Non-Mercer Kernels for SVM Object Recognition

Sabri Boughorbel[1] Jean-Philippe Tarel[2] François Fleuret[3]

sabri.boughorbel@inria.fr tarel@lcpc.fr francois.fleuret@epfl.ch

[1,3]IMEDIA Research Group
BP 105
INRIA Rocquencourt
78150 Le Chesnay, France

[2] ESE
LCPC
58 Boulevard Lefebvre
75015 Paris, France

Abstract

On the one hand, Support Vector Machines have met with significant success in solving difficult pattern recognition problems with global features representation. On the other hand, local features in images have shown to be suitable representations for efficient object recognition. Therefore, it is natural to try to combine SVM approach with local features representation to gain advantages on both sides. We study in this paper the Mercer property of matching kernels which mimic classical matching algorithms used in techniques based on points of interest. We introduce a new statistical approach of kernel positiveness. We show that despite the absence of an analytical proof of the Mercer property, we can provide bounds on the probability that the Gram matrix is actually positive definite for kernels in large class of functions, under reasonable assumptions. A few experiments validate those on object recognition tasks.

1 Introduction

Objects recognition in variable contexts under different illuminations remains one of the most challenging problem in computer vision and artificial intelligence. Compared to the efficiency of human brain, algorithms have been unable so far to demonstrate competitive performances in real usage conditions.

Among all the techniques developed during the last decades, points of interest in computer vision [17, 16, 10] and support vector machines (SVMs) in statistical learning [5, 1] have been successful in solving many real-world problems.

Points of interest based techniques combine the information provided by the responses of local filters at several highly informative locations in the picture and their global geometric configuration. Such approaches lead to compact and invariant representations. SVM algorithms search the optimal separating plane between positive and negative examples. It has the advantage of high generalization capacity from a few training examples.

In some sense, points of interest techniques and SVMs address two different issues. While the formers provide a very meaningful way to represent and compare images, the laters are able to combine several training examples into a consistent and statistically sound view-based representation.

BMVC 2004 doi:10.5244/C.18.16

A few attempts have tried recently to combine those two approaches into a common framework by building kernels on the space of feature vector sets [4, 19, 21, 12]. While kernels are usually defined on vector spaces, they have to deal in this new context with features of various size where no order is defined between components. Also, since the classical matching algorithm used in points of interest techniques are able to tackle the occluding problem and to provide invariance to the pose, it seems highly desirable that the used kernels mimic those matching algorithms.

The SVM problem is convex whenever the used kernel is a Mercer one. The convexity insures the convergence of the SVM algorithm towards a unique optimum. The uniqueness of the solution is one of the main advantages of the SVM compared to other learning approaches such as neural networks. Unfortunately, as we will see, numerous examples show that using matching algorithms as kernels for SVM are not in general Mercer.

Nevertheless the use of simple matching algorithms as kernels gives good results in practice. This experimental observation leads us to consider the positiveness of the kernel from a statistical point of view. We show that, even if kernels based on matching are not always positive, this is likely to be true for a large class of functions. Moreover, we propose a way to control the probability of positiveness by tuning kernel parameters in this class. This control can be also used with advantages in many other applications where SVM approach applies, with local and global representations as well.

In §2 we present different approaches for recognition with local features, in §3 we present our main result of bound on the Gram matrix probability to be positive definite and give experimental results in §4.

2 Recognition With Local Features

Let $\mathcal{X}$ be the space where the features are taken from ($\mathbb{R}^n$ for example). The dimension of $\mathcal{X}$ is in general finite and fixed. In the context of recognition from images, this implies that the used representation is global on the image. Therefore, in such a case, the design of a kernel for object recognition is closely related to template matching. This family of algorithms is mainly based on correlations between two images. In [3], Mercer kernels for object recognition in images are build based on such template matching algorithms. Two main drawbacks of template matching algorithms are known to be their computational cost and poorly robust results in case of occlusion.

