
# Incorporating context and geometry in kernel design for support vector machines

## *Contexte et géométrie pour la conception des noyaux*

---

Hichem Sahbi
Jean-Yves Audibert
Renaud Keriven

**2009D002**

Janvier 2009

Département Traitement du Signal et des Images
Groupe TII : Traitement et Interprétation des Images

# Incorporating Context and Geometry in Kernel Design for Support Vector Machines

## Contexte et Géométrie pour la Conception des Noyaux

**Hichem Sahbi** HICHEM.SAHBI@TELECOM-PARISTECH.FR
*CNRS LTCI UMR 5141,*
*TELECOM ParisTech, Paris, France*

**Jean-Yves Audibert** AUDIBERT@CERTIS.ENPC.FR
*Willow, ENS ULM/INRIA, Paris, France*
*Certis Lab, Ponts ParisTech, France*

**Renaud Keriven** KERIVEN@CERTIS.ENPC.FR
*Certis Lab, Ponts ParisTech, France*

### Abstract

Kernels are functions designed in order to capture resemblance between data and they are used in a wide range of machine learning techniques including support vector networks (SVMs). In their standard version, commonly used kernels such as the Gaussian, show reasonably good performance in many classification and recognition tasks in computer vision, bio-informatics and text processing. In the particular task of object recognition, the main deficiency of standard kernels, such as the convolution one, resides in the lack in capturing the right geometric structure of objects while also being invariant.

This paper introduces a novel framework for designing a new class of kernels, referred to as Context-Dependent (CDK), with the particularity of (i) incorporating the geometric structure and the spatial relationships between image primitives (such as interest points, regions, etc.) while (ii) handling invariance and satisfying the positive definiteness which is necessary in all kernel methods including SVMs. Beside these properties, the outperformance of CDKs with respect to standard context-free kernels, is shown for pattern classification problems on standard databases.

### Résumé

Les noyaux sont des fonctions qui modélisent la ressemblance entre les données et sont utilisés dans plusieurs techniques d'apprentissage telles que les machines à vecteurs de supports (SVMs). Dans leur version standard, les noyaux les plus utilisés fournissent de bonnes performances dans plusieurs taches en classification et reconnaissance, en vision par ordinateurs, en bio-informatique, etc. En reconnaissance d'objet, en particulier, les noyaux classiques ne caractérisent pas suffisamment bien les structures géométriques des objets et

manquent aussi d'invariance.

Dans ce rapport, on introduit une nouvelle classe de noyaux dite dépendente du contexte (notée "CDK") avec la particularité (i) d'inclure la structure géométrique et les relations spatiales entre les primitives des images (points d'intérêts, régions, etc.) (ii) d'être invariante aux transformations géometriques et (iii) de satisfaire la condition de Mercer. Outre les propriétés théoriques du noyau "CDK", ses performances sont meilleures par rapport aux noyaux classiques et "indépendants du contexte".

**Keywords:** Kernel Design, Statistical Machine Learning, Support Vector Machines, Graph Matching, Context-Free Kernels, Context-Dependent Kernels, Object Recognition.

## 1. Introduction

**K**ernel methods including support vector machines (SVMs), initially introduced in Boser et al. (1992), show a particular interest as they are performant and theoretically well grounded (Bishop, 2007). These methods rely on the hypothesis of the existence of (explicit or implicit) mapping functions which transfer training and test data, via kernels, from *input* spaces into high dimensional Hilbert spaces, also referred to as *feature* spaces (Shawe-Taylor and Cristianini, 2000). Kernels are symmetric, continuous, bi-variate similarity functions which take high values when input data share similar structures, appearance, behaviors,..., and should be as invariant as possible to the linear and non linear transformations. For instance, in object recognition, a kernel should take a high value *only* when two objects (such as faces) belong to the same class or have the same identity, expression, etc. and regardless their pose. A wide range of vision applications were solved using kernel methods including optical character recognition (Miyao et al., 2005), pose estimation (Ng and Gong, 1999), image retrieval (Tong and Chang, 2001) and the most studied object recognition problem (Lyu, 2005; Barla et al., 2002; Grauman and Darrell, 2007). In almost all the proposed solutions, authors use and combine, via algebraic operations, standard kernels such as the linear, the polynomial and the Gaussian (Genton, 2001). These kernels are also referred to as *holistic* as they operate on fixed length and ordered data (Swain and Ballard, 1991; Chapelle et al., 1999) and even though proved to be relatively performant, they contain no a priori knowledge about the task at hand and the expected properties of invariance.

A second generation of kernels, referred to as *local*, has recently emerged as an alternative to holistic kernels. The former handle structured data as bags or local sets, i.e., data which cannot be represented in fixed length and ordered spaces, such as interest points, regions, blocks, graphs, trees, etc.(Gartner, 2003). It is well known that both holistic and local kernels should satisfy certain properties among them the positive definiteness, low complexity for evaluation, flexibility in order to handle variable-length data and also invariance. Holistic kernels have the advantage of being simple to evaluate, discriminating but less flexible than local kernels in order to handle invariance. While the design of kernels gathering flexibility, invariance and low complexity is a challenging task; the proof of their positive definiteness is sometimes harder (Cuturi, 2005). This property also known as the Mercer condition ensures, according to Vapnik's SVM theory (Vapnik, 1998), optimal generaliza-

tion performance and also the uniqueness of the SVM solution.

Considering a database of objects (images), each one seen as a constellation of local features, for instance interest points (Schmid and Mohr, 1997; Lowe, 2004; Lazebnik et al., 2004), extracted using any suitable filter (Harris and Stephens, 1988). Again, original holistic kernels explicitly (or implicitly) map objects into fixed-length feature vectors and take the similarity as a decreasing function of any well-defined distance (Barla et al., 2002). In contrast to holistic kernels, local ones are designed in order to handle variable-length and unordered data. Two families of local kernels can be found in the literature; those based on statistical "length-insensitive" measures such as the Kullback Leibler divergence, and those which require a preliminary step of alignment. In the first family, the authors in Kondor and Jebara (2003); Moreno et al. (2003) estimate for each object (constellation of local features) a probability distribution and compute the similarity between two objects (two distributions) using the "Kullback Leibler divergence" in Moreno et al. (2003) and the "Bhattacharyya affinity" in Kondor and Jebara (2003). Only the function in Kondor and Jebara (2003) satisfies the Mercer condition and both kernels were applied for image recognition tasks. In Wolf and Shashua (2003), the authors discuss a new type of kernel referred to as "principal angles" which is positive definite. Its definition is based on the computation of the principal angles between two linear subspaces under an orthogonality constraint. The authors demonstrate the validity of their method on visual recognition tasks including classification of motion trajectory and face recognition. An extension to subsets of varying cardinality is proposed in Shashua and Hazan (2004). In this first family of kernels, the main drawback, in some methods, resides is the strong assumption about the used probabilistic models in order to approximate the set of local features which may not hold true in practice.

In the second family, the "max" kernel (Wallraven et al., 2003) considers the similarity function, between two feature sets, as the sum of their matching scores and unlike discussed in Wallraven et al. (2003) this kernel is actually not Mercer (Bahlmann et al., 2002). In Lyu (2005), the authors introduced the "circular-shift" kernel defined as a weighted combination of Mercer kernels using an exponent. The latter is chosen in order to give more prominence to the largest terms so the resulting similarity function approximates the "max" and also satisfies the Mercer condition. The authors combined local features and their relative angles in order to make their kernel rotation invariant and they show its performance for the particular task of object recognition. In Boughorbel (2005), the authors introduced the "intermediate" matching kernel, for object recognition, which uses virtual local features in order to approximate the "max" while satisfying the Mercer condition. Recently, Grauman and Darrell (2007) introduced the "pyramid-match" kernel, for object recognition and document analysis, which maps feature sets using a multi-resolution histogram representation and computes the similarity using a weighted histogram intersection. The authors showed that their function is positive definite and can be computed linearly with respect to the number of local features. Other matching kernels include the "dynamic programming" function which provides, in Bahlmann et al. (2002), an effective matching strategy for handwritten character recognition, nevertheless the Mercer condition is not guaranteed. The main drawback of these kernels resides in their sensitivity to alignment so this affects their precision while the first class of local kernels suffers from the lack of discrimination

as the underlying statistical measures do not take into account the structure of data (for instance geometry in object recognition).

