Efficient Approximations for Support Vector Machines in Object Detection

Wolf Kienzle, Gökhan Bakır, Matthias Franz, and Bernhard Schölkopf

Max Planck Institute for Biological Cybernetics
{kienzle,gb,mof,bs}@tuebingen.mpg.de
http://www.tuebingen.mpg.de
Dept. Schölkopf, Spemannstraße 38, 72076 Tübingen, Germany

Abstract. We present a new approximation scheme for support vector decision functions in object detection. In the present approach we are building on an existing algorithm where the set of support vectors is replaced by a smaller so-called reduced set of synthetic points. Instead of finding the reduced set via unconstrained optimization, we impose a structural constraint on the synthetic vectors such that the resulting approximation can be evaluated via separable filters. Applications that require scanning an entire image can benefit from this representation: when using separable filters, the average computational complexity for evaluating a reduced set vector on a test patch of size $h \times w$ drops from $O(h \cdot w)$ to $O(h+w)$. We show experimental results on handwritten digits and face detection.

1 Introduction

It has been shown that support vector machines (SVMs) provide state-of-the-art accuracies in object detection. In time-critical applications, however, they are of limited use due to their computationally expensive decision functions.

In SVMs the time complexity of a classification operation is characterized by the following two parameters. First, it is linear in the number of support vectors (SVs). Unfortunately, it is known that for noisy problems, the number of SVs can be rather large, essentially scaling linearly with the training set size [10]. Second, it scales with the number of operations needed for computing the similarity (or kernel function) between an SV and the input. When classifying $h \times w$ patches using plain gray value features, the decision function requires an $h \cdot w$ dimensional dot product for each SV. As the patch size increases, these computations become extremely expensive: the evaluation of a single 20×20 pattern on a 320×240 image at 25 frames per second already requires 660 million operations per second. For such systems to run in (or at least near) real-time, it is therefore necessary to lower the computational cost of the SV evaluations as well.

In the past, however, research towards speeding up kernel expansions has focused exclusively on the first issue, i.e., the number of expansion vectors. It has been pointed out that one can improve evaluation speed by using approximations with smaller expansion sets. In [2] Burges introduced a method that, for a given SVM, creates a set of so-called reduced set vectors (RSVs) which approximate the decision function. In the image classification domain, speedups of the order of 10 to 30 have been reported [2, 3, 5] while the full accuracy was retained.

In contrast, this work focuses on the second issue. To this end, we borrow an idea from image processing to compute fast approximations to SVM decision functions: by constraining the RSVs to be separable, they can be evaluated via separable convolutions. This works for most standard kernels (e.g. linear, polynomial, Gaussian and sigmoid) and decreases the computational complexity of the RSV evaluations from $O(h \cdot w)$ to $O(h + w)$. One of the primary target applications for this approach is face detection, an area that has seen significant progress of machine learning based systems over the last years [7, 11, 4, 6, 12, 8].

2 Unconstrained Reduced Set Construction

The current section describes the reduced set method [2] on which our work is based. To simplify the notation in the following sections, image patches are written as matrices (denoted by capital letters).

Assume that an SVM has been successfully trained on the problem at hand. Let $\{\mathbf{X}_1, \ldots \mathbf{X}_m\}$ denote the set of SVs, $\{\alpha_1, \ldots \alpha_m\}$ the corresponding coefficients, $k(\cdot, \cdot)$ the kernel function and b the bias of the SVM solution. The decision rule for a test pattern $\mathbf{X}$ reads

$$f(\mathbf{X}) = \operatorname{sgn}\left(\sum_{i=1}^{m} y_i \alpha_i k(\mathbf{X}_i, \mathbf{X}) + b\right). \tag{1}$$