2.1 Local Feature Representations

To better tackle these difficulties, local features representation were introduced for images. The idea is, first to build a detector of particular points in images, usually called *points of interest* (or key points, anchors points, salient points), and second to characterize each point by its local environment. This obtained set of features is the local feature representation of an image. This kind of representations have shown to be suitable representations of images for object recognition, see for instance [17, 10]. Indeed, the information about images has been drastically reduced, allowing faster comparison algorithms than using correlation based algorithms. There is many ways to locally characterize the environment of a point. The most used feature seems to be the so-called jet, which is the vector based on a differential characteristic's around the point of interest. However, following [16], SIFT feature seems to provide improvement to jet features.

2.2 Point Matching Algorithms

Given two images of the same object, each image being represented by a set of local features $\mathcal{X}$ and $\mathcal{X}'$, occlusion may remove several points of interest in one or in the other image. Moreover, due to complex background, extra points appear as outliers. Therefore, correspondences must be robustly established between the two sets $\mathcal{X}$ and $\mathcal{X}'$. This task can be performed by matching algorithms. Depending on the constrains enforced on matching (bijective, symmetric,...), different algorithms were derived (see [22] for a partial review). Optimal matching algorithms have to face with combinatorial explosion, and thus many matching algorithms are in practice based on heuristics.

The result of the matching algorithm is two index functions $\Phi_1(n) \in [1, ..., N_{\mathcal{X}}]$ and $\Phi_2(n) \in [1, ..., N_{\mathcal{X}'}]$ that gives the indices of the N matched pairs of feature vectors $(x_{\Phi_1(n)}, y_{\Phi_2(n)})_{1 \leq n \leq N}$, where $N_{\mathcal{X}}$ and $N_{\mathcal{X}'}$ denotes the size of sets $\mathcal{X}$ and $\mathcal{X}'$

3 Kernels for Sets

Recent works [13, 9, 12, 21] have focused on designing Mercer kernels for different kinds of structured features such as strings, DNA, graphs, trees, and sets.

Any Mercer kernel can be written as an inner product after mapping by a well-chosen $\mathbf{f}$ function [5], i.e $K(\mathbf{x}, \mathbf{x}') = \langle \mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{x}') \rangle$. Thus, Mercer kernels can be seen as a measure of dissimilarity between vectors $\mathbf{x}$ and $\mathbf{x}'$.

3.1 Known Kernels

For local feature representations, the simplest approach is to define the global dissimilarity between two sets of vectors as the sum over dissimilarities between all possible pairs of vectors. The dissimilarity of a pair of vectors is obtained by a Mercer kernel $k(\mathbf{x}_i, \mathbf{x}'_j)$. The Summation Kernel is thus:

$$K_S(\mathcal{X}, \mathcal{X}') = \sum_{i=1}^{N_{\mathcal{X}}} \sum_{j=1}^{N_{\mathcal{X}'}} k(\mathbf{x}_i, \mathbf{x}'_j) \tag{1}$$

We may use RBF kernel for k:

$$k(\mathbf{x}_i, \mathbf{x}'_j) = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}'_j\|^2}{2\sigma^2}}$$

where σ is the scale parameter. Although the numbers $N_{\mathcal{X}}$ and $N_{\mathcal{X}'}$ of vectors in the two sets $\mathcal{X}$ and $\mathcal{X}'$ are not fixed, it can be proved that K_S is a Mercer kernel [11]. However, the number of pairs is increasing more rapidly than the number of vectors. Thus using K_S, correct correspondences are swamped into bad correspondences.

In [21] a Mercer kernel is proposed for sets of vectors based on the concept of principal angles between two linear subspaces. This last approach is interesting but the proposed kernel is invariant when local feature vectors $\mathbf{x}_i$ are combined linearly. Thus, this kernel does not match our practical needs. In [12], another approach is proposed where the set is represented as a mixture of gaussian. This approach requires relatively dense sets of vectors to be meaningful, which is difficult to achieve in practice.