### 1.1 Motivation

The success of the second family of local kernels strongly depends on the quality of alignments which are difficult to obtain mainly when images contain redundant and repeatable structures. Regardless the Mercer condition, a *naive* matching kernel (such as the "max"), which looks for all the possible alignments and sums the best ones, will certainly fail and results into many false matches. Figures (1 and 2, left) illustrate the deficiency[1] of such naive context-free kernels when estimating the matching and also similarity between two groups of interest points. The highest values of the Gram matrix are clearly not concentrated in the diagonal. The Gaussian kernel, considered as context-free, is used in order to evaluate similarity between all the pairs of interest points, each one represented by its 3D RGB color attributes. Any slight perturbation of these attributes will result into unstable matching results if no context is taken into account (see Fig. 1). The same argument is supported in Schmid and Mohr (1997), for the general problem of visual features matching, about the strong spatial and geometric correlations between interest points and the corresponding close local features in the image space. This limitation also appears in closely related areas such as text analysis, and particularly string alignment. A simple example, of aligning two strings ("Sir" and "Hi Sir") using a simple similarity measure $\mathbb{1}_{\{c_1=c_2\}}$ between any two characters $c_1$ and $c_2$, shows that without any extra information about the *context* (i.e., the sub-string) surrounding each character in ("Sir" and "Hi Sir"), the alignment process results into false matches (see for instance Sahbi et al. (2008)). *Hence, it is necessary to consider the context as a part of the alignment process when designing kernels.* Our postulate states that one does not need perfect matching in order to improve the performance of kernels, but better alignment should produce better kernels.

### 1.2 Contribution

In this paper, we introduce a new kernel, called "context-dependent" ("CDK") and defined as the fixed-point of an energy function which balances a "fidelity" term and a "neighborhood" criterion. The fidelity is inversely proportional to the expectation of the Euclidean distance between the most likely aligned features (see Section 2) while the neighborhood criterion measures the spatial coherence of the alignments; given a pair of features $(f_p, f_q)$ with a high alignment quality, the neighborhood criterion is proportional to the alignment quality of all the pairs close to $(f_p, f_q)$ *but with a given spatial configuration*[2]. *The general form of "CDK" captures the similarity between any two features by incorporating their context and also geometry, i.e., the similarity of the features with exactly the same geometric configuration with respect to* $(f_p, f_q)$. Our proposed kernel can be viewed as a variant of "dynamic programming" kernel (Bahlmann et al., 2002) where instead of using the ordering assumption we consider a context (neighborhood) constraint which states that two points match if they have similar intrinsic features and if their neighbors *with exactly the same*

1. Deficiency means the poor discrimination of standard kernels due to the lack of their geometric representational power.
2. This is one of the most significant differences w.r.t. our former work.

*geometric configuration* match too. This also appears in other well studied kernels such as Fisher (Jaakkola et al., 1999), which implements the conditional dependency between data using the Markov assumption. "CDK" also implements such dependency with an extra advantage of being the fixed-point and the (sub)optimal solution of an energy function closely related to the goal of our application. This goal is to gather the properties of flexibility, invariance and mainly discrimination by allowing each local feature to consider its local geometry in the matching process. Moreover and in contrast to other existing state of the art approaches (for instance Fischler and Bolles (1981)), the proposed alignment strategy (and hence our kernel design) is model-free and is not based on any a priori alignment model such as affinity, homography or similarity which might fail in practice; for instance when objects deform. Notice also that the goal of this work is not to extend local features to be global and doing so (as in Mortensen et al. (2005); Amores et al. (2005)) makes local features less invariant, but rather to design a family of similarity kernels ("CDKs") which capture the context/geometry while also being invariant. Even though we investigated "CDKs" in the particular task of object recognition and retrieval, we can easily extend it to handle closely related areas in machine learning such as text alignment for document retrieval (Nie et al., 1999), machine translation (Sim et al., 2007) and bioinformatics (Scholkopf et al., 2004).

Given a database of images, each one seen as a constellation of local features. We use convolution kernels (see Section 2.2) in order to build similarity or Gram matrices, consisting of all the possible cross similarities of images in the database. A convolution kernel, as will be reminded in Section (2.2), is the sum of all possible cross similarities each one computed using a *minor kernel*; actually "CDK". Again, the minor kernel ("CDK") is the fixed point of an energy function which contains (i) a "fidelity" term which measures how similar are two primitives (ii) a "context/neighborhood" criterion which measures how good two primitives, with exactly the same context/geometric configuration, match and (iii) an "entropy" term which considers that without any a priori knowledge about the matching results between two primitives, the joint probability distribution of matchings should be as flat as possible so this term acts as a *regularizer*; furthermore it helps finding a direct and an easy to evaluate solution.

Notice that this work is the continuation of a previous work (Sahbi et al., 2008) with several updates and improvements:

- The notion of similarity and context is updated with finer statistics and measurements about geometry and co-occurrence (see Section 2.1).
- The theoretical results about the Mercer condition and the convergence to the fixed point are updated so now our framework provides better and loose constraints about the setting of some parameters as will be shown in Section 4.
- And, as will be shown in proposition 2, under certain circumstances about the choice of other parameters; actually adjacency matrices defined in 2.3, any Gram matrix built using "CDK" will be full rank so invertible.

In case of SVM, resp. SVR (support vector regression), the last property guarantees the existence of a classifier which separates (resp. approximates) a training set whatever its

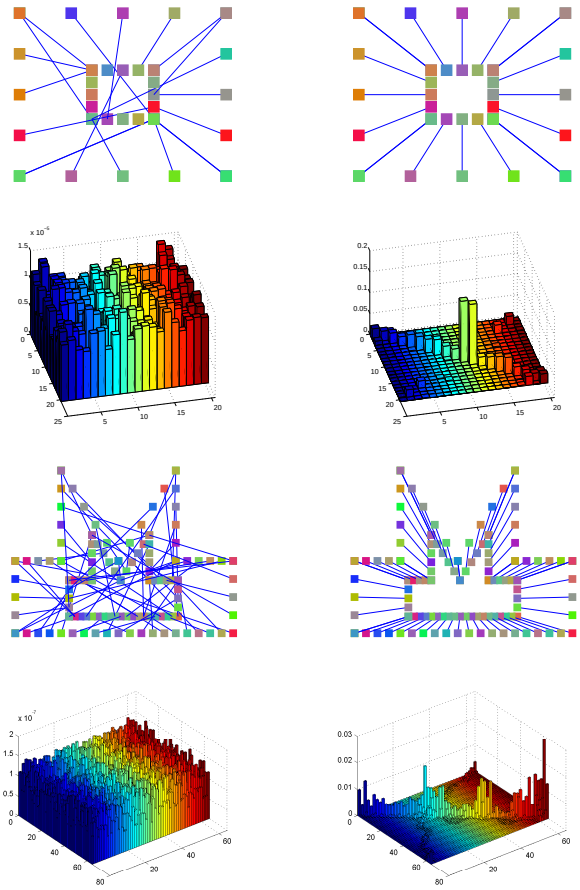

Figure 1: This figure shows a comparison of the matching results when using a naive (context-free) matching strategy and our "context-dependent" kernel matching. Second and fourth rows show the distribution of kernel values $k(x_i, x_j), i \in \mathcal{I}_p,\ j \in \mathcal{I}_q$ using a context-free kernel (left) and our "CDK" kernel (right). We can clearly see that the highest values change their locations so the matching results are now corrected.

labeling. This property is very desired and shows the discrimination power of "CDK" even though it is known that this will make the underlying VC dimension (Vapnik, 1998) infinite. Nevertheless, the actual VC dimension is bounded by the size of the training set as SVMs (resp. SVRs) find the solution in the span of training data so theoretical generalization bounds (Vapnik, 1998) are in practice not loose.