A central property of SVMs is that the decision surface induced by f corresponds to a hyperplane in the reproducing kernel Hilbert space (RKHS) associated with k [9]. The normal is given by

$$\Psi = \sum_{i=1}^{m} y_i \alpha_i k(\mathbf{X}_i, \cdot). \tag{2}$$

As the computational complexity of f scales with the number of SVs m, we can speed up its evaluation using a smaller reduced set (RS) $\{\mathbf{Z}_1, \ldots \mathbf{Z}_{m'}\}$ of size $m' < m$, i.e. an approximation to Ψ of the form

$$\Psi' = \sum_{i=1}^{m'} \beta_i k(\mathbf{Z}_i, \cdot). \tag{3}$$

To find such Ψ', i.e. the $\mathbf{Z}_i$ and their corresponding expansion coefficients β_i, we fix a desired set size m' and solve

$$\min \|\Psi - \Psi'\|^2_{\mathrm{RKHS}}. \tag{4}$$

for β_i and $\mathbf{Z}_i$. Here, $\|\cdot\|_{\mathrm{RKHS}}$ denotes the Euclidian norm in the RKHS. The resulting RS decision function f' is then given by

$$f'(\mathbf{X}) = \mathrm{sgn}\left(\sum_{i=1}^{m'} \beta_i k(\mathbf{Z}_i, \mathbf{X}) + b\right). \tag{5}$$

In practice, $\beta_i, \mathbf{Z}_i$ are found using a gradient based optimization technique. Details are given in [2].

3 Constrained Reduced Set Construction

We now describe the concept of separable filters in image processing and show how this idea can be applied to a special class of nonlinear filters, namely those used by SVMs during classification.

3.1 Linear Separable Filters

Applying a linear filter to an image amounts to a two-dimensional convolution of the image with the impulse response of the filter. In particular, if $\mathbf{I}$ is the input image, $\mathbf{H}$ the impulse response, i.e. the filter mask, and $\mathbf{J}$ the output image, then

$$\mathbf{J} = \mathbf{I} * \mathbf{H}. \tag{6}$$

If $\mathbf{H}$ has size $h \times w$, the convolution requires $O(h \cdot w)$ operations for each output pixel. However, in special cases where $\mathbf{H}$ can be decomposed into two column vectors $\mathbf{a}$ and $\mathbf{b}$, such that

$$\mathbf{H} = \mathbf{a}\mathbf{b}^\top \tag{7}$$

holds, we can rewrite (6) as

$$\mathbf{J} = (\mathbf{I} * \mathbf{a}) * \mathbf{b}^\top, \tag{8}$$

since here, $\mathbf{a}\mathbf{b}^\top = \mathbf{a} * \mathbf{b}^\top$, and since the convolution is associative. This splits the original problem (6) into two convolution operations with masks of size $h \times 1$ and $1 \times w$, respectively. As a result, if a linear filter is separable in the sense of equation (7), the computational complexity of the filtering operation can be reduced from $O(w \cdot h)$ to $O(w + h)$ per pixel by computing (8) instead of (6). Note that for this to hold, the size of the image $\mathbf{I}$ is assumed to be considerably larger than h and w.

3.2 Nonlinear Separable Filters

Due to the fact that in 2D, correlation is identical with convolution if the filter mask is rotated by 180 degrees (and vice versa), we can apply the above idea to any image filter $f(\mathbf{X}) = g(c(\mathbf{H}, \mathbf{X}))$ where g is an arbitrary nonlinear function and $c(\mathbf{H}, \mathbf{X})$ denotes the correlation between images patches $\mathbf{X}$ and $\mathbf{H}$ (both of size $h \times w$). In SVMs this amounts to using a kernel of the form

$$k(\mathbf{H}, \mathbf{X}) = g(c(\mathbf{H}, \mathbf{X})). \tag{9}$$

If $\mathbf{H}$ is separable, we may split the kernel evaluation into two 1D correlations plus a scalar nonlinearity. As a result, if the RSVs in a kernel expansion such as (5) satisfy this constraint, the average computational complexity decreases from $O(m' \cdot h \cdot w)$ to $O(m' \cdot (h + w))$ per output pixel. This concept works for many off-the-shelf kernels used in SVMs. While linear, polynomial and sigmoid kernels are defined as functions of input space dot products and therefore immediately satisfy equation (9), the idea applies to kernels based on the Euclidian distance as well. For instance, the Gaussian kernel reads

$$k(\mathbf{H}, \mathbf{X}) = \exp(\gamma(c(\mathbf{X}, \mathbf{X}) - 2c(\mathbf{H}, \mathbf{X}) + c(\mathbf{H}, \mathbf{H}))). \tag{10}$$

Here, the middle term is the correlation which we are going to evaluate via separable filters. The first term is independent of the SVs. It can be efficiently pre-computed and stored in a separate image. The last term is merely a constant scalar independent of the image data. Once these quantities are known, their contribution to the computational complexity of the decision function becomes negligible.