3.2 Matching Kernel

A better approach to compute global dissimilarity between two sets of local feature vectors extracted from two images is to take into account only dissimilarities between matched local features. Thus, the global dissimilarity is defined by:

$$K_M(\mathcal{X}, \mathcal{X}') = \sum_{n=1}^{N} k(\mathbf{x}_{\Phi_1(n)}, \mathbf{x}'_{\Phi_2(n)}) \tag{2}$$

where $\Phi_1(n) \in \{1, ..., N_{\mathcal{X}}\}$, $\Phi_2(n) \in \{1, ..., N_{\mathcal{X}'}\}$ denote the mappings between the two sets of vectors, like in Sec. 2.2. N denotes the number of matched pairs of vectors, and thus we have $N \leq \min(N_{\mathcal{X}}, N_{\mathcal{X}'})$.

The optimal mapping (Φ_1^*, Φ_2^*) is obtained as the one that maximizes the dissimilarity, so:

$$(\Phi_1^*, \Phi_2^*) = \arg \max_{\Phi_1, \Phi_2} K_M(\mathcal{X}, \mathcal{X}') \tag{3}$$

Since the exact solution of the previous problem is time consuming, as explained in Sec. 2.2, we better use an heuristic as matching algorithm. We have used the *winner-take-all* approach: at each step, the pair of points having the maximum dissimilarity is matched and the associated vectors are removed to future examinations for the next steps.

We notice that in general the mapping depends on sets $\mathcal{X}$ and $\mathcal{X}'$, and thus, contrary to [19], it is not easy to prove that $K_M(\mathcal{X}, \mathcal{X}')$ is a Mercer kernel, even when the local kernel k is Mercer. If the assertion "the $\max$ of Mercer kernels is still a Mercer kernel" was true, we could prove that $K_M(\mathcal{X}, \mathcal{X}')$ is Mercer. Unfortunately, the $\max$ of Mercer kernels is not Mercer [2], and even not conditionally positive definite [18].

A simple counter example is now presented. We consider the matrices G_1, G_2 and $G_3 = \max(G_1, G_2)$, with eigenvalues λ_i respectively, as:

$$G_1 = \begin{bmatrix} 2 & -1 & -2 \\ -1 & 2 & 3 \\ -2 & 3 & 8 \end{bmatrix} \quad \lambda_\mathbf{1} = \begin{bmatrix} 0.72 \\ 1.4 \\ 9.8 \end{bmatrix}$$

$$G_2 = \begin{bmatrix} 7 & 4 & -2 \\ 4 & 3 & -1 \\ -2 & -1 & 1 \end{bmatrix} \quad \lambda_\mathbf{2} = \begin{bmatrix} 0.29 \\ 0.68 \\ 10.02 \end{bmatrix}$$

$$G_3 = \begin{bmatrix} 7 & 4 & -2 \\ 4 & 3 & 3 \\ -2 & 3 & 8 \end{bmatrix} \quad \lambda_\mathbf{3} = \begin{bmatrix} -0.92 \\ 9.34 \\ 9.57 \end{bmatrix}$$

G_1 and G_2 are two positive definite Gram matrices but their $\max$ (G_3) is not. This suggests that in general matching algorithms are not Mercer kernels. Indeed, we have found other more complicated counter-examples to Mercer or conditional positiveness, for optimal matching algorithm.

3.3 Statistical positiveness of Kernels

In [19, 2, 6] non-Mercer kernels have been used for SVM based pattern recognition. Although performances of these kernels are good, the convergence of SVM algorithm to the unique optimum is not insured since there is no warranty that the SVM optimization problem is convex. We introduce next a new definition of kernel positiveness based on a statistical approach. This definition is general enough to include usual Mercer kernels. The advantage is that we can show that matching kernels are statistically positive definite kernels.