We consider the following organization of the paper; we first introduce in Section 2, our energy function which makes it possible to design our context-dependent kernel while integrating context/geometry and we show that this kernel satisfies the Mercer condition so we can use it for support vector machine training and other kernel methods. In Section 3 we show the application of this kernel in object recognition. We discuss in Section 4 the advantages and weaknesses of this kernel and the possible extensions in order to handle other tasks such as video retrieval. We conclude in Section 5 and we provide some future research directions.

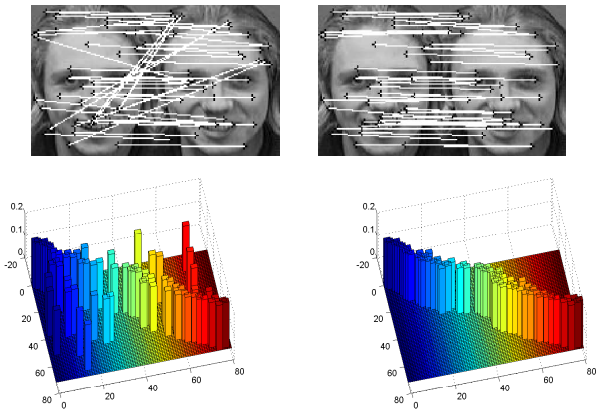

Figure 2: The caption of figure (1) is similar, but now we threshold the Gram matrices so only the maximum value of each row is set to a constant. We clearly see the improvement brought by CDK.

## 2. Incorporating Context and Geometry in Kernel Design

Let $\mathcal{S}_p = \{x_1^p, \ldots, x_n^p\}$ be the list of interest points of object $p$ (the value of $n$ may vary with the object $p$). The set $\mathcal{X}$ of all possible interest points is the union over all possible object $p$ of $\mathcal{S}_p$:

$$\mathcal{X} = \cup_p \mathcal{S}_p.$$

We consider $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ as a symmetric function which, given two interest points $(x_i^p, x_j^q)$, provides a similarity measure between them. Our goal is to design a similarity between any two objects $p$ and $q$, by designing a kernel $\mathcal{K}$ between the list of interest points $\mathcal{S}_p$ and $\mathcal{S}_q$ characterizing the objects $p$ and $q$.

### 2.1 Context and Co-occurrence

A feature refers to a local interest point $x = (\psi_g(x), \psi_f(x), \psi_o(x), \omega(x))$. The symbol $\psi_g(x) \in \mathbb{R}^2$ stands for the $2D$ coordinates of the interest-point $x$ while $\psi_f(x) \in \mathbb{R}^s$ corresponds to the descriptor of $x$ (for instance the 128 coefficients of the SIFT; Lowe (2004)). In case of SIFT, we have an extra information about the orientation of $x$ (denoted $\psi_o(x) \in [-\pi, +\pi]$) which is provided by the SIFT gradient. Finally, we use $\omega(x)$ to denote the object from which the interest point comes from, so that two interest points with the same position, descriptor and orientation are considered different when they are not in the same image (this is not surprising since we want to take into account the context of the interest point).

Let $d(x, x') = \|\psi_f(x) - \psi_f(x')\|_2$ measure the similarity between two interest point descriptors. Introduce

$$\mathcal{N}^{\theta,\rho}(x) = \{x' : \omega(x') = \omega(x), x' \neq x \text{ s.t. (i) and (ii) hold}\},$$

with

$$\frac{\rho - 1}{N_r}\ \epsilon_p \leq \|\psi_g(x) - \psi_g(x')\|_2 \leq \frac{\rho}{N_r}\ \epsilon_p, \qquad \text{(i)}$$

and

$$\frac{\theta - 1}{N_a}\pi \leq \angle\big(\psi_g(x), \psi_g(x') - \psi_g(x)\big) \leq \frac{\theta}{N_a}\pi. \qquad \text{(ii)}$$

Here $\epsilon_p$ defines a neighborhood and $\theta = 1, ..., N_a$, $\rho = 1, ..., N_r$. Notice that the definition of the neighborhood in this paper is different from the one proposed in Sahbi et al. (2008), as the latter provides only a set of neighbors $\mathcal{N}(x)$ around $x$ which are not segmented into different parts. In Sahbi et al. (2008), $\mathcal{N}_p(x) = \cup_{\theta,\rho}\mathcal{N}^{\theta,\rho}(x)$, so the new definition of neighborhoods $\{\mathcal{N}_p^{\theta,\rho}(x)\}_{\theta,\rho}$ reflects the co-occurrence of different interest points with particular geometric constraints (see Fig. 3).

### 2.2 Convolution Kernels

**Definition 1 (Subset Kernels)** *let $\mathcal{X}$ be an input space, and consider $\mathcal{S}_p, \mathcal{S}_q \subseteq \mathcal{X}$ as two finite subsets of $\mathcal{X}$. We define the similarity function or kernel $\mathcal{K}$ between $\mathcal{S}_p = \{x_i^p\}$ and $\mathcal{S}_q = \{x_j^q\}$ as $\mathcal{K}(\mathcal{S}_p, \mathcal{S}_q) = \sum_i^n \sum_j^m k\left(x_i^p, x_j^q\right)$.*

Here $k$ is symmetric and continuous on $\mathcal{X} \times \mathcal{X}$, so $\mathcal{K}$ will also be continuous and symmetric, and if $k$ is positive definite then $\mathcal{K}$ will also be positive definite (Haussler, 1999). Since $\mathcal{K}$ is defined as the cross-similarity $k$ between all the possible sample pairs taken from $\mathcal{S}_p \times \mathcal{S}_q$, it is obvious that $\mathcal{K}$ has the big advantage of not requiring any (hard) alignment between the samples of $\mathcal{S}_p$ and $\mathcal{S}_q$. Nevertheless, for a given $\mathcal{S}_p$, $\mathcal{S}_q$, the value of $\mathcal{K}(\mathcal{S}_p, \mathcal{S}_q)$ should be dominated by $\sum_i \max_j k\left(x_i^p, x_j^q\right)$, so the minor kernel $k$ should be appropriately designed.

### 2.3 Minor Kernel Design

For a finite collection of objects having each a finite number of interest points, the set $\mathcal{X}$ is finite. Provided that we put some (arbitrary) order on $\mathcal{X}$, we can view a kernel $k$ on $\mathcal{X}$ as a matrix $\mathbf{K}$ in which the "$(x, x')-$element" is the similarity between $x$ and $x'$: $\mathbf{K}_{x,x'} = k(x, x')$.

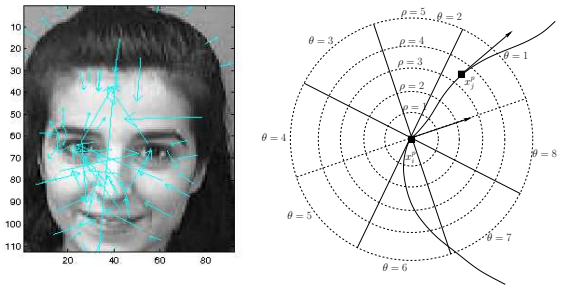

Figure 3: This figure shows the distribution of SIFT interest points (location, orientation and scales) (left) and the partitioning of the neighborhood of a given point into different sectors for orientations and bands for location and scale.