3.3 The Proposed Method

In order to compute such separable SVM approximations, we use a constrained version of Burges' method. The idea is to restrict the RSV search space to the manifold spanned by all separable image patches, i.e. the one induced by equation (7). To this end, we replace the $\mathbf{Z}_i$ in equation (3) with $\mathbf{u}_i s_i \mathbf{v}_i^\top$. This yields

$$\Psi'' = \sum_{i=1}^{m'} \beta_i k(\mathbf{u}_i s_i \mathbf{v}_i^\top, \cdot) \tag{11}$$

where, for $h \times w$ patches, $\mathbf{u}_i$ and $\mathbf{v}_i$ are $h \times 1$ and $w \times 1$ vectors of unit length, while the scale of the RSV $\mathbf{u}_i s_i \mathbf{v}_i^\top$ is encoded in the scalar s_i. Analogously to the unconstrained case (4), we solve

$$\arg \min_{\beta, \mathbf{u}, s, \mathbf{v}} \|\Psi - \Psi''\|^2_{\text{RKHS}} \tag{12}$$

via gradient decent. Note that during optimization, the unit length of $\mathbf{u}_i$ and $\mathbf{v}_i$ needs to be preserved. Instead of normalizing $\mathbf{u}_i$ and $\mathbf{v}_i$ after every step, we use an optimization technique for orthogonal matrices, where the $\mathbf{u}_i$ and $\mathbf{v}_i$ are updated using rotations rather than linear steps [1]. This allows us to perform relatively large steps, while $\mathbf{u}_i$ and $\mathbf{v}_i$ stay on the so-called Stiefel manifold which in our case is simply the unit sphere in $\mathbb{R}^h$ and $\mathbb{R}^w$, respectively. The derivation of the rotation matrix is somewhat technical. For detailed information about gradient decent on Stiefel manifolds, see [1].

4 Experiments

We have conducted two experiments: the first one shows the convergence of our approximations on the USPS database of handwritten digits [9]. Note that since this is usually considered a recognition task rather than a detection problem in the sense that we classify single patches as opposed to every patch within a larger image, this experiment can only illustrate effects on classification accuracy, not on speed. In contrast, the second part of this section describes how to speed up a cascade-based face detection system using the proposed method. Here, we illustrate the speedup effect which is achieved by using separable RSV approximations during early evaluation stages of the cascade.

4.1 Handwritten Digit Recognition

The USPS database contains gray level images of handwritten digits, 7291 for training and 2007 for testing. The patch size is 16×16. In this experiment we trained hard margin SVMs on three two-class problems, namely "0 vs. rest", "1 vs. rest" and "2 vs. rest", using a Gaussian kernel with $\sigma = 15$ (chosen according to [9], chapter 7). The resulting classifiers have 281, 80 and 454 SVs, respectively. Classification accuracies are measured via the area under the ROC curve (AUC), where the ROC curve plots the detection rate against the false positive rate for varying decision thresholds. Hence, an AUC equal to one amounts to perfect prediction, whereas an AUC of 0.5 is equivalent to random guessing.

Figure 1 shows the AUC of our approximations for RS sizes up to $m' = 32$. It further plots the performance of the unconstrained RS approximations as well as the full SVM classifier. We found that both unconstrained and constrained approximations converge to the full solution as m' grows. As expected, we need a larger number of separable RSVs than unconstrained RSVs to obtain the same classification accuracy. However, the next experiment will show that in a detection setting the accuracy is actually increased as soon as the number of required computations is taken into account.

4.2 Face Detection

We now give an example of how to speed up a cascade based face detection system using our method. The cascaded evaluation [6, 12] of classifiers has become a popular technique for building fast object detection systems. For instance, Romdhani et al. presented an algorithm that on average uses only 2.8 RSV evaluations per scanned image position. The advantage of such systems stems from the fact that during early evaluation stages, fast detectors discard a large number of the false positives [6, 12]. Hence, the overall computation time strongly depends on how much 'work' is done by these first classifiers. This suggests replacing the first stages with a separable RSV approximation that classifies equally well.