We denote here by $X_1, \ldots, X_\ell$ a family of ℓ i.i.d random variables standing for the training samples. We want to bound the probability that the Gram matrix $(K_\sigma(X_i,\ X_j))_{i,j}$ violates the Mercer condition, σ is the tuning scale hyperparameter of the kernel. A sufficient condition for the Gram matrix to be positive definite is to be diagonal dominant. We recall, that a matrix G is called diagonal dominant when for each i, we have:

$$|G_{ii}| \geq \sum_{i \neq j} |G_{ij}| \tag{4}$$

Diagonal dominance condition is an easy way to enforce Mercer condition directly on the Gram matrix. It has been used besides to derive kernels for strings [20]. In the following, McDiarmid concentration inequality [15] is recalled and a modified version is introduced. Many concentration inequalities can be found in the literature [14] like Hoeffding, Bennett ones, but they are useless in our case because of too strict assumptions with respect to independence. McDiarmid inequality will allow us to rewrite diagonal dominance condition in a statistical point of view.

Theorem. *McDiarmid inequality (1989)*
$X_1, \ldots, X_\ell$ *be independent random variables* $\in \mathcal{A}$.
Let $f : (\mathcal{A})^\ell \rightarrow \mathbb{R}$ *satisfies the bounding difference property:*

$$\sup_{\substack{x_1,\ldots,x_\ell \in \mathcal{A} \\ x_i' \in \mathcal{A}}} |f(x_1, \ldots, x_\ell) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_\ell)| \leq c_i \ , \ 1 \leq i \leq \ell$$

Thus, the following probability is bounded by:

$$\forall \epsilon > 0, P\left\{f(X_1, \ldots, X_\ell) - E\left(f(X_1, \ldots, X_\ell)\right) > \epsilon\right\} \leq e^{-\frac{2\epsilon^2}{\Sigma_{i=1}^{\ell} c_i^2}}$$

In the case of $0 \leq f(x_1, \ldots, x_\ell) < m$, where $m \ll c_i$ with a high probability but not equal to 1, McDiarmid inequality leads to a poor bound. To solve the problem, we give next an improvement of the McDiarmid inequality in the case of positive function f.

Lemma. *Modified McDiarmid inequality*
Let $M > 0$ *such that* $P\left(0 \leq f(X_1, \ldots, X_\ell) \leq M\right) = 1$,
for $0 < \eta < 1$, *let* $0 < m < M$ *such that* $P\left(0 \leq f(X_1, \ldots, X_\ell) < m\right) = \eta$ *then*
$\forall \epsilon > 0$,

$$P\left\{f(X_1, \ldots, X_\ell) - E\left(f(X_1, \ldots, X_\ell)\right) > \epsilon\right\} \leq \eta e^{-\frac{2\epsilon^2}{\ell m^2}} + (1 - \eta) e^{-\frac{2\epsilon^2}{\ell (M-m)^2}}$$

Proof.

$$\mathbb{1}_{\{0 \leq f(x_1,\ldots,x_\ell) \leq M\}} = \mathbb{1}_{\{0 \leq f(x_1,\ldots,x_\ell) < m\}} + \mathbb{1}_{\{m \leq f(x_1,\ldots,x_\ell) \leq M\}}$$

the two events are disjoint and we have :

$$\begin{array}{rcl} P\left\{\mathbb{1}_{\{f(X_1,\ldots,X_\ell) < m\}}\right\} & = & \eta \\ |f(x_1, \ldots, x_\ell) - f(x_1', \ldots, x_\ell'\}| & < & m \text{ since } 0 \leq f(x_1, \ldots, x_\ell) < m \\ P\left(\mathbb{1}_{\{m \leq f(X_1,\ldots,X_\ell) \leq M\}}\right) & = & 1 - \eta \\ |f(x_1, \ldots, x_\ell) - f(x_1', \ldots, x_\ell')| & \leq & M - m \text{ since } m \leq f(x_1, \ldots, x_\ell) \leq M \end{array}$$

We obtain the modified McDiarmid inequality by applying McDiarmid inequality on the two terms on the right:

$$\begin{aligned} P\{f(X_1,\ldots,X_\ell)-E(f(X_1,\ldots,X_\ell))>\epsilon\} &= \eta\ P\{f(X_1,\ldots,X_\ell)-E(f(X_1,\ldots,X_\ell))>\epsilon\}+ \\ &\quad +(1-\eta)\ P\{f(X_1,\ldots,X_\ell)-E(f(X_1,\ldots,X_\ell))>\epsilon\} \\ &\leq \eta e^{-\frac{2\epsilon^2}{\ell m^2}} + (1-\eta)e^{-\frac{2\epsilon^2}{\ell(M-m)^2}} \end{aligned}$$