Let $\mathbf{P}_{\theta,\rho}$ be the intrinsic adjacency matrices respectively defined as $\mathbf{P}_{\theta,\rho,x,x'} = g_{\theta,\rho}(x,x')$, where $g$ is a decreasing function of any (pseudo) distance involving $(x,x')$, *not necessarily symmetric*. In practice, we consider $g_{\theta,\rho}(x,x') = \mathbb{1}_{\{\omega(x)=\omega(x')\}} \times \mathbb{1}_{\{x'\in\mathcal{N}^{\theta,\rho}(x)\}}$. Let $\mathbf{D}_{x,x'} = d(x,x')$. We propose to use the kernel on $\mathcal{X}$ defined by solving

$$\begin{aligned} \min_{\mathbf{K}} \quad & \mathrm{Tr}\big(\mathbf{K}\ \mathbf{D}^{'}\big) \ + \ \beta\ \mathrm{Tr}\big(\mathbf{K}\ \log\ \mathbf{K}^{'}\big) \\ & - \alpha \sum_{\theta,\rho} \mathrm{Tr}\big(\mathbf{K}\ \mathbf{P}_{\theta,\rho}\ \mathbf{K}'\ \mathbf{P}'_{\theta,\rho}\big) \qquad (1) \\ \text{s.t.} \quad & \begin{cases} \mathbf{K} \geq 0 \\ \|\mathbf{K}\|_1 = 1 \end{cases} \end{aligned}$$

Here the operations $\sqrt{\ }$, log and $\leq$ are applied individually to every entry of the matrix (for instance, $\log \mathbf{K}$ is the matrix with $(\log \mathbf{K})_{x,x'} = \log k(x,x')$), $\|\cdot\|_1$ is the "entrywise" $L_1$-norm (i.e., the sum of the absolute values of the matrix coefficients) and Tr denotes matrix trace. The first term, in the above constrained minimization problem, measures the quality of matching two descriptors $\psi_f(x)$, $\psi_f(x')$. In the case of SIFT, this is considered as the distance, $d(x,x')$, between the 128 SIFT coefficients of $x$ and $x'$. A high value of $d(x,x')$ should result into a small value of $k(x,x')$ and vice-versa.
The second term is a regularization criterion which considers that without any a priori knowledge about the aligned samples, the probability distribution $\{k(x,x')\}$ should be flat so the negative of the entropy is minimized. This term also helps defining a simple solution and solving the constrained minimization problem easily. The third term is a neighborhood criterion which considers that a high value of $k(x,x')$ should imply high kernel values in the neighborhoods $\mathcal{N}^{\theta,\rho}(x)$ and $\mathcal{N}^{\theta,\rho}(x')$. This criterion makes it possible to consider the context and geometry (spatial configuration) of each sample in the matching process.
We formulate the minimization problem by adding an equality constraint and bounds which ensure a normalization of the kernel values and allow to see $\{k(x,x')\}$ as a probability distribution on $\mathcal{X}\times\mathcal{X}$. Besides, for two different objects $p$ and $q$, $\{k(x,x')/\sum_{x:\omega(x)=p,x':\omega(x')=q} k(x,x')\}$ can be seen as a probability on $\mathcal{S}_p \times \mathcal{S}_q$.

### 2.4 Solution

**Proposition 2** *Let* $\mathbf{u}$ *denote the matrix of ones and introduce*

$$\zeta = \frac{\alpha}{\beta} \sum_{\theta,\rho} \|\mathbf{P}_{\theta,\rho}\mathbf{u}\mathbf{P}'_{\theta,\rho} + \mathbf{P}'_{\theta,\rho}\mathbf{u}\mathbf{P}_{\theta,\rho}\|_\infty,$$

*where* $\|\cdot\|_\infty$ *is the "entrywise"* $L_\infty$*-norm. Provided that the following two inequalities hold*

$$\zeta \exp(\zeta) < 1 \tag{2}$$

$$\|\exp(-\mathbf{D}/\beta)\|_1 \geq 2 \tag{3}$$

*the optimization problem* (1) *admits a unique solution* $\tilde{\mathbf{K}}$*, which is the limit of the context-dependent kernels*

$$\mathbf{K}^{(t)} = \frac{G(\mathbf{K}^{(t-1)})}{\|G(\mathbf{K}^{(t-1)})\|_1},$$

*with*

$$G(\mathbf{K}) = \exp\left\{ -\frac{\mathbf{D}}{\beta} + \frac{\alpha}{\beta} \sum_{\theta,\rho} \left(\mathbf{P}_{\theta,\rho}\mathbf{K}\mathbf{P}'_{\theta,\rho} + \mathbf{P}'_{\theta,\rho}\mathbf{K}\mathbf{P}_{\theta,\rho}\right) \right\},$$

*and*

$$\mathbf{K}^{(0)} = \frac{\exp(-\mathbf{D}/\beta)}{\|\exp(-\mathbf{D}/\beta)\|_1}$$

*Besides the kernels* $\mathbf{K}^{(t)}$ *satisfy the convergence property:*

$$\|\mathbf{K}^{(t)} - \tilde{\mathbf{K}}\|_1 \leq L^t \|\mathbf{K}^{(0)} - \tilde{\mathbf{K}}\|_1. \tag{4}$$

*with* $L = \zeta \exp(\zeta)$*.*

By taking not too large $\beta$, one can ensure that (3) holds. Then by taking small enough $\alpha$, Inequality (2) can also be satisfied. Note that $\alpha = 0$ corresponds to a kernel which is not context-dependent: the similarities between neighbors are not taken into account to assess the similarity between two interest points. Besides our choice of $\mathbf{K}^{(0)}$ is exactly the optimum (and fixed point) for $\alpha = 0$.

To have partitioned the neighborhood into several cells corresponding to different degrees of proximity (as shown in Fig. 3) has lead to significant improvements of our experimental results. On a theoretical viewpoint, it allows us to consider a larger $\alpha$ (since the constraint (2) becomes easier to satisfy with partitioned neighborhood), and apparently a more positively influencing context-dependent term (last term in (1)).

**Proof** Introduce the function

$$F : \mathbf{K} \mapsto \mathrm{Tr}\big(\mathbf{K}\ \mathbf{D}'\big)\ +\ \beta\ \mathrm{Tr}\big(\mathbf{K}\ \log\ \mathbf{K}'\big) - \alpha\ \sum_{\theta,\rho} \mathrm{Tr}\big(\mathbf{K}\ \mathbf{P}_{\theta,\rho}\ \mathbf{K}'\ \mathbf{P}'_{\theta,\rho}\big).$$

This function is continuous on the compact set defined by the constraints in (1) so it admits a minimum on it. Since the function $t \mapsto t \log t$ on the real numbers has an infinite negative derivative when $t$ tends to zero, none of the $\mathbf{K}_{x,x'}$ are equal to 0 at the minimum. Since the constraint $K \geq 0$ is not active on a minimum, the minima of $F$ are obtained when the gradient of $F$ is parallel to the gradient of the active constraint $\sum_{x,x'} \mathbf{K}_{x,x'} = 1$, i.e. when there exists $\lambda' \in \mathbb{R}$ such that for any $x, x' \in \mathcal{X}$,

$$\frac{\partial F}{\partial \mathbf{K}_{x,x'}} = \lambda',$$

hence when

$$\mathbf{D} + \beta(\mathbf{u} + \log \mathbf{K}) - \alpha \sum_{\theta,\rho} \left( \mathbf{P}_{\theta,\rho} \mathbf{K} \mathbf{P}'_{\theta,\rho} + \mathbf{P}'_{\theta,\rho} \mathbf{K} \mathbf{P}_{\theta,\rho} \right) = \lambda' \mathbf{u},$$

where we recall that $\mathbf{u}$ denotes the matrix of ones. So the minimum satisfies necessarily the fixed point relation

$$\mathbf{K} = \frac{G(\mathbf{K})}{\|G(\mathbf{K})\|_1},$$

with

$$G(\mathbf{K}) = \exp\left\{ -\frac{\mathbf{D}}{\beta} + \frac{\alpha}{\beta} \sum_{\theta,\rho} \left( \mathbf{P}_{\theta,\rho} \mathbf{K} \mathbf{P}'_{\theta,\rho} + \mathbf{P}'_{\theta,\rho} \mathbf{K} \mathbf{P}_{\theta,\rho} \right) \right\}, \tag{5}$$

where the function exp is applied individually to every entry of the matrix. We will now prove the unicity of the solution of this fixed point equation (5).

