency of this objective balloons to $\mathcal{O}(D^3 + ND^2)$ as $\mathbf{h}$ *must* be solved in the spatial domain.

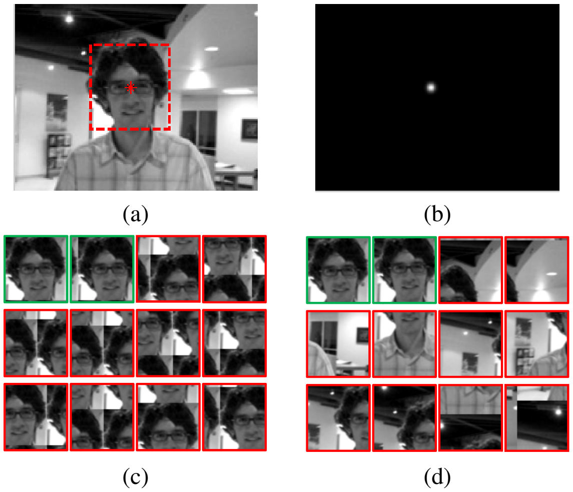
(a) (b)
(c) (d)

Figure 1: (a) An example of fixed spatial support within the image from which the peak correlation output should occur. (b) The desired output response, based on (a). (c) A subset of shifted patch examples used in a canonical correlation filter where green and red respectively denote a non-zero and zero correlation output at (b). (d) A subset of real patch examples used in our proposed correlation filter.

Our proposed approach for solving Equation 3 is introducing an auxiliary variable $\hat{\mathbf{g}}$. In this case, Equation 3 can be identically expressed as:

$$\begin{aligned} E(\mathbf{h},\hat{\mathbf{g}}) &= \frac{1}{2}\sum_{i=1}^{N}||\hat{\mathbf{y}}_i - \text{diag}(\hat{\mathbf{x}}_i)^\top \hat{\mathbf{g}}||_2^2 + \frac{\lambda}{2}||\mathbf{h}||_2^2 \\ \text{s.t.} \quad & \hat{\mathbf{g}} = \sqrt{D}\mathbf{F}\mathbf{P}^\top \mathbf{h} \ . \end{aligned} \tag{4}$$

We propose to handle the introduced equality constraints through an Augmented Lagrangian Method (ALM) [2], in particular Alternating Direction Method of Multipliers (ADMM), as detailed in our paper.

The dominant cost per iteration of the ADMM optimization process is $\mathcal{O}(T\log T)$ for FFT. There is a per-computation cost for estimating the auto- and cross-spectral energies. This cost is $\mathcal{O}(NT\log T)$ where N is the number of training images. Given that K represents the number of ADMM iterations the overall cost of the algorithm is therefore $\mathcal{O}([N+K]T\log T)$.

Key Results. Comparing our approach with a steepest descent method [3] to solve Equation 2 showed that (i) our convergence performance is largely independent to the filter size and the number of images, and (ii) relatively few iterations are required to achieve good convergence. Moreover, we evaluated our correlation filter for the task of eye detection on CMU Multi-PIE face database. The results showed that the detection performance of our approach is significantly higher than the state of the art correlation filters (around 15%). Finally, we demonstrated the superiority of the proposed method for real-time tracking (100 fps) comparing with recent leading trackers/correlation filters. The results showed the robustness of our approach against camera motion, rotation, scaling, illumination and partial occlusion.

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

[2] Stephen Boyd. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Foundations and Trends in Machine Learning*, 3:1–122, 2010. ISSN 1935-8237.

[3] MD Zeiler, Dilip Krishnan, and GW Taylor. Deconvolutional networks. *CVPR*, 2010.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.