□

Some assumptions are given on the kernel, we suppose that the kernel is constant on the diagonal, i.e. $K_\sigma(x_i, x_i) = k$, a wide class of kernel verifies this assumption, for example kernels like $K_\sigma(x, y) = g(\frac{\|x-y\|}{\sigma})$. In our case the matching kernel also belongs to this class. Another assumption is that the kernel is positive and bounded, i.e. $0 \leq K_\sigma(x, y) \leq c$ for $x,\ y \in \mathfrak{X}$. Last assumption is on the asymptotic behavior of the kernel with respect to the parameter σ, $K_\sigma(x, y) \underset{\sigma\to 0}{\longrightarrow} 0$ for $x \neq y$ which mean that the kernel vanishes at infinity.

Proposition. *Let K_σ be a kernel, with a hyperparameter σ, satisfying the following conditions: $K_\sigma(x, x) = k$, $0 \leq K_\sigma(x, y) \leq c$ and $K_\sigma(x, y) \underset{\sigma\to 0}{\longrightarrow} 0$.*
Let p_d be the probability that the Gram matrix of the kernel K_σ not being diagonal dominant. X_i with *$i = 1, \ldots, \ell$ are ℓ i.i.d random variables representing training samples:*

$$p_d = P\left\{\exists i_0,\ \sum_{\substack{j=1\\ j\neq i_0}}^{\ell} K_\sigma(X_{i_0}, X_j) > K_\sigma(X_{i_0}, X_{i_0})\right\}$$

for $0 < \eta < 1$, there exists σ such that we bound p_d as following:

$$p_d \leq \ell\ \left(\eta\ exp\left\{\frac{-2(k-(\ell-1)e_\sigma)^2}{(\ell-1)m_\sigma^2}\right\} + (1-\eta)\ exp\left\{\frac{-2(k-(\ell-1)e_\sigma)^2}{(\ell-1)(M-m_\sigma)^2}\right\}\right) \quad (5)$$

where $e_\sigma = E\left(K_\sigma(X_1, X_2)\right)$, $M = (\ell-1)c$ and m_σ is defined by
$P\left(f(X_1, \ldots, X_n) < m_\sigma\right) = \eta$

Proof. We define the function f_i as the following

$$f_i(x_1, \ldots, x_\ell) = \sum_{\substack{j=1\\ j\neq i}}^{\ell} K_\sigma(x_i, x_j)$$

since K_σ is bounded, we have $0 \leq f_i(x_1, \ldots, x_\ell) \leq M$, where $M = (\ell-1)c$. As a consequence, f_i satisfies the bounding property of modified McDiarmid inequality. The probability that Gram matrix not being diagonal dominant can be expressed as following:

there exists a line from the Gram matrix that does not satisfy dominance condition (4).

$$\begin{aligned}
p_d &= P\{\exists i_0, \textstyle\sum_{\substack{j=1\\ j\neq i_0}}^{\ell} K_\sigma(X_{i_0}, X_j) > K_\sigma(X_{i_0}, X_{i_0})\} \\
&= P\{\exists i_0, \textstyle\sum_{\substack{j=1\\ j\neq i_0}}^{\ell} K_\sigma(X_{i_0}, X_j) > k\} \\
&\leq \textstyle\sum_{n=1}^{\ell} P\{\sum_{\substack{j=1\\ j\neq n}}^{\ell} K_\sigma(X_n, X_j) > k\} \qquad \text{since union bound propriety} \\
&= \ell\, P\{\textstyle\sum_{\substack{j=1\\ j\neq i}}^{\ell} K_\sigma(X_i, X_j) > k\} \qquad \text{since } X_j \text{ are i.i.d} \\
&= \ell P\{f_i(X_1,\ldots,X_\ell)>k\} \\
&= \ell P\{f_i(X_1,\ldots,X_\ell)-E(f_i(X_1,\ldots,X_\ell))>k-E(f_i(X_1,\ldots,X_\ell))\} \\
&= \ell P\{f_i(X_1,\ldots,X_\ell)-E(f_i(X_1,\ldots,X_\ell))>k-(\ell-1)\underbrace{E(K_\sigma(X_1,X_2))}_{e_\sigma}\} \\
&= \ell P\{f_i(X_1,\ldots,X_\ell)-E(f_i(X_1,\ldots,X_\ell))>\epsilon_\sigma\} \qquad (6)
\end{aligned}$$