**Lemma 3** *Let $\mathcal{B}$ be the set of matrices with nonnegative entries and of unit $L_1$-norm, i.e., $\mathcal{B} = \{\mathbf{K} : \mathbf{K} \geq 0, \|\mathbf{K}\|_1 = 1\}$. If we have $\|\exp(-\mathbf{D}/\beta)\|_1 \geq 2$, then the function $\psi : \mathcal{B} \to \mathcal{B}$ defined as $\psi(\mathbf{K}) = G(\mathbf{K})/\|G(\mathbf{K})\|_1$ is $L$-Lipschitzian, with $L = \zeta \exp(\zeta)$, where we recall the definition $\zeta = \frac{\alpha}{\beta} \sum_{\theta,\rho} \|\mathbf{P}_{\theta,\rho} \mathbf{u} \mathbf{P}'_{\theta,\rho} + \mathbf{P}'_{\theta,\rho} \mathbf{u} \mathbf{P}_{\theta,\rho}\|_\infty$.*

As a consequence of this lemma, as soon as we have $L = \zeta \exp(\zeta) < 1$, the fixed point equation (5) admits a unique solution $\tilde{\mathbf{K}}$, and Inequality (4) holds.

**Proof** [Proof of Lemma 3] Let $\mathbf{K}_1$ and $\mathbf{K}_2$ be two matrices in $\mathcal{B}$. Introduce $\mathbf{G}_1 = G(\mathbf{K}_1)$ and $\mathbf{G}_2 = G(\mathbf{K}_2)$. We have

$$\begin{aligned}
& \|\psi(\mathbf{K}_2) - \psi(\mathbf{K}_1)\|_1 \\
= & \left\| \frac{\mathbf{G}_2}{\|\mathbf{G}_2\|_1} - \frac{\mathbf{G}_1}{\|\mathbf{G}_1\|_1} \right\|_1 \\
\leq & \left\| \frac{\mathbf{G}_2}{\|\mathbf{G}_2\|_1} - \frac{\mathbf{G}_2}{\|\mathbf{G}_1\|_1} \right\|_1 + \left\| \frac{\mathbf{G}_2}{\|\mathbf{G}_1\|_1} - \frac{\mathbf{G}_1}{\|\mathbf{G}_1\|_1} \right\|_1 \\
= & \min_{\mathbf{K}:\|\mathbf{K}\|_1=1} \left\| \mathbf{K} - \frac{\mathbf{G}_2}{\|\mathbf{G}_1\|_1} \right\|_1 + \left\| \frac{\mathbf{G}_2}{\|\mathbf{G}_1\|_1} - \frac{\mathbf{G}_1}{\|\mathbf{G}_1\|_1} \right\|_1 \\
\leq & \left\| \frac{\mathbf{G}_1}{\|\mathbf{G}_1\|_1} - \frac{\mathbf{G}_2}{\|\mathbf{G}_1\|_1} \right\|_1 + \left\| \frac{\mathbf{G}_2}{\|\mathbf{G}_1\|_1} - \frac{\mathbf{G}_1}{\|\mathbf{G}_1\|_1} \right\|_1 \\
= & \frac{2}{\|\mathbf{G}_1\|_1} \|\mathbf{G}_2 - \mathbf{G}_1\|_1 \\
\leq & \|\mathbf{G}_2 - \mathbf{G}_1\|_1, \qquad (6)
\end{aligned}$$

where the last inequality uses the assumption of the lemma. To upper bound the last difference, we use Taylor's formula. Consider $y, y'$ in $\mathcal{X}$. Let $\Delta G = \left|\mathbf{G}_2 - \mathbf{G}_1\right|$ and $\Delta K = \left|\mathbf{K}_2 - \mathbf{K}_1\right|$ be the matrices defined by $[\Delta G]_{x,x'} = \left|[\mathbf{G}_2]_{x,x'} - [\mathbf{G}_1]_{x,x'}\right|$ and $[\Delta K]_{x,x'} = \left|[\mathbf{K}_2]_{x,x'} - [\mathbf{K}_1]_{x,x'}\right|$. We have

$$\begin{aligned}
& \frac{\beta}{\alpha} \frac{\partial [G(\mathbf{K})]_{y,y'}}{\partial \mathbf{K}_{x,x'}} \\
= & \sum_{\theta,\rho} \big( [\mathbf{P}_{\theta,\rho}]_{x,y} [\mathbf{P}_{\theta,\rho}]_{x',y'} + [\mathbf{P}_{\theta,\rho}]_{y,x} [\mathbf{P}_{\theta,\rho}]_{y',x'} \big) [G(\mathbf{K})]_{y,y'}.
\end{aligned}$$

Therefore we have

$$\begin{aligned}
& \frac{\beta}{\alpha} [\Delta G]_{y,y'} \\
\leq & \sum_{\theta,\rho} \left[ \mathbf{P}'_{\theta,\rho} \Delta K \mathbf{P}_{\theta,\rho} + \mathbf{P}_{\theta,\rho} \Delta K \mathbf{P}'_{\theta,\rho} \right]_{y,y'} \|G(\mathbf{K})\|_\infty,
\end{aligned}$$

which implies

$$\begin{aligned}
& \frac{\beta}{\alpha} \|\mathbf{G}_2 - \mathbf{G}_1\|_1 \\
= & \frac{\beta}{\alpha} \sum_{y,y'} [\Delta G]_{y,y'} \\
\leq & \sum_{\theta,\rho} \mathrm{Tr}\big( \mathbf{P}'_{\theta,\rho} \Delta K \mathbf{P}_{\theta,\rho} \mathbf{u} + \mathbf{P}_{\theta,\rho} \Delta K \mathbf{P}'_{\theta,\rho} \mathbf{u} \big) \|G(\mathbf{K})\|_\infty \\
\leq & \sum_{\theta,\rho} \|\mathbf{P}_{\theta,\rho} \mathbf{u} \mathbf{P}'_{\theta,\rho} + \mathbf{P}'_{\theta,\rho} \mathbf{u} \mathbf{P}_{\theta,\rho}\|_\infty \|\Delta K\|_1 \|G(\mathbf{K})\|_\infty.
\end{aligned}$$

Now we trivially have

$$0 \leq G(\mathbf{K}) \leq \exp\left\{\frac{\alpha}{\beta}\sum_{\theta,\rho}\left(\mathbf{P}_{\theta,\rho}\mathbf{u}\mathbf{P}'_{\theta,\rho} + \mathbf{P}'_{\theta,\rho}\mathbf{u}\mathbf{P}_{\theta,\rho}\right)\right\},$$

hence we obtain

$$\|\mathbf{G}_2 - \mathbf{G}_1\|_1 \leq \zeta\|\Delta K\|_1 \exp(\zeta).$$

Plugging this inequality into (6), we get

$$\|\psi(\mathbf{K}_2) - \psi(\mathbf{K}_1)\|_1 \leq \zeta \exp(\zeta)\|\mathbf{K}_2 - \mathbf{K}_1\|_1.$$

### 2.5 Mercer Condition

A kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive (semi-)definite or is a Mercer kernel on $\mathcal{X}$, if and only if the underlying Gram matrix $\mathbf{K}$ is positive (semi-)definite. In other words, it is positive definite if and only if we have $V'\mathbf{K}V > 0$ for any vector $V \in \mathbb{R}^{\mathcal{X}} - \{0\}$. When we just have $V'\mathbf{K}V \geq 0$ for any vector $V \in \mathbb{R}^{\mathcal{X}} - \{0\}$, we just say that it is positive semi-definite. A Mercer kernel guarantees the existence of a Reproducing Kernel Hilbert Space (RKHS) (Shawe-Taylor and Cristianini, 2000) such that $k(x, x') = \langle\phi(x), \phi(x')\rangle$, where $\phi$ is an explicit (or more likely implicit) mapping function from $\mathcal{X}$ to the RKHS, and $\langle\cdot,\cdot\rangle$ is the dot kernel in the RKHS.