The full SVM was trained using our own face detection database. It consists of 19×19 gray value images, normalized to zero mean and unit variance. The training set contains 11204 faces and 22924 non-faces, the test set contains 1620

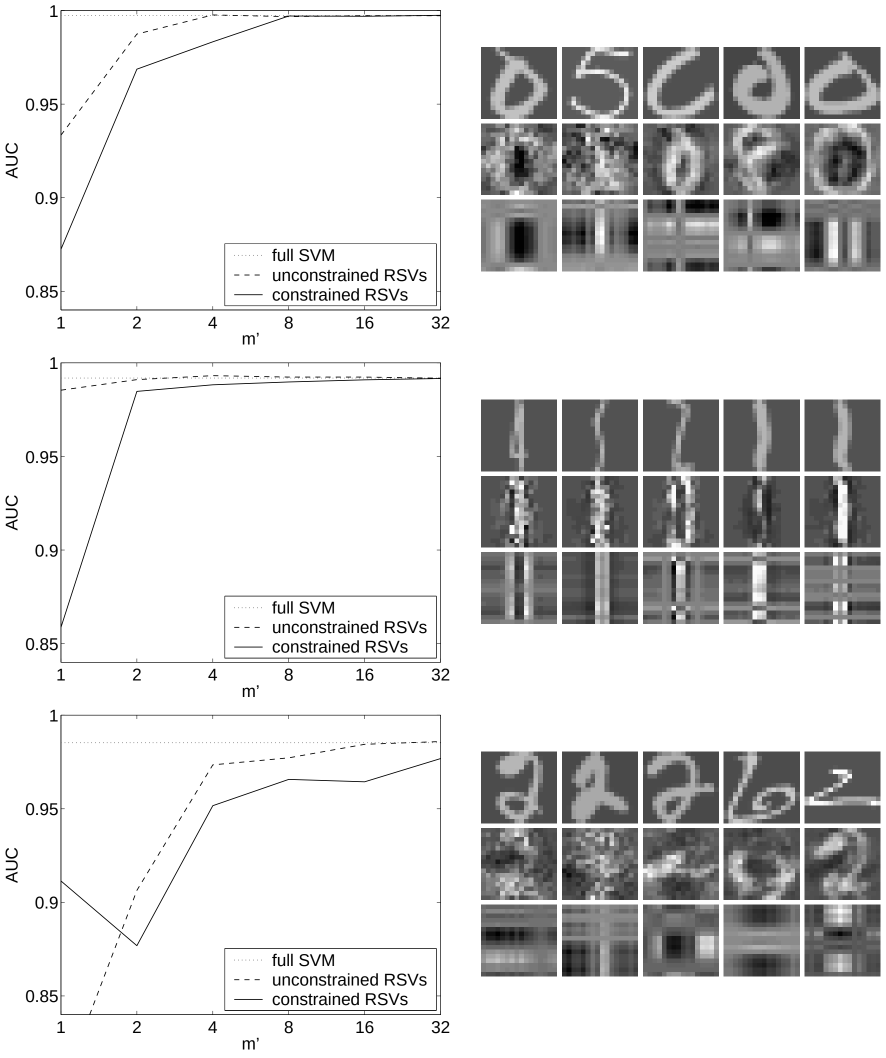

Fig. 1. Left column, top to bottom: the AUC (area under ROC curve) for the USPS classifiers "0 vs. rest", "1 vs. rest" and "2 vs. rest" , respectively. For the approximations (dashed and solid lines), the size parameter m' was varied between 1 and 32. The right column shows subsets of the corresponding expansion vectors: In each figure, the top row illustrates five (randomly selected) SVs used in the full SVM, while the middle and bottom rows shows five of the unconstrained and separable RSVs, respectively.

faces and 4541 non-faces. We used a Gaussian kernel with $\sigma = 10$, the regularization constant was set to $C = 1$. This yielded a classifier with 7190 SVs. Again, we computed RSV approximations up to size $m' = 32$, both separable and unconstrained.