where $\epsilon_\sigma = k-(\ell-1)e_\sigma$. We need to insure $\epsilon_\sigma > 0$ to apply McDiarmid inequality. We have by definition $e_\sigma = E(K_\sigma(X_1, X_2) = \int K_\sigma(x,y)dP(x,y)$. As we supposed that $K_\sigma(x,y) \underset{\sigma\rightarrow 0}{\longrightarrow} 0$, we deduce $e_\sigma \underset{\sigma\rightarrow 0}{\longrightarrow} 0$ and thus there exists a σ such that $\epsilon_\sigma > 0$.
We consider the following interval subdivision :

$$\mathbb{1}_{\{0\leq f_i(x_1,\ldots,x_\ell)\leq M\}} = \mathbb{1}_{\{0\leq f_i(x_1,\ldots,x_\ell)\leq m_\sigma\}} + \mathbb{1}_{\{m_\sigma < f_i(x_1,\ldots,x_\ell)\leq M\}}$$

For this subdivision, we now apply the modified McDiarmid inequality on equation (6) and we thus deduce (5). □

The confidence term η can be chosen close to one to insure that second term of the bound (5) is very small.
For a given η, we choose $m_\sigma > 0$ such that $\eta = P\left\{f_i(X_1,\ldots,X_n) < m_\sigma\right\}$.
We choose σ as small as needed to insure the first term of (5) to be small enough. This is always possible. Indeed, Markov inequalities implies that
$1-\eta = P\left\{f_i(X_1,\ldots,X_n) > m_\sigma\right\} \leq (\ell-1)\frac{e_\sigma}{m_\sigma}$, so we obtain that $m_\sigma \leq (\ell-1)\frac{e_\sigma}{1-\eta}$.
As shown previously, $e_\sigma \underset{\sigma\rightarrow 0}{\longrightarrow} 0$, thus $m_\sigma \underset{\sigma\rightarrow 0}{\longrightarrow} 0$, therefore the first term of (5) goes to 0. As a consequence, it is possible to enforce kernel positiveness with high probability by tuning kernel hyperparameters. If the obtained Gram matrix has a too large diagonal, generalization performance of the SVM classifier can be poor, but we can use a technique presented in [8, 7] that solve such problem.
It is known that the L_2-SVM soft margin is equivalent to adding a ridge to the Gram matrix [5]. Thus such regularization forces the Gram-matrix towards diagonal dominant ones.

4 Results

The database used in our experiments (see Fig. 4) contains about 165 images of 128× 128 pixels obtained by encrusting complex background images to original images of the same object which are taken from COIL-100 Database to generate positive samples. Similarly, negative samples are obtained by a random choice of objects images of the COIL-100 Database.

Points of interest are extracted with Harris detector. As the background contains edges, textures etc, many outliers points are extracted, so we enforce point repartition to be almost uniform. For that, the image is divided in sub-windows and a threshold on the number of points per sub-window is set-up. The performance criterion is the recognition error. It is estimated by cross-validation procedure: a proportion of images are chosen randomly for training, the remaining images are used for testing. This operation is repeated several times to obtain statistically reliable results. Matching based kernels gives good performances for points of interest representation. Tab. 1 presents comparison of the matching kernel ($\mathcal{K}_M$, (2)) with summation kernel ($\mathcal{K}_S$, (1)) . Recognition error is about 3 % for $\mathcal{K}_M$ with a configuration of 90 points per image. Comparison of matching kernel with global representation is summarized in Tab. 2, k-NN denotes k-Nearest Neighborhood classifier for $k = 2$, $\mathcal{K}_H$ denotes SVM classifier. k-NN and $\mathcal{K}_H$ are used both with global Color Histogram features. $\mathcal{K}_M$ leads to the smallest recognition error with 4.23%. The tuning hyperparameter is chosen to be the scale factor σ in (2) of local kernel k.