**Proposition 4** *The context-dependent kernels on $\mathcal{X}$ defined in Proposition* (2) *by the matrices $\tilde{\mathbf{K}}$ and $\mathbf{K}^{(t)}$, $t \geq 0$, are positive definite.*

**Proof** Let us prove that if $\mathbf{K}$ is positive semi-definite then $G(\mathbf{K})$ is also positive definite. We start by noticing that for a positive definite matrix $\mathbf{K}$ and for any matrix $\mathbf{P}$, the matrix $\mathbf{P}\mathbf{K}\mathbf{P}'$ is positive semi-definite since we have

$$V'\mathbf{P}\mathbf{K}\mathbf{P}'V = (\mathbf{P}'V)'\mathbf{K}(\mathbf{P}'V) \geq 0.$$

So the matrix $\mathbf{A} = \frac{\alpha}{\beta}\sum_{\theta,\rho}\left(\mathbf{P}_{\theta,\rho}\mathbf{K}\mathbf{P}'_{\theta,\rho} + \mathbf{P}'_{\theta,\rho}\mathbf{K}\mathbf{P}_{\theta,\rho}\right)$ is positive semi-definite. As a consequence, from (Shawe-Taylor and Cristianini, 2000, Proposition 3.12 p.42), the matrix $\sum_{i=1}^{\ell}\frac{A^i}{i!}$ is also positive semi-definite, where $A^i$ is the matrix such that $[A^i]_{x,x'} = (A_{x,x'})^i$ (that is, we consider the entrywise product, and not the matricial product). We get that $\exp(-\mathbf{D}/\beta)\sum_{i=1}^{\ell}\frac{A^i}{i!}$, and consequently $\mathbf{B} = \exp(-\mathbf{D}/\beta)\sum_{i=1}^{\infty}\frac{A^i}{i!}$, are also positive semi-definite. Since we have

$$G(K) = \exp(-\mathbf{D}/\beta) + \mathbf{B},$$

with $\mathbf{B}$ positive semi-definite and $\exp(-\mathbf{D}/\beta)$ positive definite (since it is a Gaussian kernel), we have thus proved that $G(\mathbf{K})$ is positive definite.

We now proceed by induction to prove that the functions $\mathbf{K}^{(t)}$ are positive definite. The function $\mathbf{K}^{(0)}$ is positive definite since it is a Gaussian kernel (up to a positive multiplicative factor). Since $\mathbf{K}^{(t)}$ is equal to $G(\mathbf{K}^{(t-1)})$ up to a positive multiplicative factor, we have by induction that $\mathbf{K}^{(t)}$ is a positive definite kernel. Since $\tilde{\mathbf{K}}$ is the limit of $\mathbf{K}^{(t)}$, we obtain that $\tilde{\mathbf{K}}$ is positive semi-definite. From this and the fixed point equation satisfied by $\tilde{\mathbf{K}}$, we obtain that $\tilde{\mathbf{K}}$ is positive definite.

## 3. Experiment

### 3.1 Databases and Settings

In order to show the extra-value of the "CDK" kernel with respect to standard usual ones, we evaluate the performances of support vector classifiers on different databases ranging from simple ones such as the Olivetti to more challenging such as the SmithSonian. The latter contains 35 leaf species, each one represented with $4 - 100$ examples, resulting into 1.525 images while the former is a face database of 40 persons each one contains 10 instances. We also experimented "CDK" on the standard MNIST database containing 10 digits, each one represented by $\sim 7.000$ examples (see Fig. 7). Interest points are extracted from all these databases and encoded using the usual SIFT descriptor.

Performances are measured using the Hold-Out error rate; in case of Olivetti and SmithSonian, half of data was used for training and the other half for testing while in case of MNIST only 40 examples per digit were used for training and the remaining data was used for testing. The use of this small training set is motivated by the run-time overhead due to estimating the Gram matrix. In spite of this extreme small limit in the size of the training samples, the reported results of "CDK" are significantly better compared to standard kernels as will be shown in the remainder of this section.

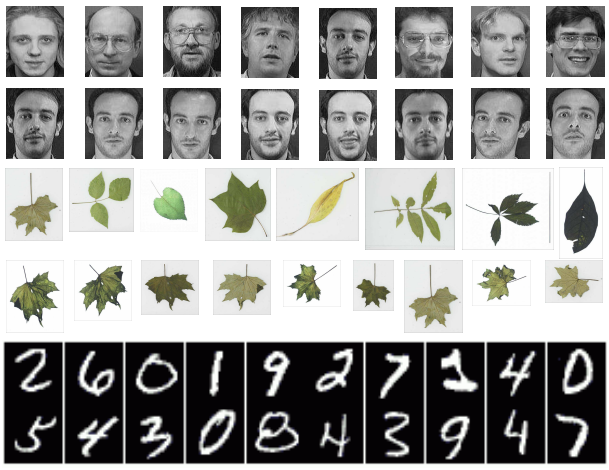

Figure 4: This figure shows samples from Smithsonian leaf, MNIST digits and Olivetti face databases.

### 3.2 Hold-Out Generalization and Comparison

We evaluate $\mathcal{K}$ (see section 2.2) and hence $\mathbf{K}^{(t)},\ t \in \mathbb{N}^+$ using three power assist settings: (i) linear $\mathbf{K}^{(0)}_{x,x'} = \langle \psi_f(x), \psi_f(x') \rangle$ (ii) polynomial $\mathbf{K}^{(0)}_{x,x'} = (\langle \psi_f(x), \psi_f(x') \rangle + 1)^2$ and (iii) Gaussian $\mathbf{K}^{(0)}_{x,x'} = \exp(-\|\psi_f(x) - \psi_f(x')\|^2/\beta)$. Our goal is to show the improvement brought when using $\mathbf{K}^{(t)},\ t \in \mathbb{N}^+$ , so we tested it against the standard context-

| Gaussian $\log\beta$ | Oliv (error+std) | MNIST (error+std) | Smith (error+std) |
|---|---|---|---|
| $-2$ | $1.62\% \pm 1.20$ | $36.94\% \pm 6.42$ | $6.65\% \pm 4.64$ |
| $-1$ | $\mathbf{1.12\% \pm 2.26}$ | $\mathbf{10.4\% \pm 6.14}$ | $\mathbf{3.49\% \pm 2.38}$ |
| 0 | $1.68\% \pm 3.41$ | $12.7\% \pm 5.09$ | $9.38\% \pm 8.85$ |
| 1 | $14.5\% \pm 12.3$ | $10.9\% \pm 6.10$ | $15.5\% \pm 13.2$ |

Table 1: This figure shows the error rate w.r.t the scale of the Gaussian kernel $\beta$.

free kernels (i.e., $\mathbf{K}^{(t)},\ t = 0$). For this purpose, we trained "one-versus-all" SVM classifiers for each class in SmithSonian, MNIST and Olivetti sets using the subset kernel $\mathcal{K}(\mathcal{S}_p, \mathcal{S}_q) = \sum_{x\in\mathcal{S}_p, x'\in\mathcal{S}_q} \mathbf{K}^{(t)}_{x,x'}$. Again, performances are reported, on different test sets, using the Hold Out error rate (see Fig. 5).

The setting of $\beta$ is performed by maximizing the performance of the Gaussian kernel as the latter corresponds the left-hand side (and the baseline form) of $\mathbf{K}^{(t)}$, i.e., when[3] $\alpha = 0$. For different databases, we found that the best performances are achieved[4] for $\beta = 0.1$ and this also guarantees condition (3). The influence (and the performance) of the right-hand side of $\mathbf{K}^{(t)}, \alpha \neq 0$ increases as $\alpha$ increases (see Table. 2 and Fig. 6), nevertheless and as shown earlier, the convergence of $\mathbf{K}^{(t)}$ to a fixed point is guaranteed only if (2) is satisfied. Therefore, it is obvious that $\alpha$ should be set to the highest possible value which also satisfies condition (2).

Diagrams in (5), show the hold out generalization errors and standard deviations resp. (from top-to-bottom) on Olivetti, SmithSonian and MNIST for different iterations; we clearly see the out-performance and the improvement of the "CDK" kernel (i.e., $\mathbf{K}^{(t)}$, $t \in \mathbb{N}^+$) with respect to the context-free kernels (i.e., $\mathbf{K}^{(0)}$.) We notice that both the Smithsonian and MNIST datasets are difficult compared to Olivetti, so the performances are worse on the former sets. Nevertheless, the improvement brought by CDK w.r.t. standard kernels is clear and consistent.