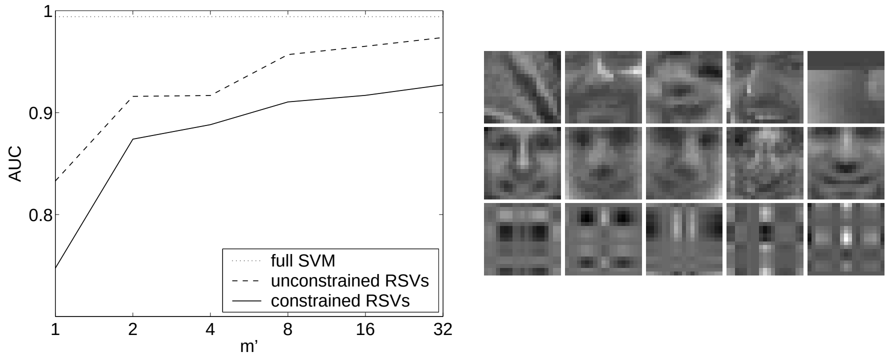

Fig. 2. Left: accuracies of the unconstrained and constrained approximations in face detection. As before, the dotted line shows the accuracy of the full SVM, whereas the dashed and solid line correspond to unconstrained and separable RSV classifiers, respectively. Right: additionally, we show a subset of the SVs of the full SVM (top row) plus five unconstrained and constrained RSVs (middle and bottom row, respectively).

The results are depicted in Figure 2. Note that for 19 × 19 patches, scanning an image with a separable RSV reduces the number of required operations to less than 11%, compared to the evaluation of an unconstrained RSV. This suggests that for our cascade to achieve the accuracy of the unconstrained $m' = 1$ classifier after the first stage, we may for instance plug in the separable $m' = 2$ version, which requires roughly 22% of the previous operations and yet classifies better (the AUC improves from 0.83 to 0.87). Alternatively, replacing the first stage with the separable $m' = 8$ classifier results in an AUC of 0.9 instead of 0.83, while the computational complexity remains the same.

5 Discussion

We have presented a reduced set method for SVMs in image processing. As our constrained RSV approximations can be evaluated as separable filters, they require much less computations than their non-separable counterparts when applied to complete images. Experiments have shown that for face detection, the degradation in accuracy caused by the separability constraint is more than compensated by the computational advantage. The approach is thus justified in terms of the expected speedup.

Another vital property of our approach is simplicity. By construction, it allows the use of off-the-shelf image processing libraries for separable convolutions. Since such operations are essential in image processing, there exist many — often highly optimized — implementations. Moreover, by directly working on gray values, separable RSVs can be mixed with unconstrained RSVs or SVs without affecting the homogeneity of the existing system. As a result, the required changes in existing code, such as for [6], are negligible.

We are currently integrating our method into a complete face detection system. Future work includes a comprehensive evaluation of the system as well as its extension to other detection problems such as component based object detection and interest operators. Furthermore, since SVMs are known to also yield good results in regression problems, the proposed method might provide a convenient tool for speeding up different types of image processing applications that require real-valued (as opposed to binary) outputs. As a final remark, note that separable filters can be applied to higher dimensional grid data as well (e.g. volumes or time sequences of images), providing further possible applications for our approach.

References

1. G .H. Bakir, A. Gretton, M.O. Franz, and B. Schölkopf. Multivariate regression via stiefel manifold constraints. *Proc. of the Pattern Recognition Symposium, DAGM*, 2004.
2. C. J. C. Burges. Simplified support vector decision rules. In L. Saitta, editor, *Proceedings of the 13th International Conference on Machine Learning*, pages 71–77, San Mateo, CA, 1996. Morgan Kaufmann.
3. C. J. C. Burges and B. Schölkopf. Improving the accuracy and speed of support vector learning machines. In M. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems 9*, pages 375–381, Cambridge, MA, 1997. MIT Press.
4. B. Heisele, T. Poggio, and M. Pontil. Face detection in still gray images. Technical Report 1687, MIT A.I. Lab, 2000.
5. E. Osuna, R. Freund, and F. Girosi. Training support vector machines: An application to face detection. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, pages 130–136, 1997.
6. S. Romdhani, P. Torr, B. Schölkopf, and A. Blake. Fast face detection, using a sequential reduced support vector evaluation. In *Proceedings of the International Conference on Computer Vision*, 2001.
7. H. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:23–38, 1998.
8. H. Schneiderman. A statistical approach to 3d object detection applied to faces and cars. *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, 2000.
9. B. Schölkopf and A. J. Smola. *Learning with Kernels.* MIT Press, Cambridge, MA, 2002.
10. Ingo Steinwart. Sparseness of support vector machines—some asymptotically sharp bounds. In Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16.* MIT Press, Cambridge, MA, 2004.
11. K. Sung and T. Poggio. Example-based learning for view-based human face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20, 1998.
12. P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, 2001.