Although matching based kernel leads to a good performance, the Mercer condition in not usually insured. Fig. 2 presents ranked eigenvalues of the Gram matrix for different values of σ. For $\sigma = 100$, there exists negative eigenvalues of the Gram matrix, this means that matching kernel is not positive definite in general. By decreasing the scale σ, eigenvalues become always positive. Fig. 3-a shows the variation of the recognition error with respect to $\log_{10}(\sigma)$. The optimal value obtained by cross-validation is $\sigma^* = 10^{-3}$ which corresponds to a positive definite Gram matrix (positive eigenvalues). Fig. 3-b represents the bound of probability p_d (5) in a logarithmic scale with respect to $\log_{10}(\sigma)$. The probability confidence is set as $\eta = 1 - 10^{-6}$. For the optimal choice of σ, the bound is very poor and useless. For $\sigma = 10^{-4}$, which corresponds to a recognition error of 8%, we have $p_d \leq 10^{-3}$. We can say for these values that Gram matrix is diagonally dominant almost sure and so it is positive definite one. By choosing $\sigma = 10^{-4}$ we loose almost 5 % of recognition error which is not too much compared to the obtained warranty of convergence of the SVM algorithm to unique solution.

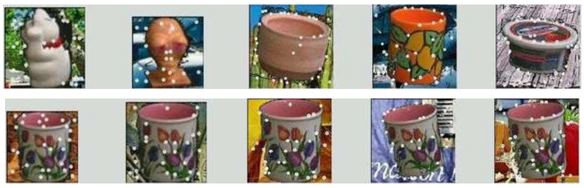


Figure 1: COIL-100 database with encrusted backgrounds, first raw: negative examples, second raw: positive examples.

Kernel	Rec. Error
$\mathcal{K}_S$	23.49%
$\mathcal{K}_M$	9.58%

(a)

Kernel	Rec. Error
$\mathcal{K}_S$	15.51 %
$\mathcal{K}_M$	3.08%

(b)

Table 1: Local feature Kernels comparison, (a): 50 pts/image, (b): 90 pts/image

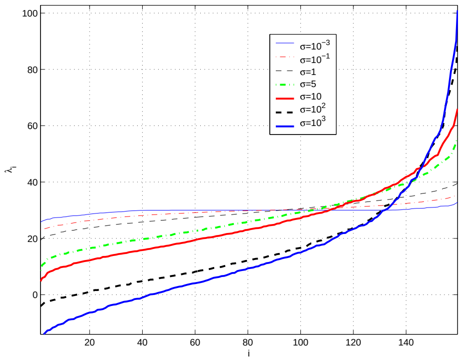
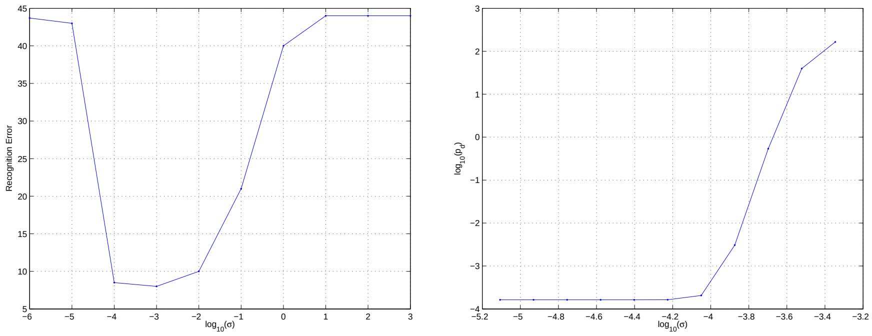


Figure 2: Eigenvalues of Gram matrix for different σ.