## 4. Discussion

Needless to say, interest points are order-less and variable-length, so it is meaningless to compare all the previous results w.r.t. *holistic* kernels. CDK clearly balances between holistic and local kernels as it allows us to have a local representation and similarity between images while taking into account the spatial configuration and the coocurence of interest points. Global kernels are very sensitive to misalignment of features due to transformations while local kernels are less sensitive to these effects but less discriminating. "CDKs" gather

3. Notice that selecting $\beta$ independently from $\alpha$ is obviously *"not sub-optimal"* for CDK but *"sub-optimal"* for the Gaussian kernel.
4. As the 128 SIFT coefficients are normalized to 1, it is not unreasonable that the optimal $\beta$ is the same for the three used test sets.

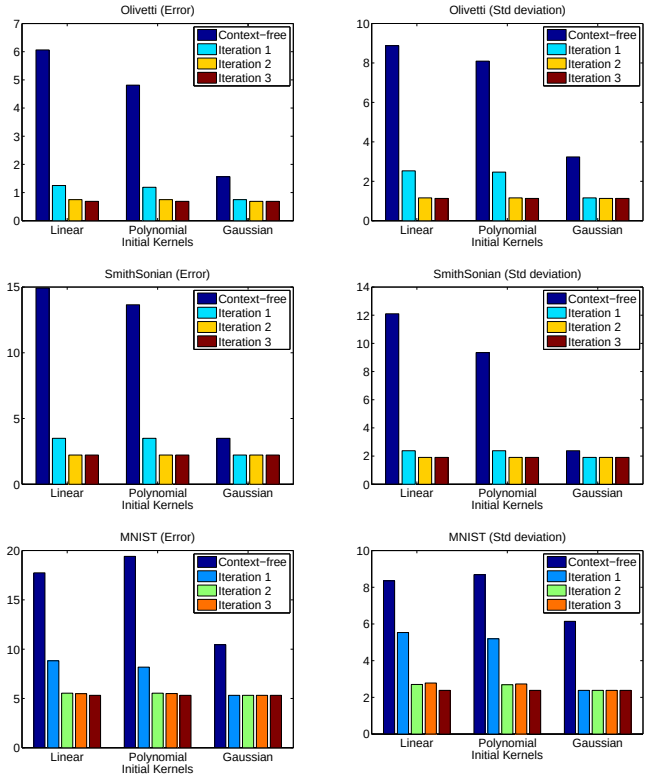


Figure 5: Error rates (left) and standard deviations (right) on resp. Olivetti, SmithSonian and MNIST test sets.($\alpha = 0.1$ and $\beta = 0.1$)

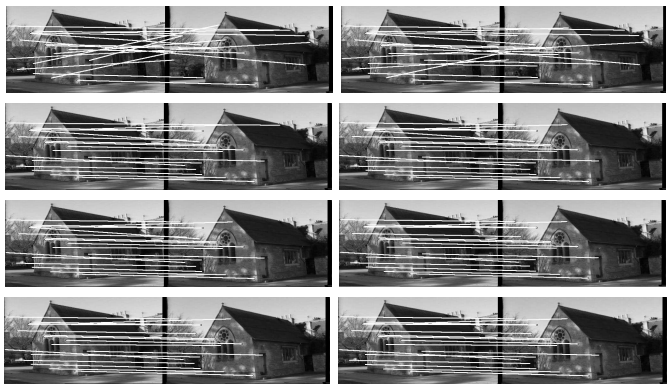

Figure 6: This figure shows the evolution of sample matches on a picture taken from 'http://www.robots.ox.ac.uk/∼vgg/hzbook/code/', for different and increasing values of $\alpha$ (resp, from top-left to bottom-right, 0, 0.1, 2.5, 3, 3.5, 4, 4.5 and 5). We clearly see that when $\alpha$ increases the matching results are better. We set $\beta = 0.1$ and $t = 1$ iteration only.

| Database<br>$\log \alpha$ | Oliv<br>(error+std) | MNIST<br>(error+std) | Smith<br>(error+std) |
|---|---|---|---|
| $-4$ | $1.12\% \pm 2.26$ | $10.4\% \pm 6.14$ | $3.49\% \pm 2.38$ |
| $-3$ | $1.12\% \pm 2.18$ | $10.3\% \pm 6.13$ | $3.49\% \pm 2.38$ |
| $-2$ | $0.93\% \pm 1.85$ | $9.69\% \pm 5.90$ | $3.49\% \pm 2.38$ |
| $-1$ | $\mathbf{0.68\% \pm 1.13}$ | $\mathbf{5.32\% \pm 2.38}$ | $\mathbf{2.22\% \pm 1.90}$ |

Table 2: This figure shows that error rate is a decreasing function of $\alpha$. Gaussian kernel is used for initialization, $\beta = 0.1$.

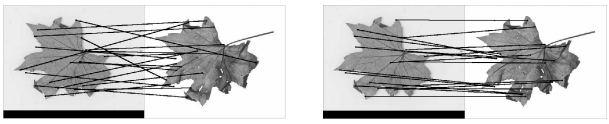

Figure 7: This figure shows samples of matches from the Smithsonian database, without CDK (left) and with CDK (right).

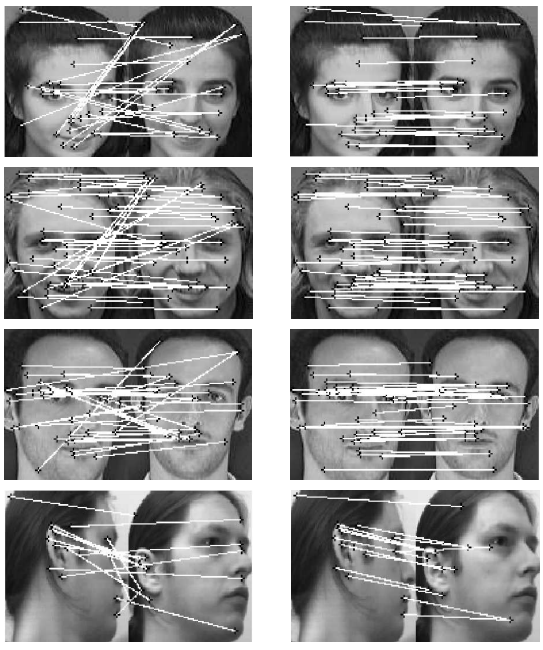

Figure 8: This figure shows samples of matching results using a context free kernel, actually Gaussian (left) and context dependent kernel (right).

these two properties while also being Mercer.

The adjacency matrices $\mathbf{P}_{\theta,\rho}$, in $\mathbf{K}$, provides the intrinsic properties and also characterizes the geometry of objects $\{\mathcal{S}_p\}$ in $\mathcal{X}$. It is easy to see that $\mathbf{P}_{\theta,\rho}$ is translation and rotation invariant and can also be made scale invariant when $\epsilon_p$ (see (i)) is adapted to the scales of $\psi_g(\mathcal{S}_p)$. It follows that the right-hand side of our kernel is invariant to any $2D$ similarity transformation. Notice, also, that the left-hand side of $\mathbf{K}^{(t)}$ may involves similarity invariant descriptors $\psi_f(.)$ (for instance SIFTs), so $\mathbf{K}^{(t)}$ (and also $\mathcal{K}$) is similarity invariant.

The out-performance of our kernel comes essentially from the inclusion of the context; in almost all cases, one iteration was sufficient in order to improve the performance of the Gaussian kernel, and couple iterations for the other context-free kernels. On the one hand, this corroborates the fact that the Gaussian kernel provides state of the art performance, when its parameters are well chosen (see section 3.2), and on the other hand, its performance can be consistently improved by including the intrinsic properties (i.e., geometry and context) of objects. Even though tested only on visual object recognition, our "CDK" kernel can be extended to many other pattern analysis problems such as bioinformatics, speech, text, etc. For instance, in text analysis and particularity machine translation (Sim et al., 2007), the design of a similarity kernel between words in two different languages, can be achieved using any standard dictionary (for instance WordNet). Of course, the latter defines similarity between any two words $(w_e, w_f)$ independently from their bilingual training text (or bitext), i.e., the phrases where $(w_e, w_f)$ might appear and this results into bad translation performances. A better estimate of similarity between two words $(w_e, w_f)$, can be achieved using their context i.e., the set of words which cooccure frequently with $(w_e, w_f)$ (Koehn et al., 2003).

Other extensions of this work may include:

- The interpolation of "CDK" on unseen data points, when training (and test) data are added incrementally.
- The refinement of the adjacency matrices which are currently estimated using scale and orientation of the SIFT features. According to our experiments, the orientation and scale are reasonably precise but might not be sufficiently consistent through scenes, even when SIFT points are repeatable; so one may start with a coarse estimation of the adjacency matrices and consider their refinement as a part of the minimization process (see equation 1) using expectation maximization.
- And also the application of the method in video search by considering instead of "geometric" context, the "spatio-temporal adjacency" and context between frames in video sequences.
- Finally, one current limitation of our "CDK" kernel $\mathbf{K}^{(t)}$ resides in its evaluation complexity. Assuming $\mathbf{K}^{(t-1)}$ known, for a given pair $x$, $x'$, the worst complexity is $O\left(\max(N^2, s)\right)$, where $s$ is the dimension of $\psi_f(x)$ and $N = \max_{x,p,\theta,\rho} \#\{\mathcal{N}_p^{\theta,\rho}(x)\}$. It is clear enough that when $N < \sqrt{s}$, the complexity of evaluating our kernel is strictly equivalent to that of usual kernels such as the linear. Nevertheless, the worst

case ($N \gg \sqrt{s}$) makes our kernel evaluation prohibitive and this is mainly due to the right-hand side of $\mathbf{K}^{(t)}_{x,x'}$ which requires the evaluation of kernel sums in a hypercube of dimension 4. A simple and straightforward generalization of the integral image (see for instance Viola and Jones (2001)) will reduce this complexity to $O(s)$. An other possibility for speed up, mainly in image retrieval, considers the fact that this kernel is Mercer, so it might be used in order to define a metric and exploit lossy (or not) acceleration techniques based on pivots or kd-trees.

## 5. Conclusion

We introduced in this paper a new type of kernels referred to as context-dependent. Its strength resides in the improvement of the alignments between interest points which is considered as a preliminary step in order to increase the robustness and the precision of object recognition. We have also shown that our kernel is Mercer and applicable to SVM learning. This is achieved for object and shape recognition problems and has better performance than SVM with context-free kernels.

The proposed approach, even though presented for kernel design, might be straightforwardly seen as graph matching. Indeed, one may define graph adjacency matrices and use exactly the same energy as (1) in order to derive similarity between nodes belonging to two different graphs. Obviously, matches correspond to pairs which maximize kernel values.

Future work includes the comparison of our kernel with other context-free kernels and its application in scene understanding.

## References

J. Amores, N. Sebe, and P. Radeva. Fast spatial pattern discovery integrating boosting with constellations of contextual descriptors. *IEEE International Conference on Computer Vision and Pattern Recognition*, 2005.

C. Bahlmann, B. Haasdonk, and H. Burkhardt. On-line handwriting recognition with support vector machines, a kernel approach. *IWFHR*, pages 49–54, 2002.

A. Barla, F. Odone, and A. Verri. Hausdorff kernel for 3d object acquisition and detection. *In Proceedings of the European conference on Computer vision LNCS 2353*, pages 20–33, 2002.

C.M. Bishop. Pattern recognition and machine learning. *Springer*, 2007.

B. Boser, I. Guyon, and V. Vapnik. An training algorithm for optimal margin classifiers. *In Fifth Annual ACM Workshop on Computational Learning Theory, Pittsburgh*, pages 144–152, 1992.

S. Boughorbel. *Kernels for Image Classification with Support Vector Machines*. PhD thesis, PhD. Thesis, Faculte d'Orsay, 2005.

O. Chapelle, P. Haffner, and V. Vapnik. Svms for histogram-based image classification. *Transaction on Neural Networks*, 10(5), 1999.

M. Cuturi. Etude de noyaux de semigroupe pour objets structures dans le cadre de l'apprentissage statistique. *PhD thesis GÃ©ostatistique, ENSMP*, 2005.

M.A. Fischler and R.C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Comm. of the ACM*, 24: 381–395, 1981.

T. Gartner. A survey of kernels for structured data. *Multi Relational Data Mining*, 5(1): 49–58, 2003.

G. Genton. Classes of kernels for machine learning: A statistics perspective. *journal of machine learning research*, 2(12):299–312, 2001.

K. Grauman and T. Darrell. The pyramid match kernel: Efficient learning with sets of features. *Journal of Machine Learning Research (JMLR)*, 8:725–760, 2007.

C. Harris and M. Stephens. A combined corner and edge detector. *Alvey Vision Conference*, pages 147–151, 1988.

D. Haussler. Convolution kernels on discrete structures. *Technical Report UCSC-CRL-99-10, University of California in Santa Cruz, Computer Science Department, July*, 1999.

Tommi Jaakkola, Mark Diekhans, and David Haussler. Using the fisher kernel method to detect remote protein homologies. *ISMB*, pages 149–158, 1999.

Philipp Koehn, Franz Joseph Och, and Daniel Marcu. Statistical phrase-based translation. *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL)*, 2003.

R. Kondor and T. Jebara. A kernel between sets of vectors. *In proceedings of the 20th International conference on Machine Learning*, 2003.

S. Lazebnik, C. Schmid, and J. Ponce. Semi-local affine parts for object recognition. *In British Machine Vision Conference (BMVC)*, 2004.

D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

S. Lyu. Mercer kernels for object recognition with local features. *In the proceedings of the IEEE Computer Vision and Pattern Recognition*, 2005.

H. Miyao, M. Maruyama, Y. Nakano, and T. Hananoi. Off-line handwritten character recognition by svm on the virtual examples synthesized from on-line characters. *Proceedings of the Eighth International Conference on Document Analysis and Recognition*, pages 494–498, 2005.

P. Moreno, P. Ho, and N. Vasconcelos. A kullback-leibler divergence based kernel for svm classfication in multimedia applications. *In Neural Information Processing Systems*, 2003.

E.N. Mortensen, H. Deng, and L. Shapiro. A sift descriptor with global context. *In IEEE International Conference on Computer Vision and Pattern Recognition*, pages 184–190, 2005.

J. Ng and S. Gong. Multi-view face detection and pose estimation using a composite support vector machine across the view sphere. *RATFG-RTS*, 1999.

Jian-Yun Nie, Michel Simard, Pierre Isabelle, and Richard Durand. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 1999.

H. Sahbi, J.Y. Audibert, J. Rabarisoa, and R. Keriven. Context dependent kernel design for object matching and recognition. *in IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

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

B. Scholkopf, K. Tsuda, and J.-P. Vert. Kernel methods in computational biology. *MIT Press*, 2004.

A. Shashua and T. Hazan. Algebraic set kernels with application to inference over local image representations. *In Neural Information Processing Systems (NIPS)*, 2004.

John Shawe-Taylor and Nello Cristianini. Support vector machines and other kernel-based learning methods. *Cambridge University Press*, 2000.

K.C. Sim, W. Byrne, M. Gales, H. Sahbi, and P. Woodland. Consensus network decoding for statistical machine translation system combination. *In the 32nd International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2007.

M.J. Swain and D.H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.

S. Tong and E. Chang. Support vector machine active learning for image retrieval. *MULTIMEDIA '01: Proceedings of the ninth ACM international conference on Multimedia*, pages 107–118, 2001.

Vladimir N. Vapnik. Statistical learning theory. *A Wiley-Interscience Publication*, 1998.

P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. *In CVPR*, 2001.

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

L. Wolf and A. Shashua. Learning over sets using kernel principal angles. *Journal of Machine Learning Research*, 4:913–931, 2003.

Dépôt légal : 2009 – 1er trimestre
Imprimé à l'Ecole Nationale Supérieure des Télécommunications – Paris
ISSN 0751-1345 ENST D (Paris) (France 1983-9999)

**TELECOM ParisTech**

Institut TELECOM - membre de ParisTech

46, rue Barrault - 75634 Paris Cedex 13 - Tél. + 33 (0)1 45 81 77 77 - www.telecom-paristech.frfr

**Département TSI**