(a)

(b)

Figure 3: (a) Recognition error with respect to σ, (b) bound on p_d with respect to σ

Methods	k-NN	$\mathcal{K}_H$	$\mathcal{K}_M$
Rec. Error	7.42 %	6.58 %	4.23%

Table 2: Global features (k-NN, $\mathcal{K}_H$, 64 bins) vs Local feature ($\mathcal{K}_M$, 70 pts/image)

5 Conclusion

We have presented in this paper the use of matching kernels for SVMs in the context of object recognition. Despite the absence of an analytical proof of the Mercer property for such a kernel and the actual existence of counter examples, we have shown that we can choose kernel hyperparameter such that its Gram matrix is nevertheless positive definite with a very high probability. This criterion can be also used with advantages in many other applications where SVM approach applies, with local and global representations as well.

References

[1] Schölkopf B. and Smola A. *Learning with kernels*. MIT University Press Cambridge, 2002.

[2] C. Bahlmann, B. Haasdonk, and H. Burkhardt. On-line handwriting recognition with support vector machines—a kernel approach. In *Proc. of the 8th IWFHR*, pages 49–54, 2002.

[3] A. Barla, F. Odone, and A. Verri. Hausdorff kernel for 3d object acquisition and detection. In *In Proceedings of the European Conference on Computer Vision, LNCS 2353*, page p. 20 ff, 2002.

[4] B. Caputo and Gy. Dorko. How to combine color and shape information for 3d object recognition: kernels do the trick. In *Advances in Neural Information Processing Systems*, 2002.

[5] N. Cristianini and J. Shawe-Taylor. *Support Vector Machines*. Cambridge University Press, 2000.

[6] D. DeCoste and B. Schölkopf. Training invariant support vector machines. In *In Machine Learning 46*, pages 161–190, 2002.

[7] B. Schölkopf et al. A kernel approach for learning from almost orthogonal patterns. In *Proceedings of the 13th European Conference on Machine Learning*, 2002.

[8] J. Weston et al. Dealing with large diagonals in kernel matrices. In *Spring Lecture Notes in Computer Science 243*.

[9] T. Gaertner. A survey of kernels for structured data. In *IGKDD Explorations*, 2003.

[10] V. Gouet, P. Montesinos, D., and Pelé. Stereo matching of color images using differential invariants. In *Proceedings of the* IEEE *International Conference on Image Processing*, Chicago, Etats-Unis, 1998.

[11] D. Haussler. Convolution kernels on discrete structures. In *Technical Report UCS-CRL-99-10*, 1999.

[12] R. Kondor and T. Jebara. A kernel between sets of vectors. In *Proceedings of the ICML*, 2003.

[13] R. I. Kondor and J. Lafferty. Diffusion kernels on graphs and other discrete structures. In *Proceedings of the ICML*, 2002.

[14] G. Lugosi. On concentration of measure inequalities. *Lecture Notes*, 1998.

[15] C. McDiarmid. On the method of bounded differences. *London Mathematical Society Lecture Note Series*, 141(5):148–188, 1989.

[16] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. In *CVPR*, 2003.

[17] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5):530–535, 1997.

[18] B. Schölkopf. The kernel trick for distances. In *NIPS*, pages 301–307, 2000.

[19] C. Wallraven, B. Caputo, and A. Graf. Recognition with local features: the kernel recipe. In *ICCV*, pages 257–264, 2003.

[20] C. Watkins. Dynamic alignment kernels. In *Technical Report CSD-TR-98-11*, 1999.

[21] L. Wolf and A. Shashua. Kernel principal angles for classification machines with applications to image sequence interpretation. In *CVPR*, 2003.

[22] Z. Zhang, R. Deriche, O. Faugeras, and Q. Luong. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *Artificial Intelligence*, 78(1-2):87–119, 1995.